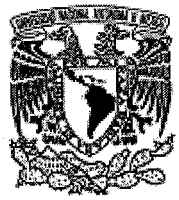


UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

MÉTODO DE CLASIFICACIÓN DE ÁRBOLES EN *POTIMIRIM MEXICANA* (CRUSTACEA: CARIDEA): CAMARÓN HERMAFRODITA DE LAS COSTAS DE VERACRUZ, MÉXICO

T E S I S

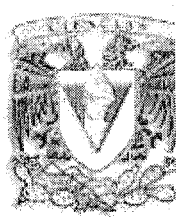
QUE PARA OBTENER EL GRADO DE:

MATEMÁTICO

P R E S E N T A :

GERMÁN HERNÁNDEZ GARCÍA

T U T O R E S :



M. EN A. P. MA. DEL PILAR ALONSO REYES
M. EN C. JOSÉ LUIS BORTOLINI ROSALES

2009

ÍNDICE

Se agradece al proyecto: “Adaptaciones ecofisiológicas de langostinos: implicaciones para su conservación”, PAPIIT IN208702, DGAPA-UNAM, Responsable: Dr. Fernando Álvarez Noguera.

Por proporcionar la información morfométrica de *Potimirim mexicana* para la realización de ésta investigación.

Resumen

Entre abril de 2003 y mayo de 2004, se realizaron 5 colectas en donde se capturaron 191 organismos pertenecientes a *Potimirim mexicana* (Crustacea: Caridea), en el Río Máquinas de la Estación de Biología Tropical Los Tuxtlas de la Universidad Nacional Autónoma de México ubicada en el Estado de Veracruz, México. Estos organismos son hermafroditas protándricos secuenciales, es decir, los organismos eclosionan de los huevos sexualmente indeterminados y el primer tipo de organismo sexualmente diferenciado que aparece en la población es un macho que posteriormente se diferencia en una hembra, observamos entonces, organismos indeterminados, machos y hembras en la población. De la totalidad de ésta, se obtuvieron medidas de diferentes estructuras del cuerpo: ocho para los 191 organismos, Largo del caparazón (LC); Diámetro del ojo (ED); Largo del telson (LT), Largo del segundo segmento del primer (L1ERPLEO) y segundo (L2DOPLEO) pleópodos, Largo del mero (LME) y carpo (LCA) del tercer pereiópodo; Longitud total (LTO); y finalmente, para el subconjunto de machos, el largo (AML) y ancho (AMW) del *appendix masculina*, que es una estructura asociada a la morfología de machos, característica sexual secundaria y que está localizada en el segundo pleópodo.

Con base en estas morfometrías, se analizaron los organismos por el método de árboles de clasificación y regresión (CART, por sus siglas en ingles). Este genera particiones recursivas binarias del espacio de medidas, asignando una clase a cada vector de medidas. La regla para detener la división es el de podado y consiste en desarrollar un árbol lo mas grande que se pueda y podarlo ascendentemente, complementándose con el criterio de costo-complejidad. Es decir, se define la complejidad para cada subárbol generado por la poda y se elige la mínima. El criterio de división es el de costo por clasificación errónea, que penaliza el número total de individuos erróneamente clasificados, por medio de una función de impureza en los nodos terminales que es minimizada. Para el análisis y obtención de resultados y gráficas de la base de datos generada, se utilizó STATISTICA 6.0, software creado por STATISTICA Enterprise Systems (Dar información técnica del programa: marca, versión, etc.). Este paquete estadístico proporciona tres tipos de medidas (Ji-cuadrada, Gini y G-cuadrada) y dos elecciones para las probabilidades *a*

priori, estimadas e iguales. Solamente se eligieron las dos primeras medidas y se combinaron con los dos tipos de probabilidades *a priori*.

Como resultado, se obtuvieron 6 árboles de clasificación y se compararon entre ellos para elegir sólo uno, con base en los nodos generados, organismos clasificados erróneamente y por el número de variables utilizadas. Adicionalmente, se obtuvieron en el proceso, las variables más relevantes para la clasificación y una función para poder clasificar nuevos organismos.

Abstract

In the period between April of 2003 and May of 2004, five samplings were performed where 191 organisms of *Potimirim mexicana* (Crustacea: Caridea), were captured, in the Maquinas River of the Tropical Biology Station, the Tuxtlas belonging to Universidad Nacional Autonoma de Mexico located in the State of Veracruz, Mexico. These organisms are sequential protandry hermaphrodite, that is, the sexually undetermined organisms are born of the egg and the first sexually mature organism that appears in the population is a male who later sex change in a female, we observed then, undetermined, male and females organisms in the population. From the totality of this one (sample), eight measures of different structures from the body were obtained: carapace length (LC); eye diameter (ED); telson length (LT), length of the second segment of first (L1ERPLEO) and second pleopod (L2DOPLEO); merus (LME) and carpus length (LCA) of third pereiópod; total length (LTO); and finally, for the subgroup of males, the length (AML) and width (AMW) of appendix masculine, this is a structure associated to the morphology of males, secondary sexual characteristic and that is located in the second pleopod.

With base in these measures, the organisms by the method classification and regression trees were analyzed (CART, by its abbreviations in English). This it generate binary recurrent partitions of the space of measures, assigning a class to each vector of measures. The stop-splitting rule is the pruning process and the first step is to grow a very large tree by letting the splitting procedure continue until all terminal nodes are small or pure or contain only identical measurement vectors, then pruning upward and complementing with the cost-complexity criterion. That is, the complexity for each subtree is defined and the minimal one is chosen. The splitting rule is misclassification cost that it penalizes the total number misclassified individuals. The idea is define an impurity function having certain desirable properties, then at any current terminal node, choose that split which most reduces the impurity function. For the analysis and obtaining of results and graphs of the data base generated, STATISTICA 6.0, software created by STATISTICA Enterprise Systems was used. This statistical package provides three types of measures (Ji-squared, G-squared and Gini) and two elections for *a priori* probabilities, estimated

and equal. The two first measures were only chosen and they were combined both with types of probabilities *a priori*.

Like result, six classification trees were obtained and they were compared among them to only choose one, with base in the generated nodes, organisms misclassified cost and by the number of used variables. Additionally, they were obtained in the process, variables for the classification and a function to be able to classify new organisms or samples.

Introducción

La clasificación es uno de los primeros problemas que aparecen en la actividad científica y constituye un proceso esencial en casi cualquier actividad humana, de una manera tal que en la resolución de problemas y en la toma de decisiones, la primera parte de la tarea consiste, por lo tanto, en clasificar.

Cuando uno desea clasificar un tipo (elemento, espécimen) en un cierto grupo, comenzando con valores de una serie de parámetros medidos u observados, y esa clasificación tiene cierto grado de incertidumbre llegara a ser necesario el uso de la estadística que, de hecho, mide la incertidumbre.

Particularmente, en la Biología es de suma importancia la clasificación de los individuos, incluso en un mismo grupo, debido a que pueden existir confusiones o puede tratarse de una nueva clase.

De las distintas formas para realizar clasificaciones, se encuentra utilizar datos morfométricos y con ellos determinar si un individuo pertenece a cierto grupo o descartar su pertenencia sin lugar a dudas.

En el caso que se trata en este trabajo *Potimirim mexicana*, existen complicaciones debido a la presencia del hermafroditismo y de los individuos indeterminados por su falta de madurez sexual.

La estadística es la rama de las matemáticas que desarrolló técnicas especializadas en la clasificación. Utiliza el análisis de clasificación, ya sea por medio del análisis de conglomerados o por el análisis de discriminante y en este caso, por el análisis de árboles de clasificación, el cual comienza con una serie de valores de los parámetros de los individuos, y por medio de un clasificador, determina a que grupo pertenece el espécimen, proporcionando una regla de clasificación para el presente grupo, así como para futuros grupos o individuos.

Una clasificación parte de un conjunto X cuyos elementos se desean catalogar. Se trata de obtener sucesivas particiones, que están organizadas en diferentes conjuntos, donde cada partición es formada por clases disjuntas. (Breiman *et al*. 1984)

El presente trabajo une el interés biológico de clasificar con la técnica de árboles en *Potimirim mexicana*, y con ello se generan distintos árboles de clasificación basados en las medidas Ji-cuadrada y de Gini, donde se determina cuáles variables morfométricas son las significativas para la clasificación.

Así, los objetivos que se persiguen son:

Objetivo general

Caracterizar por medio de medidas morfométricas la población de *Potimirim mexicana* colectada en el Río Máquinas en el Estado de Veracruz, México durante 2003 y 2004.

Objetivos particulares

- Determinar las variables morfométricas más importantes en la discriminación.

- Mostrar el mejor modelo de discriminación a través de la técnica de árboles de clasificación.

- Encontrar (exhibir proporcionar) una función que clasifique a un nuevo individuo de *Potimirim mexicana*.

El trabajo se elaboró con tres capítulos; el primero “Antecedentes” cumple el cometido de proporcionar la información biológica sobre el organismo que se estudia y formar el marco teórico que sustenta el análisis biológico de *Potimirim mexicana*.

El capítulo dos muestra el material y los métodos empleados en la tesis, se indica cómo y dónde se realizó la colecta y que variables fueron medidas; también, se explica la teoría matemática de árboles de clasificación.

En el capítulo tres se desarrollan los modelos de los árboles de clasificación con estadísticos como Ji-cuadrada y Gini y se determina un análisis de cuál árbol resultó mejor.

Por ultimo se presentan las conclusiones, anexos y la literatura consultada para elaborar este trabajo.

1. Antecedentes

1.1 Características biológicas

El nombre genérico *Potimirim*, propuesto por Holthuis (1954) fue adoptado por Villalobos (1959), en el trabajo en donde revisó los caracteres diagnósticos del género; haciendo una revisión de las especies hasta entonces conocidas: *P. glabra, P. mexicana* y *P. potimirim* (Barros y Fontoura, 1996).

Las especies incluidas hasta el momento en este género son: *P. americana*, *P. brasiliana*, *P. glabra*, *P. mexicana* y *P. potimirim* que han tenido su presencia registrada solamente para el Continente Americano (Barros y Fontoura, 1996).

Según Ortmann (1897), casi todos los miembros de la familia Atyidae se encuentran limitados a cuerpos de agua dulce. Los camarones del género *Potimirim*, se localizan preferentemente en las orillas de los ríos de aguas claras con corrientes rápidas, entre rocas y raíces de plantas acuáticas, en profundidades menores a 0.50 m; en ocasiones algunos organismos son llevados hacia el mar, presos por las raíces de plantas acuáticas (Barros y Fontoura, 1996).

No existen especies marinas conocidas del género *Potimirim* Villalobos (1959), Fryer (1977) y Molina (1987) (Barros y Fontoura, 1996).

Los organismos de este grupo pueden presentar migraciones y grandes fluctuaciones de densidad poblacional como resultado de variaciones de disponibilidad de alimentos en función de las corrientes del curso del agua Covich *et al.* (1991), (Barros y Fontoura, 1996).

Los sexos separados son observados en la especie de carideos, sin embargo, las especies hermafroditas con cambio de sexo han sido reportados en aproximadamente 30 (Bauer, 2000). El fenómeno de la protandria, es una madurez sexual primero como un macho para después cambiar a hembra (Charnov, 1979).

Entonces, en una distribución de frecuencia de una población, los individuos reproductivos más pequeños son machos y los más grandes son hembras.

El género *Potimirim* ha sido observado en ambos litorales mexicanos; *P. glabra* a lo largo de las costas del Pacifico, desde el México meridional hasta América Central; *P. mexicana*, está distribuida en los llanos costeros de el Golfo de México, desde el Río Soto La Marina, Tamaulipas hasta el Río Grijalva, Tabasco (Villalobos, 1959).

P. glabra exhibe dimorfismo sexual, evidente en la distribución de frecuencia del tamaño. Bajo 8.0 Mm. de la distancia total más especímenes son indiferenciados.

El primer organismo maduro sexualmente que aparece con características sexuales externas como el primer y segundo pleópodos, son los machos con longitudes entre 8.0 y 15.0 Mm. Las hembras aparecen después con longitudes extendiéndose desde 10.0 hasta 26.0 Mm., aunque en algunas poblaciones de hembras maduras sexualmente han sido reportadas de los 6.2 Mm. de longitud total (Martínez, 2003).

Machos de *P. mexicana* adquieren su madurez sexual comenzando en 8.0 Mm. mientras que las hembras comienzan a aparecer con 11.0 Mm. de longitud total (Luna, 1989).

El mecanismo reproductivo de *P. glabra* y *P. mexicana* según en estudios anteriores, sugiere que la estrategia de este tipo de hermafroditismo, corresponde al modelo de ventaja del tamaño. Ghiselin (1969), propuso tres modelos para explicar bajo que condiciones se desarrolla y favorece el fenómeno del hermafroditismo en animales, estos son:

1) el de baja densidad,
2) el de ventaja de tamaño y
3) el del gen disperso.

El primero, es la explicación clásica del fenómeno del hermafroditismo a partir de la teoría de la selección natural, en donde un organismo por algún atributo (baja

densidad poblacional o baja movilidad de los organismos reproductores) reduce las oportunidades de aparearse con otros individuos de la especie. Los hermafroditas, entonces, no tienen ese problema.

El segundo modelo, habla de la ventaja de la talla de alguno de los individuos. Bajo el supuesto de que las funciones reproductivas están descargadas de mejor manera sobre uno de los sexos y ello incrementa el potencial reproductivo de la especie. Dentro de este modelo se encuentran un gran número de ejemplos del hermafroditismo secuencial, en donde los individuos se reproducen más eficientemente como miembros de uno de los dos sexos los cuales presentan diferentes tallas. En este modelo, no están influenciados atributos como la densidad poblacional.

El tercer modelo, denominado del gen disperso o ambiente genético Mayer (1954), está basado en el supuesto de que las limitaciones de dispersión pueden afectar a la estructura poblacional. Esto es que bajo ciertas condiciones, la disponibilidad de individuos para aparearse está restringida por falta de movilidad, mientras que los hermafroditas salvan estas dificultades.

1.2 Diagnosis del género *Potimirim* HOLTHIUS (1954); VILLALOBOS (1959).

Atyidae con rostro más bien corto, desprovisto de dientes en la parte superior, con dientes por debajo. Espinas supraorbitales ausentes; espinas antenal y pterigostomiana presentes en las hembras, la ultima reducida o ausente en los machos. Ojos con cornea no muy extendida. Sin exopodios en la base de los pereiópodos. “Sin artrobranquia en los perieópodos del primer par, sólo una pleurobranquia en la somita correspondiente”. Epipoditos en las bases de los primeros tres o cuatro pares de pereiópodos. Quelas delgadas con largos mechones de cerdas en las puntas de los dedos, “muchas de estas dentadas con el extremo bífido y los dentículos en una sola hilera a lo largo de la cerda”. Palma presente aunque extremadamente corta. Carpo de los primeros pereiópodos excavado anteriormente. Carpo de los segundos pereiópodos más largo que ancho. “Órgano

sexual masculino generalmente presente en el carpopodio de los pereiópodos tercero y cuarto". *Appendix masculina* diferenciado. Habitan aguas dulces o "salobres".

Existen diferencias morfológicas importantes que permiten separar a *Potimirim* en dos grupos naturales Villalobos (1959).

- Con epipodito en los cuartos pereiópodos (aparato epipodial completo).
  Potimirim glabra (Kingsley).
  Potimirim brasiliana nov. Sp.
- Sin epipodito en los cuartos pereiópodos (aparato epipodial incompleto).
  Potimirim mexicana (Saussure, 1858).
  Potimirim potimirim (Müller, 1881).

Holthius (1955) establece una diferencia esencial que permite separar a *Potimirim glabra* de *Potimirim mexicana*, este se basa en la presencia de epipodito en el coxopodio del periópodo del cuarto par en la primera especie; este dato tiene antecedentes en un trabajo de Bouvier (1909), con el cual es posible separar los dos grupos naturales a los que ya se hizo referencia.

El trabajo de Villalobos (1959) permite ofrecer una característica más, referente a la ausencia de pleurobranquia en la última somita torácica de los machos de aquellas especies que sólo presentan epipodito hasta el tercer par de pereiópodos.

1.3 Diagnosis de *Potimirim mexicana* VILLALOBOS (1959)

Rostro con dos o tres dientes en la quilla ventral; el de las hembras alcanzando con su ápice el tercio medio o el anterior del segundo artejo del pedúnculo antenular.

Caparazón del macho sin espina pterigostomiana. Macho sin pleurobranquia en la última somita torácica. Pincel de los quelípedos con cerdas dentadas y éstas con su extremo distal bífido.

Carpopodio del segundo pereiópodo del macho de la misma longitud que el borde inferior del propodio. Meropodio de los pereiópodos tercero y cuarto del macho esbelto en su segundo tercio distal. Carpopodio de los cuartos pereiópodos del macho sin espina subdistal en el lado externo.

Carpopodio de los pereiópodos tercero y cuarto del macho con dos espinas, una a cada lado de la región inferoproximal y dos zonas de pequeños tubérculos acompañando a dichas espinas.

Coxopodio de los cuartos pereiópodos sin epipodito. Meropodio de los quintos pereiópodos del macho con tres espinas. *Appendix masculina* de los pleópodos segundos con una muy ligera escotadura en su borde posterior, separando el lóbulo superior del lóbulo medio; índice de diámetros 58.75; *appendix interna* de los mismos pleópodos, corto, con ocho a doce *uncinuli* o ganchitos *retinaculares*. Borde distal del telson sin cerda mediana corta.

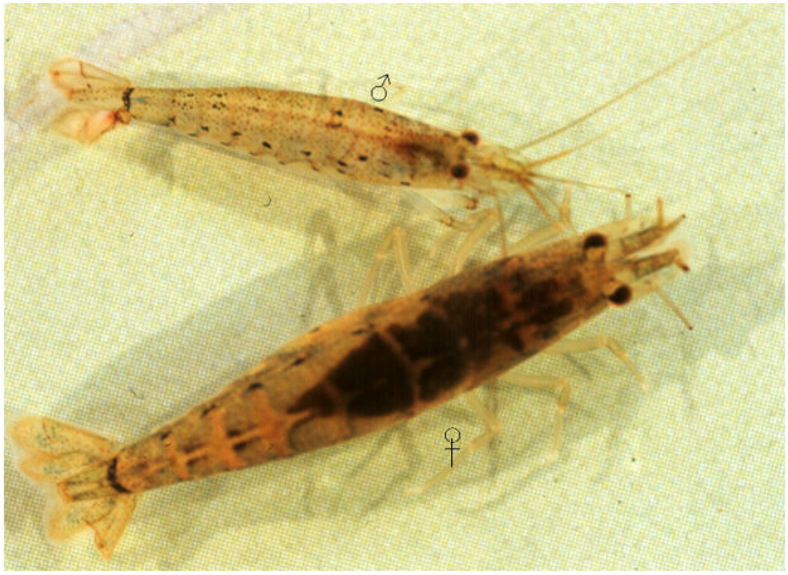


Figura 1. *Potimirim mexicana* (Obsérvese la mayor talla de la hembra con relación al macho).

1.4 Morfometrías de *Potimirim mexicana*

Con referencia a trabajos previos de morfometrías de organismos del género *potimirim*, se encuentra el de Villalobos (1959).

1.4.1 Morfometrías del macho

Es de talla más pequeña y de aspecto general más esbelto que las hembras. La longitud del caparazón proyectada sobre el abdomen, alcanza hasta la tercera somita. (Figura 1).

El rostro rebasa ligeramente el borde proximal del segundo artejo del pedúnculo antenular, pero en los individuos jóvenes es más corto. La quilla ventral del rostro presenta dos o tres dentículos, el posterior se dispone aproximadamente al principio del tercio anterior; el ápice rostral es muy agudo, dirigido hacia delante o ligeramente levantado.

La espina suborbital es aguda y de contorno triangular; no existe espina pterigostomiana, pues el borde anteroinferior del caparazón es redondeado; la altura posterior del escudo céfalo torácico es de la mitad de su longitud.

El abdomen es esbelto y en los ejemplares fijados, generalmente se nota una inflexión brusca al nivel de la tercera somita abdominal. Las regiones pleurales son mucho menos amplias que en las hembras; las pleuras de las somitas I, III, IV y V son muy semejantes en anchura, la de la somita II es el doble de la longitud del terguito de la somita I.

El telson es moderadamente ancho en su base y los bordes laterales son ligeramente convexos; el borde posterior es semicircular; a partir de los dos tercios distales de la superficie dorsal se destacan cinco pares de espinulas, dispuestas longitudinalmente en dos series que divergen hacia los ángulos laterodistales del telson; en posición submarginal y en el extremo distal, hay un proceso espinoforme triangular y aplanado, cuyo vértice coincide con el borde del telson; a cada lado de esta espina hay una serie de cuatro cerdas muy delgadas ordenadas paralelamente

al borde posterior del telson; en cada ángulo posteroexterno del telson se articula un largo proceso espinoforme fuertemente quitinizado, ligeramente recurvado, ancho en su base, con forma de punta de flecha en el ápice y cuya longitud es poco menos que la mitad de la de las cerdas plumosas medianas que bordean posteriormente al telson; estas ultimas son en número normal de ocho, repartidas simétricamente a uno y otro lado de la línea media; cada cerda presenta a lo largo de los dos tercios distales, líneas transversas equidistantes que le dan una apariencia articulada; no existe cerda mediana, pero si alguna llega a ocupar ese sitio por propia simetría, es exactamente de la misma naturaleza y longitud que las otras.

El pereiópodo del primer par, presenta un epipodito, una mastigobranquia y una pleurobranquia. El basipodio es el artejo mas corto de apéndice, su longitud mayor es la mitad de la del isquiopódio.

El isquiopódio es un artejo medianamente largo y esbelto, su longitud equivale a la de la porción dactilar del propodio. El meropodio es el artejo de mayor longitud del apéndice, la cual equivale a seis veces la anchura mayor del mismo artejo.

El carpopodio es mucho mas corto que el del pereiópodo del segundo par; tiene forma subcónica, truncado en el ápice o extremo distal y con la base sesgada que corresponde al extremo proximal; la longitud del borde inferior es un tercio mayor que la del borde superior; la cara anterior del artejo es cóncava y en ella se adosa perfectamente la región palmar del propodio que es convexa; en el borde anterosuperior se destacan varias cerdas, una de ellas muy gruesa y otras tres delgadas y cortas.

El propodio o dedo inmóvil de la quela es robusto en su base; la región palmar corresponde a un cuarto de su longitud total y la forma de la superficie posterior, se amolda a la cavidad anterior del carpopodio; la superficie superior de este artejo es acanalada y desde el último cuarto distal presenta abundantes cerdas; las proximales dispuestas a uno y otro lado del borde mientras que las distales se insertan además sobre la superficie dorsal; entre estas últimas hay algunas cuya longitud es ligeramente mayor que la de la región dactilar del propodio, y otras que

presentan una serie de estructuras dentiformes en la mayor parte de su longitud, quedando libres de ellas la región proximal y la distal, esta última se separa en parte de la región dentada y además presenta abundantes pelos cortos.

El dactilopodio o dedo inmóvil es ligeramente estrecho subproximalmente, pero su grosor se acentúa hacia la región distal, en donde también existen numerosas cerdas, que junto con las del dedo opuesto forman el pincel de la quela; entre las cerdas del dedo inmóvil también hay cerdas dentadas.

El pereiópodo del segundo par, es mucho mas largo que el anterior y salvo el coxopodio, todos sus artejos son esbeltos. Presenta un epipodito, una mastigobranquia y una pleurobranquia.

El basipodio muestra una cerda larga y desnuda en el ángulo inferodistal. Este mismo carácter se repite en el isquiopodio, pero el punto de inserción de la cerda está ligeramente desplazado hacia la posición subdistal. El meropodio es subcilindrico, excepto en la superficie articular, la cual equivale aproximadamente a un tercio de su longitud; el borde superior de este artejo muestra una serie de seis a siete pequeñas cerdas equidistantes repartidas.

El carpopodio es subcilíndrico, mas ensanchado en el extremo anterior, ligeramente cóncavo en el perfil superior y recto en el inferior; la cara distal es cóncava y ahí se ajusta perfectamente la parte posterior del propodio; en el borde superodistal del artejo se destaca una cerda robusta y cónica, acompañada de otras mas pequeñas; en cuanto a las proporciones del carpopodio, la mas importante es la que se refiere al borde inferior, cuya longitud es igual a la del borde inferior del propodio, desde la articulación con el carpopodio hasta el extremo distal del dedo.

El propodio tiene forma de cuña, la parte que corresponde con la región palmar de la quela, es corta y su anchura es tres y media veces menor que la longitud del borde inferior; en el tercio terminal se insertan las cerdas que forman parte del pincel del dedo, las mas largas de las cuales llegan a tener una longitud igual a la del borde

inferior del artejo; entre las cerdas existen algunas dentadas, de aspecto y tamaño semejantes a las descritas para el pereiópodo del primer par.

El dactilopodio o dedo móvil de la quela es muy esbelto en el extremo proximal y robusto en el distal, por lo cual, en perfil, se asemeja a un basto; las cerdas más largas que forman el pincel son iguales en longitud al borde superior del carpopodio y algunas de ellas son dentadas.

El pereiópodo del tercer par es el más robusto de todos los apéndices; le corresponde un epipodito, una mastigobranquia y una pleurobranquia. Tanto el basipodio como el isquipodio presentan una cerda larga subterminal en el borde inferior.

El meropodio es largo y fuerte, recto en el borde superior, convexo en el inferior; este ultimo con cuatro espinas cónicas, inclinadas hacia delante e insertas en la mitad distal, la última de las cuales colocada en la región inferolateral; el borde anterosuperior de este artejo presenta dos o tres cerdas espinoformes dirigidas distalmente.

El carpopodio es la mitad de la longitud del meropodio, es robusto en el extremo distal y esbelto en el proximal; el borde superior se prolonga anteriormente en una especie de escama; en la región posteroinferior de este mismo artejo, se destaca un órgano que sin duda tiene cierto papel en la cópula, puesto que es privativo del macho; esta formado por una zona del mismo borde inferior con abundantes tubérculos muy cortos, que se prolonga hasta la mitad de la longitud del artejo; muy cerca del borde articular proximal y en cada zona lateroinferior se inserta una espina; la interna con su borde lateral axilar oblicuamente estriado y la externa completamente lisa; además en la cara interna hay otra pequeña espina dirigida distalmente y una pequeña placa semicircular con su borde inferior libre y dentado.

El propodio es muy esbelto, cilíndrico, ligeramente más grueso hacia el extremo distal, con sus caras superior e inferior provistas de pequeñas espinas, pero la de los bordes distales superior e inferior son muy desarrolladas; la longitud de este artejo equivale a 1.25 veces la del carpopodio.

El dactilopodio es corto, su borde superior es un cuarto de la longitud del borde superior del propodio; el borde inferior está armado de fuertes espinas, recurvadas hacia atrás, con excepción de la terminal que es casi recta y cónica; el número total de espinas es de ocho.

El pereiópodo del cuarto par presenta una setobranquia o mastigobranquia y en la somita correspondiente una pleurobranquia, pero carece de epipodito; el basipodio tiene un mechón de cerdas en posición subterminal en el borde inferior.

El isquiopodio aun conserva la cerda que se muestra en los pereiópodos del segundo y tercer par; el meropodio es el mayor de todos los artejos, su aspecto es muy semejante al del pereiópodo del tercer par y con cuatro espinas insertas en la mitad distal del borde inferior, la proximal está ligeramente separada de las otras.

El carpopodio tiene el mismo órgano sexual que el apéndice anterior y la longitud de este artejo cabe una y media veces en la del propodio. Este último tiene un aspecto semejante al del tercer par de pereiópodos, pero de longitud proporcionalmente mayor, es decir, equivalente a 1.5 veces la longitud del carpopodio.

El dactilopodio es un cuarto de la longitud del propodio, está provisto de siete espinas ligeramente recurvadas y además una muy pequeña en la parte posterior de la serie.

El pereiópodo del quinto par no presenta epipodito, setobranquia ni pleurobranquia. Sólo el isquiopodio muestra un mechón de cerdas en la región subdistal del borde inferior y su longitud mayor es aproximadamente la mitad del borde superior del meropodio. Este último artejo es relativamente mas corto que el de los pereiópodos

anteriores, presenta sólo tres espinas en el borde inferior; el borde superodistal está provisto de tres o cuatro espinas, una de ellas más larga que las otras.

El carpopodio es casi la mitad de la longitud de propodio y muestra una espina en la región subproximal y otra en la subdistal del borde inferior, pero carece del órgano sexual característico de los dos pares de pereiópodos anteriores.
El propodio alcanza en este apéndice la categoría de artejo mayor, es aproximadamente el doble de la longitud del carpopodio y presenta su borde inferior armado de unas ocho espinas robustas, cuya longitud se incrementa a medida que son mas distales; en el borde anteroinferior se destacan dos espinas semejantes a las anteriores pero de una longitud mucho mayor, el borde superior muestra unas seis cerdas y otras tres insertas en el borde articular distal.

El dactilopodio es un cuarto de la longitud del propodio; en el borde inferior presenta una serie pectinada de espinas, cuyo número varia entre 22 y 23, y además el diente terminal con que finaliza el artejo.

El pleópodo del primer par presenta un exopodio laminar oblongo, orlado de cerdas plumosas; el endopodio, en cambio, es ancho en su base y esbelto en el ápice, su longitud es dos tercios la del exopodio, su borde está provisto de cerdas rígidas y desnudas, cuya longitud disminuye hacia el extremo distal; el borde interno, en cambio, presenta proximalmente una seis cerdas delgadas, después cuatro o cinco un poco mas gruesas pero de mayor longitud, el total de las cerdas del borde externo se distribuye en toda la mitad proximal de éste; en la región terminal del endopodio hay una zona restringida con siete a diez *uncinuli*.

En el pleópodo del segundo par el *appendix masculina* es el órgano más conspicuo; es laminar, con forma aproximada de un triángulo rectángulo cuya base es dos tercios de la altura.

Con el objeto de facilitar la comparación de este órgano entre las distintas especies, se decidió obtener de él un índice, el cual es el resultado de la siguiente formula: i =diámetro anteroposterior X 100 entre diámetro superoinferior. Se ha marcado la

disposición de estos diámetros con dos líneas rectas llamadas SID y APD. El índice 58.75 que se obtiene para el *appendix masculina* de *Potimirim mexicana* permite distinguir de inmediato una diferencia positiva con otras especies.

El borde posterior está armado de espinas que se disponen en él en una serie lineal; los bordes laterales de dichas espinas están provistos de pequeños dentículos, con excepción de las distales y proximales, que solo los tienen en uno de sus bordes; esta serie de espinas se desorganiza en el extremo distal y parte del borde anterior del órgano; en la cara externa del *appendix masculina* hay una zona provista de espinas largas y desnudas; mientras que en la cara interna sólo se encuentra la *appendix interna*, que es una pequeña prolongación del contorno oblongo con nueve a doce *uncinuli* en la región internodistal.

Los urópodos rebasan ampliamente la longitud del telson; el protopodito es muy agudo en su porción terminal.

El exopodio es un poco mas ancho que el endopodio; en la línea articular distal se cuentan hasta diez y nueve espinas, de las cuales la extrema externa es muy larga; en el borde externo de la sección anterior hay una serie submarginal de cerdas, la cual termina insensiblemente antes del ángulo posteroexterno y al mismo tiempo se acerca cada vez mas al borde externo de esta pieza.

El endopodio es agudo en su porción terminal y las cerdas del borde externo se inician desde la articulación proximal.

1.4.2 Morfometrías de la hembra

Los ejemplares femeninos presentan un tamaño muy regular, aproximadamente de 18 Mm. por termino medio; casi el 90% con huevecillos en distintos estados de desarrollo, aún ejemplares pequeños con una longitud de caparazón de 4.498 Mm.

El rostro es agudo, recto, llegando a alcanzar con su ápice, el borde articular distal del segundo artejo del pedúnculo antenular; la quilla ventral muestra de uno a cuatro dientes, pero el término medio se mantiene entre 2 y 3.

La espina antenal del caparazón es aguda, con el vértice muy ligeramente dirigido hacia arriba y hacia afuera; la espina pterigostomiana es presente, mas aguda que la antenal.

La altura posterior del caparazón es más o menos la mitad de la longitud total de aquél; pero la altura anterior es menos de un tercio la distancia total del caparazón.

La longitud del tergum de la sexta somita abdominal es ligeramente mayor que la del quinto y el ángulo pleural posterior de este último segmento es recto y el vértice agudo.

El telson es ligeramente angosto en el extremo distal, por lo que sus bordes laterales son tenuemente convergentes en esa dirección; la anchura anterior es casi el doble de la posterior; el contorno distal es redondeado, con una apófisis aplanada sobre la línea media longitudinal que sobresale francamente del borde; submarginalmente muestra de 20 a 25 cerdas delgadas, desnudas y dirigidas hacia atrás; por debajo del borde se articulan 11 cerdas plumosas, sin que se puedan distinguir entre ellas una mediana, porque la apófisis mencionada, situada en la línea media, separa las cerdas en dos grupos: el del lado derecho con cinco cerdas e izquierdo con seis.

Sobre la superficie dorsal del telson hay dos filas de apófisis espinosas, dispuestas en forma paralela en la parte anterior y divergentes en la posterior, cada una formada por seis; pero en los ángulos laterodistales del telson hay una espina que indudablemente pertenece a la misma estirpe que las anteriores, pero que ha alcanzado un gran desarrollo: entre las espinas de uno y otro ángulo quedan enmarcadas las cerdas plumosas del borde posterior del telson; su longitud es apenas un tercio de la de las cerdas plumosas.

La anténula coincide en su extremo distal con la espina del escafocerito o escama antenal; el estilocerito presenta una forma de contorno muy semejante a la del macho, lo mismo en lo que se refiere a la forma y disposición de las cerdas del borde externo.
La escama antenal, cuyo índice es de 25, es más ancha que la del macho, su borde interno está orlado de cerdas y es recto hasta el ápice; la espina de la escama es muy aguda; la longitud del escafocerito rebasa francamente el extremo distal del pedúnculo antenular.

El artejo distal el tercer maxilípedo es recto, pero proporcionalmente de mayor longitud que en las otras especies. No existen diferencias en el macho en lo que se refiere a las estructuras branquiales del primero y segundo maxilípedos.

El primer par de pereiópodos presenta un epipodito, una mastigobranquia y le corresponde una pleurobranquia; el carpopodio es proporcionalmente más corto que el del macho. El propodio es una y media veces mayor que el carpo; la región palmar es mucho mas corta que la del pereiópodo del segundo par; el pincel de cerdas del dactilopodio es casi de la misma longitud que este artejo.

El segundo par de pereiópodos presenta un epipodito, una mastigobranquia y también le corresponde una pleurobranquia. El carpopodio es más corto y menos esbelto que el del macho y la quela tiene un aspecto más grueso, siendo el pincel de cerdas de menor longitud que ella.

El pereiópodo del tercer par presenta un epipodito, una mastigobranquia y le corresponde una pleurobranquia. El meropodio es el artejo de mayor longitud del apéndice, presenta cuatro espinas en el borde inferior y en la mitad distal del artejo, las dos espinas medias están más juntas entre si.

El carpopodio es relativamente mas corto que en el macho y a la vez más esbelto; su longitud es exactamente la mitad de la del meropodio; en el tercio proximal del borde inferior no existe el carácter que distingue a los machos, es decir, no muestra las dos espinas ni la zona tuberculada; en cambio hay dos procesos espinoformes,

el proximal mas desarrollado que el distal; muy cerca del borde superior, se destacan dos pequeñas espinas en el lado externo, dispuestas a cierta distancia una de otra e insertas precisamente en la mitad proximal del artejo; en el lado interno hay tres espinas, las dos primeras ligeramente posteriores a las relativas del lado externo; la última, de inserción subdistal; en el ángulo inferodistal se destaca otra espina y por último, en la superficie externa y cerca del borde articular distal hay un gran proceso espinoforme, el mas conspicuo del artejo, que rebasa con su ápice el borde articular.

El propodio es una y media veces mayor que el carpopodio, es regular en grosor, con dos filas de espinas en el borde inferior y otras espinas en las superficies externa e interna; el borde superior sólo con algunas cerdas.

El dactilopodio es alargado y angosto; la uña terminal sólo ligeramente recurvada y los procesos espinoformes del borde inferior, en número de seis, casi rectos, de tamaño decreciente hacia el extremo proximal; la longitud de este artejo es igual a 0.20 de la longitud del artejo inmediato anterior.

El pereiópodo del cuarto par carece de epipodito, pero le corresponde una mastigobranquia y una pleurobranquia, salvo que los artejos son mas cortos, la disposición de las espinas es semejante a la ya descrita para el apéndice anterior.

El pereiópodo del quinto par no tiene mastigobranquia ni epipodito, pero su somita correspondiente conserva la pleurobranquia. Es importante hacer notar que las hembras muy jóvenes aun no desarrollan dicha pleurobranquia.

En este par de apéndices el meropodio es proporcionalmente mas corto que en los dos anteriores y su longitud es 1.5 veces mayor que la del carpopodio; presenta dos espinas en el borde inferior, la primera de ellas pequeña e inserta casi en la mitad de la longitud del artejo; la segunda al final del segundo tercio; además hay otro proceso espinoforme muy largo y agudo dispuesto submarginalmente en la superficie externa del artejo y aproximadamente entre el quinto y el último sexto de la longitud.

En la cara externa del carpopodio se observan cuatro espinas, tres de ellas pequeñas insertas como sigue: una submarginal con relación al borde inferior y a la vez subproximal, otra cerca del ángulo inferodistal y la tercera cerca del borde articular distal en el lado externo; por último, la cuarta espina es un proceso muy desarrollado semejante en forma y disposición a los ya citados para el carpopodio de los pereiópodos tercero y cuarto.

El propodio es el doble en longitud que el carpopodio, con ocho espinas en el borde inferior, más o menos equidistantemente repartidas, la distal muy larga; sobre la línea media de la cara externa y en su mitad proximal, hay otras cinco espinas, y en la cara interna ocho espinas mas que no se arreglan en una serie lineal.

El dactilopodio es la quinta parte de la longitud del propodio; tiene forma más o menos oblonga con la serie pectinada de espinas en el borde inferior, las cuales insensiblemente van apareciendo más quitinizadas y anchas hacia el extremo distal.

El pleópodo del primer par presenta un exopodito relativamente corto, de forma oblonga y orlado de cerdas. El endopodito es mas largo, es angosto en su base, ancho en el segundo tercio y después angosto y acintado; el borde externo está parcialmente orlado de cerdas, pues se disponen regularmente en sólo los dos tercios proximales, el resto es desnudo; el borde interno tiene un grupo de cerdas revertido hacia la base e inmediatamente otro con disposición normal; finalmente el endopodio remata en su porción apical con tres o cuatro cerdas.

1.5 Morfometrías de uso común en camarones

Lo más común entre una variedad de medidas del cuerpo en camarones son longitud del caparazón, longitud total y peso húmedo (Primavera *et al.*, 1998).

Acuaculturistas y practicantes de la industria cuyos beneficios dependen de la biomasa comúnmente registran el peso corporal; considerando que, los taxonomistas, ecólogos y otros profesionales del sector de la investigación prefieren

medidas de longitud las cuales son mas fácilmente medidas en el campo y no están sujetas a variaciones amplias (Primavera *et al.*, 1998).

Las relaciones que permiten intercoversiones entre los varios parámetros de peso y longitud son necesarias, por ejemplo para comparar los parámetros de desarrollo (que especie tiene tasa de crecimiento mas rápida) Dall *et al.*, (1990), especialmente para especies comercialmente importantes (Primavera *et al.*, 1998).

Los rasgos morfométricos utilizados para describir el desarrollo generalmente incluyen longitud del caparazón, longitud total y peso húmedo. Se ha argumentado que una definición exacta de la longitud del caparazón es el indicador mas preciso del tamaño del cuerpo. (Dall et al., 1990).

Aunque el peso del camarón es comúnmente registrado para propósitos de cultivo y administración (por ejemplo, estimaciones de la tasa de crecimiento, cociente de conversión alimentaría, peso de la cosecha y productividad), la aplicación de las relaciones morfométricas podría ser una simple alternativa para estimar el peso corporal de las medidas de longitud que son menos variables y mas fácilmente medidas en el campo. (Cheng and Chen, 1990; Primavera *et al.*, 1998).

Relaciones morfométricas de longitud y peso han sido determinadas principalmente en adultos de varias especies, y esto ha resultado en extrapolaciones erróneas para individuos juveniles. (Dall *et al.*, 1990).

Primavera *et al*. (1998) encontraron que la etapa de la vida (edad), fue la diferencia mas impresionante para la relación longitud-longitud y longitud-peso en especies cultivadas.

Existe una necesidad de investigar la relación longitud-peso para un rango de tamaño más grande, incluyendo camadas más grandes y crías utilizadas en criaderos; y para comparar relaciones entre tamaños de grupos, entre sexos y fuentes (salvaje y cultivo), para determinar como las relaciones cambian con el tamaño del camarón o edad (Peixoto *et al.*, 2004).

2. Materiales y métodos

2.1 Colecta

Se realizaron cinco colectas durante 2003 y 2004, en el Río Máquinas que se encuentra en la Estación de Biología Tropical de Los Tuxtlas en el Estado de Veracruz, México. Se colectaron un total de 191 organismos en tres localidades.

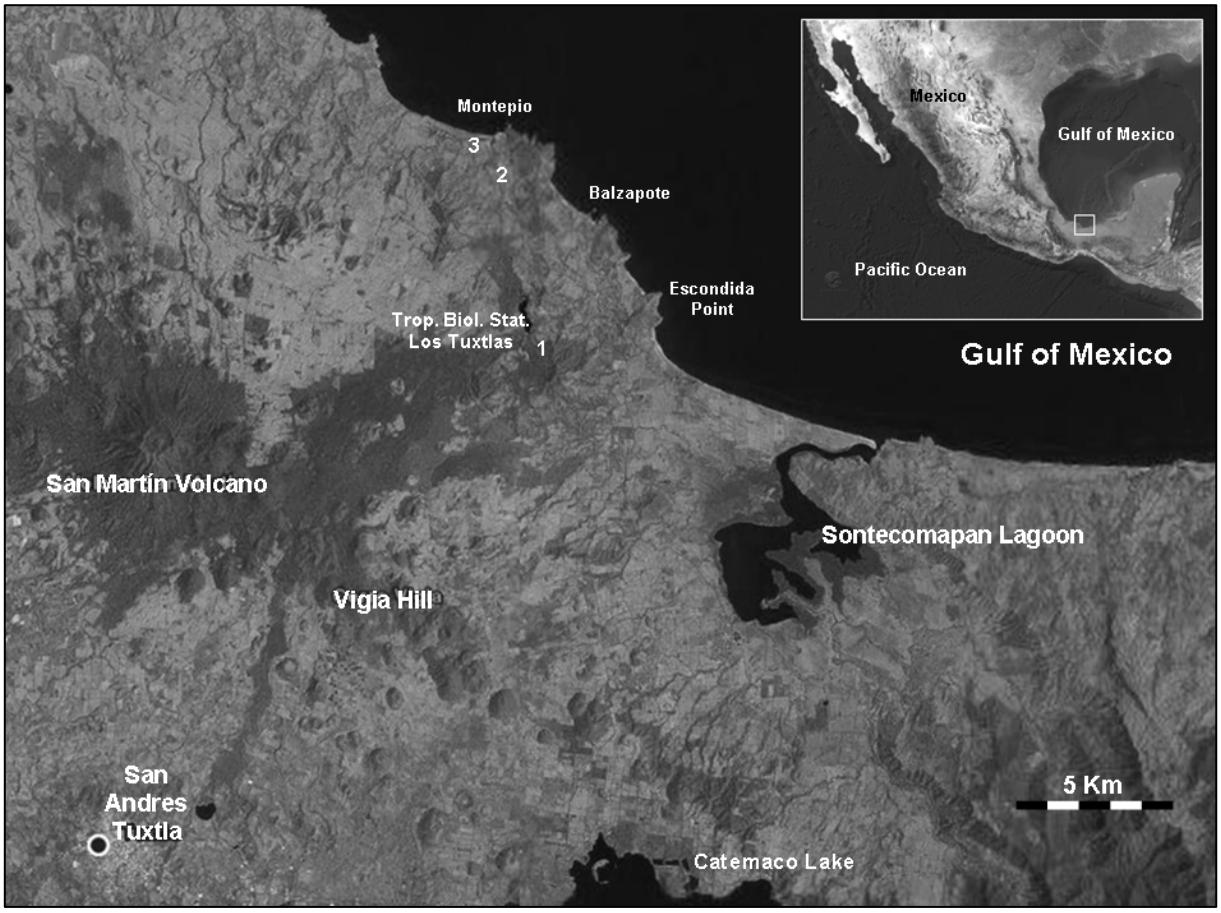


Figura 2. Ubicación de las Estaciones de muestreo. 1. - N 18° 35´ 01.2´´ - W 95° 04´ 41.5´´. 99 m. s. n. m.; 2. - N 18° 37´ 16.1´´ - W 95° 05´ 26.6´´. 22 m. s. n. m. y 3. - N 18° 38´ 32.8´´ - W 95° 05´ 49.7´´. 0 m. s. n. m.

2.2 Área de estudio

La Estación de Biología Tropical de Los Tuxtlas de la Universidad Nacional Autónoma de México (UNAM) se encuentra en la zona próxima al volcán San Martín,

18 Km. al norte de Catemaco entre los 95˚ 04´ y 95˚ 09´ de longitud Oeste y los 18˚ 34´ y 18˚ 36´ de latitud Norte, a una altitud media sobre el nivel del mar entre los 150 y los 750 m. s. n. m. (Figura 2).

2.3 Obtención de los datos morfométricos

Para la totalidad de los organismos se tomaron ocho medidas de diferentes estructuras del cuerpo las cuales son: largo del caparazón (LC); diámetro del ojo (ED); largo del telson (LT), largo del segundo segmento del primer (L1ERPLEO) y segundo (L2DOPLEO) pleópodos, largo del mero (LME) y carpo (LCA) del tercer pereiópodo; longitud total (LTO); y finalmente, en machos, el largo (AML) y ancho (AMW) del *appendix masculina*, localizado en el segundo pleópodo. (Figura 3).

2.4 Base de datos

A continuación se muestran las medidas morfométricas de tres individuos.

	LC	ED	LT	L1ERPLEO	L2DOPLEO	LME	LCA	LTO
Organismo 1	2.744	0.642	1.729	0.767	1.003	1.864	0.718	12.64
Organismo 2	3.341	0.751	2.132	1.023	1.442	2.349	0.956	16.08
Organismo 3	4.150	0.795	2.435	1.195	1.653	2.573	1.029	19.07

Tabla 2.1 Las medidas morfométricas están dadas en mm.

2.5. Árboles de clasificación

2.5.1 Clasificadores como particiones

Los días en el valle de México son clasificados acorde a los niveles de ozono:

Clase 1: Sin alerta (Ozono Bajo).

Clase 2: Primer estado de alerta (Ozono Moderado).

Clase 3: Segundo estado de alerta (Ozono Alto).

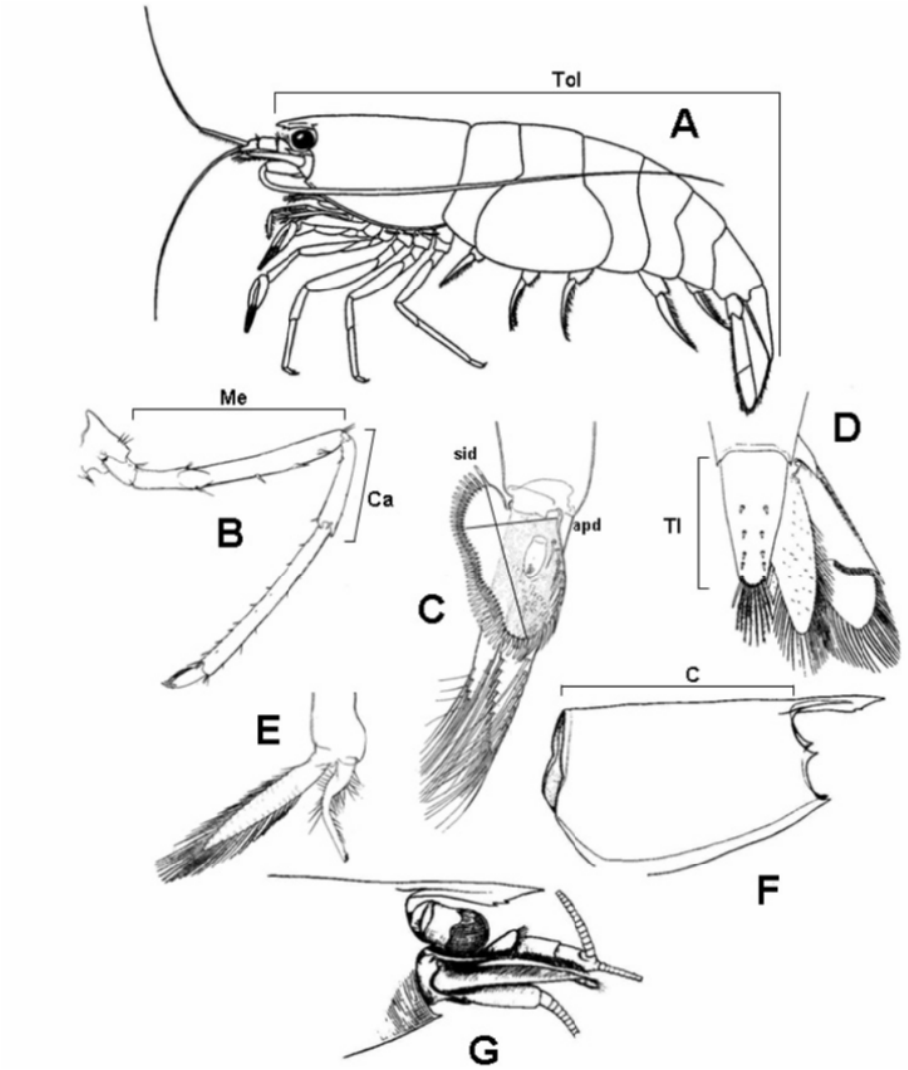


Figura 3. Estructuras anatómicas de *Potimirim* sp. A, *Potimirim glabra* (Anatomía lateral y largo total ToL); B, Quinto pereiópodo; C, segundo pleópodo con *appendix masculina*, sid, diámetro supra-inferior; apd, diámetro antero-posterior; D, telsón y urópodos; E, Primer pleópodo; F, Vista lateral del caparazón; G, Vista lateral del rostro. A tomado de Fryer, 1977; B-G tomados de Villalobos, 1959.

Durante un día, las medidas son hechas sobre varias variables meteorológicas, tales como temperatura, humedad, condiciones atmosféricas superiores, y sobre los niveles actuales de un número de contaminantes que viajan por el aire. El propósito es predecir la clasificación del siguiente día.

En problemas de este tipo, el objetivo es el mismo: dado un conjunto de medidas de un caso u objeto, se quiere encontrar un camino sistemático para predecir en que clase está el objeto.

Para dar una formulación más precisa, se toman las medidas $x_1, x_2, ..., x_n$; donde, por ejemplo, x_1 es la temperatura, x_2 es la humedad, se definen las medidas $(x_1, x_2, ..., x_n)$ hechas sobre un caso, como el *vector de medidas* $\underline{x}$ correspondiente al caso. Tomando el *espacio de medida* X a ser definido como conteniendo todos los posibles vectores de medida.

Por ejemplo, en el estudio del ozono en el Valle de México, X es un espacio de 400 dimensiones tal que la primera coordenada x_1 (temperatura) tiene rango sobre todos los valores enteros de -50 hasta 50 grados Celsius, la segunda coordenada, humedad, podría estar definida en un rango continuo desde 0 hasta 100.

Puede haber un número de diferentes definiciones de X. Lo importante es que cualquiera tenga la propiedad que el vector de medidas $\underline{x}$ correspondiente a cualquier caso que se desee clasificar, pertenezca a X.

Supóngase que los casos u objetos caen (o pertenecen) dentro de J clases. Numerandolas $1, 2, ..., J$ y sea C el conjunto de las clases; esto es, $C = \{1, 2, ..., J\}$.

Una manera sistemática de predecir la pertenencia a una clase es una regla que asigna una clase en C a cada vector de medidas $\underline{x}$ en X. Esto es, dado cualquier $\underline{x} \in X$ la regla asigna una de las clases $\{1, 2, ..., J\}$ a $\underline{x}$.

DEFINICIÓN 2.1. Un clasificador o regla de clasificación es una función $d(\,)$ definida sobre X tal que para cada $\underline{x}$, $d(\underline{x})$ es igual a uno de los números $1, 2, ..., J$.

Otra manera de ver a un clasificador es definir A_j como el subconjunto de X en el cual $d(\underline{x}) = j$; esto es,

$$A_j = \{\underline{x} : d(\underline{x}) = j\}$$

Los conjuntos $A_1, A_2, ..., A_J$ son disjuntos y $X = \bigcup_j A_j$. Así, el A_j forma una partición de X.

DEFINICIÓN 2.2. Un clasificador es una partición de X en J subconjuntos disjuntos $A_1, A_2, ..., A_J$, $X = \bigcup_j A_j$ tal que para cada $\underline{x} \in A_j$ la clase asignada es j.

2.5.2 Uso de datos en la construcción de clasificadores

Los clasificadores no son construidos caprichosamente, están basados en experiencia pasada.

En la construcción sistemática de un clasificador, la experiencia pasada está resumida por un *aprendizaje de la muestra.* Éste consiste de los datos medidos en N observaciones en el pasado junto con su actual clasificación.

DEFINICIÓN 2.3. Una muestra de aprendizaje consiste en un conjunto de datos $(\underline{x}_1, j_1), (\underline{x}_2, j_2), ..., (\underline{x}_N, j_N)$ sobre N casos, donde $\underline{x}_n \in X$ y $j_n \in \{1, 2, ..., J\}$, $n = 1, 2, ..., N.$ El aprendizaje de la muestra está denotado por $L;$ es decir,

$$L = \{(\underline{x}_1, j_1), (\underline{x}_2, j_2), ..., (\underline{x}_N, j_N)\}$$

Se distinguen dos tipos generales de variables que pueden aparecer en el vector de medida.

DEFINICIÓN 2.4. Una variable es llamada ordenada o numérica si sus valores medidos son números reales. Una variable es nominal si toma valores en un conjunto finito no teniendo algún orden natural.

DEFINICIÓN 2.5. Si todos los vectores de medida $\underline{x}_n$ tienen la misma dimensión, se dice que los datos tienen estructura estándar.

En el proyecto del ozono en el Valle de México, un conjunto fijo de variables es medido en cada caso (o día); los datos tienen estructura estándar.

2.5.3 El propósito del análisis de clasificación

Dependiendo del problema, el propósito básico de un estudio de clasificación puede ser producir un clasificador exacto (preciso) o descubrir la estructura predictiva del problema. Si se está interesa en lo último, entonces se intentará conseguir una comprensión de qué variables o interacciones entre ellas dominan el fenómeno, esto es, dar una simple caracterización de las condiciones (en términos de las variables medibles $\underline{x} \in X$) que determinen cuando un objeto está en una clase y no en otra.

Estos dos objetivos no son excluyentes. Más a menudo, el fin serán ambos, precisión y comprensión. Algunas veces uno u otro tendrá más énfasis.

Un importante criterio para un buen procedimiento de clasificación es no sólo producir clasificadores exactos (dentro de los límites de los datos) sino proveer señales y comprensión dentro de la estructura predictiva de las observaciones.

Muchas de las técnicas estadísticas disponibles actuales fueron diseñadas para pequeños conjuntos de datos teniendo estructura estándar con todas las variables de un mismo tipo; la suposición subyacente fue que el fenómeno era homogéneo, esto es, que la misma relación entre variables se cumple (se tiene) sobre todo el espacio de medida. Esto trae como consecuencia modelos donde sólo algunos parámetros fueron necesarios para hallar los efectos de los factores envueltos.

Con conjuntos de datos grandes conteniendo muchas variables, más estructuras serán distinguidas y una variedad de diversas aproximaciones pueden intentarse.

Lo que hace a un conjunto de datos interesante no es sólo su tamaño, sino su complejidad, donde complejidad puede incluir consideraciones tales como:

- alta dimensionalidad,
- una mezcla de tipos de datos,
- estructura de datos no estándar.

Tal vez el mayor reto sea la homogeneidad; esto es, las diversas relaciones que se cumplen entre variables en diferentes partes del espacio de medida.

Junto con conjuntos de datos complejos viene “la maldición de la dimensionalidad”. La dificultad es que a más alta dimensionalidad, los puntos están más dispersos y separados.

A menos que se haga la suposición fuerte de que las variables son independientes, el número de parámetros usualmente necesarios para especificar una distribución M dimensional crece más rápido que su cardinalidad. Para poner esto de otra manera, *la complejidad de un conjunto de datos se incrementa rápidamente con el aumento de dimensionalidad.*

En respuesta a la dimensionalidad cada vez mayor en los conjuntos de datos, los procedimientos multivariados, contienen una cierta clase de proceso de reducción de la misma. La elección de variables de manera gradual (stepwise), la selección de un subconjunto de variables en regresión y el análisis de discriminante son ejemplos para tal fin.

2.5.4 Estimación de la precisión

Dado un clasificador, esto es, dada una función $d(\underline{x})$ definida sobre X tomando valores en $C,$ se denota por $R^*(d)$ su verdadera tasa de clasificación errónea. La pregunta en esta sección es: ¿cuál es la verdadera tasa de clasificación errónea? y ¿cómo puede ser estimada?

Una manera de ver cuan preciso es un clasificador es probarlo sobre subsecuentes casos cuya correcta clasificación haya sido observada anteriormente.

- El valor de $R^*(d)$ puede ser determinado en esta forma: usando $L,$ construir $d,$ después, extraer otro conjunto de casos muy grande de la misma población donde L fue tomado.

- Verificar que se dio una correcta clasificación de cada uno de esos casos, e incluso encuéntrese la clasificación predicha usando $d(\underline{x})$. La proporción de clasificación errónea por d es el valor de $R^*(d)$.

Para hacer preciso el concepto precedente, es necesario un modelo de probabilidad.

Se define el espacio $X\ C$ como el conjunto de todas las parejas $(\underline{x}, j)$ donde $\underline{x} \in X$ y j es la clase etiquetada, $j \in C$. Sea $P(A, j)$ una medida de probabilidad sobre $X \times C, \quad A \subset X, \quad j \in C$. La interpretación de $P(A, j)$ es que un caso extraído aleatoriamente de la población relevante tiene probabilidad $P(A, j)$, y su vector de medida $\underline{x}$ está en A y su clase es j. Asumiendo que el aprendizaje de la muestra L consiste de N casos $(\underline{x}_1, j_1),(\underline{x}_2, j_2),...,(\underline{x}_N, j_N)$ extraídos independientemente al azar de la distribución $P(A, j)$. Se construye $d(\underline{x})$ usando L, entonces definiendo a $R^*(d)$ como la probabilidad de que d clasificará erróneamente una nueva muestra extraída de la misma distribución que L.

DEFINICIÓN 2.6. Sea $(\underline{x}, Y), \ \underline{x} \in X, \ Y \in C,$ una nueva muestra de la distribución de probabilidad $P(A, j),$ es decir,

- $P(\underline{x} \in A, Y = j) = P(A, j),$
- $(\underline{x}, Y)$ es independiente de L.

Entonces se define $R^*(d) = P(d(\underline{x}) \neq Y)$.

En la determinación de la probabilidad $P(d(\underline{x}) \neq Y),$ el conjunto L es considerado fijo.

Una notación más precisa es $P(d(\underline{x}) \neq Y \mid L),$ la probabilidad de clasificar erróneamente la nueva muestra dada la muestra de aprendizaje L.

Este modelo debería ser aplicado cautelosamente ya que puede haber dependencia entre los datos.

En problemas actuales, sólo los datos en L están disponibles con baja perspectiva de conseguir una muestra grande adicional de casos por clasificar, entonces L debe ser utilizada para construir $d(\underline{x})$ y para estimar $R*(d)$. Así $R^*(d)$ es llamado e*stimador interno*.

Tres tipos de estimadores internos serán de interés. El primero, menos preciso, y más comúnmente usado es el *estimador de resustitución*.

Después de que el clasificador d es construido los casos en L son probados a través del mismo. La proporción de observaciones clasificadas erróneamente es el estimador por resustitución. Para poner esto en forma de ecuación:

DEFINICIÓN 2.7. Se define la función indicadora $I(.)$ como 1 si la afirmación dentro del paréntesis es cierta, 0 en otro caso.

Entonces el estimador por resustitución es:

$$R(d) = \frac{1}{N}\sum_{1}^{N} I\left(d(\underline{x}_n), j_n\right) \tag{2.1}$$

El problema con el estimador por resustitución es que se calcula usando los mismos datos para construir d, en lugar de usar otra muestra independiente. Todos los procedimientos de clasificación, directos o indirectos, intentan minimizar $R(d)$. Usando el valor de $R(d)$ como un estimador de $R*(d)$ puede dar un cuadro excesivamente optimista de la precisión de d.

La segunda propuesta es la *prueba de estimación de la muestra.* Aquí las observaciones en L son divididas en dos conjuntos L_1 y L_2. Sólo los casos en L_1 son utilizados para construir d y los datos en L_2 son utilizados para estimar $R*(d)$. Si N_2 es el número de elementos en L_2, entonces la prueba de estimación de la muestra, $R^{ts}(d)$ está dada por:

$$R^{ts}(d) = \frac{1}{N} \sum_{(\underline{x}_n, j_n) \in L_2} I\left(d(\underline{x}_n) \neq j\right) \tag{2.2}$$

En este método, se necesita tener cuidado para poder considerar los elementos en L_2 como independientes de los casos en L_1 y extraídos de la misma distribución.

El procedimiento más comúnmente usado para asegurarse esas propiedades es extraer L_2 aleatoriamente de L. Frecuentemente L_2 es tomado como 1/3 de los casos de L, pero no se sabe de una justificación teórica para esta división (Breiman Leo *et all*, 1984).

Esta prueba tiene la desventaja que reduce efectivamente el tamaño de la muestra. En una división tal, sólo 2/3 de los datos son utilizados para construir d, y 1/3 para estimar $R^*(d)$. Si el tamaño de muestra es grande por ejemplo 5000 elementos, esto es una dificultad menor, y el estimador es equilibrado y eficiente.

Para muestras pequeñas, otra propuesta, llamada V *doble validación cruzada*, es preferida. Los casos en L son divididos aleatoriamente en V subconjuntos de igual tamaño tanto como sea posible. Se denotan esos subconjuntos por $L_1, L_2, ..., L_V$. Se asume que el procedimiento para construir el clasificador puede ser aplicado a cualquier subconjunto. Para cada subconjunto se aplica el procedimiento utilizado para una muestra de aprendizaje L. Entonces $d_V(\underline{x})$ será el clasificador resultante. Como ninguno de los casos en L_V ha sido utilizado en la construcción de d_V, una prueba de estimación de la muestra para $R^*(d_V)$ es:

$$R^{ts}(d_V) = \frac{1}{N_V} \sum_{(\underline{x}_n, j_n) \in L_V} I\left(d(\underline{x}) \neq j_n\right) \tag{2.3}$$

Donde $N_V = \frac{N}{V}$ es el número de casos en L_V.

Ahora utilizando el mismo procedimiento, se construye el clasificador d usando la totalidad de L.

Para V grande, cada uno de los clasificadores v es construido usando una muestra de aprendizaje de tamaño $N\left(1-\frac{1}{V}\right)$ casi tan grande como L. La suposición básica de la validación cruzada es que el procedimiento tiende a estabilizar los cálculos. Esto es, que los clasificadores $d_v,\ v=1,2,...,V,$ construidos cada uno utilizando casi la totalidad de $L,$ tiene tasas de error de clasificación $R^*(d_V)$ cercanamente igual a $R^*(d)$. Guiándose por esta heurística, se define el estimador de la V doble validación cruzada como:

$$R^{cv}(d)=\frac{1}{V}\sum_{v=1}^{V}R^{ts}\left(d^{(v)}\right) \tag{2.4}$$

Un caso particular del planteamiento anterior sería la propuesta $N-$doble validación cruzada en donde el estimador se construye dejando una observación afuera. Para cada $n,\ n=1,2,...,N,$ el $n-ésimo$ caso se pone a un lado y el clasificador es construido utilizando los otros $N-1$ casos. Entonces el $n-ésimo$ caso es usado como muestra de prueba de un sólo caso y $R^*(d)$ es estimado por la fórmula anterior.

La validación cruzada es modesta con los datos. Cada observación en L es utilizada para construir $d,$ y cada caso es usado exactamente una vez en una prueba de la muestra.

2.5.5 La regla de Bayes y procedimientos actuales de clasificación

La alternativa que ha sido más utilizada en la construcción de clasificadores es la del concepto de la regla de Bayes. Si los datos son extraídos de una distribución de probabilidad $P(A,j),$ entonces la forma de este criterio más precisa puede ser dada en términos de $P(A,j).$ Este criterio es llamado la regla de Bayes y es denotado por $d_B(\underline{x}).$

Con mayor precisión, supóngase que $(\mathbf{X},Y)$ con $\mathbf{X}\in X,$ y $Y\in C,$ es un muestreo aleatorio de una distribución de probabilidad $P(A,j)$ en $C;$ i.e., $P(\mathbf{X}\in A,Y=j)=P(A,j).$

DEFINICIÓN 2.8. $d_B(\underline{x})$ es una regla de Bayes si para cualquier otro clasificador $d_B(\underline{x}),\ P(d_B(\underline{x}) \neq Y) \leq P(d(\underline{x}) \neq Y).$ Entonces la tasa de clasificación errónea es

$$R_B = P(d_B(\underline{x}) \neq Y) \leq P(d(\underline{x}) \neq Y)$$

Para ilustrar como $d_B(\underline{x})$ puede ser obtenida de $P(A, j),$ se da su forma en un caso especialmente importante.

DEFINICIÓN 2.9. Se define la clase de probabilidades *a priori* $\pi(j),\ \ j = 1, 2, ..., J,$ como:

$$\pi(j) = P(Y = j)$$

y la distribución de probabilidad de la $j - ésima$ clase del vector de medidas por:

$$P(A \mid j) = \frac{P(A, j)}{\pi(j)}$$

SUPOSICIÓN 2.1. X es un espacio euclidiano de dimensión M y para cada j, $j = 1, 2, ..., J,\ \ P(A \mid j)$ tiene la densidad de probabilidad $f_j(\underline{x});$ i.e., para conjuntos $A \subset X,$

$$P(A \mid j) = \int_A f_j(\underline{x}) d\underline{x}$$

entonces,

TEOREMA 2.1. Bajo la suposición 2.1 la regla de Bayes está definida por:

$$d_B(\underline{x}) = j \text{ sobre } A_j = \{\underline{x} : f_j(\underline{x})\pi(j) \max_i f_i(\underline{x})\pi(i)\} \qquad (2.5)$$

y la tasa de clasificación errónea de Bayes es:

$$R_B = 1 - \int \max_j [f_j(\underline{x})\pi(j)] d\underline{x} \qquad (2.6)$$

Aunque d_B es llamada la regla de Bayes, es incluso reconocible como un criterio de máxima verosimilitud: Clasificar $\underline{x}$ con esa j para la cual $f_j(\underline{x})\pi(j)$ es máxima. Como una característica menor, obsérvese que en (2.5) no se define únicamente el $d_B(\underline{x})$ en puntos $\underline{x}$ tales que $\max_j f_j(\underline{x})\pi(j)$ es alcanzado por dos o más diferentes j. En esta situación, se define $d_B(\underline{x})$ arbitrariamente como cualquiera de las j que maximiza.

La prueba del Teorema (2.1) es inmediata. Para cualquier clasificador d, bajo la suposición (2.1),

$$\begin{aligned} P(d(\mathsf{X}) = Y) &= \sum_{j=1}^{J} P(d(\mathsf{X}) = j \mid Y = j)\pi(j) \\ &= \sum_{j=1}^{J} \int_{\{d(\underline{x})=j\}} f_j(\underline{x})\pi(j)\,d\underline{x} \\ &= \int \left[\sum_{j=1}^{J} I(d(\underline{x}) = j) f_j(\underline{x})\pi(j) \right] d\underline{x} \end{aligned}$$

Para un valor fijo de $\underline{x}$

$$\sum_{j=1}^{J} I(d(\underline{x}) = j) f_j(\underline{x})\pi(j) \le \max_j \left[f_j(\underline{x})\pi(j) \right]$$

y se alcanza la igualdad si $d(\underline{x})$ es igual a j para el cual $f_j(\underline{x})\pi(j)$ es un máximo.

Por lo tanto, la regla d_B dada en el teorema tiene la propiedad que para cualquier otro clasificador d,

$$P(d(\mathsf{X}) = Y) \le P(d_B(\mathsf{X}) = Y) = \int \max_j \left[f_j(\underline{x})\pi(j) \right] d\underline{x}$$

2.5.6 Árboles de clasificación estructurados

Los árboles de clasificación estructurado o árboles de clasificación estructurados binarios, son construidos por divisiones repetidas de subconjuntos de X en dos subconjuntos descendientes, comenzando con X mismo. Este proceso está dibujado para un árbol hipotético con seis clases en la Figura 4.

En ésta Figura X_2 y X_3 son disjuntos, con $X = X_2 \cup X_3$. Similarmente, X_4 y X_5 son disjuntos con $X_4 \cup X_5 = X_2$ y $X_6 \cup X_7 = X_3$. Aquellos subconjuntos que no están divididos, en este caso $X_6, X_8, X_{10}, X_{11}, X_{12}, X_{14}, X_{15}, X_{16}, X_{17}$ son llamados subconjuntos terminales. Esto será indicado por una caja rectangular; subconjuntos no terminales son representados por círculos.

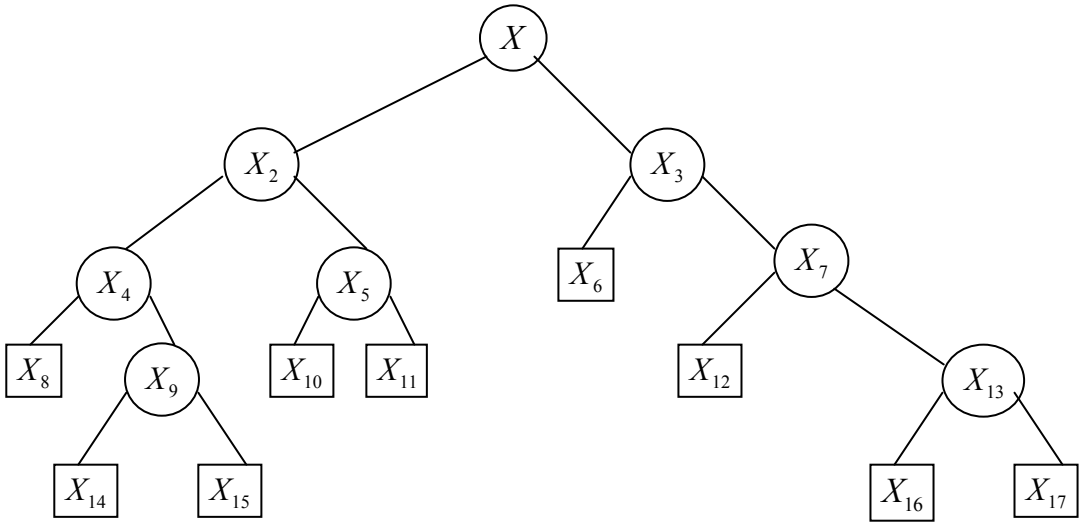


Figura 4

Los subconjuntos terminales forman una partición de X. Cada subconjunto terminal está asociado con una etiqueta de clase. Puede haber dos o más subconjuntos terminales con la misma etiqueta de clase, la partición correspondiente al clasificador es obtenida por poner juntos todos los subconjuntos terminales correspondientes a la misma clase. Así,

$$A_1 = X_{15} \qquad A_2 = X_{11} \cup X_{14}$$
$$A_3 = X_{10} \cup X_{16} \qquad A_4 = X_6 \cup X_{17}$$
$$A_5 = X_8 \qquad A_6 = X_{12}$$

Las divisiones son formadas por condiciones sobre las coordenadas de $\underline{x} = (x_1, x_2, ..., x_M)$. Por ejemplo, la división 1 de X en X_2 y X_3 podría ser de la forma

$$\begin{aligned} X_2 &= \{\underline{x} : x_4 \le 7\}, \\ X_3 &= \{\underline{x} : x_4 > 7\} \end{aligned} \tag{2.7}$$

La división de X_3 en X_6 y X_7 podría ser de la forma

$$\begin{aligned} X_6 &= \{\underline{x} \in X_3 : x_3 + x_5 \le -2\}, \\ X_7 &= \{\underline{x} \in X_3 : x_3 + x_5 > -2\} \end{aligned}$$

El clasificador de árbol predice una clase para el vector de medida $\underline{x}$ en esta forma: de la definición de la primera división, se decide si $\underline{x}$ desciende a X_2 o X_3. Por ejemplo, si (2.7) es utilizado, $\underline{x}$ se dirige a X_2 si $x_4 \le 7$, y a X_3 si $x_4 > 7$. Si $\underline{x}$ se coloca en X_3, entonces por la definición de la división de X_3 se determina si $\underline{x}$ desciende a X_6 o X_7.

Cuando $\underline{x}$ finalmente llega dentro de un conjunto terminal, su clase es proporcionada por la etiqueta asignada al conjunto terminal.

Cualquier nodo t es un subconjunto de X, el nodo raíz es definido como $t_1 = X$, subconjuntos terminales serán nodos terminales (nodos hoja) y subconjuntos no terminales serán nodos no terminales (nodos rama).

La construcción entera de un árbol entonces, se centra alrededor de tres elementos:

1. La selección de las divisiones.
2. La decisión de cuando declarar un nodo terminal o continuar dividiéndolo.
3. La asignación de cada nodo terminal a una clase.

El problema es cómo usar los datos contenidos en L para determinar las divisiones, los nodos terminales y sus asignaciones. Resulta que la asignación de clases es simple. La dificultad estriba en es encontrar buenas divisiones y en conocer cuando detener la división.

2.5.7 Construcción de un árbol de clasificación

El primer problema en la construcción de un árbol es como usar L para determinar la división binaria de X en piezas cada vez más pequeñas. La idea fundamental es seleccionar cada partición de un subconjunto de tal forma que los datos de los subconjuntos descendientes sean más "puros" que las observaciones en el subconjunto progenitor (generador).

Una vez que una buena división de un nodo es encontrada, entonces una búsqueda es hecha para desarrollar particiones buenas de los nodos descendientes.

Para finalizar el desarrollo de un árbol, una regla para terminar con el proceso de crecimiento de un árbol puede ser la siguiente, cuando un nodo es alcanzado de tal forma que no es posible un decrecimiento significante en la impureza, entonces el nodo no será dividido y se convertirá en un nodo terminal.

El árbol es utilizado como un clasificador en una forma obvia. Si un objeto, cuya clase se desconoce, se analiza de acuerdo a un árbol de clasificación y resulta que su posición final coincide con un nodo terminal, por ejemplo el $j,$ entonces dicho elemento es etiquetado como $j.$

2.5.8 Desarrollo inicial de un árbol

En la muestra de aprendizaje L para un problema con J clases, sea N_j el número de casos en la clase $j.$ A menudo las probabilidades *a priori* $\{\pi(j)\}$ son tomadas como las proporciones $\frac{N_j}{N}.$ Pero el cociente de la muestra de aprendizaje puede no reflejar las proporciones esperadas en casos futuros. De cualquier forma, el conjunto de $\{\pi(j)\}$ es estimado con los datos o bien proporcionado por el analista.

En un nodo t, sea $N(t)$ el número total de casos en L con $\underline{x}_n \in t$, y $N_j(t)$ el número de casos de la clase j en t. La proporción de los casos de la clase j en L que cae dentro de t es $\frac{N_j(t)}{N_j}$. Para un conjunto *a priori* dado, $\pi(j)$ es interpretado como la probabilidad de que un caso en la clase j estará presente en el árbol. Por lo tanto se toma:

$$p(j,t) = \pi(j)\frac{N_j(t)}{N_j} \tag{2.8}$$

Como el estimador por resustitución para la probabilidad de que un caso estará en la clase j y caiga dentro del nodo t.

El estimador por resustitución $p(t)$ de la probabilidad que cualquier caso caiga dentro del nodo t está definido por:

$$p(t) = \sum_{j=1}^{J} p(j,t) \tag{2.9}$$

El estimador por resustitución de la probabilidad de que un caso esté en la clase j dado que cayó en el nodo t está definido por:

$$p(j \mid t) = \frac{p(j,t)}{p(t)} \tag{2.10}$$

y satisface:

$$\sum_{j=1}^{J} p(j \mid t) = 1$$

Cuando $\pi(j) = \frac{N_j}{N}$, entonces $p(j \mid t) = \frac{N_j(t)}{N_j}$, así $p(j \mid t)$ es la proporción relativa de casos en la clase j en el nodo t.

Obsérvese que p minúscula denotará probabilidades estimadas y P mayúscula a probabilidades teóricas.

Los cuatro elementos necesarios en el procedimiento inicial de crecimiento del árbol son:

1. Un conjunto Q de preguntas binarias de la forma $¿\underline{x} \in A?,\ A \subset X.$
2. Un criterio de bondad de la división $\phi(s,t)$ que puede ser evaluado para cualquier división s de cualquier nodo $t.$
3. Una regla de suspensión de la división.
4. Una regla de asignación para cada nodo terminal a una clase.

2.5.8.1 El conjunto Q de preguntas

El conjunto Q de preguntas binarias genera un conjunto S de divisiones s de cada nodo $t.$

Aquellos casos en t que responden “si” se alojan en el nodo descendente izquierdo t_L con una proporción p_L y aquellos con respuesta “no” al descendiente derecho t_R con proporción p_R (Figura 5).

Entonces la bondad de la división está definida como el decrecimiento en impureza

$$i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

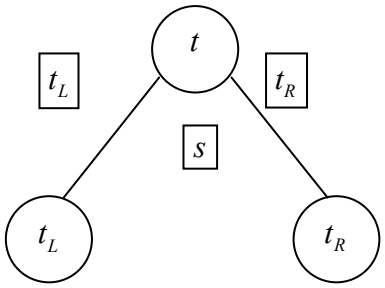


Figura 5.

Para cada nodo intermedio t, la partición seleccionada es la división s^* la cual maximiza $\phi(s,t)$. De hecho, si la pregunta es $¿\underline{x} \in A?$, entonces $t_L = t \cap A$ y $t_R = t \cap A^c$, donde A^c es el complemento de A en X.

Si los datos tienen estructura estándar, la clase Q de preguntas puede ser estandarizada. Se asume que el vector de medidas tiene la forma $\underline{x} = (x_1, x_2, ..., x_M)$, donde M es la dimensionalidad y las variables pueden ser ordinales o categóricas. El conjunto Q de preguntas estandarizadas está definido como sigue:

1. Cada división depende del valor de solo una variable.
2. Para cada variable ordenada x_i, Q incluye todas las preguntas de la forma $¿x_i \le c?$ para toda $c \in (-\infty, \infty)$.
3. Si x_i es nominal, tomando valores en $\{b_1, b_2, ...b_l\}$, entonces Q incluye todas las preguntas de la forma $¿x_i \in S?$ siendo S un subconjunto de $\{b_1, b_2, ...b_l\}$.

Las divisiones en 2 y 3 para todas las M variables constituyen el conjunto estandarizado.

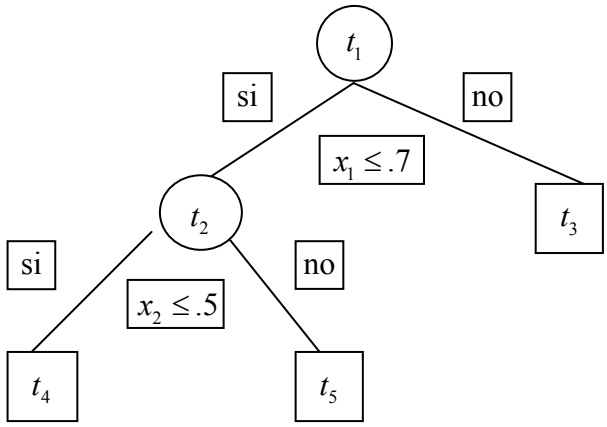
No existe un número infinito de distintas divisiones de los datos. Por ejemplo, si x_1 es ordenada, y se tiene una muestra de tamaño m, entonces el vector columna es $(x_{1,1}, x_{1,2}, ..., x_{1,m})$. Hay a lo más m-1 diferentes divisiones generadas por el conjunto de preguntas $\{x_1 \le c_j\}$ $j = 1, 2, ..., m-1$ donde c_j es tomada como el punto medio entre valores de datos consecutivos distintos de x_1.

Para cada nodo el algoritmo construcción de un árbol busca a través de las variables una por una la mejor partición, comenzando con x_1 y continuando hasta x_n. Entonces compara las divisiones de variables y selecciona la mejor.

Cuando los datos son de dimensión fija y tienen sólo variables ordenadas, otra forma de ver el procedimiento del árbol estructurado es como una partición recursiva del espacio de los datos en rectángulos.

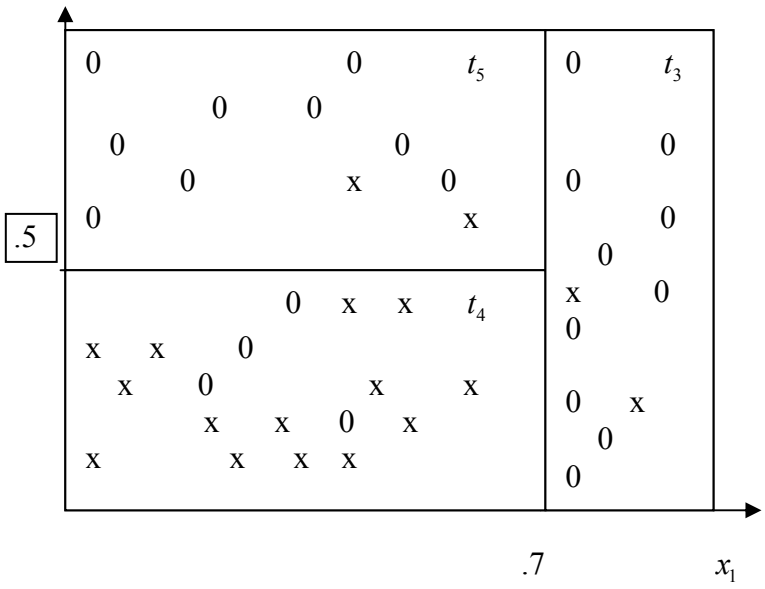
Como ejemplo se considera un problema de un árbol con dos clases utilizando datos consistiendo de dos variables ordenadas $x_1,\ x_2$ con $0 \le x_i \le 1,\ i = 1,2.$ supóngase que el diagrama de árbol se ve como el de la Figura 6. Una forma equivalente de ver este árbol es que divide el cuadrado unitario como se muestra en la Figura 7.

Para este punto de vista geométrico, el procedimiento de construcción del árbol particiona recursivamente X en rectángulos tales que la población dentro de cada rectángulo se convierte cada vez más en una clase homogénea.



$clase\ 1 = \{t_3, t_5\},\ clase\ 2 = \{t_4\}$

Figura 6



$0 = clase\ 1,\ x = clase\ 2.$

Figura 7

2.5.8.2 La regla de división y suspensión de la división

La bondad del criterio de división fue originalmente derivada de una función de impureza.

DEFINICIÓN 2.10. Una función de impureza es una función ϕ definida sobre el conjunto de todas las $J-uplas$ de números $(p_1, p_2, ..., p_J)$ satisfaciendo $p_j \geq 0$, $j = 1, 2, ..., J$, $\sum_{j=1}^{J} p_j = 1$ con las propiedades

i. ϕ es un máximo solamente en el punto $\left(\frac{1}{j}, \frac{1}{j}, ..., \frac{1}{j}\right)$,

ii. ϕ alcanza su mínimo solamente en los puntos e_i que constan de un cero en todos los lugares excepto en el $i-ésimo$ en donde hay un uno.

iii. ϕ es una función simétrica de $p_1, p_2, ..., p_j$.

DEFINICIÓN 2.11. Dada una función de impureza ϕ, se define la medida de la impureza $i(t)$ de cualquier nodo t como

$$i(t) = \phi\big(p(1|t), p(2|t), ..., p(J|t)\big)$$

Si una división s de un nodo t envía una proporción p_R de los casos en t a t_R y la proporción p_L a t_L, se define el decrecimiento en impureza como

$$i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

entonces se elige la bondad de la división $\phi(s,t)$ como $i(s,t)$.

La impureza del nodo es más grande cuando todas las clases son igualmente mezcladas juntas en el nodo, y más pequeña cuando el nodo contiene sólo una clase.

Supóngase que se han hecho algunas divisiones y llegado a una colección actual de nodos terminales. El conjunto de divisiones usado, junto con el orden con el cual fueron utilizados, determina lo que se llama un árbol binario T.

Se denota el actual conjunto de nodos terminales por $\bar{T}$; se elige $\mathrm{I}(t) = i(t)\,p(t)$, y se define la impureza del árbol por

$$\mathrm{I}(T) = \sum_{t \in \bar{T}} \mathrm{I}(t) = \sum_{t \in \bar{T}} i(t)\,p(t)$$

No es difícil de observar que seleccionando las divisiones que maximizan $i(s,t)$ es equivalente a elegir esas particiones que minimizan impureza total del árbol $\mathrm{I}(T)$. Seleccionando cualquier nodo $t \in \bar{T}$ y utilizando una bifurcación s, se divide el nodo en t_L y t_R. El nuevo árbol T' tiene impureza

$$\mathrm{I}(T') = \sum_{\bar{T}-\{t\}} \mathrm{I}(t) + \mathrm{I}(t_L) + \mathrm{I}(t_R)$$

El decrecimiento en impureza del árbol es

$$\mathrm{I}(T) - \mathrm{I}(T') = \mathrm{I}(t) - \mathrm{I}(t_L) - \mathrm{I}(t_R)$$

Esto depende sólo del nodo t y la división s. Por lo tanto, maximizando el decrecimiento en impureza del árbol por divisiones sobre t es equivalente a maximizar la expresión

$$\mathrm{I}(s,t) = \mathrm{I}(t) - \mathrm{I}(t_L) - \mathrm{I}(t_R) \qquad (2.11)$$

Se definen las proporciones $p_L,\ p_R$ de la población del nodo t que van a t_L y t_R, respectivamente, por

$$p_L = \frac{p(t_L)}{p(t)},\ \ p_R = \frac{p(t_R)}{p(t)}$$

entonces

$$p_L + p_R = 1$$

y (2.11) puede ser rescrito como

$$\begin{aligned} \mathrm{I}(s,t) &= \left[i(t) - p_L i(t_L) - p_R i(t_R) \right] p(t) \\ &= i(s,t)\, p(t) \end{aligned}$$

Debido a que $\mathrm{I}(s,t)$ difiere de $i(s,t)$ por el factor $p(t)$, la misma partición s^* maximiza ambas expresiones. Así, el procedimiento de la selección de la división puede ser imaginado como un repetido intento para minimizar la impureza del árbol total.

La regla inicial de suspensión de la división es simple. Se elige un umbral $\beta > 0$ y se declara el nodo t terminal si

$$\max_{s \in S} \mathrm{I}(s,t) < \beta \qquad (2.12)$$

2.5.8.3 La regla de asignación de clases y las estimaciones por resustitución

Se supone que un árbol T ha sido construido y tiene nodos terminales $\tilde{T}$.

DEFINICIÓN 2.12. Una regla de asignación de clase determina una clase $j \in \{1, 2, ..., J\}$ a cada nodo terminal $t \in \tilde{T}$. La clase asignada al nodo $t \in \tilde{T}$ es denotada por $j(t)$.

Para cualquier conjunto de probabilidades *a priori* y regla de asignación de clase $j(t)$,

$$\sum_{j \neq j(t)} p(j \mid t)$$

Es la estimación por resustitución de la probabilidad de clasificación errónea dado que un caso cae en el nodo t. Se toma como la regla de asignación de clase $j^*(t)$ la regla que minimiza esta estimación.

DEFINICIÓN 2.13. La regla de asignación de la clase $j^*(t)$ está dada por: si $p(j \mid t) = \max_i p(i \mid t)$, entonces $j^*(t) = j$. Si el máximo es alcanzado para dos o más clases diferentes, asignar $j^*(t)$ arbitrariamente a cualquiera de las clases con la propiedad.

Utilizando esta regla, se obtiene

DEFINICIÓN 2.14. La estimación por resustitución $r(t)$ de la probabilidad de clasificación errónea, dado que una clase cae dentro del nodo t, es

$$r(t) = 1 - \max_j p(j \mid t)$$

Se denota $R(t) = r(t)\, p(t)$.

Entonces la estimación por resustitución para la tasa total de clasificación errónea $R^*(T)$ del clasificador del árbol T es

$$R^*(T) = \sum_{t \in \tilde{T}} R(t)$$

Hasta ahora, ha sido tácitamente hecha la suposición de que el costo o pérdida en clasificar erróneamente a un objeto de la clase j como uno de la clase i fue la misma para toda $i \neq j$. En algunos problemas de clasificación esto no es una consideración realista. Por lo tanto, se introduce un conjunto de costos por clasificación errónea $C(i \mid j)$, donde

DEFINICIÓN 2.15. $C(i \mid j)$ es el costo por clasificar erróneamente un objeto de la clase j como un objeto de la clase i y satisface

i. $C(i \mid j) \geq 0, \ i \neq j$

ii. $C(i \mid j) = 0, \ i = j$

Dado un nodo t con probabilidades estimadas del nodo $p(j \mid t), \ j = 1, 2, ..., J,$ si un objeto aleatoriamente seleccionado de clase desconocida cae dentro de t y es seleccionado como clase $i,$ entonces la esperanza estimada del costo de clasificación errónea es

$$\sum_j C(i \mid j) p(j \mid t)$$

Una regla natural de asignación del nodo es seleccionar i para minimizar esta expresión. Por lo tanto,

DEFINICIÓN 2.16. Elegir $j^*(t) = i_0$ si i_0 minimiza $\sum_j C(i \mid j) p(j \mid t);$ se define la estimación por resustitución $r(t)$ de la esperanza del costo por clasificación errónea, dado el nodo $t,$ por

$$r(t) = \min_i \sum_j C(i \mid j) p(j \mid t)$$

y se define la estimación por resustitución

$$R(T) = \sum_{t \in \tilde{T}} r(t) p(t) = \sum_{t \in \tilde{T}} R(T)$$

Donde $R(t) = r(t) p(t).$

Obsérvese que en el caso del costo unitario por clasificación errónea, $C(i \mid j) = 1, \ i \neq j,$

$$\sum_j C(i \mid j) p(j \mid t) = 1 - p(i \mid t)$$

Y el criterio del costo mínimo se reduce a la regla dada en la definición (2.16).

En adelante, se elige $j^*(t)$ como la regla de asignación de clase sin preocupación adicional.

Una importante propiedad de $R(T)$ es que más de una división en cualquier camino, $R(T)$ se convierte en la más pequeña. Más precisamente, si T' es obtenida de T por dividir en cualquier camino a un nodo terminal de T, entonces

$$R(T') \leq R(T)$$

Poniéndolo en otra forma:

PROPOSICIÓN 2.1. Para cualquier división de un nodo t en t_L y t_R,

$$R(T) \geq R(t_L) + R(t_R)$$

La demostración es directa, pero se deja para más adelante (Proposición 3.4).

2.5.8.4 Combinación de variables

Una deficiencia en árboles usando la estructura estándar es que todas las divisiones son sobre una variable, es decir, todas las particiones son perpendiculares a los ejes coordenados. En situaciones donde la estructura de la clase depende de la combinación de variables, el procedimiento estándar del árbol descubrirá pobremente la estructura. Por ejemplo, considere el problema de dos clases con dos variables ilustrado en la Figura 8.

El análisis de discriminante funcionaría mucho mejor sobre este conjunto de datos. El procedimiento estándar del árbol dividiría muchas veces en un intento por aproximar el hiperplano separado por rectángulos. Sería difícil observar las estructuras lineales de los datos por examinar los resultados del árbol.

En problemas donde se sospeche una estructura lineal, el conjunto de divisiones permitidas está extendido a todas las combinaciones lineales de la forma $\sum_{1}^{n} a_i x_i \leq c.$

Un algoritmo está desarrollado para buscar a través de tales divisiones en un esfuerzo por encontrar la que maximiza la bondad del criterio de división.

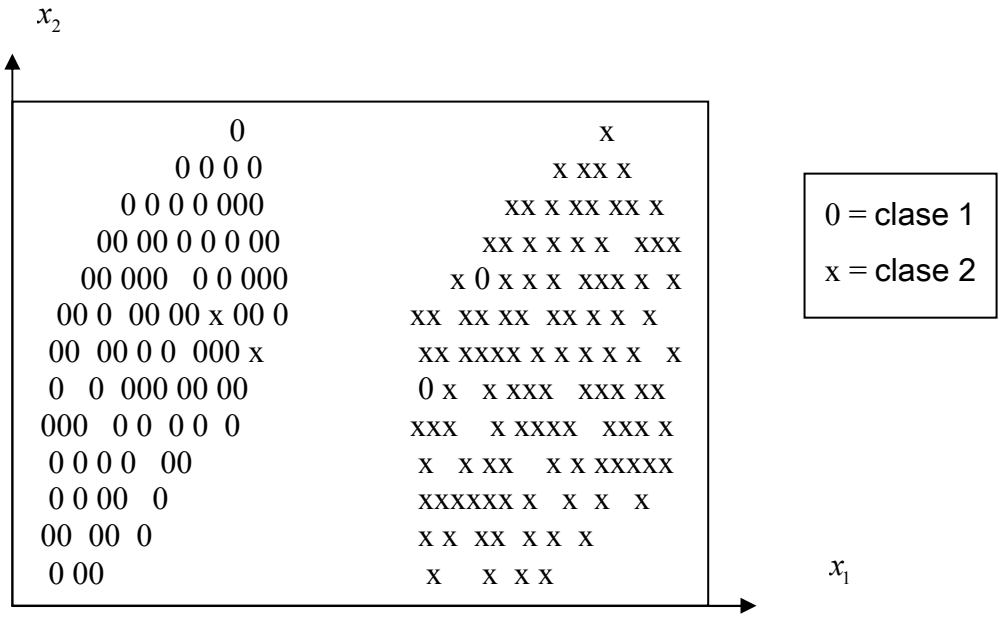


Figura 8

2.5.8.5 Las ventajas de la exactitud de la estructura de árbol

Como ha sido delineado en las secciones previas, el árbol de clasificación estructurado es un procedimiento recursivo e iterativo que requiere las especificaciones de sólo algunos elementos tales como:

1. El conjunto Q de preguntas.
2. Una regla para la selección de la mejor división para cualquier nodo.
3. Un criterio para elegir el árbol del tamaño correcto.

Tiene el potencial para ser una herramienta de clasificación poderosa y flexible. En particular:

1. Puede ser aplicado a cualquier estructura de datos a través de la apropiada formulación del conjunto de preguntas Q. En observaciones con estructura estándar se manejan variables ordenadas y nominales en una forma simple y natural.
2. La clasificación final tiene una forma simple la cual puede ser compactamente almacenada y que clasifica eficientemente datos nuevos.
3. Hace uso poderoso de la información condicional en el manejo de relaciones no homogéneas. Una vez que un nodo es dividido en t_L y t_R, entonces para esos dos nodos es individualmente buscada la división más significativa.
4. Selecciona las variables automáticamente por etapas y reduce la complejidad. Por esto se entiende que como una parte intrínseca del desarrollo de la estructura del árbol, una búsqueda es hecha para cada nodo para la división más significativa. En este sentido se asemeja más a un procedimiento por etapas más que a un método del mejor subconjunto. Para cada estado intenta extraer la información sobresaliente de la parte del espacio donde esta trabajando.
5. Proporciona, sin esfuerzo adicional, no sólo una clasificación, incluso una estimación de la probabilidad de clasificación errónea para un objeto.
6. En datos con estructura estándar es invariante bajo todas las transformaciones monótonas de variables ordenadas individuales. Por ejemplo, si x_1 es una variable ordenada en un problema estándar y la transformación $x_1^{'} = x_1^3$ es hecha, la división óptima será la misma no obstante si x_1 o $x_1^{'}$ es utilizada. Si la división óptima sobre x_1 es $x_1 \leq 3,$ entonces la división óptima utilizando $x_1^{'}$ será $x_1^{'} \leq 27,$ y los mismos casos irán a la derecha y a la izquierda.
7. Es extremadamente robusta con respecto a una clasificación errónea. En el espacio de medida X los árboles tiene una propiedad de robustez similar a la mediana. Una observación con un sólo dato tendrá peso 1 entre los casos con N datos. En la evaluación de cualquier división, se cuenta esencialmente cuantos casos de cada clase descienden a la derecha y cuantos van a la izquierda. Otro tipo de error que ocurre frecuentemente es la etiquetación errónea de algunos casos en el conjunto de aprendizaje. Esto puede tener un efecto desastroso sobre la discriminación lineal. El método de árbol

estructurado otra vez asigna un peso a cada punto sólo como uno entre N y no es apreciablemente afectado por algunos puntos etiquetados erróneamente.

8. El resultado del procedimiento del árbol proporciona fácil comprensión e interpretación de la información con relación a la estructura predicativa de los datos.

2.5.9 Árboles de tamaño correcto y estimadores honestos. Regla de suspensión de la división

Esta sección está enfocada en dos principales resultados: conseguir el árbol de tamaño correcto T y obtener estimaciones más precisas de la probabilidad verdadera de la clasificación errónea o de la esperanza verdadera del costo por la clasificación errónea $R^*(T)$.

En cada etapa la estructura de árbol hace una optimización sobre el número de posibles divisiones de los datos. Si sólo los estimadores de resustitución son utilizados, los resultados generalmente son demasiadas divisiones, los árboles son mucho más grandes que lo permitido por los datos, y un estimador de resustitución $R(T)$ que es sesgado hacia abajo.

Por ejemplo, si la división es llevada al punto donde cada nodo terminal contiene sólo un elemento, entonces cada nodo es clasificado por el dato que contiene, y el estimador de resustitución da una tasa cero del error de clasificación.

En general, más divisiones resultan en valores bajos del estimador de resustitución $R(T)$. Al respecto, el procedimiento de árbol es similar a cada etapa de la regresión lineal, en la cual el estimador R^2 se incrementa con cada variable incorporada, alentando la entrada de variables que no tienen poder predictivo cuando es probada sobre muestras independientes extraídas de la misma distribución.
De hecho, en cada etapa de la simulación de la regresión se ha mostrado que más allá de cierto punto, la entrada de variables adicionales causará el decrecimiento de R^2 ajustado. La situación es similar con los árboles. Un árbol demasiado grande

tendrá una alta tasa en el error de clasificación verdadero que el árbol de tamaño "correcto".

Por otra parte, un árbol demasiado pequeño no utilizará algo de la información de la clasificación disponible en L, otra vez resultando en una alta tasa de error de clasificación verdadero que el árbol de tamaño "correcto".

La selección de árboles excesivamente grandes y el uso de estimaciones por resustitución inexactas han permitido muchas de las pasadas críticas a los procedimientos de generación de un árbol estructurado (Breiman, Leo *et all* 1984).

El trabajo fue centrado en encontrar reglas apropiadas de suspensión, esto es, en encontrar un criterio para declarar un nodo terminal. Ninguna de las propuestas con anterioridad fue generalmente aceptada.

Finalmente, una conclusión fue alcanzada, buscar la regla de suspensión correcta fue la forma equivocada de mirar el problema un procedimiento más satisfactorio fue encontrado, consistente de dos elementos clave y que son:

1. Podar en lugar de detener. Desarrollar un árbol que es mucho más grande y podarlo hacia arriba en la "forma correcta" hasta que finalmente se llegue al nodo raíz.
2. Utilizar estimaciones más precisas de $R^*(T)$ para seleccionar el árbol de tamaño correcto de entre los subárboles podados.

Este nuevo marco lleva inmediatamente a dos preguntas: Cómo hacer una poda ascendente en la forma correcta y cómo pueden ser conseguidas mejores estimaciones de $R^*(T)$.

Cuando un árbol es podado ascendentemente, la tasa estimada del error por clasificación primero decrece lentamente, alcanza un mínimo gradual y entonces se incrementa rápidamente hacia el número de nodos terminales.

2.5.10 Podado

Para fijar la notación, la estimación por resustitución del costo total del error por clasificación $R^*(T)$ está dado por

$$R^*(T) = \sum_{t \in \tilde{T}} r(t)\, p(t)$$
$$= \sum_{t \in \tilde{T}} R(t)$$

Donde $R(T)$ y $R(t)$ son los costos del error por clasificación del árbol y del nodo respectivamente.

El primer paso es hacer crecer un árbol muy grande $T_{\max}$ para permitir que el procedimiento de división continúe hasta que todos los nodos terminales sean pequeños o puros o contengan solamente vectores de medida idénticos.

Aquí, puro significa que los casos en los nodos están todos en una sola clase. Con tiempo ilimitado de cómputo, la mejor forma de crecimiento para este árbol inicial sería continuar dividiendo hasta que cada nodo terminal contenga exactamente un caso de la muestra.

El tamaño del árbol inicial no es crítico mientras sea bastante grande. Si se comienza con el árbol más grande posible $T'_{\max}$ o con uno más pequeño, pero aún suficientemente grande $T_{\max}$, el proceso de podado producirá los mismos subárboles en el siguiente sentido: si el corte comienza con $T'_{\max}$ produce un subárbol contenido en $T_{\max}$, entonces el podado comenzando con $T_{\max}$ producirá exactamente el mismo subárbol.

El método convenido adoptado para el crecimiento inicial de un árbol suficientemente grande $T_{\max}$ especifica un número $N_{\min}$ y continúa dividiendo hasta que cada nodo terminal es puro o satisface $N(t) \le N_{\min}$ o contiene sólo vectores de medida idénticos. Generalmente, $N_{\min}$ ha sido elegido en 5, ocasionalmente en 1.

Comenzando con el árbol más grande $T_{\max}$ y selectivamente podando hacia arriba produce una sucesión de subárboles de $T_{\max}$ eventualmente colapsando a el árbol $\{t_1\}$ consistente de el nodo raíz.

Para definir el proceso de podado más precisamente, se llamará un nodo t' que se sitúa más abajo en el árbol un *descendiente* de un nodo más alto t si existe una trayectoria descendiente conectando el árbol que lleva de t a t'. Entonces t incluso es llamado un antecesor de t'. Así, en la Figura 9 $t_4,\ t_5,\ t_8,\ t_9,\ t_{10}$ y t_{11} son todos nodos descendientes de $t_2,$ pero no t_6 y t_7.

Similarmente, $t_4,\ t_2$ y t_1 son antecesores de $t_9,$ pero t_3 no es antecesor de t_9.

DEFINICIÓN 2.17. Una rama T_t del árbol T con nodo raíz $t \in T$ consiste del nodo raíz y todos sus descendientes de t en T.

La rama T_{t_2} es ilustrada en la Figura 10.

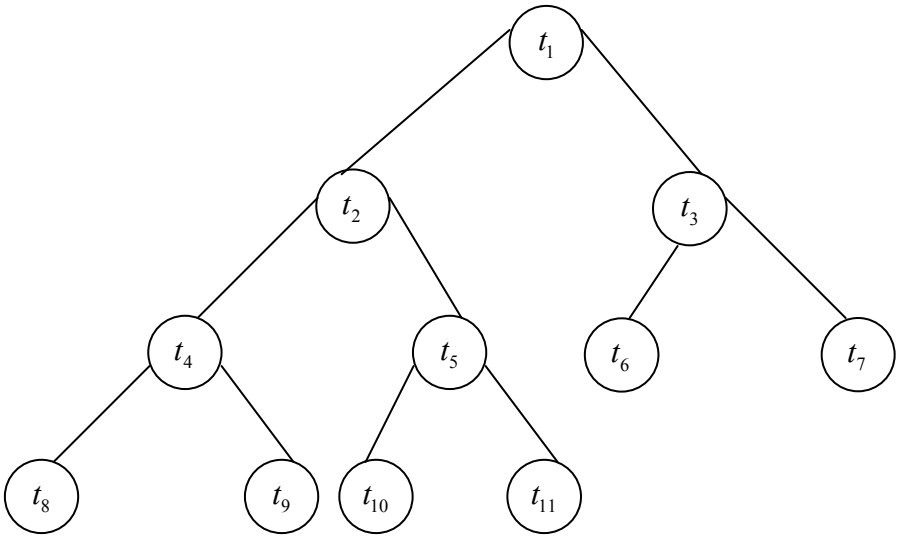


Figura 9

DEFINICIÓN 2.18. El podado de una rama T_t de un árbol T consiste en borrar (separar) de T todos los nodos descendientes de $t,$ esto es, cortar todo T_t excepto su nodo raíz. El árbol podado de esta forma será denotado por $T - T_t$.

El árbol podado $T - T_{t_2}$ es mostrado en la Figura 11.

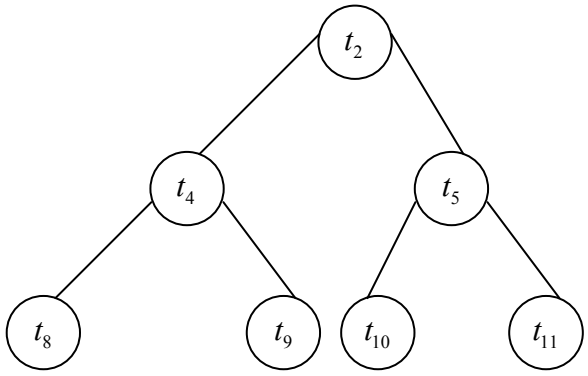


Figura 10

DEFINICIÓN 2.19. Si T' es obtenido de T por sucesivos podados de las ramas, entonces T' es llamado un subárbol podado de T y denotado por $T' \prec T.$ (obsérvese que T' y T tienen el mismo nodo raíz).

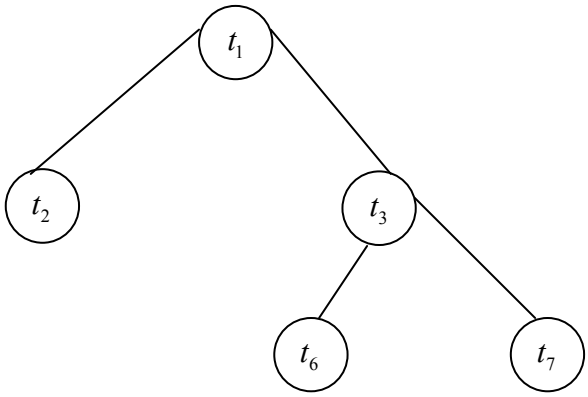


Figura 11

Aún para un tamaño moderado de $T_{\max}$ conteniendo, por ejemplo, de 30 a 40 nodos, existe un número extremamente grande de subárboles y una cantidad aún mayor de formas distintas de podar ascendentemente hacia $\{t_1\}.$ Un procedimiento "selectivo" de podado es necesario, esto es, una selección de un número razonable de subárboles, decreciendo en tamaño, tal que, cada subárbol seleccionado es el "mejor" subárbol en el rango de su tamaño.

La palabra *mejor* indica el uso de algún criterio para juzgar cuan bueno un subárbol T es. Aunque conociendo que $R(T)$ carece de precisión como un estimador de

$R^*(T)$, es el criterio más natural para usar en la comparación de diferentes subárboles del mismo tamaño.

Sin importar cómo $T_{\max}$ fue construido, que criterio de división fue empleado, el proceso selectivo de podado comienza con el árbol inicial dado $T_{\max}$, calculando $R(T)$ para cada nodo $t \in T_{\max}$, y progresivamente podando $T_{\max}$ hacia arriba a su nodo raíz tal que en cada estado de podado, $R(T)$ es tan pequeño como sea posible.

Aquí está un simple ejemplo de tal proceso de podado selectivo. Supóngase que $T_{\max}$ tiene L nodos terminales. Entonces constrúyase una sucesión de árboles cada vez más pequeños

$$T, T_1, T_2, ..., \{t_1\}$$

Como sigue: para cada valor de H, $1 \le H < L$, considérese la clase $_H$ de todos los subárboles de $T_{\max}$ teniendo $L-H$ nodos terminales. Seleccione T_H como el subárbol en $_H$ el cual minimiza $R(T)$; esto es,

$$R(T_H) = \min_{T \in T_H} R(T)$$

Poniéndolo de otra forma, T_H es el subárbol con costo mínimo teniendo $L-H$ nodos. Este es un procedimiento intuitivamente atractivo y puede ser eficientemente implementado. Sin embargo, tiene algunas desventajas.

Quizá la más importante es que la sucesión de subárboles no es anidada, esto es, T_{H+1} no es necesariamente un subárbol de T_H. Los nodos que fueron previamente cortados pueden repetirse. En resumen, la sucesión de subárboles no está formada por un podado progresivo ascendente.

2.5.11 Podado por costo-complejidad mínimo

La idea detrás del podado de costo-complejidad mínimo es la siguiente:

DEFINICIÓN 2.20. Para cualquier subárbol $T \prec T_{\max}$, se define su complejidad $\left|\tilde{T}\right|$, como el número de nodos terminales en T. Sea $\alpha \geq 0$ un número real llamado el parámetro de complejidad y se define la medida de costo-complejidad $R_\alpha(T)$ como

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right|$$

Así, $R_\alpha(T)$ es una combinación lineal del costo del árbol y su complejidad. Si se considera a α como el costo por complejidad por nodo terminal, $R_\alpha(T)$ está formado por sumar al costo por error de clasificación del árbol una penalización por complejidad.

Ahora, para cada valor de α, se encuentra que el subárbol $T(\alpha) \prec T_{\max}$ el cual minimiza $R_\alpha(T)$, es decir;

$$R_\alpha\left(T(\alpha)\right) = \min_{T \prec T_{\max}} R_\alpha(T)$$

Si α es pequeña, la señalización por tener un número grande de nodos terminales es pequeña y $T(\alpha)$ será grande.

Por ejemplo, si $T_{\max}$ es tan grande que cada nodo terminal contiene sólo un caso, entonces cada caso es clasificado correctamente; $R(T_{\max}) = 0$, así que $T_{\max}$ minimiza $R_0(T)$. Como la señalización α por nodos terminales aumenta, los subárboles minimizados $T(\alpha)$ tendrán algunos nodos terminales.

Finalmente, para α suficientemente grande, el subárbol minimizado $T(\alpha)$, consistirá solamente del nodo raíz, y el árbol $T_{\max}$ habrá sido completamente podado. Aunque α corre en un continuo de valores, hay a lo más un número finito de subárboles $T_{\max}$. Así, el proceso de podado produce una sucesión finita de subárboles T_1, T_2, T_3, ... con progresivamente, algunos nodos terminales. Debido al aspecto finito, lo que sucede es que si $T(\alpha)$ es el árbol minimizado para algún valor dado de α, entonces continúa siendo el mínimo cuando α se incrementa hasta que

un salto en el punto α' es alcanzado, y un nuevo árbol $T(\alpha')$ llega a ser minimizado y continúa así hasta el próximo salto α''.

Existe un problema que es el de la unicidad que se centra alrededor de una apropiada definición y una prueba de que el objeto definido realmente exista. La inclusión e implementación efectiva entonces se sigue de una reexaminación cercana del mecanismo del podado por costo-complejidad mínimo.

DEFINICIÓN 2.21. El más pequeño subárbol minimizado $T(\alpha)$ para el parámetro de complejidad α es definido por las condiciones

i. $R_\alpha(T(\alpha)) = \min_{T \prec T_{\max}} R_\alpha(T)$
ii. Si $R_\alpha(T) = R_\alpha(T(\alpha))$, entonces

Esta definición rompe empates en costo-complejidad mínimo por seleccionar el más pequeño minimizador de R_α. Obviamente, si tal subárbol existe, debería ser único.

La pregunta es la existencia. Por ejemplo, supóngase que hay exactamente dos minimizadores T y T' de R_α, pero ninguno de los dos contiene al otro. Entonces $T(\alpha)$, como fue definida anteriormente no existe. Sin embargo,

PROPOSICIÓN 2.2. Para cada valor de α, existe el más pequeño subárbol minimizado.

El salto puntual para el podado no es $T_{\max}$ pero si $T_1 = T(0)$. Esto es, T_1 es el subárbol más pequeño de $T_{\max}$ satisfaciendo

$$R(T_1) = R(T_{\max})$$

Para obtener T_1 de $T_{\max}$, sea t_L y t_R cualesquiera dos nodos terminales en $T_{\max}$ obtenidos de una división del nodo t intermedio ascendente. Se recuerda de la proposición 3.1 que $R(t) \geq R(t_L) + R(t_R)$. Si $R(t) = R(t_L) + R(t_R)$, entonces podar t_L y

t_R. Hay que continuar este proceso hasta que no sea posible cortar más. El árbol resultante es T_1.

Para cualquier rama T_t de T_1, se define $R(T_t)$ por

$$R(T_t) = \sum_{t' \in \tilde{T}_t} R(t')$$

Donde $\tilde{T}_t$ es el conjunto de nodos terminales de T_t.

PROPOSICIÓN 2.3. Para t cualquier nodo no terminal de T_1,

$$R(t) > R(T_t)$$

Comenzando con T_1, el corazón del podado costo-complejidad mínimo está en el entendimiento que trabaja por corte de acoplamiento más débil.

Para cualquier nodo t que pertenece a T_1, se denota por $\{t\}$ la subrama de T_t consistente del nodo (único) $\{t\}$.

Se elige $R_\alpha(\{t\}) = R(t) + \alpha$.

Para cualquier rama T_t, se define $R_\alpha(T_t) = R(T_t) + \alpha|\tilde{T}_t|$. Tan grande como $R_\alpha(T_t) < R_\alpha(\{t\})$,

La rama T_t tiene un costo-complejidad más pequeño que el nodo único $\{t\}$. Pero para algún valor crítico de α, los dos costo-complejidad llegan a ser iguales. En este punto la subrama $\{t\}$ es más pequeña que T_t, tiene el mismo costo-complejidad, y es por lo tanto preferible. Para encontrar este valor crítico de α, se resuelve la desigualdad

$$R_\alpha(T_t) < R_\alpha(\{t\})$$

Obteniendo

$$\alpha < \frac{R(t) - R(T_t)}{\left|\tilde{T}_t - 1\right|} \tag{2.13}$$

Por la proposición 2.3 el valor crítico del lado derecho de (2.13) es positivo. Entonces se define una función $g_1(t),\ t \in T_1,$ por

$$g_1(t) = \begin{cases} \dfrac{R(t) - R(T_t)}{\left|\tilde{T}_t - 1\right|}, & t \notin \tilde{T}_1 \\ +\infty, & t \in \tilde{T}_1 \end{cases}$$

Entonces se define el acoplamiento más débil $\bar{t}_1$ en T_1 como el nodo tal que

$$g_1(\bar{t}_1) = \min_{t \in T_1} g_1(t)$$

Y se asigna

$$\alpha_2 = g_1(\bar{t}_1)$$

El nodo $\bar{t}_1$ es el acoplamiento más débil en el sentido que, como el parámetro α se incrementa, es el primer nodo tal que $R_\alpha(\{t\})$ llega a ser igual a $R_\alpha(T_t)$. Entonces $\{\bar{t}_1\}$ será preferible $T_{\bar{t}_1}$, y α_2 es el valor de α en el cual la igualdad ocurre.

Se define un nuevo árbol T_2 tal que $T_2 \prec T_1$ por podar y separar la rama $T_{\bar{t}_1}$, esto es,

$$T_2 = T_1 - T_{\bar{t}_1}$$

Ahora, usando T_2 en lugar de T_1, se encuentra el acoplamiento más débil en T_2.
Más precisamente, sea T_{2t} la parte de la rama de T_t la cual está contenida en T_2, se define

$$g_2(t)=\begin{cases}\dfrac{R(t)-R(T_{2t})}{\left|\tilde{T}_{2t}-1\right|}, & t\in T_2 \quad t\notin \tilde{T}_2\\ +\infty, & t\in \tilde{T}_2\end{cases}$$

Y $\bar{t}_2 \in T_2$, α_3 dado por

$$g_2\left(\bar{t}_2\right)=\min_{t\in T_2} g_2(t)$$

$$\alpha_3 = g_2\left(\bar{t}_2\right)$$

Repitiendo el procedimiento se define

$$T_3 = T_2 - T_{\bar{t}_2}$$

Y encontrando el acoplamiento más débil $\bar{t}_3$ en T_3 y el correspondiente valor del parámetro α_4; ahora formar T_4 y repetir otra vez.

Si para cualquier estado existe una multiplicidad del acoplamiento más débil, por ejemplo, si

$$g_k\left(\bar{t}_k\right)=g_k\left(\bar{t}'_k\right)$$

Entonces se define

$$T_{k+1}=T_k-T_{\bar{t}_k}-T_{\bar{t}'_k}$$

Continuando de esta forma, se obtiene una sucesión decreciente de subárboles

$$T_1 \succ T_2 \succ T_3 \succ \ldots \succ \{t_1\}$$

La conexión con el podado por costo-complejidad mínimo está dado por:

TEOREMA 2.2. $\{\alpha_k\}$ es una sucesión creciente, esto es, $\alpha_k < \alpha_{k+1},\ k \geq 1,$ donde $\alpha_1 = 0.$ Para $k \geq 1,\ \alpha_k \leq \alpha \leq \alpha_{k+1},\ T(\alpha) = T(\alpha_k) = T_k.$

Este teorema describe como trabaja el podado por costo-complejidad mínimo. Comienza con $T_1,$ encuentra el acoplamiento más débil de la rama $T_{t_1},$ y poda el árbol para obtener T_2 cuando α alcanza $\alpha_2.$ Ahora encuentra el acoplamiento más débil para la rama T_{t_2} en T_2 y lo poda para obtener T_3 cuando α alcanza $\alpha_3;$ y así sucesivamente.

2.5.12 El mejor subárbol: un problema de estimación

El método de podado discutido en la sección previa resulta en una sucesión decreciente de subárboles $T_1 \succ T_2 \succ ... \succ \{t_1\},$ donde $T_k = T(\alpha_k),\ \alpha_1 = 0.$ El problema ahora es reducido a seleccionar uno de esos como el árbol de tamaño óptimo.

Si la estimación por resustitución $R(T_k)$ es usada como un criterio, el árbol más grande T_1 sería seleccionado. Pero si uno tiene una estimación "honesta" $\tilde{R}(T_k)$ del costo del error por clasificación, entonces el mejor subárbol T_{k_0} debería estar definido como el subárbol que minimiza $\tilde{R}(T_k);$ i.e.,

$$\tilde{R}(T_k) = \min_k \tilde{R}(T_k) \tag{2.14}$$

Los resultados obtenidos en esta sección es la construcción de estimaciones relativamente insesgadas del verdadero costo del error por clasificación $R^*(T_k).$ Dos métodos de estimación son discutidos: uso de una prueba de muestra independiente y validación-cruzada. De las dos, el uso de una prueba de muestra independiente es computacionalmente más eficiente y es preferida cuando la muestra de aprendizaje contiene un gran número de casos. Como un subproducto útil da estimaciones relativamente insesgadas del costo del error por clasificación. Validación-cruzada es computacionalmente más cara, pero hace más efectivo el uso de todos los casos y da información útil considerando la estabilidad de la estructura de árbol.

Para estudiar el sesgo o el error estándar de una estimación, es necesario un modelo de probabilidad. Se asumirá en esta sección el modelo usado anteriormente: Los casos en L son N extracciones independientes de una distribución de probabilidad $P(A, j)$ sobre X C, y $(\ \ ,Y)$ es un muestreo aleatorio con distribución $P(A, j)$, independiente de L. Si no existe costo variable del error por clasificación, se sabe que $R^*(d)$ está definida como

$$R^*(d) = P(d(\mathrm{X}) \neq Y)$$

En el caso general, con costo variable del error por clasificación $C(i \mid j)$,

DEFINICIÓN 2.22. Se define:

a) $Q^*(i \mid j) = P(d(\ \ ) = i \mid Y = j)$ tal que $Q^*(i \mid j)$ es la probabilidad de que un caso en j sea clasificado dentro de i por d.

b) $R^*(j) = \sum_i C(i \mid j) Q^*(i \mid j)$ como el costo esperado del error por clasificación para los elementos de la clase j.

c) $R^*(d) = \sum_j R^*(j) \pi(j)$ como el costo esperado del error por clasificación para el clasificador d.

Ambos prueba de la muestra y validación-cruzada proveen estimaciones de $Q^*(i \mid j)$ y $R^*(j)$, así como de $R^*(d)$. La idea básica de los procedimientos es que $Q^*(i \mid j)$ puede ser estimado usando simples cuentas de casos del error por clasificación. Entonces $R^*(j)$, $R^*(T_k)$ son estimados a través de las definiciones 2.22 b) y c). Adicionalmente, los errores estándar pueden ser calculados asumiendo un simple modelo binomial para la estimación de $Q^*(i \mid j)$.

2.5.12.1 Estimaciones de la prueba de la muestra

Se selecciona un número fijo $N^{(2)}$ de casos aleatoriamente de L para constituir la prueba de la muestra L_2. El restante L_1 integra la nueva muestra de aprendizaje.

El árbol $T_{\max}$ es desarrollado utilizando sólo L_1 y se poda ascendentemente para dar la sucesión $T_1 \succ T_2 \succ ... \succ \{t_1\}$. Esto es, la secuencia $\{T_k\}$ de árboles es construida y los nodos terminales asignados a una clasificación sin aún haber visto cualquiera de los casos en L_2.

Ahora se toman las observaciones en L_2 y se insertan en T_1. Cada árbol en T_k asigna una clasificación predicha a cada caso en L_2. Debido a que la verdadera clase para cada elemento en L_2 es conocida, el costo del error por clasificación de T_k operando con L_2 puede ser calculado. Esto produce la estimación $R^{ts}(T_k)$.

Se denota por $N_j^{(2)}$ el número de casos de la clase j en L_2. Para T cualquiera de los árboles $T_1, T_2, ...$, sea $N_{ij}^{(2)}$ la cantidad de observaciones de la clase j en L_2 cuyo clasificación predicha por T es la clase i.

La estimación básica es obtenida por elegir

$$Q^{ts}(i \mid j) = \frac{N_{ij}^{(2)}}{N_j^{(2)}}$$

Esto es, $Q^{*}(i \mid j)$ es estimada como la proporción de la prueba de la muestra de los casos en la clase j que el árbol T clasifica como i (se elige $Q^{ts}(i \mid j) = 0$ si $N_j^{(2)} = 0$).

Utilizando la Definición 2.22. b) da la estimación

$$R^{ts}(j) = \sum_i C(i \mid j) Q^{ts}(i \mid j)$$

Si las probabilidades *a priori* $\{\pi(j)\}$ son dadas o estimadas, la definición 2.22. c) indican la estimación

$$R^{ts}(T) = \sum_j R^{ts}(j) \pi(j) \qquad (2.15)$$

Si las probabilidades *a priori* son estimadas, se usa L_2 para hacerlo, entonces como

$$\pi(j)\frac{N_j^{(2)}}{N^2}.$$

En este caso, (2.15) se simplifica a

$$R^{ts}(T)=\frac{1}{N^{(2)}}\sum_{i,j}C(i\mid j)N_{ij}^{(2)} \qquad (2.16)$$

Esta última expresión (2.16) tiene una simple interpretación. Se calcula el costo del error por clasificación para cada caso en L_2 descendiendo por T y entonces tomar el porcentaje.

En el costo unitario, $R^{ts}(j)$ es la proporción de casos probados del error por clasificación de la clase $j,$ con las probabilidades *a priori* estimadas $R^{ts}(T)$ es la proporción total de los casos probados del error por clasificación de $T.$

Utilizando el modelo de probabilidad asumido, es fácil mostrar que las estimaciones de $Q^{ts}(i\mid j)$ son sesgados sólo si $N_j^{(2)}=0.$ Para cualquier razonable distribución del tamaño de muestra, la probabilidad de que $N_j^{(2)}=0$ es tan pequeña que esas estimaciones pueden ser tomadas como insesgadas. En consecuencia, son los estimadores de $R^{ts}(T).$ De hecho, en el caso de las probabilidades *a priori* estimadas, hay cancelación y $R^{ts}(T)$ es exactamente insesgado.

Los estimadores de la prueba de la muestra pueden ser usados para seleccionar el tamaño del árbol correcto T_{k_0} por la regla

$$R^{ts}\left(T_{k_0}\right)=\min_k R^{ts}\left(T_k\right)$$

2.5.12.2 Estimaciones de la validación-cruzada

Al menos que el tamaño de muestra en L sea verdaderamente grande, la validación-cruzada es el método de estimación preferido.

En la validación-cruzada V doble, la muestra de aprendizaje original L es dividida por una selección aleatoria dentro de V subconjuntos, L_v, $v = 1, 2, ..., V$, cada uno conteniendo el mismo número de casos (tan cercano como sea posible).

La $v - ésima$ muestra de aprendizaje es

$$L^{(v)} = L - L_v, \quad v = 1, 2, ..., V$$

Tal que $L^{(v)}$ contiene la fracción $\frac{(V-1)}{V}$ del total de los casos. Imaginando que V es razonablemente grande. Usualmente V es tomado como $10,$ para que cada muestra de aprendizaje $L^{(v)}$ contenga $\frac{9}{10}$ de los casos.

En la V doble validación cruzada, V árboles auxiliares se desarrollan junto con el principal desplegado sobre $L.$ El $v - ésimo$ árbol auxiliar es desarrollado usando el aprendizaje de la muestra $L^{(v)}.$ Se comienza por hacer crecer V árboles excesivamente grandes $T_{\max}^{(v)}, v = 1, 2, ..., V,$ así como $T_{\max},$ utilizando el criterio que las divisiones continúan hasta que los nodos son puros o tiene menos casos que $N_{\min}.$

Para cada valor del parámetro de complejidad $\alpha,$ sea $T(\alpha),$ $T^{(v)}(\alpha),$ $v = 1, 2, ..., V,$ el subárbol correspondiente por costo-complejidad de $T_{\max},$ $T_{\max}^{(v)}.$ para cada $v,$ los árboles $T_{\max}^{(v)},$ $T^{(v)}(\alpha)$ han sido construidos sin haber observado los casos en $L_V.$

Así, los elementos en L_V pueden servir como una prueba de la muestra independiente para el árbol $T^{(v)}(\alpha).$

Se coloca L_V por debajo del árbol $T_{\max}^{(v)}$, $v = 1, 2, ..., V$. Se fija el valor del parámetro de complejidad α. Para cada valor de v, i, j, se define $N_{ij}^{(v)}$ como el número de casos en la clase j en L_v clasificados como i por $T^{(v)}(\alpha)$, y se elige

$$N_{ij} = \sum_v N_{ij}^{(v)}$$

Así, N_{ij} es el número total de casos probados en la clase j clasificados como i. Cada elemento en L aparece en una y sólo una prueba de la muestra L_v. Por lo tanto, el número total de casos en la clase j es N_j, el número de observaciones en la clase j en L.

La idea es que para V grande, $T^{(v)}(\alpha)$ debería tener aproximadamente la misma exactitud de clasificación que $T(\alpha)$. Por lo tanto, se hace el paso fundamental de estimar $Q^*(i \mid j)$ por $T^{(v)}(\alpha)$ como

$$Q^{cv}(i \mid j) = \frac{N_{ij}}{N} \tag{2.17}$$

Para las $\pi(j)$ proporcionadas o estimadas, se elige

$$R^{cv}(j) = \sum_i C(i \mid j) Q^{cv}(i \mid j)$$

Y se selecciona

$$R^{cv}(T(\alpha)) = \sum_j R^{cv}(j) \pi(j) \tag{2.18}$$

Si las probabilidades a priori son datos estimados, elegir $\pi(j) = \frac{N_j}{N}$. Entonces la expresión anterior se convierte en

$$R^{cv}(T(\alpha)) = \frac{1}{N} \sum_{i,j} C(i \mid j) N_{ij} \tag{2.19}$$

En el costo unitario por caso la expresión de arriba representa la proporción de la prueba que elige casos del error por clasificación.

La implementación es simplificada por el hecho de que aunque α puede variar continuamente, los árboles por costo-complejidad mínimo que se desarrollaron sobre L son iguales a T_k para $\alpha_k \le \alpha \le \alpha_{k+1}$. Se selecciona

$$\alpha_k^{'} = \sqrt{\alpha_k \alpha_{k+1}}$$

De tal forma que $\alpha_k^{'}$ es el punto medio geométrico del intervalo tal que $T(\alpha) = T_k$. Entonces considerar

$$R^{cv}(T_k) = R^{cv}\left(T\left(\alpha_k^{'}\right)\right)$$

Donde el lado derecho está definido por (2.18) esto es, $R^{cv}(T_k)$ es la estimación obtenida por poner la prueba de la muestra L_V a través de los árboles $T^{(V)}\left(\alpha_k^{'}\right)$. Para el árbol del nodo raíz $\{t_1\}$, $R^{cv}\left(\{t_1\}\right)$ está seleccionado como igual al costo por resustitución $R\left(\{t_1\}\right)$.

Ahora la regla para seleccionar el árbol de tamaño correcto es: Se elige el árbol T_{k_0} tal que

$$R^{cv}\left(T_{k_0}\right) = \min_k R^{cv}\left(T_k\right)$$

Entonces usar $R^{cv}\left(T_{k_0}\right)$ como una estimación para el costo del error por clasificación.

Preguntas concernientes al sesgo exacto de R^{cv} son difíciles de determinar, debido a que la formulación probabilística es muy compleja. Un punto es bastante claro. Las estimaciones de la validación-cruzada para el costo del error por clasificación para árboles se desarrollaron sobre una fracción de los datos $\frac{V-1}{V}$. Debido a que los árboles auxiliares por validación-cruzada se desarrollaron sobre muestras pequeñas, ellos tienden a ser menos exactos.

Las estimaciones por validación-cruzada, entonces, tienden a ser conservadores en la dirección de sobreestimar los costos del error por clasificación.

La muestra de aprendizaje L_1 es usada para construir T. Sea la prueba de la muestra L_2 una extracción independiente de la misma distribución fundamental de L_1 pero independiente de L_1. La estimación $R^{ts}(T)$ es la proporción de casos en L_2 clasificados erróneamente por T. La probabilidad p^* de que cualquier caso es clasificado erróneamente es $R^*(T)$. Así, se tiene N_2 ensayos independientes binomiales con probabilidad de éxito p^* en cada ensayo donde es estimada como la proporción p de éxitos. Claramente, $E(p) = p^*$, así p es insesgado. Adicionalmente,

$$Var(p) = \frac{p^*(1-p^*)}{N_2}$$

Así, la estimación del error estándar de p es calculada por $\left[\frac{p(1-p)}{N_2}\right]^{\frac{1}{2}}$, llevando la estimación del error estándar para $R^{st}(T)$ como

$$SE\left(R^{ts}(T)\right) = \left[\frac{R^{ts}(T)\left(1-R^{ts}(T)\right)}{N_2}\right]^{\frac{1}{2}}$$

2.5.13 Reglas de división. Reduciendo el costo por clasificación errónea

En la sección 2.5.8.2 un marco fue dado para generar las reglas de división. La idea fue definir una función de impureza $\phi(p_1, p_2, ..., p_J)$ teniendo ciertas propiedades deseables (Definición 2.10). Se define la función de impureza $i(t)$ como $\phi\left(p(1|t), p(2|t), ..., p(J|t)\right)$, se selecciona $\mathrm{I}(t) = i(t)\,p(t)$, y se define la impureza del árbol $\mathrm{I}(T)$ como

$$\mathrm{I}(T) = \sum_{t \in \tilde{T}} \mathrm{I}(t)$$

Entonces en cualquier nodo terminal actual, se elige qué división reduce más $\mathrm{I}(T)$ o, equivalentemente, maximiza

$$\mathrm{I}(s,t) = \mathrm{I}(t) - \mathrm{I}(t_L) - \mathrm{I}(t_R)$$

o

$$i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

Dentro de este marco parece más natural tomar la impureza del árbol como $R(T)$, la estimación por resustitución para el costo esperado por clasificación errónea. La mejor división entonces sería la que más reduce la tasa estimada por clasificación errónea. Esto es equivalente a definir $i(t)$ como igual a $r(t)$, donde

$$\begin{aligned} r(t) &= \min_i \sum_j C(i \mid j)\, p(j \mid t) \\ &= 1 - \max_j p(j \mid t) \end{aligned} \qquad \text{(Costo unitario por caso).}$$

Entonces la mejor división de t maximiza

$$r(t) - p_L r(t_L) - p_R r(t_R) \tag{2.20}$$

o, equivalentemente, maximiza

$$R(t) - R(t_L) - R(t_R) \tag{2.21}$$

La función correspondiente de la impureza del nodo (costo unitario) es

$$\phi(p_1, p_2, ..., p_J) = 1 - \max_j p_j$$

Esta función tiene todas las propiedades deseables listadas en la definición 2.10.

PROPOSICIÓN 2.4. Para cualquier división de t en t_L y t_R,

$$R(t) \geq R(t_L) + R(t_R)$$

Y la igualdad se cumple si $j^*(t) = j^*(t_L) = j^*(t_R)$.

Demostración. Obsérvese que

$$\begin{aligned} R(t) &= \sum_j C\left(j^*(t) \mid j\right) p(j \mid t) \\ &= \sum_j C\left(j^*(t) \mid j\right)\left[p(j,t_L) + p(j,t_R)\right] \end{aligned}$$

o

$$\begin{aligned} R(t) &= R(t_L) - R(t_R) \\ &= \sum_j C\left(j^*(t) \mid j\right) p(j,t_L) - \min_i \sum_j C(i \mid j)\, p(j,t_L) + \\ &+ \sum_j C\left(j^*(t) \mid j\right) p(j,t_R) - \min_i \sum_j C(i \mid j)\, p(j,t_R) \end{aligned}$$

El lado derecho es ciertamente no negativo e igual a cero bajo las condiciones $j^*(t) = j^*(t_L) = j^*(t_R)$.

Ahora supóngase que se tiene un problema de dos clases y con probabilidades *a priori* iguales y están en el nodo t la cual tiene una preponderancia de casos de clase 1. Es concebible que cada división de t produce nodos t_L y t_R donde ambos tienen mayoría de clase 1. Entonces $R(t) - R(t_L) - R(t_R) = 0$ para todas las divisiones en S y un número pequeño de mejores divisiones.

El segundo defecto es más difícil de cuantificar. En suma, la reducción de la tasa de clasificación errónea no parece ser un buen criterio para el procedimiento del desarrollo multipasos total del árbol.

Del ejemplo de la Figura 12, supóngase que el nodo superior es el nodo raíz y se asumen probabilidades *a priori* iguales. La primera división lleva a un árbol en el cual 200 casos son clasificados erróneamente y

$$R(T) = \frac{200}{800} = .025$$

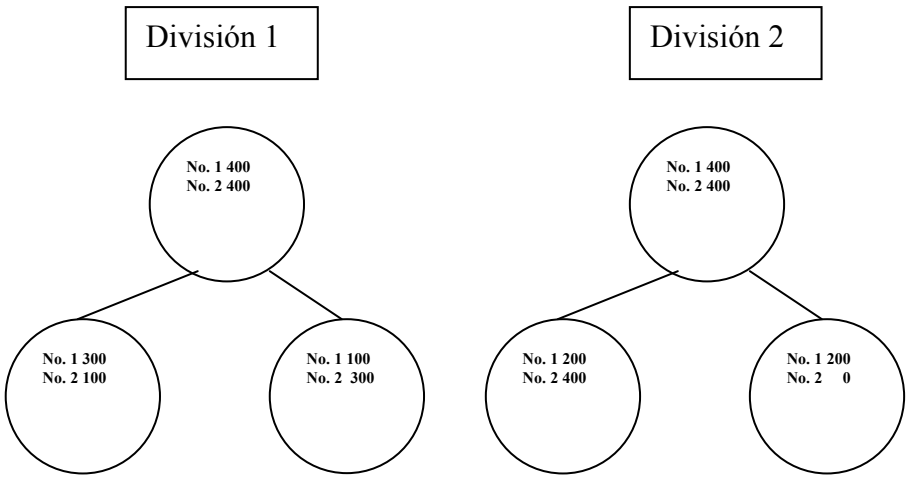


Figura 12

En la segunda división, 200 casos son incluso clasificados erróneamente y $R(T) = .25.$

Aún pensando que ambas divisiones están dando iguales tasas por el criterio $R(T),$ la segunda partición es probablemente más deseable en términos del futuro crecimiento del árbol. La primera división produce dos nodos, ambos teniendo $r(t) = .25.$ Estos probablemente necesitarán más bifurcaciones para obtener un árbol con un valor bajo de $R(T).$ El segundo árbol tiene un nodo con $r(t) = .33,$ el cual tendrá que ser dividido otra vez sobre las mismas bases. Pero incluso tiene un nodo con $\frac{1}{4}$ de los casos para el cual $r(t) = 0,$ este nodo es terminal.

Además del problema de degeneración, entonces, el criterio $R(T)$ no parece que recompense con divisiones apropiadas, que sean más deseables en el contexto del crecimiento continuo del árbol. Otros ejemplos muestran que este comportamiento puede ser aun más pronunciado cuando el número de clases consigue ser grande.
Este problema es en gran parte causado por el hecho de que la estructura de desarrollo del árbol esta basada en un procedimiento de optimización de un paso.

2.5.14 El problema multiclases: costo unitario

Dos diferentes criterios han sido adoptados para utilizarse en el problema multiclases con costos unitarios. Estos provienen de diferentes acercamientos hacia la generalización del criterio de las dos clases y son llamados

- El criterio de Gini
- El criterio Twoing.

2.5.14.1 El criterio de Gini

El concepto de un criterio dependiendo de una medida de impureza de un nodo ha sido introducido. Dado un nodo t con probabilidad de clase estimada $p(j|t)$, $j=1,2,...,J,$ una medida de impureza del nodo t

$$i(t)=\phi\left(p(1|t),p(2|t),...,p(J|t)\right)$$

es definida y una búsqueda de la impureza es hecha para la división que más reduce al nodo, o equivalentemente al árbol.

La función original seleccionada fue

$$\phi(p_1,p_2,...,p_J)=-\sum_j p_j \log p_j$$

El índice de diversidad de Gini fue adoptado, éste tiene la forma

$$i(t)=\sum_{j\neq i} p(j|t)\,p(i|t) \tag{2.22}$$

Incluso puede ser escrito como

$$i(t)=\left(\sum_j p(j|t)\right)^2-\sum_j p^2(j|t)=1-\sum_j p^2(j|t) \tag{2.23}$$

En el problema de las dos clases el índice se reduce a

$$i(t) = 2p(1|t)\,p(2|t)$$

El índice de Gini tiene una interesante interpretación. En lugar de utilizar la regla de la pluralidad para clasificar objetos en un nodo $t,$ usa el criterio que asigna un objeto seleccionado aleatoriamente del nodo a la clase i con probabilidad $p(i|t).$ La probabilidad estimada de que el elemento esté actualmente en la clase j es $p(j|t),$ por lo tanto, la probabilidad estimada de clasificación errónea bajo esta regla es el índice de Gini

$$\sum_{j\neq i} p(j|t)\,p(i|t)$$

Otra interpretación está en términos de las varianzas. En un nodo $t,$ se asignan a todos los objetos de la clase j el valor 1 y a los otros elementos el valor 0.

Entonces la varianza de la muestra de esos valores es $p(j|t)\big(1-p(j|t)\big).$ Si esto es repetido para las J clases y la varianza es sumada, el resultado es

$$\sum_j p(j|t)\big(1-p(j|t)\big) = 1-\sum_j p^2(j|t)$$

Finalmente, obsérvese que el índice de Gini considerado como una función $\phi(p_1, p_2, ..., p_J)$ es un polinomio cuadrático con coeficientes no negativos. Por lo tanto es cóncavo en el sentido que para $r+s=1,\ r\geq 0,\ s\geq 0,$

$$\phi\left(rp_1+sp_1^{'}, rp_2+sp_2^{'}, ..., rp_J+sp_J^{'}\right) \geq r\phi(p_1, p_2, ..., p_J) + s\phi\left(p_1^{'}, p_2^{'}, ..., p_J^{'}\right)$$

Esto asegura que para cualquier división $s,$

$$i(s,t)\geq 0$$

Realmente, es estrictamente cóncava, así que $i(s,t)=0$ si sólo si $p(j\mid t_L)=p(j\mid t_R)=p(j\mid t),\ j=1,2,...,J.$

2.5.14.2 El criterio Twoing

El segundo acercamiento al problema de multiclases adopta una estrategia diferente. Se denota la clase de clases por $C,$ i.e.,

$$C=\{1,2,...,J\}$$

Para cada nodo, se separan las clases dentro de dos superclases,

$$C_1=\{j_1,j_2,...,j_n\},$$
$$C_2=C-C_1$$

Llamando a todos los casos cuya clase está en C_1 objetos de la clase 1, y poniendo todos los casos en C_2 dentro de la clase 2.

Para cualquier partición s del nodo, se calcula $i(s,t)$ como si fuera un problema de dos clases. Realmente $i(s,t)$ depende de la selección de $C_1,$ así la notación

$$i(s,t,C_1)$$

es utilizada. Ahora se encuentra la división $s^*(C_1)$ la cual maximiza $i(s,t,C_1),$ entonces, se encuentra la superclase C_1^* la cual maximiza $i(s^*(C_1),t,C_1).$ La división utilizada sobre el nodo es $s^*(C_1^*).$

La idea es entonces, para cada nodo, seleccionar que conglomeración de clases dentro de dos superclases para que sea considerado como un problema de dos clases, el mayor decrecimiento en la impureza del nodo es realizado.

Esta aproximación al problema tiene una ventaja significativa: Proporciona divisiones estratégicas e informa al usuario de clases similares. Para cada nodo, clasifica las

clases dentro de esos dos grupos los cuales en algún sentido son más diferentes y arroja al usuario los resultados de un agrupamiento óptimo C_1^*, C_2^* así como la mejor división s^*.

La palabra estratégica es utilizada en el sentido que cerca de lo más alto del árbol, este criterio intenta agrupar gran cantidad de clases que son similares en alguna característica. Cerca de la parte inferior del árbol intenta aislar casos únicos. Para ilustrar, supóngase que en un problema de cuatro clases, originalmente las clases 1 y 2 fueron agrupadas juntas y separadas de las clases 3 y 4, resultando en un nodo con membresía

Clases:	1	2	3	4
No. de casos:	50	50	3	1

Entonces en la siguiente división de este nodo, el potencial más grande para el decrecimiento de la impureza sería separar la clase 1 de la clase 2.

El reconocimiento de palabras es un ejemplo de un problema en el cual Twoing podría funcionar efectivamente. Proporcionando, por ejemplo, 100 palabras, la primera división podría separar palabras monosilábicas de multisilábicas. Futuras divisiones podrían aislar aquellos grupos de palabras teniendo otras características en común.

Aunque el criterio Twoing parece más deseable con un gran número de clases, en estas situaciones que tienen una aparente desventaja en la eficiencia computacional. Por ejemplo, con J clases, existen 2^{J-1} divisiones distintas de C en dos superclases. Para $J = 10, \quad 2^{J-1} = 1000.$ Sin embargo, el siguiente resultado muestra, que el criterio Twoing puede ser reducido a una regla total, consiguiendo la misma eficiencia que el criterio de Gini.

TEOREMA 2.3. Bajo el criterio de las dos clases $p(1|t)\,p(2|t),$ para una división dada $s,$ una superclase $C_1(s)$ que maximiza

$$i(s,t,C_1)$$

es

$$C_1(s)=\{j: p(j|t_L)\geq p(j|t_R)\}$$

y

$$\max_{C_1} i(s,t,C_1)=\frac{p_L p_R}{4}\left[\sum_j |p(j|t_L)-p(j|t_R)|\right]^2$$

COROLARIO 2.1. Para cada nodo t y división s de t en t_L y $t_R,$ se define la función del criterio Twoing $\phi(s,t)$ por

$$\phi(s,t)=\frac{p_L p_R}{4}\left[\sum_j |p(j|t_L)-p(j|t_R)|\right]^2$$

Entonces la mejor división con el criterio Twoing $s^*(C_1^*)$ está dada por la división s^* la cual maximiza $\phi(s,t)$ y C_1^* está dada por

$$C_1^*=\{j: p(j|t_L^*)\geq p(j|t_R^*)\}$$

Donde $t_L^*,\ t_R^*$ son los nodos dados por la división s^*.

2.5.15 Probabilidad *a priori* y costos por clasificación errónea

Los parámetros que pueden ser elegidos en la estructura del árbol de clasificación incluyen las probabilidades *a priori* $\{\pi(j)\}$ y el costo variable por clasificación errónea $\{C(i|j)\}$.

2.5.15.1 Elección de probabilidades *a priori*

Las probabilidades *a priori* son un útil conjunto de elementos de información y una selección inteligente y ajuste de ellas pueden ayudar en la construcción de un deseable árbol de clasificación.

En algunos estudios, el conjunto de datos puede ser muy desequilibrado entre sus clases. Las probabilidades *a priori* pueden ser usadas para ajustar las tasas individuales por clasificación errónea de la clase en cualquier dirección deseada.

Si la elección inicial de las probabilidades *a priori* produce resultados cuestionables, se sugiere el desarrollo de algunos árboles exploratorios usando diferentes probabilidades *a priori*.

2.5.15.2 Costos variables de la clasificación errónea vía Gini

En general, si los costos variables por clasificación errónea $\{C(i \mid j)\}$ son especificados entonces la pregunta que se presenta es cómo incorporar esos costos dentro de la regla de división. Para el índice de Gini existe una simple extensión. Otra vez, se considera la regla de clasificación subóptima la cual, en el nodo t, asigna un objeto desconocido dentro de la clase j con probabilidad estimada $p(j \mid t)$. obsérvese que el costo esperado estimado utilizando esta regla es

$$\sum_{j,i} C(i \mid j)\, p(i \mid t)\, p(j \mid t) \tag{2.24}$$

Esta expresión es utilizada como la medida de Gini de la impureza del nodo $i(t)$ para los costos variables por la clasificación errónea.

En el problema de las dos clases, (2.24) se reduce a

$$\big(C(2 \mid 1) + C(1 \mid 2)\big)\, p(1 \mid t)\, p(2 \mid t)$$

Esto señala una dificultad, en la forma en la cual el índice de Gini trata con los costos variables. Los coeficientes de $p(i|t)p(j|t)$ en (2.24) es $\left(C(2|1)+C(1|2)\right)$. El índice por lo tanto depende sólo de la simetría y no ajusta apropiadamente a costos altamente no simétricos.

Otro problema más teórico, es que $i(t)$ definido por (2.24) no es necesariamente una función cóncava de las $\{p(i|j)\}$, y así $i(s,t)$ podría concebiblemente ser negativa para algunas o para todas las divisiones en S.

2.5.16 Árboles con probabilidad de clase vía Gini

2.5.16.1 Bases y marco

En algunos problemas, dado un vector de medida $\underline{x}$, se busca una estimación de la probabilidad que un caso esté en la clase j, $j=1,2,....J$.

Más precisamente, en términos del modelo de probabilidad dado en la sección 2.5.4, se supone que los datos son extraídos de una distribución de probabilidad

$$P(A,j)=P(\mathsf{X}\in A, Y=j)$$

Entonces se quiere construir estimaciones para las probabilidades

$$P(j|\underline{x})=P(Y=j|\mathsf{X}=\underline{x}),\quad j=1,2,...,J$$

En otras palabras, dado que se observa $\underline{x}$, estimar la probabilidad de que el caso esté en la clase j, $j=1,2,...,J$.

Para este tipo de problema, en lugar de construir reglas de clasificación, se desea desarrollar criterios del tipo

$$\mathsf{d}(\underline{x})=\left(d(1|\underline{x}),d(2|\underline{x}),...,d(J|\underline{x})\right)$$

Con $d(j \mid \underline{x}) \geq 0,\ \ j = 1,2,...,J,$ y $\sum_j d(j \mid \underline{x}) = 1,$ para toda $\underline{x}.$

Tales reglas serán llamadas estimadores de probabilidad de clase.

Obviamente, el mejor estimador para este problema, el cual se llamará el estimador de Bayes y se denota por $_B,$ es

$$\mathsf{d}_B(\underline{x}) = \left(P(1 \mid \underline{x}), P(2 \mid \underline{x}), ..., P(J \mid \underline{x})\right)$$

Una pregunta crítica es cómo medir la precisión de un estimador de la probabilidad para una clase arbitraria, para tal caso se tiene

DEFINICIÓN 2.23. La precisión de un estimador de la probabilidad para una clase se define por el valor

$$E\left[\sum_j \left(P(j \mid \mathsf{X}) - \mathsf{d}(j \mid \mathsf{X})\right)^2\right]$$

Sea $\ , \ Y$ en $X \ \ C$ teniendo la misma distribución que $P(A, j)$ y se definen nuevas variables $z_j, \ \ j = 1,2,...,J,$ por

$$z_j = \begin{cases} 1 & Y = j \\ 0 & Y \neq j \end{cases}$$

Entonces

$$\begin{aligned} E(z \mid \mathsf{X} = \underline{x}) &= P(Y = j \mid \mathsf{X} = \underline{x}) \\ &= P(j \mid \underline{x}) \end{aligned}$$

Sea $\mathsf{d}(\underline{x}) = \left(d(1 \mid \underline{x}), d(2 \mid \underline{x}), ..., d(J \mid \underline{x})\right)$ cualquier estimador de la probabilidad de la clase.

DEFINICIÓN 2.24 El error cuadrático medio ECM $R^*(\ \tilde{}\ )$ de se define como

$$E\left[\sum_j \left(z_j - d(j \mid \mathsf{X})\right)^2\right]$$

Así, el ECM de es simplemente la suma de sus errores cuadráticos medios como un predictor de las variables $z_j,\ \ j = 1,2,...,J.$

La identidad clave es

PROPOSICIÓN 2.5. Para cualquier estimador de la probabilidad de la clase ,

$$R^*(\mathsf{d}) - R^*(\mathsf{d}_B) = E\left[\sum_j \left(P(j \mid \mathsf{X}) - d(j \mid \mathsf{X})\right)^2\right] \qquad (2.25)$$

2.5.16.2 Desarrollo y podado de árboles con probabilidad de la clase

Se asume que un árbol T ha sido desarrollado sobre una muestra de aprendizaje $(\underline{x}_n, j_n),\ \ n = 1,2,...,N,$ usando una regla de división no especificada y tiene el conjunto de nodos terminales $\tilde{T}.$

Asociado con cada nodo terminal t están las estimaciones por resustitución $p(j \mid t),$ $j = 1,2,...,J,$ para la probabilidad condicional de estando en la clase j dado el nodo $t.$

La forma natural de utilizar T como un estimador de la probabilidad de la clase es definiendo: si $\underline{x} \in t,$ entonces

$$\mathsf{d}(\underline{x}) = \left(p(1 \mid t), p(2 \mid t), ..., p(J \mid t)\right)$$

Estirando la notación, o T serán usados para denotar este estimador, dependiendo cual es más apropiado.

Para cada caso $(\underline{x}_n, j_n)$ en la muestra de aprendizaje, define J valores $\{z_{n,i}\}$ por

$$z_{n,i} = \begin{cases} 1 & j_n = i \\ 0 & j_n \neq i \end{cases}$$

Entonces la estimación por resustitución de $R(T)$ de $R^*(T)$ puede ser formado por este razonamiento: para todo $(\underline{x}_n, j_n)$ con $\underline{x} \in t,\ j_n = j,$

$$\begin{aligned} \sum_i \left(z_{n,i} - d(i,\underline{x})\right)^2 &= \left(1 - p(j\,|\,t)\right)^2 + \sum_{i \neq j} p^2(i\,|\,t) \\ &= 1 - 2p(j\,|\,t) + S \end{aligned}$$

Donde $$S = \sum_i p^2(i\,|\,t)$$

Entonces se pone

$$\begin{aligned} R(\mathrm{d}) &= \sum_{t \in \tilde{T}} \sum_j \left(1 - 2p(j\,|\,t) + S\right) p(j,t) \\ &= \sum_{t \in \tilde{T}} \sum_j \left(1 - 2p(j\,|\,t) + S\right) p(j\,|\,t)\, p(t) \end{aligned} \qquad (2.26)$$

Evaluando la suma sobre j en la última expresión da

$$R(\mathrm{d}) = \sum_{t \in \tilde{T}} (1 - S)\, p(t) \qquad (2.27)$$

y $$1 - S = 1 - \sum_j p^2(j\,|\,t)$$

Es exactamente el índice de la diversidad de Gini (2.23). Así desarrollando un árbol por el uso de la regla de división de Gini reduce al mínimo continuamente la estimación por resustitución $R(T)$ por el ECM. En consecuencia, el uso de la regla de partición de Gini es adoptada como la mejor estrategia para la construcción de un árbol con probabilidad de la clase.

La mayor diferencia entre el desarrollo de árboles de clasificación utilizando la regla de Gini y los árboles con probabilidad de clase está en el podado y el proceso de selección. Los árboles de clasificación son podados utilizando el criterio $R(T)+\alpha\left|\tilde{T}\right|$, donde

$$R(T)=\sum_{t\in\tilde{T}} r(t)\,p(t) \tag{2.28}$$

Y $r(t)$ es el costo por clasificación errónea dentro del nodo.

Los árboles por probabilidad de la clase son podados hacia arriba utilizando $R(T)+\alpha\left|\tilde{T}\right|$, *pero con* $r(t)$ *el índice de diversidad de Gini dentro del nodo*.

Sea $T_{\max}$ el desarrollo como antes, y podando hacia arriba, obteniendo la sucesión $T_1 \succ T_2 \succ \ldots \succ \{t_1\}$. Obtener las estimaciones de la prueba de la muestra $R^{ts}(T)$ de $R^{*}(T)$ para T cualquiera de los T_k, seguir con todos los $N_j^{(2)}$ para los casos de la clase j en la prueba de la muestra hacia abajo del árbol T. Se define

$$R_j^{st}(T)=\frac{1}{N_j^{(2)}}\sum_{n,i}\left(z_{n,i}-d\left(i\mid \underline{x}_n\right)\right)^2$$

Donde la suma es sobre los casos de la prueba de la muestra $N_j^{(2)}$. Entonces se pone

$$R^{st}(T)=\sum_j R^{st}(T)\pi(j) \tag{2.29}$$

Si las probabilidades *a priori* son datos estimados, las estimaciones para la prueba de la muestra de ellos son usados en la expresión anterior.

Si $T_1, T_2, \ldots, T_V$ son los V árboles para la validación-cruzada asociados con T, se denota por $\tilde{(v)}$, $v=1,2,\ldots,V$, los estimadores de la probabilidad de la clase correspondiente. Se define

$$R_j^{cv}(T) = \frac{1}{N_j} \sum_v \sum_{n,i} \left(z_{n,i} - d^v(i \mid \underline{x}) \right)^2 \qquad (2.30)$$

Donde la suma interior es sobre todos los casos de la clase j en la $v-ésima$ prueba de la muestra L_v. Ahora se pondrá

$$R^{cv}(T) = \sum_j R_j^{cv}(T) \pi(j) \qquad (2.31)$$

Si las probabilidades *a priori* son datos estimados, las estimaciones del total de la muestra de aprendizaje se utilizan para estimar las $\pi(j)$ en (2.31).

2.5.17 Árboles de regresión

2.5.17.1 Regresión por mínimos cuadrados

En regresión, un caso consiste de datos $(\underline{x}, y)$ donde $\underline{x}$ pertenece a un espacio de medida X y y es un número real. La variable y es generalmente llamada la respuesta o variable dependiente. Las variables en $\underline{x}$ son referidas como las variables independientes, las variables predictoras, etc.

Una regla de predicción o predictor es una función $d(\underline{x})$ definida sobre X tomando valores reales; esto es, $d(\underline{x})$ es una función de valor real sobre X.

Análisis de regresión es el término genérico girando alrededor de la construcción de un predictor $d(\underline{x})$ partiendo de una muestra de aprendizaje L. La construcción de un predictor puede tener dos propósitos: primero, predecir la variable respuesta correspondiente a un futuro vector de medida tan exacto como sea posible; segundo, comprender la relación estructural entre las variables medidas y de respuesta.

Se supone que una muestra de aprendizaje L consistente de N casos $(\underline{x}_1, j_1), (\underline{x}_2, j_2), ..., (\underline{x}_N, j_N)$ fue utilizada para construir un predictor $d(\underline{x})$. Entonces la

pregunta que surge es cómo medir la precisión de este predictor. Si se tuviera una prueba de la muestra muy grande $\left(\underline{x}_1^{'}, j_1^{'}\right),\left(\underline{x}_2^{'}, j_2^{'}\right),...,\left(\underline{x}_{N_2}^{'}, j_{N_2}^{'}\right)$ de tamaño N_2, la precisión de $d\left(\underline{x}\right)$ podría ser medida como el porcentaje de error

$$\frac{1}{N_2}\sum_{n=1}^{N_2}\left|y_n^{'} - d\left(\underline{x}_n^{'}\right)\right|$$

En $d\left(\underline{x}_n^{'}\right)$ como un predictor de $y_n^{'}$, $n = 1, 2, ..., N_2$. Pero por razones que tienen que ver con hacer los cálculos fáciles, la medida de la precisión clásicamente utilizada en regresión es el porcentaje de error cuadrado,

$$\frac{1}{N_2}\sum_{n=1}^{N_2}\left(y_n^{'} - d\left(\underline{x}_n^{'}\right)\right)^2$$

La metodología girando alrededor de esta medida es la regresión por mínimos cuadrados.

Para definir precisión en el sentido del error cuadrático medio, un marco teórico es necesario. Se asume que el vector aleatorio $\left(\mathsf{X}, Y\right)$ y la muestra de aprendizaje son independientemente extraídas de la misma distribución base.

DEFINICIÓN 2.25. Se define el error cuadrático medio $R^*\left(d\right)$ del predictor d como

$$R^*\left(d\right) = E\left(Y - d\left(\mathsf{X}\right)\right)^2$$

Esto es, $R^*\left(d\right)$ es el error cuadrático esperado utilizando $d\left(\mathsf{X}\right)$ como predictor de Y donde la esperanza es tomada manteniendo L fija.

Previamente, $R^*\left(d\right)$ fue utilizado para denotar la tasa de clasificación errónea del clasificador d. Se utiliza la misma notación aquí en el contexto de regresión para tener una notación uniforme para la medida de la precisión de un predictor, no obstante si se está prediciendo una etiqueta de clase o una respuesta ordenada.
Utilizando la definición anterior, el predictor óptimo (o Bayes) tiene una forma simple.

PROPOSICIÓN 2.6. El predictor d_B el cual minimiza $R^*(d)$ es

$$d_B(\underline{x}) = E(Y \mid \mathsf{X} = \underline{x})$$

En otras palabras, $d_B(\underline{x})$ es la esperanza condicional de la respuesta, dado que el vector de medida es $\underline{x}$. La superficie $y = d(\underline{x})$ es a menudo referida como la superficie de regresión de Y sobre .

LEMA 2.1. La constante a la cual minimiza $E(Y-a)^2$ es $E(Y)$.

2.5.17.2 Medidas del error y sus estimaciones

Dada una muestra de aprendizaje consistente de $(\underline{x}_1, j_1),(\underline{x}_2, j_2),...,(\underline{x}_N, j_N)$ nuevamente se desea utilizar L para construir un predictor $d(\underline{x})$ y estimar su error $R^*(d)$. Existen varias formas de estimar R^*. La general (y peor) es la estimación por resustitución

$$R(d) = \frac{1}{N}\sum_n \left(y_n - d(\underline{x}_n)\right)^2 \tag{2.32}$$

Las estimaciones de la prueba de la muestra $R^{ts}(d)$ son obtenidas por dividir aleatoriamente L en L_1 y L_2 y utilizando L_1 para construir d y L_2 para formar

$$R^{ts}(d) = \frac{1}{N_2}\sum_{(\underline{x}_n, y_n)\in L_2} \left(y_n - d(\underline{x}_n)\right)^2 \tag{2.33}$$

La estimación de V validación-cruzada $R^{cv}(d)$ se produce de dividir L en V subconjuntos $L_1, L_2, ..., L_V$, cada uno conteniendo (tan cercano como sea posible) el mismo número de clases. Para cada v, $v = 1, 2, ..., V$, se aplica el mismo procedimiento de construcción que para la muestra de aprendizaje $L - L_v$, obteniendo el predictor $d^{(v)}(\underline{x})$. Entonces se selecciona

$$R^{cv}(d) = \frac{1}{N}\sum_v \sum_{(\underline{x}_n, y_n)\in L_v} \left(y_n - d^{(v)}(\underline{x}_n)\right)^2 \tag{2.34}$$

El análisis razonado para R^{ts} y R^{cv} es el mismo que el discutido en la sección 2.5.4.

En clasificación, la tasa de clasificación errónea tiene una natural e intuitiva interpretación. Pero el error cuadrático medio de un predictor no.

Adicionalmente, el valor de $R^*(d)$ depende sobre la escala en la cual la respuesta es medida. Por esas razones, una medida normalizada de precisión la cual remueve la escala de dependencia es a menudo utilizada. Sea $\mu = E(Y).$ Entonces

$$R^*(\mu) = E(Y-\mu)^2$$

Es el error cuadrático medio utilizando la constante μ como un predictor de $Y,$ el cual es incluso la varianza de $Y.$

DEFINICIÓN 2.26. El error cuadrático medio relativo $RE^*(d)$ en $d(\underline{x})$ como un predictor de Y es

$$RE^*(d) = \frac{R^*(d)}{R^*(\mu)}$$

La idea aquí es que μ es el predictor de línea de fondo para Y si nada es conocido acerca de Entonces juzgando el funcionamiento de cualquier predictor d basado sobre por comparar su error cuadrático medio al de $\mu.$

El error relativo es siempre no negativo. Es generalmente, pero no siempre, menor a 1. Predictores más sensibles $d(\underline{x})$ son más precisos que $\mu,$ y $RE^*(d) < 1.$ Pero en ocasiones, algún procedimiento de construcción puede producir un pobre predictor d con $RE^*(d) \geq 1.$

Sean

$$\bar{y} = \frac{1}{N}\sum_n y_n \text{ y } R(\bar{y}) = \frac{1}{N}\sum_n (y_n - \bar{y})^2$$

Entonces el estimador de resustitución $RE(d)$ para $RE^*(d)$ es $\frac{R(d)}{R(\bar{y})}$. Utilizando la prueba de la muestra, la estimación es $RE^{ts}(d) = \frac{R^{ts}(d)}{R^{ts}(\bar{y})}$. La estimación por validación-cruzada $RE^{cv}(d)$ es $\frac{R^{cv}(d)}{R(\bar{y})}$.

En la regresión lineal por mínimos cuadrados, Si $d(\underline{x})$ es el mejor predictor lineal, entonces la cantidad

$$1 - RE(d)$$

es llamada la proporción de la varianza explicada por d. Adicionalmente, puede evidenciarse que si ρ es la correlación de la muestra entre los valores y_n y $d(\underline{x}_n)$, $n = 1, 2, ..., N,$ entonces

$$\rho^2 = 1 - RE(d).$$

En general, $R(d)$ no es una varianza y no tiene sentido referir $1 - RE(d)$ como la proporción de la varianza explicada. Tampoco es igual a el cuadrado de la correlación de la muestra entre los valores de y_n y $d(\underline{x}_n)$.

Por lo tanto, se prefiere el uso de la terminología *error relativo* y las estimaciones de $RE^*(d)$ como una medida de precisión en lugar de $1 - RE^*(d)$.

2.5.17.3 Estimaciones del error estándar

Sea la muestra de aprendizaje consistente de N_1 casos independientemente seleccionada de una distribución de probabilidad base, y supóngase que es utilizada para construir un predictor $d(\underline{x}_n)$. La prueba de la muestra consiste de N_2 casos extraídos independientemente de la misma distribución y se denota por $(X_1, Y_1), (X_2, Y_2), ..., (X_{N_2}, Y_{N_2})$. Entonces una estimación insesgada para $R^*(d)$ es

$$R^{ts} = \frac{1}{N_2}\sum_{n=1}^{N_2}\left(Y_n - d\left(\mathbf{X}_n\right)\right)^2 \qquad (2.35)$$

Debido a que los términos individuales en la expresión anterior son independientes (conservando la muestra de aprendizaje fija), la varianza de R^{ts} es la suma de las desviaciones de los términos individuales. Todos los casos tienen la misma distribución, así que la varianza de cada término es igual a la dispersión del primer término. Así, la desviación estándar de R^{ts} es

$$\frac{1}{\sqrt{N_2}}\left\{E\left(Y_1 - d\left(\mathbf{X}_1\right)\right)^4 - \left[E\left(Y_1 - d\left(\mathbf{X}_1\right)\right)^2\right]^2\right\}^{\frac{1}{2}}$$

Ahora se utilizan las estimaciones de momentos de la muestra

$$E\left(Y_1 - d\left(\mathbf{X}_1\right)\right)^4 \cong \frac{1}{N_2}\sum_{n=1}^{N_2}\left(Y_n - d\left(\mathbf{X}_n\right)\right)^4$$

y

$$E\left(Y_1 - d\left(\mathbf{X}_1\right)\right)^2 \cong \frac{1}{N_2}\sum_{n=1}^{N_2}\left(Y_n - d\left(\mathbf{X}_n\right)\right)^2 = R^{ts}$$

Estimando el error estándar de R^{ts} por

$$SE\left(R^{ts}\right) = \frac{1}{\sqrt{N_2}}\left[\frac{1}{N_2}\sum_{n=1}^{N_2}\left(Y_n - d\left(\mathbf{X}_n\right)\right)^4 - \left(R^{ts}\right)^2\right]^{\frac{1}{2}}$$

Debido a que el cuarto momento de la muestra puede ser altamente variable, menos crédito debería ser dado al error estándar en regresión que en clasificación. La medida RE^{*} es un cociente, y sus estimaciones RE^{ts} y RE^{cv} son estimadores de cocientes con fórmulas más complicadas para sus errores estándar.

Proporcionalmente, el error estándar de RE^{st}, pueden ser más grandes que esos de R^{ts}. Esto es por que la variabilidad de $\frac{R^{ts}(d)}{R^{ts}\left(\bar{y}\right)}$ es afectada por la variabilidad de

ambos, el numerador y el denominador y por la interacción entre ellas. A veces, la variabilidad del denominador es el factor dominante.

2.5.17.4 Métodos de regresión actuales

En regresión, los predictores han sido generalmente construidos utilizando una vía paramétrica. La suposición es hecha por

$$E\left(Y \mid \mathrm{X} = \underline{x}\right) = d\left(\underline{x}, \theta\right)$$

Donde d tiene forma funcional conocida dependiente de $\underline{x}$ y un conjunto finito de parámetros $\theta = \left(\theta_1, \theta_2, ...\right)$. Entonces θ es estimada como ese valor parametral $\hat{\theta}$ el cual minimiza $R\left(d\left(\underline{x}, \hat{\theta}\right)\right)$; esto es,

$$R\left(d\left(\underline{x}, \hat{\theta}\right)\right) = \min_{\theta} R\left(d\left(\underline{x}, \theta\right)\right)$$

Por ejemplo, en regresión lineal las suposiciones son que $\underline{x} = \left(x_1, x_2, ..., x_M\right)$ con $x_1, x_2, ..., x_M$ variables ordenadas y que

$$E\left(Y \mid \mathrm{X} = \underline{x}\right) = b_0 + b_1 x_1 + ... + b_M x_M$$

Donde los coeficientes $b_0, b_1, ..., b_M$ deben ser estimados. Asumiendo adicionalmente que el término del error es normalmente distribuido $N\left(0, \sigma^2\right)$ e independientemente de caso a caso llevando a una teoría inferencial.

Pero el objetivo es sobre conjuntos de datos cuya dimensionalidad requiere de alguna clase de selección de variables. En regresión lineal la práctica común es utilizar una elección por etapas o un algoritmo de subconjuntos mejores. Debido a que la selección de variables invalida el modelo inferencial, los métodos de regresión por fases o subconjuntos mejores tienen que ser vistos como herramientas heurísticas de análisis de datos.

Sin embargo, la regresión lineal con selección de variables puede ser una herramienta más poderosa y flexible que el análisis de discriminante en clasificación. Las suposiciones necesarias para una buena ejecución son mucho menos rigurosas, su comportamiento ha sido extensamente explorado, las herramientas de diagnóstico para comprobar la bondad del ajuste se están convirtiendo en disponibles, y constantemente se están robusteciendo los programas que hay disponibles.

Por lo tanto, los árboles estructurados por regresión como una alternativa para la regresión lineal debe ser observada en alguna forma diferente que el árbol estructurado por clasificación y utilizado en aquellos problemas donde sus características distintivas son deseables.

2.5.18 Árbol estructurado por regresión

Un predictor de árbol estructurado es similar a un clasificador de árbol estructurado. El espacio X es dividido por una sucesión de particiones binarias dentro de nodos terminales (Figura 13). En cada nodo terminal $t,$ el valor predicho de la respuesta $y(t)$ es constante.

Debido a que el predictor $d(\underline{x})$ es constante sobre cada nodo terminal, el árbol puede ser imaginado como una estimación del histograma de la superficie de regresión (Figura 14).

Comenzando con una muestra de aprendizaje $L,$ tres elementos son necesarios para determinar un árbol predictor:

1. Una forma para seleccionar una partición para cada nodo intermedio.
2. Una regla para determinar cuando un nodo es terminal.
3. Una regla para asignar un valor $y(t)$ para cada nodo terminal $t.$

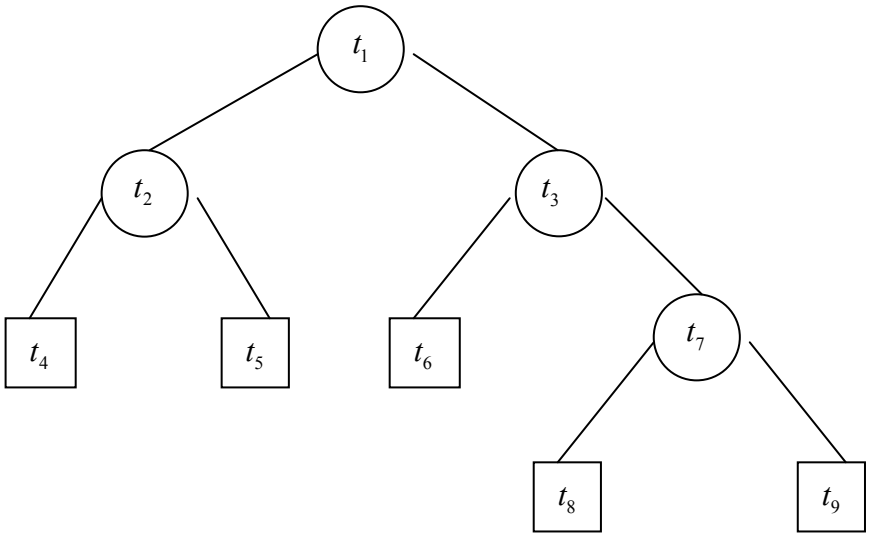


Figura 13

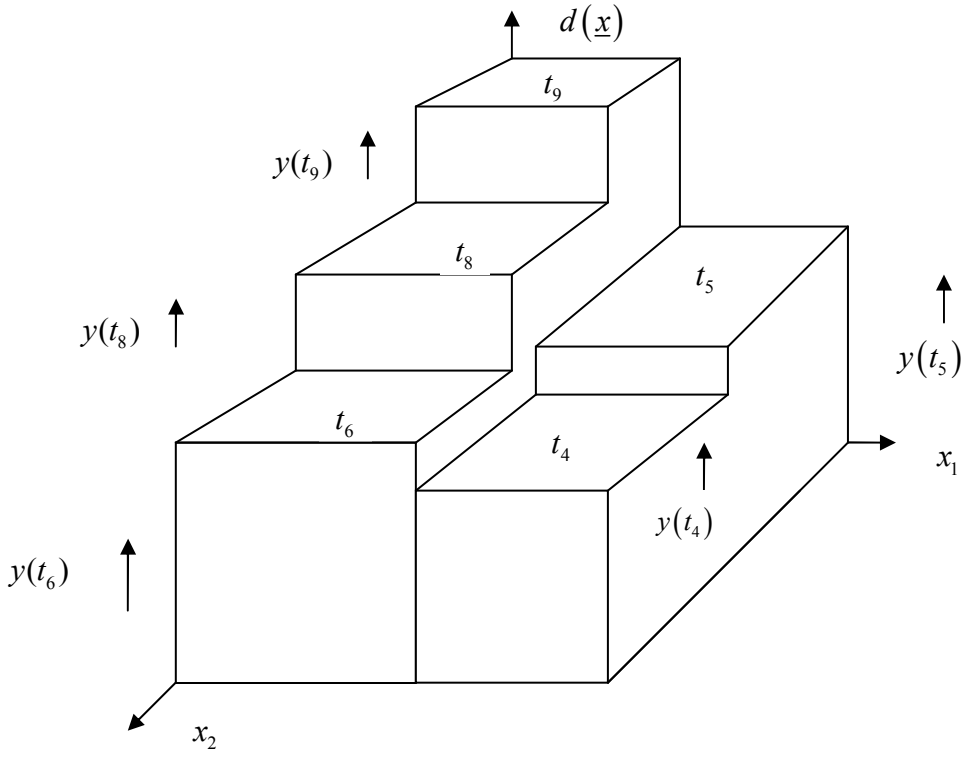


Figura 14

Resulta, como en la clasificación, que el resultado de la regla de asignación del nodo es lo más fácil de resolver.

Se comienza con la estimación por resustitución para $R^*(d)$, esto es,

$$R(\ )\ R(d) = \frac{1}{N}\sum_{n}\left(y_n - d\left(\underline{x}_n\right)\right)^2$$

Entonces se elige $y(t)$ para minimizar $R(d)$.

PROPOSICIÓN 2.7. El valor de $y(t)$ que optimiza $R(d)$ es el porcentaje de y_n para todos los casos $(\underline{x}_n, y_n)$ cayendo en t; esto es, la $y(t)$ mínima es

$$\bar{y}(t) = \frac{1}{N(t)}\sum_{\underline{x}_n \in t} y_n,$$

Donde la suma es sobre todas las y_n tales que $\underline{x}_n \in t$ y $N(t)$ es el número total de casos en t.

La demostración de la proposición 2.7 se basa en el valor que minimiza $\sum_{n}(y_n - a)^2$ el cual es

$$a = \frac{1}{N}\sum_{n} y_n$$

Similarmente, para cualquier subconjunto $y_{n'}$, de las y_n, el valor que minimiza $\sum_{n'}\left(y_{n'} - a\right)^2$ es el promedio de las $y_{n'}$.

De ahora en adelante, se elegirá el valor predicho en cualquier nodo t para ser $\bar{y}(t)$. Entonces utilizando la notación $R(t)$ en lugar de $R(d)$,

$$R(T) = \frac{1}{N}\sum_{t\in\tilde{T}}\sum_{\underline{x}_n \in t}\left(y_n - \bar{y}(t)\right)^2 \qquad (2.36)$$

Eligiendo

$$R(T) = \frac{1}{N}\sum_{\underline{x}\in t}\left(y_n - \bar{y}(t)\right)^2$$

Así (2.36) puede ser escrito como

$$R(T) = \sum_{t \in \tilde{T}} R(t) \qquad (2.37)$$

Estas expresiones tienen interpretaciones simples. Para cada nodo t, $\sum_{\underline{x} \in t} \left(y_n - \bar{y}(t) \right)^2$ está en la suma de cuadrados dentro del nodo. Esto es, está en las desviaciones cuadradas totales de y_n en t de su promedio.

Dado cualquier conjunto de divisiones S de un nodo terminal actual $t \in \tilde{T}$,

DEFINICIÓN 2.27. La mejor partición s^* de t es esa división en S en la cual mas decrece a $R(T)$.

Para cualquier división s de t en t_L y t_R, sea

$$R(s,t) = R(t) - R(t_L) - R(t_R)$$

Eligiendo la mejor partición s^* para ser una división tal que

$$R(s^*,t) = \max_{s \in S} R(s,t)$$

Así, un árbol por regresión es formado por divisiones iterativas de los nodos de tal forma que maximizan el decrecimiento en $R(T)$. En los árboles de clasificación, la elección de las mejores divisiones para ser las que minimizando la tasa de clasificación errónea de resustitución tuvieron propiedades indeseables. Criterios alternativos tuvieron que ser encontrados.

El criterio de regresión natural es el más sólido, no existen problemas similares en la definición de la mejor partición para ser la que minimice la medida del error por resustitución. Utilizando esta regla, la mejor partición para un nodo es esa división sobre las variables $\bar{x}$ la cual separa exitosamente el alto valor de la respuesta de las cantidades bajas.

Una forma alternativa del criterio es el siguiente. Sea $p(t)=\frac{N(t)}{N}$ la estimación por resustitución para la probabilidad de que un caso elegido aleatoriamente de la distribución teórica base caiga dentro del nodo t. Se define

$$s^2(t)=\frac{1}{N(t)}\sum_{\underline{x}\in t}\left(y_n-\bar{y}(t)\right)^2 \tag{2.38}$$

Así que $R(t)=s^2(t)\,p(t),$ y

$$R(T)=\sum_{t\in\tilde{T}}s^2(t)\,p(t) \tag{2.39}$$

Se observa que $s^2(t)$ es la varianza de la muestra de los valores y_n en el nodo t. Entonces la mejor división de t que minimiza el peso de la varianza

$$p_L s^2(t_L)+p_R s^2(t_R)$$

donde p_L y p_R representan la proporción de casos en t que van a la izquierda y a la derecha respectivamente.

2.5.19 Podado y estimación

2.5.19.1 Podado

Un nodo fue declarado terminal si

$$\max_s R(s,t)\le\beta R(t_1)$$

Entonces las estimaciones por resustitución $R(T)$ o $1-RE(T)$ fueron usadas como medidas de la precisión. Las dificultades son las mismas que en la clasificación. Los árboles desarrollados no son de tamaño correcto y las estimaciones son

excesivamente optimistas. La medida del error $R(t)$ tiene la propiedad que para cualquier división de t en $t_L,\ t_R$,

$$R(t) \geq R(t_L) + R(t_R)$$

Otra vez, entre más divisiones son hechas, mejor parece $RE(T)$.

El método usado para elegir un árbol es exactamente el mismo como el utilizado para seleccionar un árbol de clasificación. Primero, un árbol grande $T_{\max}$ es desarrollado por divisiones sucesivas tales que minimizan $R(T)$, hasta que para cada $t \in \tilde{T}_{\max}$, $N(t) \leq N_{\min}$. Generalmente, $N_{\min}$ es elegido como 5. Esos árboles iniciales son usualmente mucho más grandes que los árboles iniciales de clasificación.

En clasificación, las divisiones se detienen si el nodo es puro o $N(t) \leq N_{\min}$. El correspondiente, pero menos frecuentemente satisfecha, condición de pureza para árboles de regresiones que todos los valores de y en un nodo son el mismo.

Se define la medida del error-complejidad $R_\alpha(T)$ como

$$R_\alpha(T) = R(T) + \alpha\left|\tilde{T}\right|$$

Ahora el podado por error-complejidad mínimo es hecho exactamente como el podado por costo-complejidad mínimo en clasificación. El resultado es una sucesión decreciente de árboles

$$T_1 \succ T_2 \succ \ldots \succ \{t_1\}$$

Con $T_{\max} \succ = T_1$ y una correspondiente sucesión creciente de valores de α

$$0 = \alpha_1 < \alpha_2 < \ldots$$

tales que para $\alpha_k \le \alpha \le \alpha_{k+1}$, T_k es el subárbol más pequeño de $T_{\max}$ minimizando $R_\alpha(T)$.

2.5.19.2 Estimaciones de $R^*(T_k)$ y $RE^*(T_k)$

Para seleccionar el árbol de tamaño correcto de la sucesión $T_1 \succ T_2 \succ ... \succ \{t_1\}$, estimadores honestos de $R(T_k)$ son necesarios. Para conseguir estimaciones de la prueba de la muestra, los casos en L son divididos aleatoriamente dentro de una muestra de aprendizaje L_1 y una prueba de la muestra L_2. La muestra de aprendizaje L_1 es utilizada para desarrollar la sucesión $\{T_k\}$ de árboles podados, $d_k(\bar{x})$ denota el correspondiente predictor para el árbol T_k. Si L_2 tiene N_2 casos, se define

$$R^{ts}(T_k) = \frac{1}{N_2} \sum_{(\underline{x}_n, y_n) \in L_2} \left(y_n - d_k(\underline{x}_n)\right)^2$$

En la práctica, se ha utilizado generalmente validación-cruzada excepto con conjuntos grandes de datos. En V validación-cruzada L es dividida aleatoriamente dentro de $L_1, L_2, ..., L_V$ tales que cada submuestra L_v, $v = 1, 2, ..., V$, tiene el mismo número de casos (tan cercano como sea posible).

Sea $L^{(v)} = L - L_v$, la $v-ésima$ muestra de aprendizaje y se repite el desarrollo del árbol y el procedimiento de podado utilizando $L^{(v)}$. Para cada v, esto produce los árboles $T^{(v)}(\alpha)$ los cuales son los árboles por error-complejidad mínima para el valor del parámetro α.

El desarrollo y podado utilizando el total de L, produce las sucesiones $\{T_k\}$ y $\{\alpha_k\}$. Se define $\alpha_k' = \sqrt{\alpha_k \alpha_{k+1}}$. Se denota por $d_k^{(v)}(\underline{x})$ a el predictor correspondiente a el árbol $T^{(v)}(\alpha_k')$. Las estimaciones por validación-cruzada $R^{cv}(T_k)$ y $RE^{cv}(T_k)$ son dadas por

$$R^{cv}(T_k) = \frac{1}{N} \sum_{v=1}^{V} \sum_{(\underline{x}_n, y_n) \in L_v} \left(y_n - d_k^{(v)}(\underline{x})\right)^2$$

y

$$RE^{cv}\left(T_k\right)=\frac{R^{cv}\left(T_k\right)}{R\left(\overline{y}\right)}$$

2.5.19.3 Selección del árbol

En regresión, la sucesión $T_1 \succ T_2 \succ ... \succ \{t_1\}$ tiende a ser más grande que en clasificación.

El proceso de poda en árboles de regresión usualmente elimina solo dos nodos terminales a la vez. Esto contrasta con el planteamiento de clasificación, donde las ramas más grandes son podadas, resultando en una sucesión más pequeña de subárboles. El mecanismo es ilustrado por el siguiente ejemplo de clasificación hipotético. La Figura 15 ilustra una rama actual de un árbol comenzando desde un nodo intermedio. Hay dos clases con iguales probabilidades *a priori* y los números en los nodos son las poblaciones en las clases.

Si los dos nodos terminales de la extrema izquierda son podados, el resultado es ilustrado en la Figura 16. Hay un total de 100 clasificaciones erróneas en esta rama. Pero en el nodo superior, hay también 100 clasificaciones erróneas. Por lo tanto, el nodo superior, por si mismo, es una rama más pequeña teniendo la misma tasa de clasificación errónea como la configuración de tres nodos en la Figura 16.

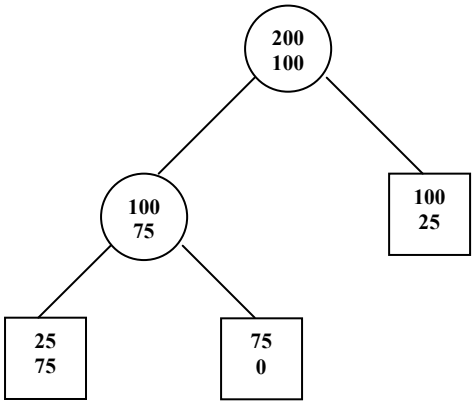


Figura 15

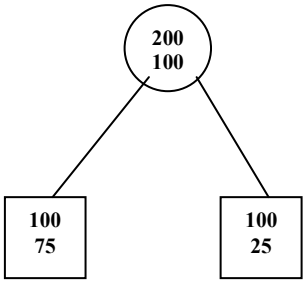


Figura 16

En consecuencia, si los dos nodos de extrema izquierda en la Figura 15 fueran podados, la rama entera seria podada.

En el procedimiento de clasificación, una división casi siempre decrece la impureza $\mathrm{I}(t),$ pero por lo que apenas ha sido ilustrado, pudiera no decrecer $R(T).$

En árboles de regresión, una división casi siempre decrece $R(T).$ Debido a que se está podando ascendentemente sobre $R(T),$ en general sólo dos nodos terminales a la vez serán podados.

No sólo es la sucesión de árboles generalmente más grande en regresión, también el valle conteniendo el valor mínimo $R^{cv}\left(T_k\right)$ tiende a ser aplanado y más ancho.

En armonía con la filosofía de seleccionar el árbol más pequeño conmensurado con la precisión, la regla SE1 fue usada. Esto es, el T_k seleccionado fue el árbol más pequeño tal que

$$R^{cv}\left(T_k\right) \leq R^{cv}\left(T_{k_0}\right) + SE$$

Donde

$$R^{cv}\left(T_{k_0}\right) = \min_k R^{cv}\left(T_k\right)$$

Y SE es la estimación del error estándar para $R^{cv}\left(T_{k_0}\right).$

2.5.20 Resultados de validación cruzada

2.5.20.1 El problema del árbol pequeño

La carencia de precisión en los árboles pequeños fue percibida en el procedimiento de clasificación, pero en grado inferior. En árboles de regresión el efecto puede ser más severo.

En general, siempre que existe un árbol T_k en la sucesión del árbol principal que es óptimo sobre un rango estrecho comparativo de α (α_k, α_{k+1}), se puede esperar que unos de los árboles por validación-cruzada tengan algunos nodos terminales más que el árbol principal. La estimación de la validación-cruzada RE^* entonces será sesgada ascendentemente hacia los valores RE^* correspondientes a T_{k+1}. Si el valor de T_{k+1} es considerablemente más grande que el de T_k, el sesgo puede ser mas grande. Así, el efecto es más pronunciado en los árboles más pequeños donde RE^* es incrementado rápidamente.

Las fuentes del sesgo en árboles pequeños son árboles por validación-cruzada que tienen algunos nodos terminales que el correspondiente árbol principal y prueba de la muestra no balanceada.

Lo anterior podría ser remediado por seleccionar los árboles por validación-cruzada para tener, tan cercano como sea posible, el mismo número de nodos terminales como el árbol principal; esto último por estratificar los casos por sus valores de y y seleccionando la prueba de las muestras combinando muestras separadas de cada estrato. Ambos han sido intentados.

Utilizando árboles por validación-cruzada con el mismo número de nodos como el árbol principal y estratificando los conjuntos de prueba reduce el sesgo en las estimaciones. Sin embargo, un remarcado sesgo ascendente permanece. Hasta cierto punto, esto parece ser dependiente del conjunto de datos. Cuando otros principios fueron utilizados para generar datos del mismo modelo, el efecto del árbol pequeño estuvo generalmente presente pero no tan pronunciado.

Para cualquier tasa, el árbol SE1 siempre ha estado en el rango donde las estimaciones por validación cruzada son razonablemente precisas.

2.5.21 Árboles con estructura estándar

Como en clasificación, se dice que los datos tienen estructura estándar si el espacio de medida X es de dimensión fija M, $\underline{x} = (x_1, x_2, ..., x_M)$, donde las variables pueden ser ordenadas o nominales.

El conjunto estándar de divisiones entonces consiste de todas las divisiones de la forma $¿x_m < c?$ en variables ordenadas y $¿x_m \in S?$ para variables nominales donde S es cualquier subconjunto de las categorías.

Variables categóricas en regresión con estructura estándar pueden ser manejadas utilizando lo siguiente. Si $x_m \in \{b_1, b_2, ..., b_L\}$ es nominal, entonces para cualquier nodo t, se define $\bar{y}(b_l)$ como el promedio sobre todas las y_n en el nodo tal que la $m-ésima$ coordenada de $\bar{x}_n$ es b_l. Ordenándolas de tal forma que

$$\bar{y}b_{l_1} \leq \bar{y}b_{l_2} \leq \bar{y}b_{l_L}$$

PROPOSICIÓN 2.8. La mejor división en x_m en el nodo t es una de las $L-1$ divisiones

$$x_m \in \{b_{l_1}, b_{l_2}, ..., b_{l_h}\}, \quad h = 1, 2, ..., L-1$$

Esto reduce la búsqueda para el mejor subconjunto de categorías de $2^{L-1}-1$ a $L-1$ subconjuntos.

Si la estructura de los datos es clara, entonces el árbol tiende a proporcionar un dibujo estable de la estructura. Inestabilidades en el árbol reflejan variables correlacionadas, reglas de predicción alternativas y ruido.

2.5.22 Desviación estándar dentro del nodo

Para ilustrar un punto, se asume que los datos son generados de un modelo de la forma

$$Y = g(\mathrm{X}) + \varepsilon$$

Donde ε es un término del error aleatorio independiente de con media cero.

Se denota la varianza de ε para $\mathrm{X} = \underline{x}$ por $\sigma^2(\underline{x})$. Si $\sigma^2(\underline{x})$ es constante, entonces el modelo es llamado homocedástico. En general, la varianza del error es distinta en diferentes partes del espacio de medida.

La carencia de homocedasticidad puede tener un desafortunado efecto en las estructuras de árbol. Por ejemplo, en un nodo dado t, la varianza interior del nodo $s^2(t)$ puede ser más grande comparada a otros nodos, aunque $\bar{y}(t)$ es una buena aproximación a la superficie de regresión en t. Entonces la búsqueda será para una división con ruido en una variable, resultando en un valor de $p_L s^2(t_L) + p_R s^2(t_R)$ considerablemente inferior que $s^2(t)$. Si tal partición puede ser encontrada, entonces puede ser conservada incluso cuando el árbol es podado ascendentemente.

Debido a que las divisiones sucesivas intentan minimizar las varianzas interiores del nodo, las estimaciones por resustitución $s^2(t)$ tenderán a ser levemente insesgadas.

Esto no es válido y no es incluso soportado heurísticamente. En primer lugar, como ha sido observado, las $s^2(t)$ son generalmente sesgadas. Un ajuste en las mismas, similar al de $r(t)$ en clasificación, ha sido explorado, pero la mejora fue marginal. Segundo, el error dominante a menudo no es el sesgo descendente de $s^2(t)$, sino que por el contrario $\bar{y}(t)$ es una estimación pobre de la media verdadera del nodo. Finalmente, el uso de $\bar{y}(t) \pm 2s(t)$ implícitamente invoca la suposición injustificable de que la distribución de los valores de y del nodo es normal.

Si los nodos terminales son más pequeños, más sesgadas son las estimaciones de $\bar{y}(t)$ y de $s^2(t)$. En situaciones donde es importante que esas estimaciones individuales del nodo sean precisas, es recomendable elegir árboles pequeños que tienen nodos terminales más grandes.

3. Resultados

Los 191 casos se dividieron por medio de observación directa en: indeterminados 43 casos, machos 64 y hembras 84, dado que se desea determinar el sexo (variable dependiente) es decir, clasificar a los individuos por el sexo (Objetivos). Para los individuos indeterminados fue asignado el número 0 para el sexo, para los machos le correspondió 1 y para las hembras el 2, es decir, C = {0, 1, 2}. La muestra de aprendizaje consistió de toda la muestra N = 191.

Las variables predictoras fueron 9: largo del caparazón (LC); diámetro del ojo (ED); largo del telsón (LT), largo del segundo segmento del primer pleópodo (L1ERPLEO) y segundo pleópodo (L2DOPLEO); largo del mero (LME) y largo del carpo LCA) del tercer pereiópodo y longitud total (LTO). De estás se determinará cuales son las que mejor describen el sexo.

Todo el análisis se llevó a cabo con el paquete Statistica 6. Se desarrollaron 6 árboles de clasificación por este método, debido a que Statistica 6 permitía elegir entre 3 tipos de medidas: Medida de Gini, Ji-cuadrada y G-cuadrada. Se eligieron las 2 primeras, esto para poder comparar entre ambos árboles, ya que sólo se está interesado en la Medida de Gini. También, es posible elegir entre probabilidades *a priori* iguales y estimadas.

3.1 Medida de Gini, probabilidades *a priori* estimadas

3.1.1 Probabilidades *a priori* y árbol de clasificación

Lo primero que se calcula son las probabilidades *a priori*. Son calculadas como la proporción de elementos observados en cada clase de la muestra de aprendizaje entre el total de la misma. La Tabla 3.1 muestra como fueron calculadas las probabilidades *a priori* estimadas.

Clase	Probabilidades *a priori*	Individuos en cada clase observada
0	$\pi_0 = 43/191 = 0.225131$	43
1	$\pi_1 = 64/191 = 0.335079$	64
2	$\pi_2 = 84/161 = 0.439791$	84

Tabla 3.1 Clase observada y probabilidades *a priori* para medida de GIni y probabilidades *a priori* estimadas. Las probabilidades *a priori* son estimadas de los datos. Muestra de aprendizaje N = 191

El árbol de clasificación generado por medio de la Medida de Gini y probabilidades *a priori* estimadas, es el representado en la Figura 17.

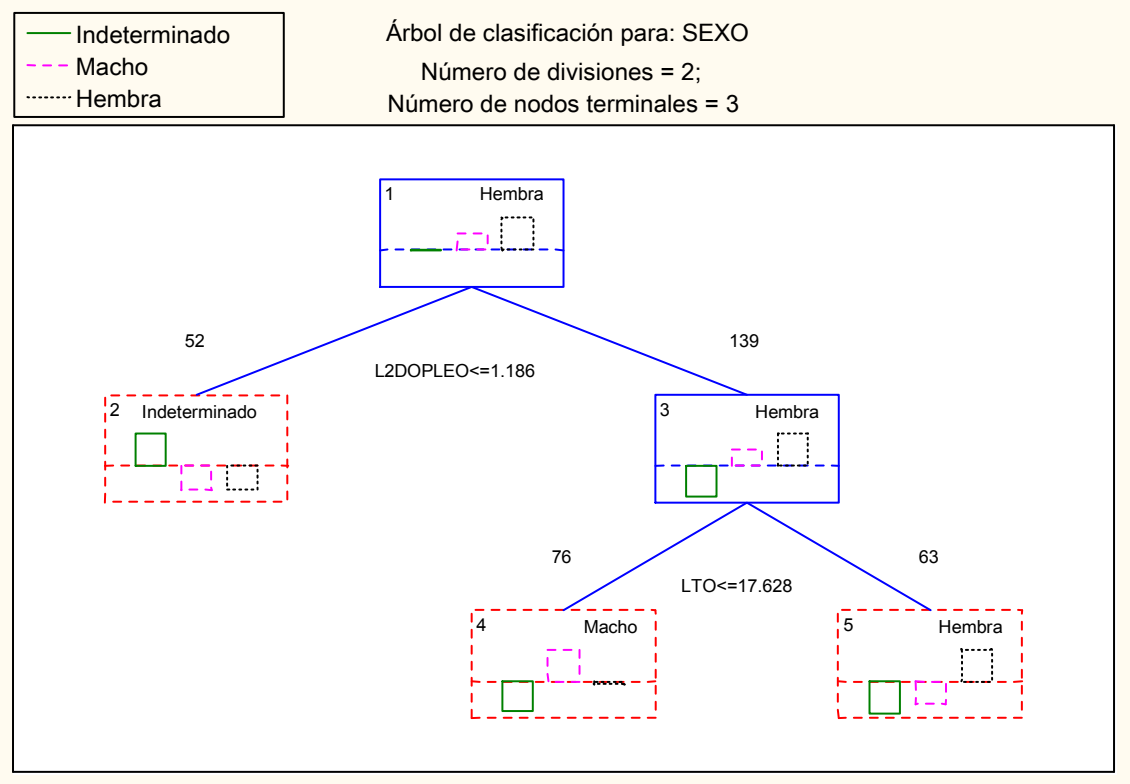


Figura 17. Árbol de clasificación para SEXO.
Medida de Gini y probabilidades *a priori* estimadas

El nodo raíz es dividido por medio de la variable longitud del segundo pleópodo (L2DOPLEO). Si la longitud indicada es menor o igual a 1.186 Mm., entonces son enviados al nodo izquierdo como indeterminados y si es mayor al nodo derecho como hembras. Esta primera partición produce los nodos 2 y 3 y se obtienen 52 individuos indeterminados y 139 hembras. El nodo 2 es terminal Figura 17.

El nodo 3 es dividido en los nodos 4 y 5. Ambos son nodos terminales. La partición es hecha por medio de la variable largo total (LTO). Si la medida de la longitud total del espécimen es menor o igual a 17.628 Mm., entonces desciende al nodo izquierdo como machos y al nodo derecho como hembras. El procedimiento envía 76 especímenes al nodo 4 como machos y 63 al nodo 5 como hembras.

Este árbol es el más pequeño posible, debido a que existen 3 clases y son particiones binarias.

3.1.2 Estructura del árbol de clasificación

En la Tabla 3.2 se muestra cómo fueron producidos los nodos descendientes y terminales a partir del nodo raíz.

Nodo	Rama izquierda	Rama derecha
1	2	3
2		
3	4	5
4		
5		

Tabla 3.2. Nodos descendientes.
Nodos descendientes para cada nodo
Medida de Gini y probabilidades *a priori* estimadas

El nodo raíz fue dividido en los nodos 2 y 3. El nodo 2 es terminal y el 3 fue dividido en los nodos 4 y 5. Ambos son terminales.

Es de suma importancia conocer cómo fueron repartidos en cada nodo los individuos, por ejemplo, para obtener la cantidad de individuos clasificados erróneamente. En la Tabla 3.3 se muestra la composición de cada nodo por individuos de cada clase observada.

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2
1	2	3	43	64	84
2			41	5	6
3	4	5	2	59	78
4			2	51	23
5			0	8	55

Tabla 3.3. Composición de clases. Nodos descendientes y número de individuos en clase observada para cada nodo. Medida de Gini y probabilidades *a priori* estimadas

Comenzando en el nodo raíz (nodo 1) con el total de los elementos (191 individuos) divididos en 43 de la clase observada 0, de la clase observada 1, 64 elementos y 84 de la clase observada 2.

La partición del nodo raíz envía al nodo 2 a 41 elementos de la clase observada 0, de la clase observada 1, 5 especímenes y 6 de la clase observada 2 para un total de 52 individuos. El nodo 3 esta compuesto por 2 elementos de la clase 0; de la clase 1, 59 especímenes y 78 unidades de la 2 para un total de 139 individuos.

La partición del nodo 3 envía al nodo 4 a los 2 últimos elementos de la clase observada 0; de la clase 1, 51 especímenes y 23 unidades de clase observada 2, para un total de 76 individuos; y al nodo 5, ocho elementos de la clase observada 1 y 55 de la 2, para un total de 63 individuos. Ambos nodos son terminales.

Para cada nodo es asignada una clase predicha. Esto es mostrado en la Tabla 3.4. Para el nodo terminal 2, la clase predicha es 0 con un total de 52 elementos; para el nodo terminal 4, la clase predicha es 1 con un total de 76 unidades y para el nodo terminal 5 la clase predicha es 2 con 63 individuos.

Las variables y las constantes elegidas para las divisiones son mostradas en la Tabla 3.5.

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2	Clase predicha
1	2	3	43	64	84	2
2			41	5	6	0
3	4	5	2	59	78	2
4			2	51	23	1
5			0	8	55	2

Tabla 3.4. Clases predichas. Nodos descendientes y número de individuos por clase observada y clase predicha por cada nodo. Medida de GIni y probabilidades *a priori* estimadas.

Nodo	Constante de división	Variable de división
1	-1.1860	L2DOPLEO
2		
3	-17.6279	LTO
4		
5		

Tabla 3.5. Condiciones de división. Condiciones de división para cada nodo. Medida de Gini y probabilidades *a priori* estimadas

En el nodo raíz se elige la longitud del segundo pleópodo para realizar la partición, así como la constante 1.1860 Mm. Si el espécimen tiene una distancia en su segundo pleópodo menor a la constante de división se envía al nodo 2 y si es mayor desciende al nodo 3. Esto genera la primera partición.

Para la segunda partición, en el nodo 3 es seleccionada la distancia total como variable de división y el valor 17.6279 Mm. Nuevamente, si es menor su longitud total al valor de la constante de clasificación, es enviado al nodo descendiente 4 y si es mayor al nodo 5.

Por último, en la Tabla 3.6 se resume toda la información de la sección.

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2	Clase predicha	Constante de división	Variable de división
1	2	3	43	64	84	2	-1.1860	L2DOPLEO
2			41	5	6	0		
3	4	5	2	59	78	2	-17.6279	LTO
4			2	51	23	1		
5			0	8	55	2		

Tabla 3.6. Estructura del árbol
Nodos descendientes, individuos en clase observada,
clase predicha y condiciones de división para cada nodo
Medida de Gini y probabilidades *a priori* estimadas.

3.1.3 Grado de importancia para las variables predictoras

Esta sección esta destinada a conocer el grado de importancia de cada variable en la clasificación. Seleccionando sólo las variables que determinan la clasificación del sexo, es posible determinar mejoras al procedimiento. La Figura 18 muestra el histograma con el nivel de jerarquía de cada variable en la selección.

Variable	Grado
LC	91
L1ERPLEO	67
L2DOPLEO	84
LTO	100
LCA	34
LME	32
ED	44
LT	71

Tabla 3.7. Grado de importancia para las variables predictoras
Basadas sobre divisiones univariadas
0 = Baja importancia; 100 = Alta importancia
Medida de Gini y probabilidades *a priori* estimadas

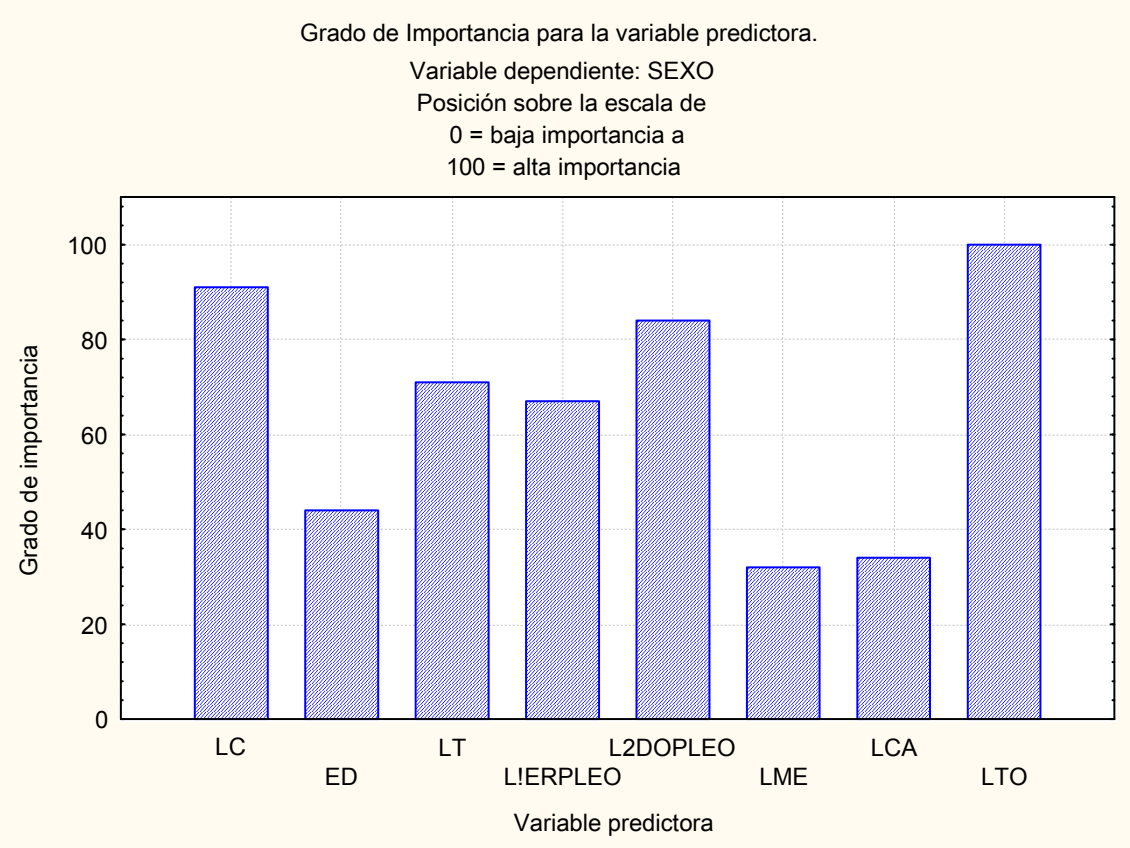


Figura 18. Medida de Gini y probabilidades *a priori* estimadas

La variable más importante es la distancia total del individuo con la mas alta escala, después la longitud del caparazón con 91 puntos. En tercer lugar esta la longitud del segundo pleópodo con 84 lugares y en cuarto la largo del telsón con 71 unidades (Tabla 3.7).

3.1.4 Sucesión de costos

Para determinar cual árbol de clasificación es el elegido, se calculan los costos por validación cruzada y por resustitución. En la Figura 19, se observa que el criterio de elección para el árbol de clasificación es aquel que se encuentra en inflexión de la curva. El árbol que se elige es el marcado con * debido a que es el que cumple con esta característica.

Examinando la Tabla 3.8 se observa que no se elige el árbol con valores menores (árbol número 6).

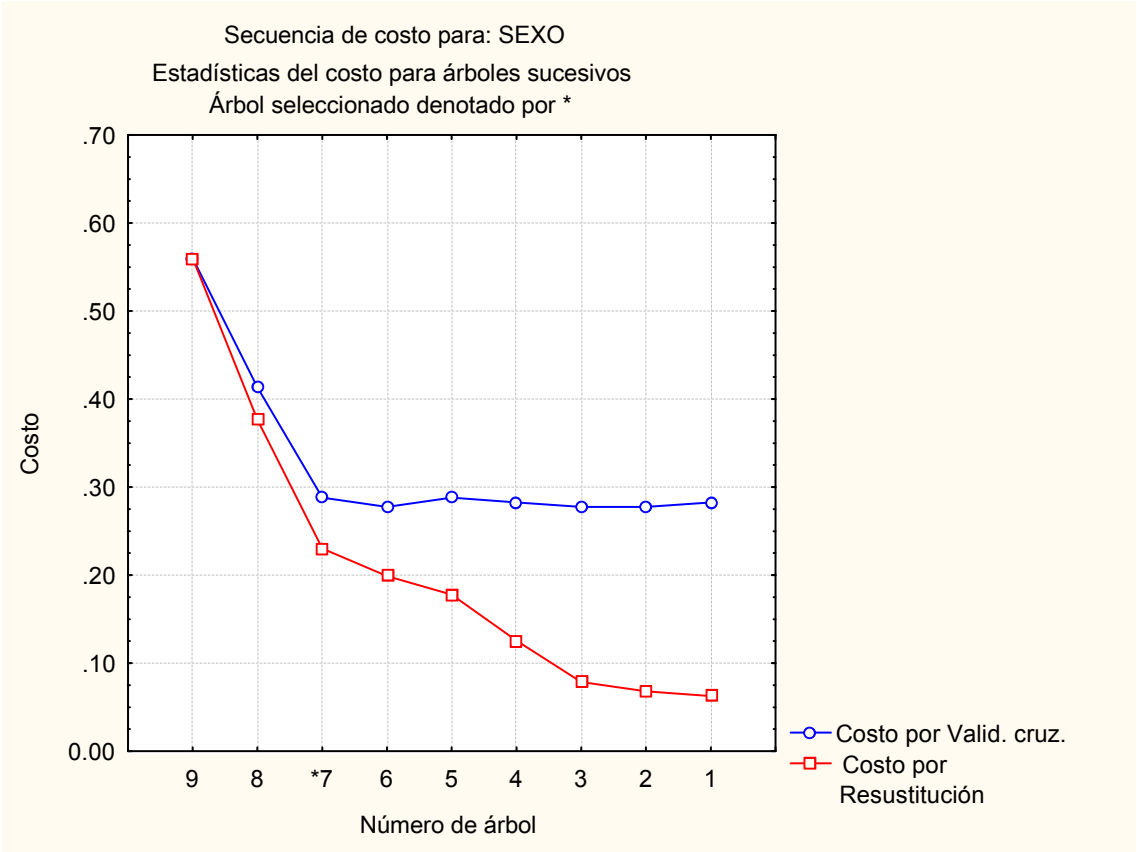


Figura 19. Medida de Gini y probabilidades *a priori* estimadas CV validación-cruzada

Número de árbol	Nodos terminales	Validación cruzada	Error estándar	Resustitución	Complejidad del nodo
1	20	0.287958	0.032764	0.062827	0.000000
2	18	0.282723	0.032584	0.068063	0.002618
3	16	0.282723	0.032584	0.078534	0.005236
4	10	0.282723	0.032584	0.125654	0.007853
5	5	0.287958	0.032764	0.178010	0.010471
6	4	0.277487	0.032399	0.198953	0.020942
***7**	3	0.287958	0.032764	0.230366	0.031414
8	2	0.413613	0.035635	0.376963	0.146597
9	1	0.560209	0.035915	0.560209	0.183246

Tabla 3.8. Sucesión de árboles. Estadísticas para árboles sucesivos.

Árbol seleccionado denotado por *

Medida de Gini y probabilidades *a priori* estimadas

En su lugar, es elegido el árbol con valores menores y que se encuentra en la parte más suave de la curva.

De hecho, el árbol 5 tiene los mismos valores que el árbol 7, pero sólo este último se encuentra en la parte suave de la curva. Para poder discernir entre el árbol 5 y 7 se puede observar en la misma tabla que el número de nodos terminales es distinto. Lo anterior produciría un costo más alto si se elige el primero de los considerados.

3.1.5 Número de individuos por clase observada

Para empezar a apreciar cuan preciso es el árbol por medio de la clasificación errónea, se necesita conocer como fueron clasificados los individuos en las clases predichas. Las siguientes tres secciones se avocan a este resultado.

Se empezará comparando el número de individuos en cada clase observada en cada nodo terminal. Se apoyará el análisis en la siguiente tabla.

Nodo	Clase 0	Clase 1	Clase 2
2	41	5	6
4	2	51	23
5	0	8	55

Tabla 3.9. Numero de individuos en clase observada por nodo terminal.
Medida de Gini y probabilidades *a priori* estimadas

El nodo terminal 2 contiene 41 elementos de la clase observada 0, 5 de la clase observada 1 y 6 de la clase observada 2, en total contiene 52 elementos.

El nodo terminal 4 contiene 2 individuos de la clase observada 0, 51 elementos de la clase observada 1 y 23 de la clase observada 2; para un total de 76 elementos.

El nodo terminal 5 contiene 0 individuos de la clase observada 0, 8 de la clase observada1 y 55 de la clase observada 2; para un total de 56 elementos.

Lo anterior se representa en el histograma de tercera dimensión de la Figura 20 y en la gráfica de dos dimensiones de la Figura 21.

En la Figura 20 la clase observada 0 es verde, la clase observada 1 es roja y la clase observada 2 es azul. A simple vista se observa que la clase observada 0 es la mejor clasificada, la peor clasificada es la clase observada 2.

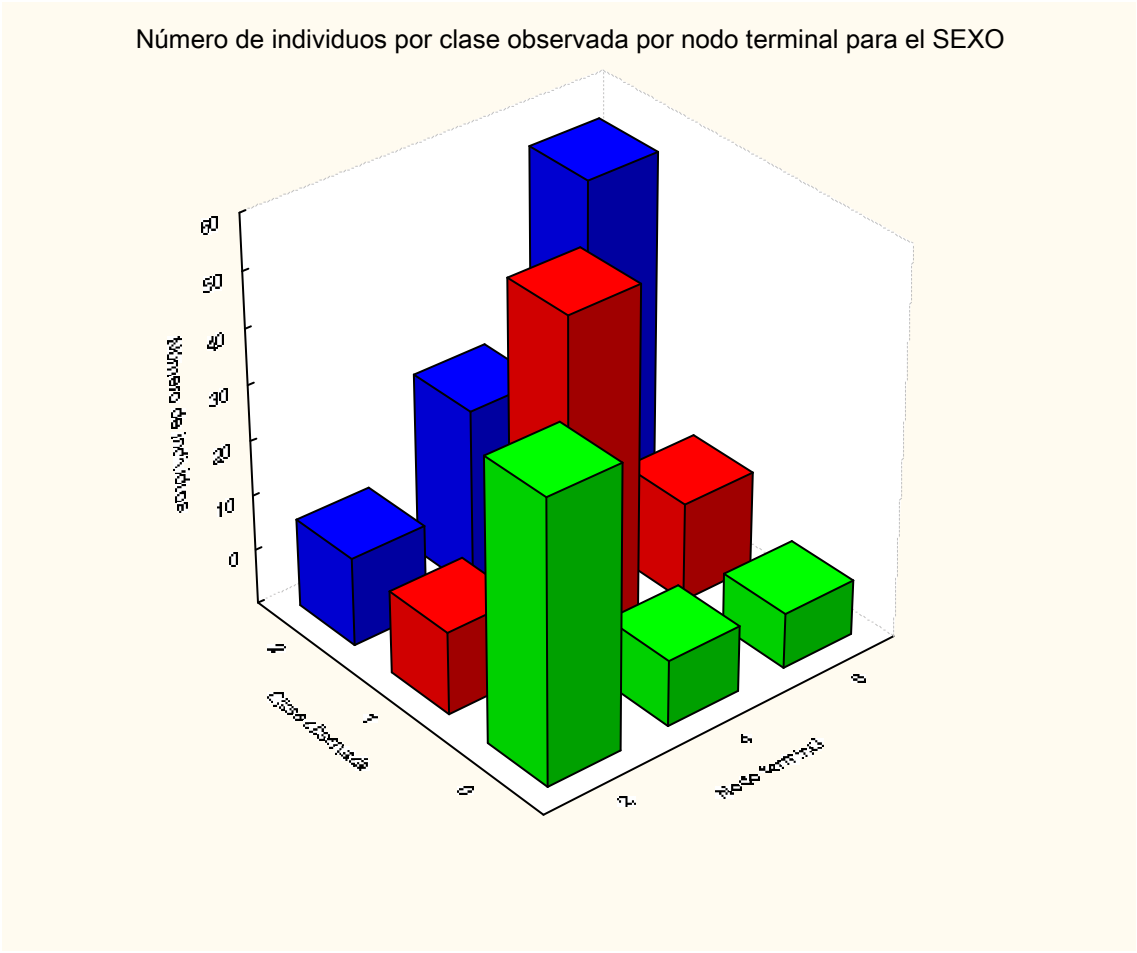


Figura 20. Medida de Gini y probabilidades *a priori* estimadas

Como puede observarse, en la diagonal está el mayor número de individuos de estas clases. Estos especímenes, como se determinará en las próximas secciones, son los clasificados correctamente.

No es deseable que fuera de la diagonal existan columnas demasiado grandes.

En la clase observada 2 y el nodo terminal 4 de la Figura 20, se presenta este fenómeno, debido a que estos elementos están erróneamente clasificados.

Lo anterior, se aprecia mejor en la Figura 21, donde, en tonalidades de verde, explica los nodos donde se encuentran el mayor número de individuos clasificados erróneamente.

La Figura 21 muestra una gráfica de 2 dimensiones o una vista aérea de la grafica de 3 dimensiones de la Figura 20.

El tono verde representa al número de individuos clasificados erróneamente. Un tono más intenso simboliza menos individuos.

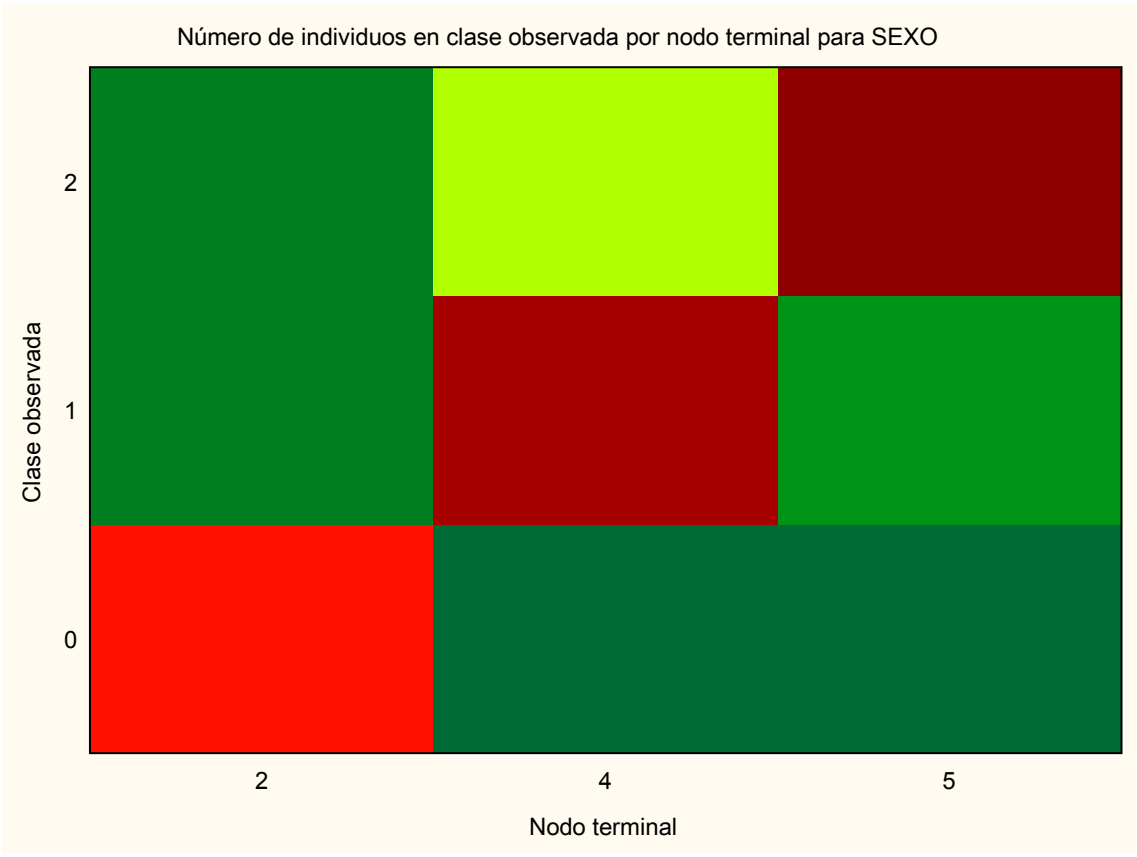


Figura 21.Medida de GIni y probabilidades *a priori* estimadas

Así, el tono amarillo (clase observada 2, intersección nodo terminal 4) representa el mayor número de individuos erróneamente clasificados (23).

El tono que sigue en matiz verde, es el de la clase observada 1 intersección el nodo terminal 5, con 8 individuos.

El siguiente es el correspondiente a las clases observadas 1 y 2 con el nodo terminal 2. El tono es el mismo porque sólo difieren en una unidad como puede comprobarse en la Tabla 5.9.

El último segmento de verde es el más intenso. Pertenece a la clase observada 0 con los nodos terminales 4 y 5.

Observando la Tabla 5.9 nuevamente, se concluye que es la clase mejor clasificada (41 individuos clasificados correctamente y 2 erróneamente).

3.1.6 Clase observada contra clase predicha

Con ayuda de la siguiente tabla se obtiene la primera comparación. En la diagonal se observan los individuos correctamente clasificados.

Clase	Clase 0	Clase 1	Clase 2
0	41	5	6
1	2	51	23
2	0	8	55

Tabla 3.10. Número de individuos en clase observada por clase predicha.
Medida de Gini y probabilidades *a priori* estimadas
Predicha (renglón) x observada (columna) matriz
Muestra de Aprendizaje N = 191

Fuera de la diagonal se observan los individuos erróneamente clasificados. A continuación se explica la Tabla 3.10 comenzando con la diagonal.

El número de individuos en la clase observada 0, clasificados correctamente en la clase predicha 0 es de 41 unidades. Ésta clase es la que mejor se clasifica, porque sólo 2 elementos no son reconocidos por este procedimiento como clase 0.

La proporción de individuos de la clase observada 0, clasificados correctamente en la clase predicha 0 es 0.95%, esto es calculado de la siguiente forma. Se divide el número de elementos clasificados correctamente (41), entre el número total en la clase observada 0 (43).

El número de unidades en la clase observada 1, clasificados correctamente en la clase predicha 1 es de 51. El porcentaje de individuos de la clase observada 1, clasificados correctamente en la clase predicha 1 es de 79.68%. Lo anterior es calculado dividiendo el número de individuos correctamente clasificados (51) entre la cantidad de elementos en la clase observada 1 (64).

El número de individuos en la clase observada 2, clasificados correctamente en la clase predicha 2 es de 55 unidades. La proporción con respecto al número de individuos en la clase observada 2 es de 65.47%.

Si se suman las tres cantidades de la diagonal principal, se obtiene el número total de individuos clasificados correctamente en la muestra. Realizando esta adición se obtienen 147 unidades clasificados correctamente de un total de 191. La proporción de elementos clasificados correctamente en la muestra es de 76.96%.

En la gráfica de tercera dimensión de la Figura 22 se observan las clases divididas por colores, la clase observada 0 es verde, la 1 es roja y la 2 es azul.

En la diagonal se colocan los individuos correctamente clasificados y fuera de esta los erróneamente clasificados.

El mayor porcentaje de clasificación correcta en las clases, pertenece a la clase observada 0, seguida de la clase observa 1 y al último la clase observada 2.

Para la Figura 23, se aprecia el error más grande en el tono verde limón y el matiz de verde más oscuro representa la mejor clasificación.

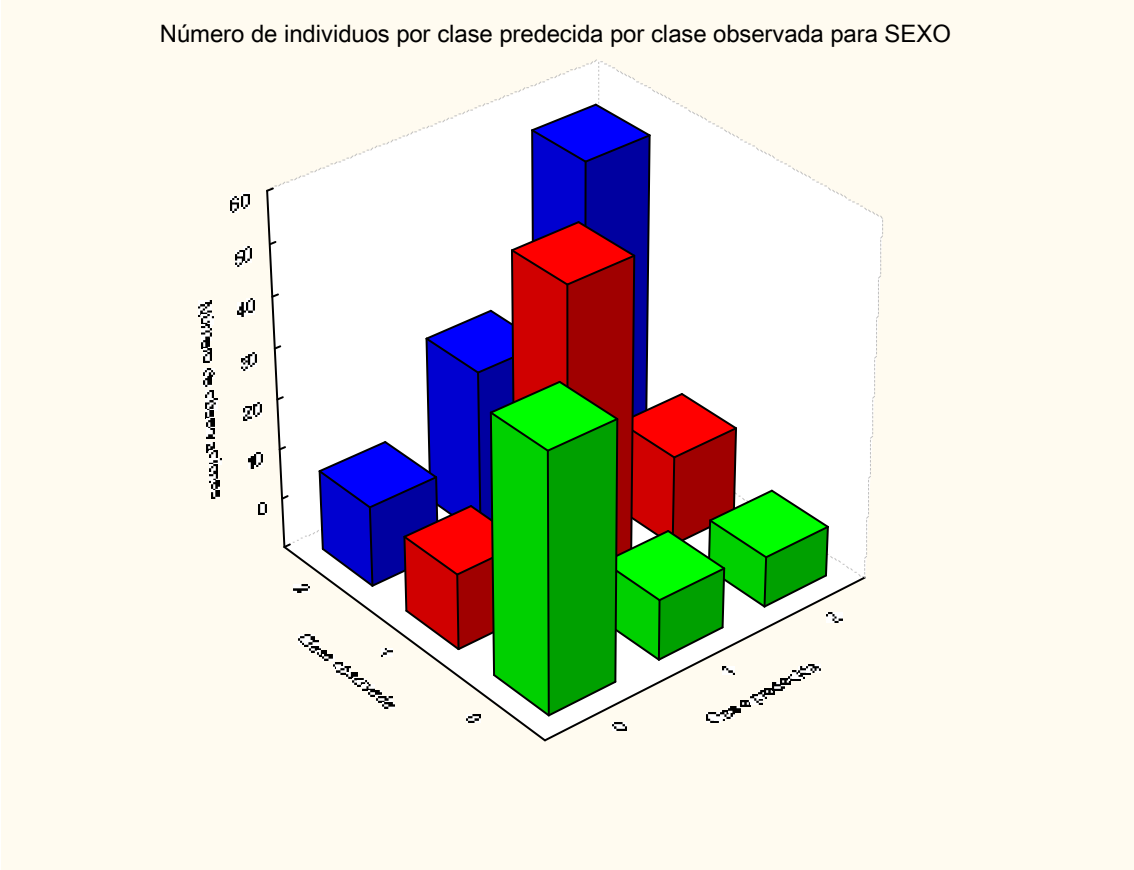


Figura 22. Medida de Gini y probabilidades *a priori* estimadas

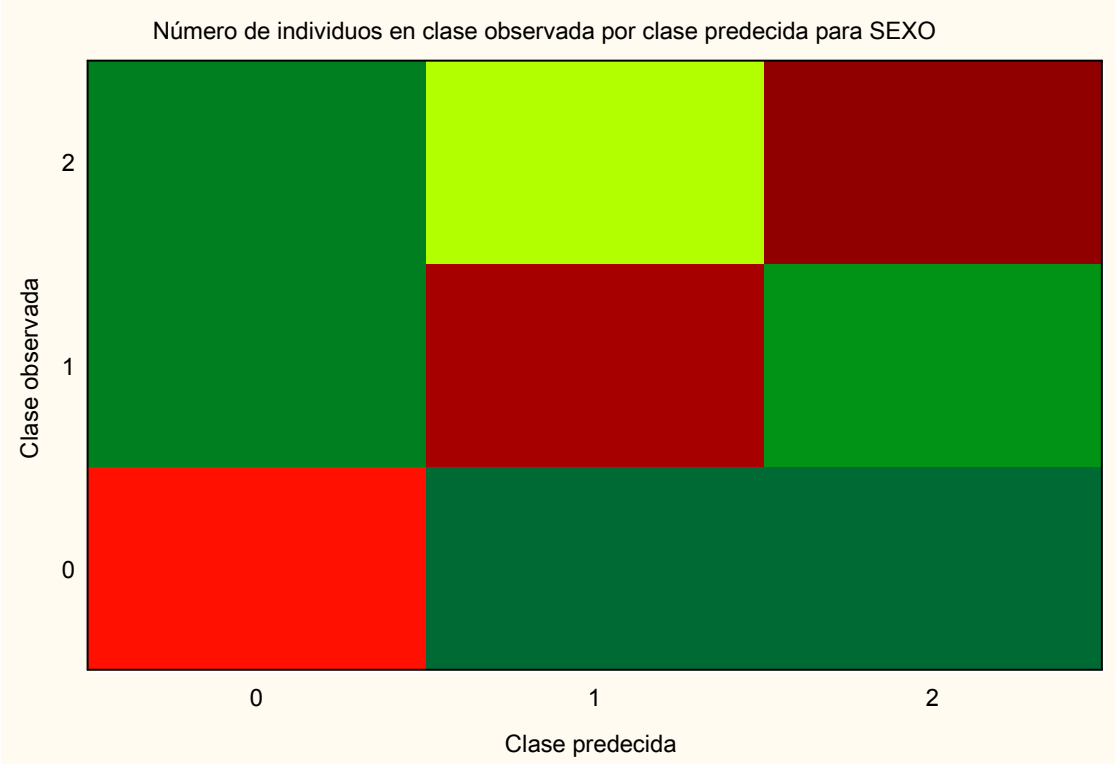


Figura 23. Medida de Gini y probabilidades *a priori* estimadas

3.1.7 Clasificación errónea

Para complementar el análisis, se necesita conocer el error de clasificación por clase y el total. En la siguiente tabla se encuentran los individuos que son clasificados erróneamente.

Clase	Clase 0	Clase 1	Clase 2
0		5	6
1	2		23
2	0	8	

Tabla 3.11. Matriz de clasificación errónea para la muestra de aprendizaje.
Medida de Gini y probabilidades *a priori* estimadas
Predicha (renglón) x observada (columna) matriz
Muestra de Aprendizaje N = 191

El número de individuos en la clase observada 0, clasificados erróneamente en la clase predicha 1 es 2 individuos y no hay individuos en la clase observada 0 clasificados erróneamente como clase predicha 2.

Por tanto, el porcentaje de individuos que pertenecen a la clase observada 0 y están erróneamente clasificados es 4.65%. Éste es calculado como la suma de los elementos erróneamente clasificados en clase predicha 1 más el número de unidades clasificados erróneamente en la clase predicha 2. Ésta suma es dividida entre el total de elementos en la clase observada 0. Este porcentaje es el mismo que el de los individuos clasificados erróneamente en la clase predicha 1, que pertenecen a la clase observada 0 debido a que los individuos erróneamente clasificados en la clase predicha 2 que pertenecen a la clase observada 0 es de cero al igual que su porcentaje.

El número de individuos en la clase observada 1, clasificados erróneamente en la clase predicha 0 es de 5 unidades y el número de elementos en la clase observada 1 clasificados erróneamente en la clase predicha 2 es de 8 unidades.

Por tanto, la proporción de individuos que pertenecen a la clase observada 1 y están erróneamente clasificados es 20.31%. El porcentaje es calculado como la suma de los porcentajes de los elementos erróneamente clasificados en la clase predicha 0 (5 unidades) dividido entre el total de la clase observada 1 (64 elementos) siendo igual a 7.81% más el porcentaje de individuos erróneamente clasificados en la clase predicha 2 (8 elementos) dividido entre el total de la clase observada 1 (64 unidades) siendo igual a 12.5%

El número de individuos en la clase observada 2, clasificados erróneamente en la clase predicha 0 es de 6 unidades y el número de individuos en la clase observada 2 clasificados erróneamente en la clase predicha 1 es de 23 elementos.

La proporción de individuos que pertenecen a la clase observada 2 y están erróneamente clasificados es 34.52%. Este porcentaje es muy grande si se ve como más de la tercera parte de los elementos de la clase observada 2. Lo anterior es debido a que existen 23 unidades clasificados erróneamente en la clase predicha 1 que pertenecen a la clase observada 2

La proporción total de individuos erróneamente clasificados que pertenecen a la clase observada 2, es calculado como la suma de los porcentajes de los elementos erróneamente clasificados en la clase predicha 0 (6 unidades) dividido entre el total de la clase observada 1 (64 especímenes) siendo igual a 7.81% más el porcentaje de individuos erróneamente clasificados en la clase predicha 2 (8 elementos) dividido entre el total de la clase observada 1 (64 especímenes) siendo igual a 12.5%

El número total de individuos clasificados erróneamente en la muestra de 191 elementos es de 44 elementos y el porcentaje de individuos en la muestra clasificados erróneamente: 23.03%

Si se desea obtener un mejor árbol de clasificación el umbral a vencer es: 23.03% en el error de clasificación o, como fue obtenido en la sección anterior se debe de clasificar correctamente al menos a 76.96% de la muestra.

3.2 Medida de Gini, probabilidades *a priori* iguales

3.2.1. Probabilidades *a priori* y árbol de clasificación

Las probabilidades en cada caso son .333333 debido a que son iguales y son mostradas en la Tabla 3.12. Este método no toma en cuenta la distribución de los individuos. Asigna la misma probabilidad a cualquiera.

Clase	Probabilidades *a priori*	Individuos en cada clase
0	0.333333	43
1	0.333333	64
2	0.333333	84

Tabla 3.12. Probabilidades a priori. Las probabilidades son iguales. Muestra de aprendizaje N = 191
Medida de Gini y probabilidades *a priori* iguales

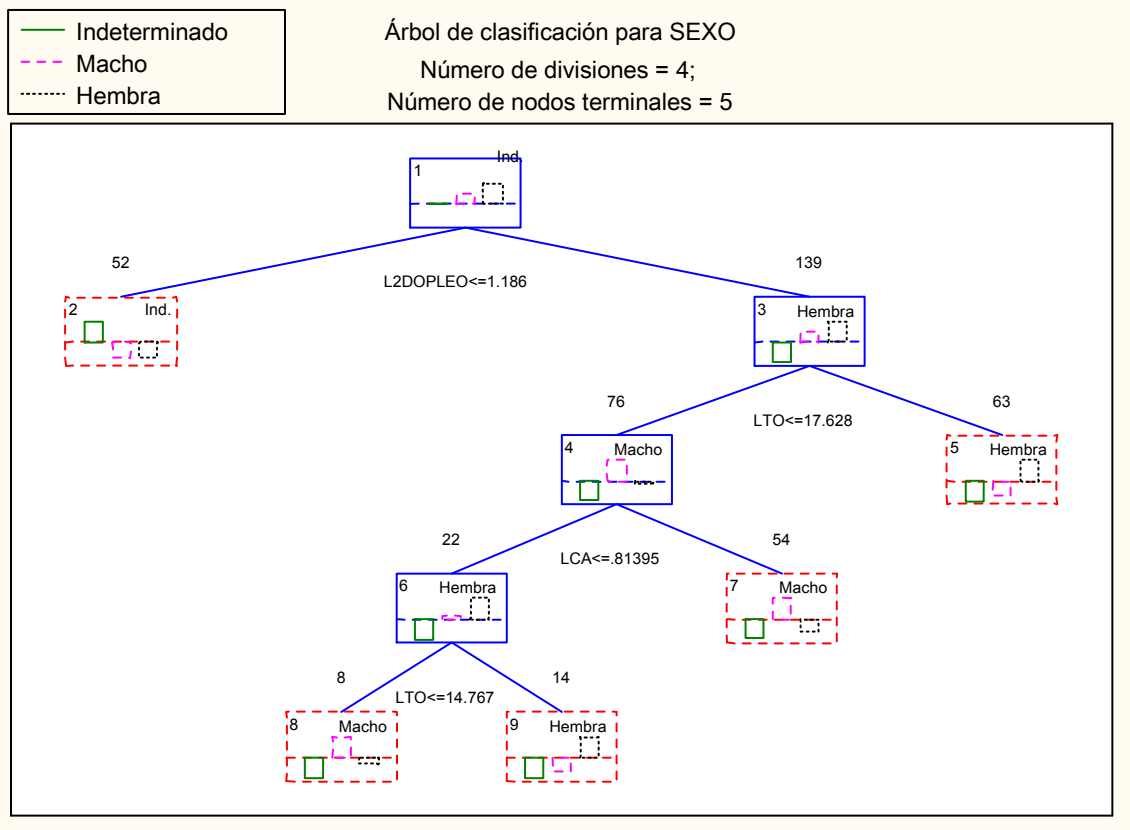


Figura 24. Medida de Gini y probabilidades *a priori* iguales.

El árbol de clasificación generado por la medida de Gini y probabilidades *a priori* iguales, contiene 5 nodos terminales 2, 5, 7, 8 y 9 y cuatro divisiones (Figura 24).

Para realizar la primera partición, el nodo raíz es dividido por la variable longitud del segundo pleópodo (L2DOPLEO).

Si la longitud del espécimen es menor o igual a 1.186 Mm., entonces es enviado al nodo descendiente 2 y se clasifican como indeterminados. El nodo 2 es terminal y contiene 52 elementos de la muestra.

Si la longitud del segundo pleópodo es mayor al valor anterior, los elementos se envían al nodo descendiente 3 y se clasifican como hembras. El nodo 3 contiene 139 elementos.

Si se observa, esta es la misma partición del árbol de clasificación de la sección 5.1 (Medida de Gini y probabilidades *a priori* estimadas).

El nodo 3 está dividido en los nodos 4 y 5, esta partición fue hecha por medio de la variable largo total y el valor 17.628 Mm., si la longitud total del espécimen es menor al valor mencionado, es enviado al nodo descendiente 4 y al nodo 5 en otro caso.

El nodo descendiente 5 es terminal y los elementos en este nodo son clasificados como hembras. Contiene 63 individuos.

El nodo 4 es particionado por medio de la variable longitud del carpo y el valor de la división es .81395. Esto genera los nodos 6 y 7.

En el nodo descendiente 6 hay 22 elementos y en el nodo 7 hay 54 elementos. Este último es terminal y los individuos en este nodo son clasificados como machos.

La última partición es realizada en el nodo 6 y produce los nodos 8 y 9. Es hecha por medio de la variable longitud total nuevamente, pero con un valor distinto. Si el largo

total del miembro de la muestra es menor a 14.767 Mm., entonces es enviado al nodo 8 y al nodo 9 en otro caso. Ambos nodos son terminales.

El nodo 8 contiene 8 unidades clasificadas como machos y al nodo 9 pertenecen 14 elementos clasificados como hembras.

3.2.2 Estructura del árbol de clasificación

Para conocer la partición de los nodos, los descendientes y los terminales a partir del nodo raíz se utiliza la Tabla 3.13.

El nodo raíz se particiona en los nodos 2 y 3. El nodo 2 es terminal y el 3 se divide en los nodos 4 y 5. El 5 es terminal y el 4 es bifurcado en los nodos 6 y 7. El 7 es terminal y el 6 es dividido en los nodos 8 y 9. Ambos son terminales.

Nodo	Rama izquierda	Rama derecha
1	2	3
2		
3	4	5
4	6	7
5		
6	8	9
7		
8		
9		

Tabla 3.13. Nodos descendientes y nodos terminales
Medida de GIni y probabilidades *a priori* iguales

Comenzando en el nodo raíz con los 191 individuos de la muestra, divididos en clase observada 0 con 43 elementos, clase observada 1 con 64 unidades y clase observada 2 con 84 especímenes. En la Tabla 3.14 se determina la composición para cada nodo. La primera división envía 52 individuos al nodo 2 y 139 al 3.

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2
1	2	3	43	64	84
2			41	5	6
3	4	5	2	59	78
4	6	7	2	51	23
5			0	8	55
6	8	9	0	8	14
7			2	43	9
8			0	6	2
9			0	2	12

Tabla 3.14. Composición de clases. Nodos descendientes y número de individuos en clase observada para cada nodo. Medida de Gini y probabilidades *a priori* iguales

El nodo 2 contiene 41 de 43 individuos de la clase observada 0 y 5 unidades de la 1 y 6 de la 2. Esto produce un total de 52 elementos en este nodo que es terminal.

El nodo 3 contiene 139 individuos, 2 de la clase observada 0, 59 de la clase observada 1 y 78 de la clase observada 2. El nodo es dividido en los nodos 4 y 5, con 76 y 63 elementos respectivamente.

El nodo 4 está compuesto por 2 elementos de la clase observada 0, 51 de la 1 y 23 de la 2 para un total de 76 individuos. Al realizar la división, envía 22 especímenes al nodo 6 y 54 al nodo 7.

El nodo 5 no contiene elementos de la clase observada 0. 8 especímenes de la clase observada 1 y 55 de la 2 lo componen para un total de 63 unidades.

El nodo 6 contiene 22 elementos, 8 de la clase observada 1 y 14 de la clase 2. Este nodo se divide y envía 8 elementos al nodo 8 y 14 al nodo 9.

El nodo 7 es descendiente del nodo 4 y contiene 54 elementos. Lo componen los 2 últimos especímenes de la clase observada 0, 43 de la clase 1 y 9 de la clase 2.

El nodo 8 es compuesto por 6 elementos de la clase observada 1 y por 2 de la clase 2 y tiene 8 unidades. Este nodo es terminal.

El nodo 9 contiene 2 elementos de la clase observada 1 y por 12 de la clase 2 para un total de 14 unidades. Este nodo es terminal.

Para poder comparar si los individuos de la muestra fueron clasificados correctamente, se debe proporcionar su clasificación predicha o la clase predicha para cada nodo. Esto sólo es importante para los nodos terminales (Tabla 3.15).

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2	Clase predicha.
1	2	3	43	64	84	0
2			41	5	6	0
3	4	5	2	59	78	2
4	6	7	2	51	23	1
5			0	8	55	2
6	8	9	0	8	14	2
7			2	43	9	1
8			0	6	2	1
9			0	2	12	2

Tabla 3.15.Clases predichas. Nodos descendientes y número de individuos por clase observada y clase predicha por cada nodo. Medida de Gini y probabilidades *a priori* iguales.

El nodo 2 tiene asignada una clase predicha 0 con un total de 52 individuos. Para los nodos 5 y 9 la clase predicha es 2, con un total de 77 elementos.

Para los nodos 7 y 8 la clase predicha es1 con un total de 62 elementos.

Los criterios de partición son mostrados en la siguiente tabla. Ésta presenta las constantes de división para los nodos, así como las variables elegidas para la división en cada nodo.

Nodo	Constante de división	Variable de división
1	-1.1860	L2DOPLEO
2		
3	-17.6279	LTO
4	-0.8140	LCA
5		
6	-14.7674	LTO
7		
8		
9		

Tabla 3.16. Condiciones de división. Condiciones de división para cada nodo.
Medida de Gini y probabilidades *a priori* iguales

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2	Clase predicha.	Constante de división	Variable de división
1	2	3	43	64	84	0	-1.1860	L2DOPLEO
2			41	5	6	0		
3	4	5	2	59	78	2	-17.6279	LTO
4	6	7	2	51	23	1	-0.8140	LCA
5			0	8	55	2		
6	8	9	0	8	14	2	-14.7674	LTO
7			2	43	9	1		
8			0	6	2	1		
9			0	2	12	2		

Tabla 3.17. Estructura de árbol. Nodos descendientes. Número de individuos en clase observada, clase predicha y condiciones de división para cada nodo.
Medida de Gini y probabilidades *a priori* iguales.

La primera división fue realizada con el segundo pleópodo como variable de partición en el nodo raíz. La segunda bifurcación utilizó la longitud total para el nodo 3. La tercera partición el largo del carpo para el nodo 4.

Esta división y esta variable resultaron muy importantes, debido a que continúa la clasificación de un árbol como el de la sección 5.1 y la hace más exacta. La información anterior se resume en la Tabla 3.17.

3.2.3 Grado de importancia para las variables predictoras

La Figura 25 describe la importancia de cada variable en la clasificación de los individuos. Con base a ésta y a la Tabla 3.18 se eligen las más importantes y se elaboran nuevos árboles para poder comparar con los anteriores.

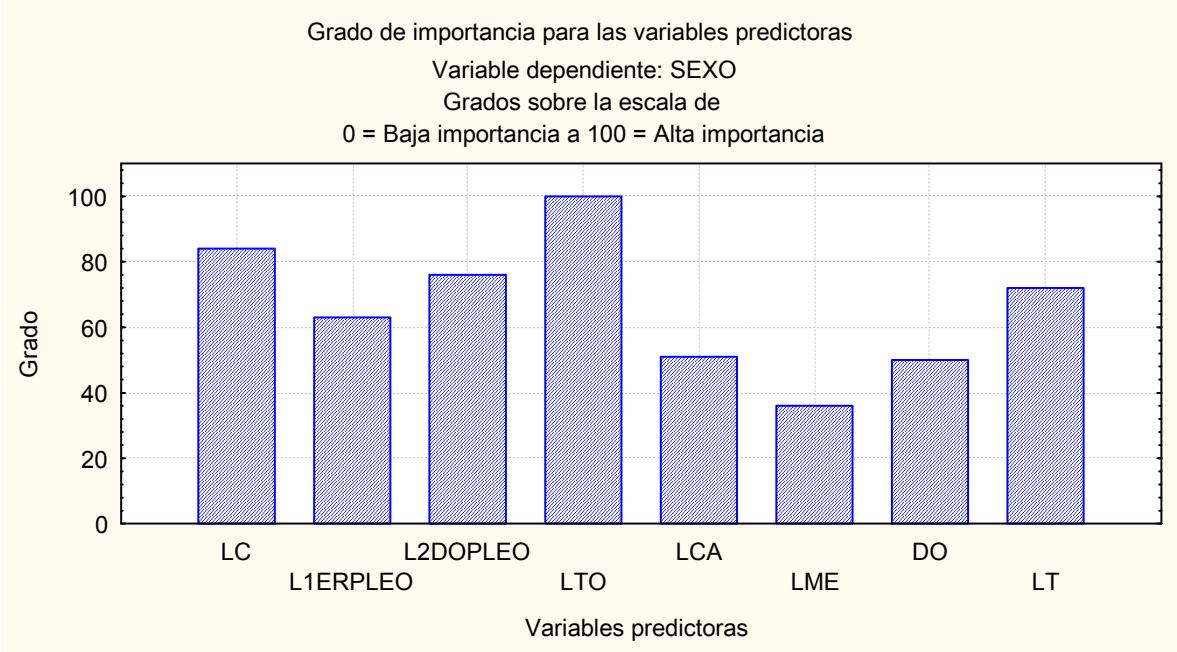


Figura 25. Medida de Gini y probabilidades *a priori* iguales.

Para esta combinación de medidas y probabilidades se obtiene que la variable que mejor clasifica es la longitud total con la más alta jerarquía, seguida de la longitud del caparazón con 84 puntos. El segundo pleópodo está en tercer lugar con 76 elementos y en cuarta posición la longitud del telsón, con 72 unidades.

3.2.4 Sucesión de costos

La selección del árbol de clasificación óptimo es por medio del cálculo de los costos por validación cruzada y por resustitución. En este caso, es elegido el de menor costo.

Variable	Grado de importancia
LC	84
L1ERPLEO	63
L2DOPLEO	76
LTO	100
LCA	51
LME	36
ED	50
LT	72

Tabla 3.18. Medida de Gini y probabilidades *a priori* iguales.
Basado en divisiones univariadas. 0 = Baja importancia; 100 =Alta importancia

De la Tabla 3.19 se desprende que el costo por validación cruzada y el error estándar en el árbol de clasificación con 10 y 5 nodos terminales, son los menores de la misma, por lo tanto es el que se elige. También cumple con la condición de pertenecer al punto de inflexión de la gráfica (Figura 26).

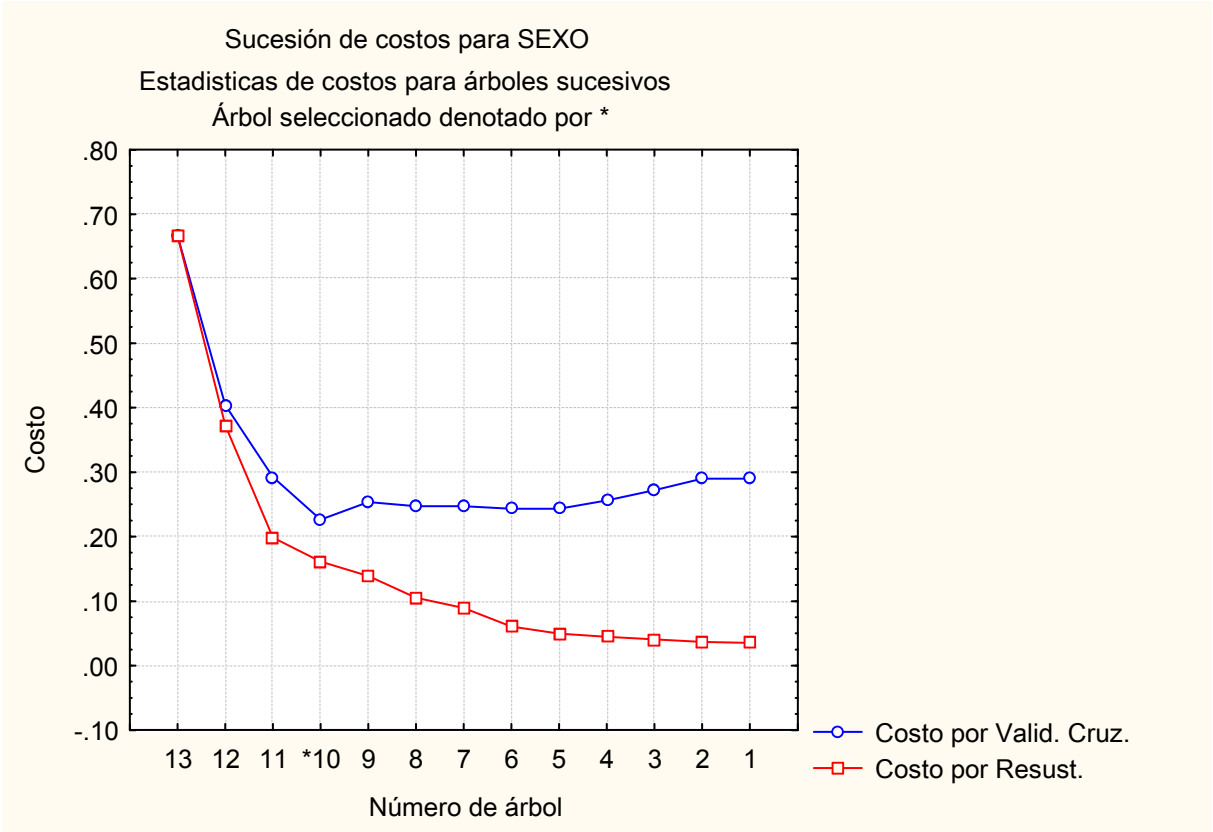


Figura 26. Árbol seleccionado denotado por * Medida de Gini y probabilidades *a priori* iguales.

3.2.5 Número de Individuos por clase observada

En esta sección se describirá, por medio de la Tabla 3.20 y de dos gráficas cómo fueron clasificados o divididos los individuos de cada clase observada en los nodos terminales. Lo anterior, para empezar a discernir cual árbol de clasificación es mejor y cuantos especímenes fueron correcta y erróneamente seleccionados.

El nodo terminal 2 contiene 41 elementos de la clase observada 0 correctamente clasificados y 11 erróneamente clasificados, 5 de la clase observada 1 y 6 de la clase 2.

El nodo terminal 5 contiene 55 elementos correctamente clasificados de la clase observada 2 y 8 seleccionados erróneamente pertenecientes a la clase observada 1.

Número de árbol	Nodos terminales	Validación cruzada	Error estándar	Resustitución	Complejidad del nodo
1	25	0.290069	0.033565	0.035466	0.000000
2	24	0.290069	0.033565	0.036706	0.001240
3	22	0.271900	0.032753	0.040675	0.001984
4	21	0.256027	0.032338	0.044643	0.003968
5	20	0.243067	0.031590	0.049851	0.005208
6	18	0.243067	0.031590	0.060516	0.005332
7	13	0.247035	0.031706	0.089286	0.005754
8	11	0.247035	0.031706	0.105159	0.007937
9	7	0.253484	0.031937	0.138889	0.008433
***10**	5	0.226260	0.029775	0.161089	0.011100
11	3	0.292480	0.030934	0.198292	0.018601
12	2	0.404087	0.030715	0.372647	0.174355
13	1	0.666667	0.000000	0.666667	0.294020

Tabla 319. Sucesión de árboles. Estadísticas para árboles sucesivos.

Árbol seleccionado denotado por * Medida de Gini y probabilidades *a priori* iguales.

Nodo	Clase 0	Clase 1	Clase 2
2	41	5	6
5	0	8	55
7	2	43	9
8	0	6	2
9	0	2	12

Tabla 3.20. Número de individuos en clase observada por nodo terminal.
Medida de Gini y probabilidades *a priori* iguales.

El nodo terminal 2 contiene 41 elementos de la clase observada 0 correctamente clasificados y 11 erróneamente clasificados, 5 de la clase observada 1 y 6 de la clase 2.

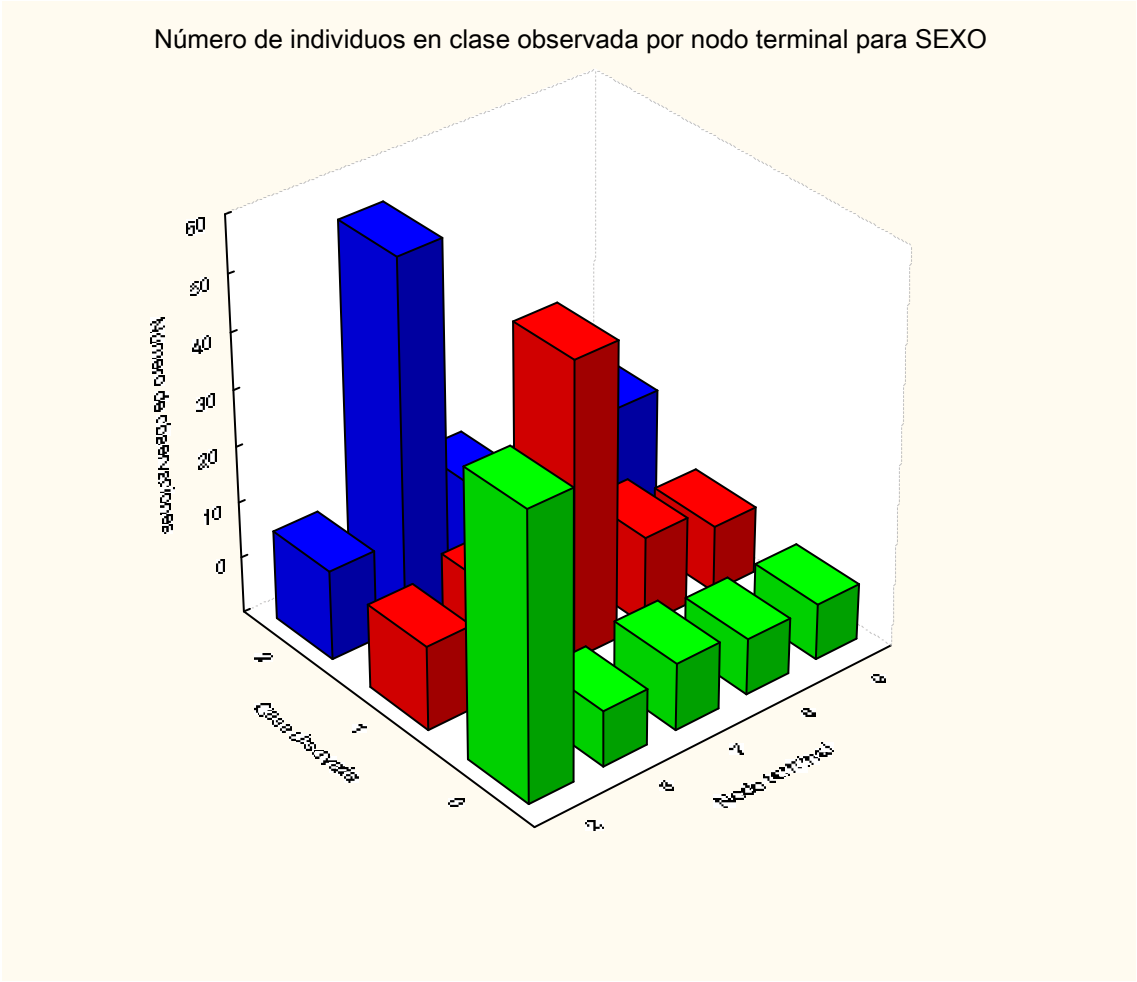


Figura 27. Medida de Gini y probabilidades *a priori* iguales.

El nodo terminal 5 contiene 55 elementos correctamente clasificados de la clase observada 2 y 8 seleccionados erróneamente pertenecientes a la clase observada 1. El nodo 5 no contiene elementos de la clase observada 0.

El nodo terminal 7 contiene 43 elementos correctamente clasificados de la clase observada 1 y 11 seleccionados erróneamente, 2 de la clase observada 0 y 9 de la 2.

El nodo terminal 8 contiene 6 elementos correctamente clasificados de la clase observada 1 y 2 erróneamente clasificados de la clase observada 2.

El nodo terminal 9 contiene 12 elementos correctamente seleccionados de la clase observada 2 y 2 erróneamente clasificados de la clase observada 1.

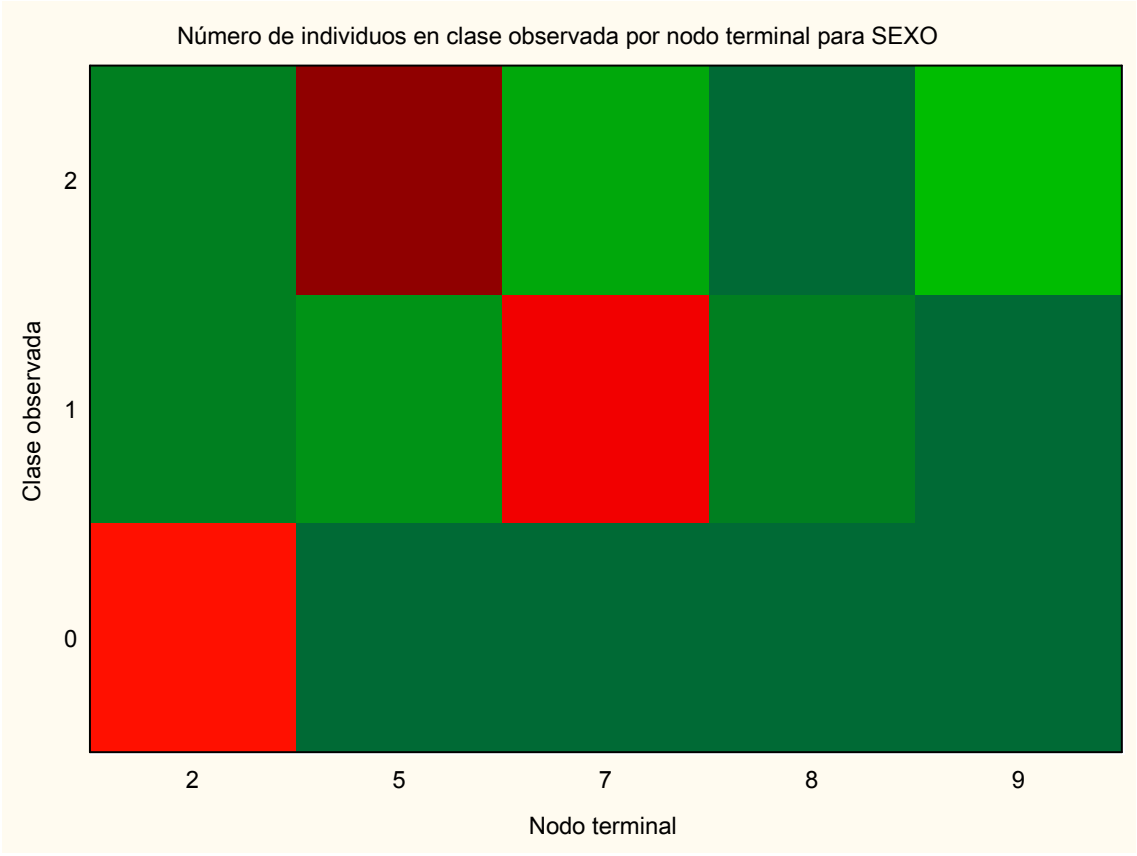


Figura 28. Medida de Gini y probabilidades *a priori* iguales.

Lo anterior se representa en una gráfica de tercera dimensión y en una de 2 dimensiones.

En la Figura 27 se dividen las clases observadas por colores. Verde para la clase observada 0, rojo para la 1 y azul para la 2. Se observa también, como están divididas las clases en los nodos terminales.

Para la Figura 28, existen 3 tonos de verde que representan la clasificación errónea. Los matices en rojo representan la selección correcta de los individuos.

El tono más claro de verde lo tienen las siguientes intersecciones. Clase observada 1 y nodo 5 con 8 elementos, clase observada 2 y nodo 7 con 9 individuos, clase observada 2 y nodo 9 con 12 unidades.

El tono intermedio de verde lo tienen las siguientes intersecciones. Clase observada 1 y nodo 8 con 6 elementos, clase observada 1 y nodo 2 con 5 unidades y clase observada 2 y nodo 2 con 6 individuos.

El tono más oscuro de verde lo tienen 5 intersecciones. Clase observada 0 y nodo 5 con 0 elementos; clase observada 0 y nodo 7 con 2 individuos, clase observada 0 y nodo 8 con 0 especímenes, clase observada 0 y nodo 9 con 0 unidades y clase observada 1 con nodo 9 con 2 individuos.

De igual manera, los tonos más oscuros de rojo representan una mejor clasificación correcta que los tonos claros.

3.2.6 Clase observada contra clase predicha

Si se desea conocer el grado de error al clasificar, en esta sección y en la siguiente se resumirá este resultado.

Si se observa la siguiente tabla, la diagonal indica los elementos correctamente seleccionados y fuera de la diagonal los erróneamente clasificados.

Si se compara el número de individuos correctamente clasificados con sus totales se encontrará la efectividad del método, es decir:

Clase	Clase 0	Clase 1	Clase 2
0	41	5	6
1	2	49	11
2	0	10	67

Tabla 3.21. Número de individuos en clase observada por clase predicha
Predicha (renglón) x observada (columna) matriz
Muestra de aprendizaje N = 191
Medida de Gini y probabilidades *a priori* iguales

El número de individuos en la clase observada 0, clasificados correctamente en la clase predicha 0 es de 41 unidades.

Ésta clase es la que mejor se clasifica, porque sólo 2 elementos no son reconocidos por este procedimiento como clase 0.

La proporción de individuos de la clase observada 0, clasificados correctamente en la clase predicha 0 es 0.95%, esto es calculado de la siguiente forma, se divide el número de elementos clasificados correctamente (41), entre la cantidad total en la clase observada 0 (43).

El número de individuos en la clase observada 1, clasificados correctamente en la clase predicha 1 es de 49. El porcentaje de elementos de la clase observada 1, clasificados correctamente en la clase predicha 1 es de 76.56%. Lo anterior es calculado dividiendo la cantidad de individuos correctamente clasificados (49) entre el numero de especímenes en la clase observada 1 (64).

El número de individuos en la clase observada 2, clasificados correctamente en la clase predicha 2 es de 67 individuos. La proporción con respecto al número de individuos en la clase observada 2 es de 79.76%.

Sumando las tres cantidades de la diagonal principal, se obtiene el número total de individuos clasificados correctamente. Realizando esta suma se obtienen 157 especímenes de la muestra clasificados correctamente de un total de 191. La proporción de individuos clasificados correctamente en la muestra es de 82.19%.

Al igual que en la Figura 29, en la Figura de arriba se representa por medio de la gráfica de tercera dimensión y por colores, cómo fueron divididas las clases observadas en clases predichas.

Para la clase observada 0 es el color verde, para la1 el rojo y para la 2 el azul
En la gráfica de la Figura 30 se aprecian solamente 2 tonos de verde. Tono claro, clase observada1 y clase predicha 2 con 11 elementos y clase observada 2 y clase predicha 1 con 10 elementos.

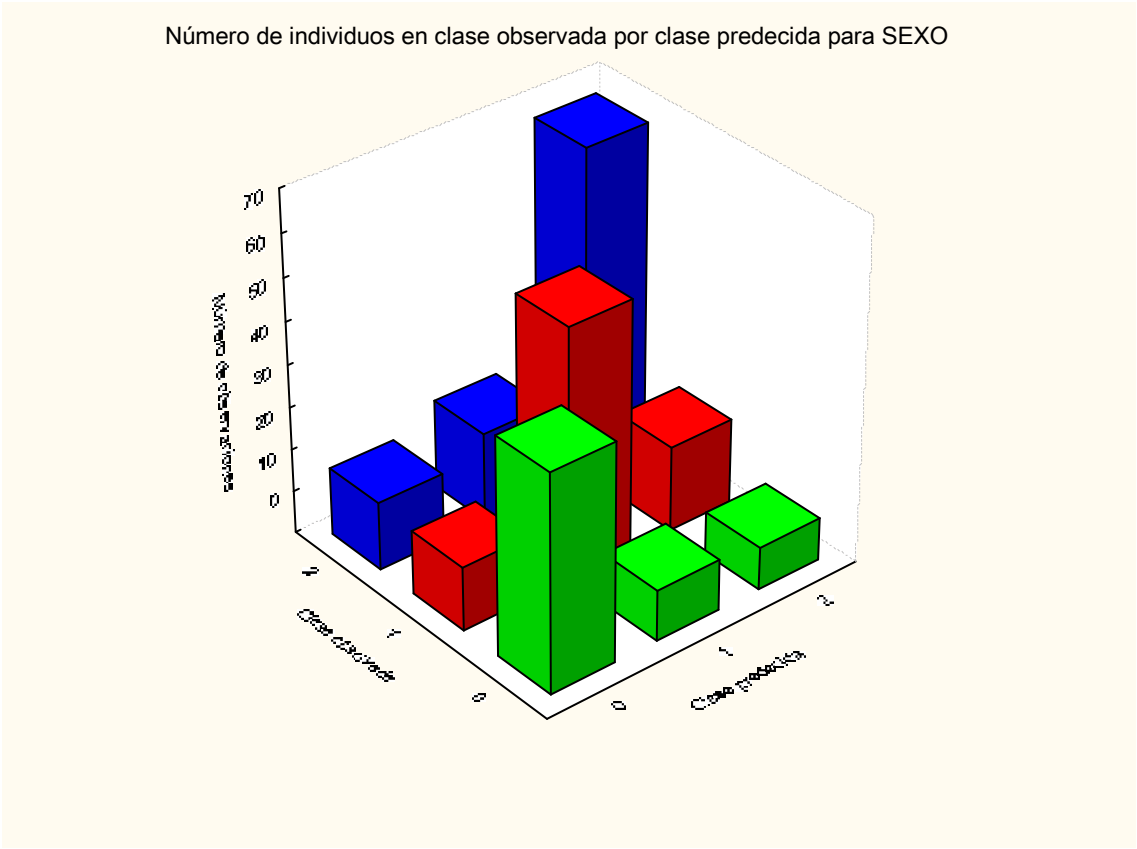


Figura 29. Medida de Gini y probabilidades *a priori* iguales.

Figura 30. Clase observada contra clase predicha
Medida de Gini y probabilidades *a priori* iguales.

Al igual que en la Figura 29, en la Figura de arriba se representa por medio de la gráfica de tercera dimensión y por colores, cómo fueron divididas las clases observadas en clases predichas. Para la clase observada 0 es el color verde, para la1 el rojo y para la 2 el azul.

En la gráfica de la Figura 30 se aprecian solamente 2 tonos de verde. Tono claro, clase observada1 y clase predicha 2 con 11 elementos y clase observada 2 y clase predicha 1 con 10 elementos.

El tono verde más intenso está en la clase observada 0 y la clase predicha 1 con 2 elementos, clase observada 0 y clase predicha 2 con 0 elementos, clase observada 1 y clase predicha 0 con 5 unidades y clase observada 2 y clase predicha 0 con 6 individuos.

3.2.7 Clasificación errónea

Observando la tabla de abajo se puede obtener las proporciones por clasificar erróneamente individuos en distintas clases.

Clase	Clase 0	Clase 1	Clase 2
0		5	6
1	2		11
2	0	10	

Tabla 3.22. Matriz de clasificación errónea para la muestra de aprendizaje.
Predicha (renglón) x observada (columna) matriz
Muestra de aprendizaje N = 191
Medida de Gini y probabilidades *a priori* iguales

El número de individuos en la clase observada 0, clasificados erróneamente en la clase predicha 1 es de 2 unidades y no hay elementos en la clase observada 0 clasificados erróneamente como clase predicha 2.

La proporción de individuos que pertenecen a la clase observada 0 y están erróneamente clasificados es 4.65%. Éste es calculado como la suma de los elementos erróneamente clasificados en clase predicha 1 más el número de especimenes clasificados erróneamente en la clase predicha 2. Ésta suma es dividida entre el total de unidades en la clase observada 0 (43).

Este porcentaje es el mismo que el de los individuos clasificados erróneamente en la clase predicha 1, que pertenecen a la clase observada 0 debido a que los individuos erróneamente clasificados en la clase predicha 2 que pertenecen a la clase observada 0 es de cero al igual que su porcentaje.

La cantidad de individuos en la clase observada 1, clasificados erróneamente en la clase predicha 0 es de 5 unidades y el número de elementos en la clase observada 1 clasificados erróneamente en la clase predicha 2 es de 10 especímenes.

La proporción de individuos que pertenecen a la clase observada 1 y están erróneamente clasificados es 23.43%. El porcentaje es calculado como la suma de los porcentajes de los especímenes erróneamente clasificados en la clase predicha 0 (5 unidades) dividido entre el total de la clase observada 1 (64 elementos) siendo igual a 7.81% más el porcentaje de individuos erróneamente clasificados en la clase predicha 2 (10 unidades) dividido entre el total de la clase observada 1 (64 elementos) siendo igual a 15.62%

El número de individuos en la clase observada 2, clasificados erróneamente en la clase predicha 0 es de 6 unidades y la cantidad de especímenes en la clase observada 2 seleccionados erróneamente en la clase predicha 1 es de 11 elementos.

El porcentaje de individuos que pertenecen a la clase observada 2 y están erróneamente clasificados es 20.23%. El porcentaje es calculado como la suma de las proporciones de los especímenes erróneamente seleccionados en la clase predicha 0 (6 unidades), dividido entre el total de la clase observada 2 (84 elementos) siendo igual a 7.14% más la proporción de individuos erróneamente elegidos en la clase predicha 2 (11 sujetos) dividido entre el total de la clase observada 2 (84 especímenes) siendo igual a 13.09%

El número total de individuos clasificados erróneamente en la muestra de 191 elementos es de 34 unidades y el porcentaje de individuos en la muestra seleccionados erróneamente: 17.8%

3.3 Combinación de variables y otras medidas

Con base a los resultados obtenidos en las secciones 3.1 y 3.2, se realizan 4 nuevos árboles de clasificación. El primero intercambia la medida de Gini por una medida Ji-cuadrada y mantiene las probabilidades *a priori* iguales. Los siguientes sólo utilizan las variables más importantes con las medidas de Gini y Ji-cuadrada con probabilidades *a priori* iguales. Las variables importantes y que cambian los árboles de clasificación manteniendo fijas las medidas y las probabilidades son: longitud

total, longitud del segundo pleópodo y longitud del carpo. No se exponen todos los resultados de estos árboles. La totalidad de los resultados están en el ANEXO 1.

3.3.1 Medida Ji-cuadrada probabilidades *a priori* iguales

Las probabilidades *a priori* son iguales para todas las clases 0.333333. El árbol con estos parámetros es el mostrado en la siguiente figura.

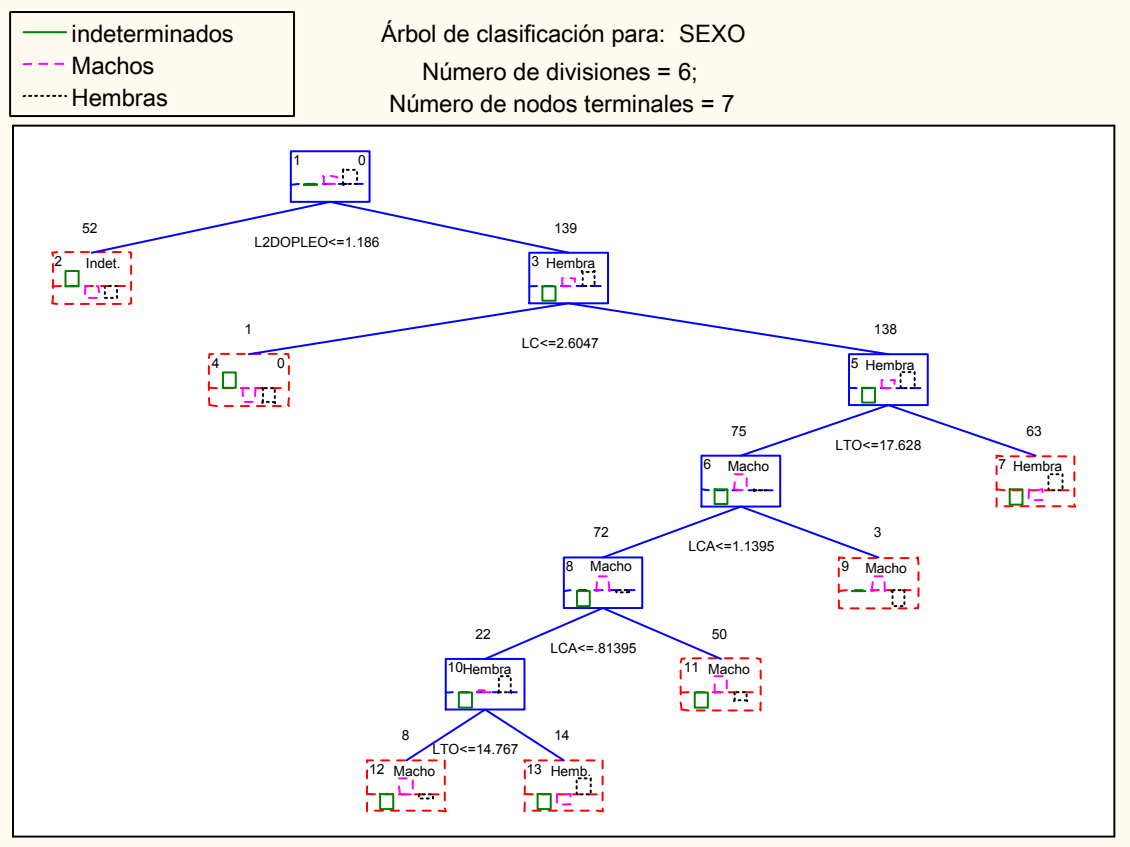


Figura 31. Medida Ji-cuadrada y probabilidades *a priori* iguales

Totalidad de variables

Clase	Clase 0	Clase 1	Clase 2
0	42	5	6
1	1	49	11
2	0	10	67

Tabla 3.23. Número de individuos en clase observada por clase predicha

Predicha (renglón) x observada (columna) matriz

Muestra de aprendizaje N = 191

Medida Ji-cuadrada, probabilidades *a priori* iguales y totalidad de variables.

El árbol de clasificación contiene 13 nodos, 6 particiones y 7 nodos terminales. Es más costoso en término de cálculos.

El número de individuos clasificados correctamente es de 158 y se obtiene sumando la diagonal principal en la tabla anterior. Por lo tanto el porcentaje de clasificación correcta es 82.72% y es calculado dividiendo el número de individuos seleccionados correctamente entre la cantidad total de la muestra.

El número de individuos clasificados erróneamente es de 33 y es calculado sumando los elementos fuera de la diagonal principal en la Tabla 3.23. Por lo tanto el porcentaje de clasificación errónea es 17.27% y es calculado dividiendo el número de especímenes seleccionados erróneamente entre la cantidad total en la muestra.

3.3.2 Medida de Gini, probabilidades *a priori* estimadas y variables importantes

Las probabilidades *a priori* son como las de la Tabla 3.1. El árbol de clasificación es el mostrado en la Figura 32.

Con 4 nodos terminales es un buen candidato para clasificar correctamente a *Potimirim mexicana* debido a que sólo tiene un nodo adicional al del árbol mínimo. Los resultados están contenidos en la Tabla 3.24. Los porcentajes son bastante buenos pero no los mejores.

Sin embargo, sólo contiene un nodo adicional con respecto al árbol de clasificación mínimo.

El número de individuos clasificados correctamente es de 153 y se obtiene sumando la diagonal principal en la tabla anterior. El porcentaje de clasificación correcta es 80.1% y es calculado dividiendo el número de individuos clasificados correctamente entre la cantidad total en la muestra.

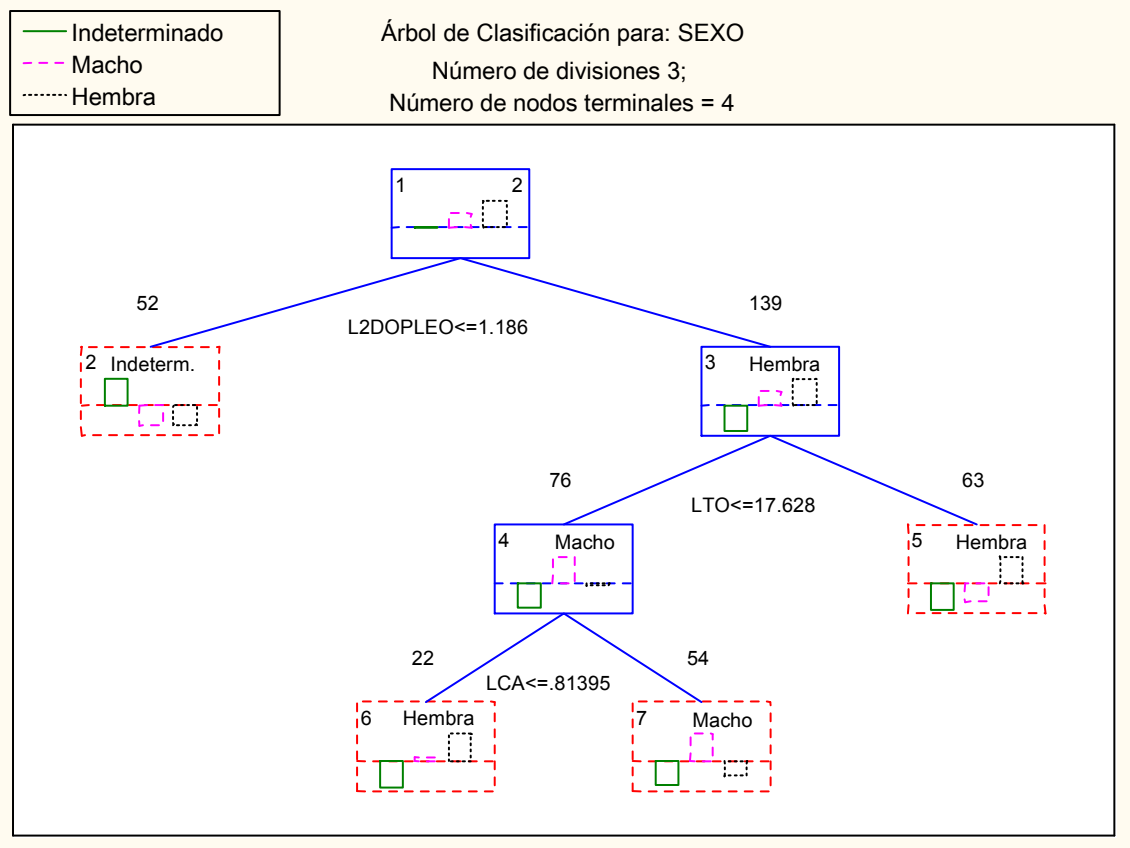


Figura 32. Árbol de Clasificación para SEXO.

Medida de Gini probabilidades *a priori* estimadas, variables importantes.

Clase	Clase 0	Clase 1	Clase 2
0	41	5	6
1	2	43	9
2	0	16	69

Tabla 3.24. Número de individuos en clase observada por clase predicha

Predicha (renglón) x observada (columna) matriz

Muestra de Aprendizaje N = 191

Medida de Gini y probabilidades *a priori* estimadas variables importantes

La cantidad de individuos clasificados erróneamente es de 38 y es calculado sumando los elementos fuera de la diagonal principal en la Tabla 3.24. Por lo tanto el porcentaje de clasificación errónea es 19.89% y es obtenido dividiendo la cantidad de especímenes seleccionados erróneamente entre el número total en la muestra.

3.3.3 Medida de Gini probabilidades *a priori* iguales y variables importantes

Esta selección de variables, de probabilidades *a priori* y de medida genera un árbol de clasificación similar al de la sección 3.2.

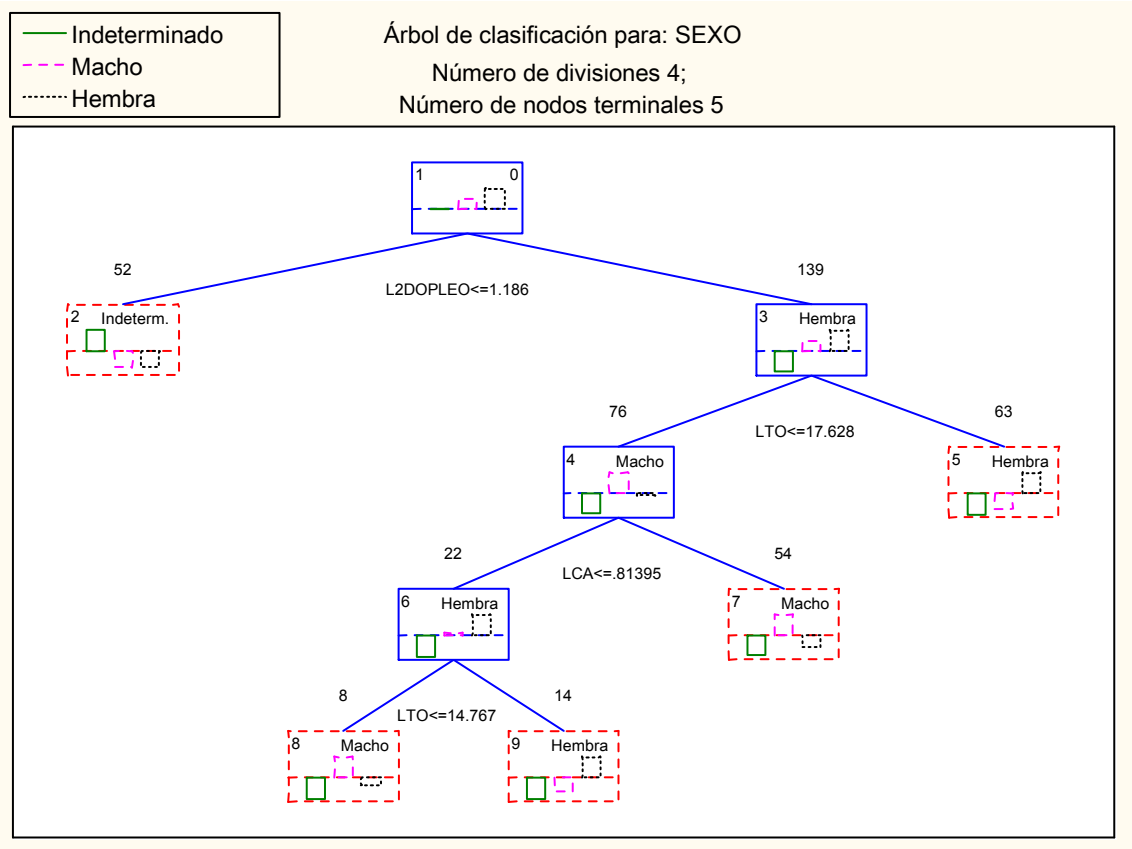


Figura 33. Árbol de clasificación para SEXO.

Medida de Gini probabilidades *a priori*, iguales y variables importantes

La diferencia con el árbol de clasificación de la sección 3.2 es que sólo se utilizaron 3 variables.

Clase	Clase 0	Clase 1	Clase 2
0	41	5	6
1	2	49	11
2	0	10	67

Tabla 3.25. Número de individuos en clase observada por clase predicha

Predicha (renglón) x observada (columna) matriz

Muestra de Aprendizaje N = 191

Medida de Gini y probabilidades *a priori* iguales variables importantes

La cantidad de individuos clasificados correctamente es de 157 y se obtiene sumando la diagonal principal en la tabla anterior. El porcentaje de clasificación correcta es 82.19% y es obtenido dividiendo el número de especímenes seleccionados correctamente entre la cantidad total en la muestra.

El número de individuos clasificados erróneamente es de 34 y es calculado sumando los elementos fuera de la diagonal principal en la tabla anterior. Por lo tanto el porcentaje de clasificación errónea es 17.8% y es obtenido dividiendo la cantidad de individuos clasificados erróneamente entre el número total de elementos en la muestra.

3.3.4 Medida Ji cuadrada, probabilidades *a priori* iguales y variables importantes

El árbol de clasificación desarrollado en esta sección contiene 11 nodos y 6 son terminales (Figura 34).

La segunda partición emplea una constante de división menor y esto genera una forma distinta en los nodos de los otros árboles de clasificación. De hecho emplea 3 particiones eligiendo a la longitud total con distintos valores.

Este árbol de clasificación es el que produce el número menor de individuos erróneamente clasificados (Tabla 3.26). Sin embargo, es el segundo en nodos producidos.

El número de individuos clasificados correctamente es de 159 y se obtiene sumando la diagonal principal en la tabla anterior. El porcentaje de clasificación correcta es 83.24% y se obtiene dividiendo la cantidad de especímenes selecionados correctamente entre el total de la muestra.

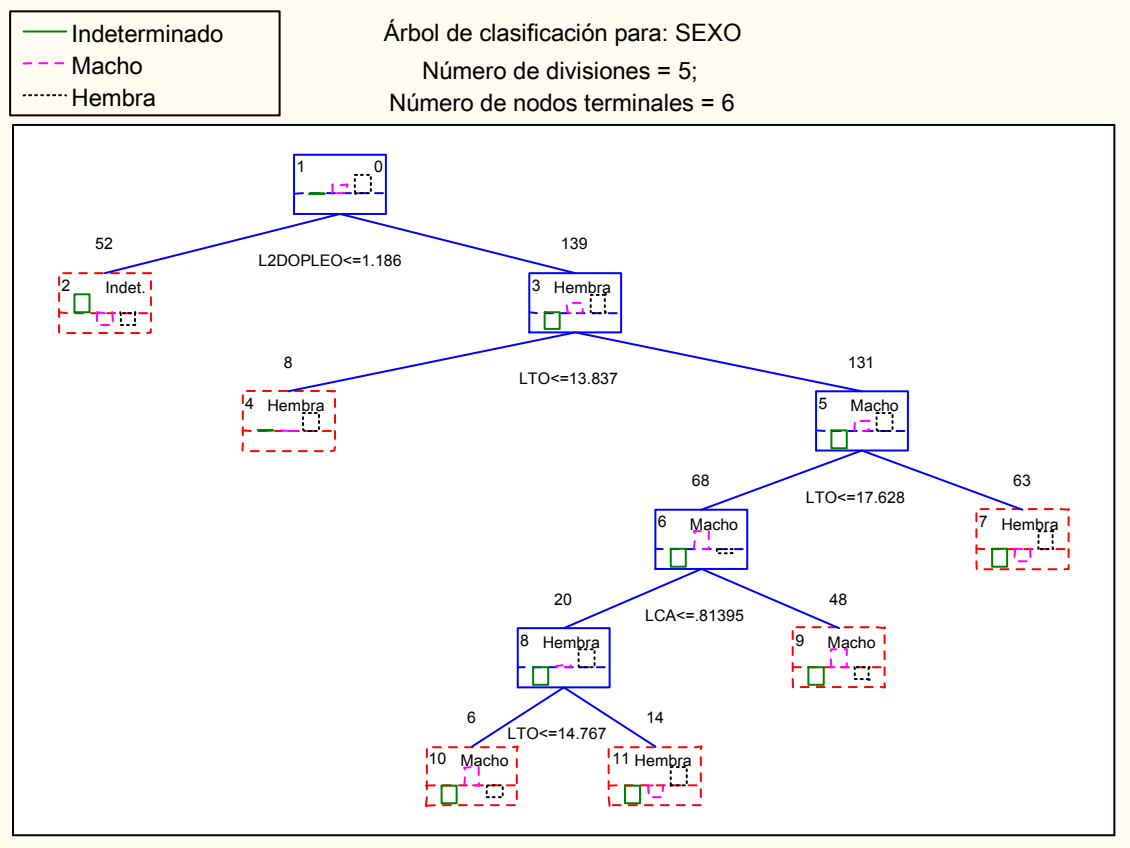


Figura 34. Árbol de clasificación para SEXO.

Medida Ji-cuadrada, probabilidades *a priori* iguales y variables importantes.

Clase	Clase 0	Clase 1	Clase 2
0	41	5	6
1	0	47	7
2	2	12	71

Tabla 3.26. Número de individuos en clase observada por clase predicha

Predicha (renglón) x observada (columna) matriz

Muestra de aprendizaje N = 191

La cantidad de individuos clasificados erróneamente es de 32 y es calculado sumando los elementos fuera de la diagonal principal en la Tabla 3.26. Por lo tanto el porcentaje de clasificación errónea es 16.75% y es computado dividiendo la cantidad de especimenes clasificados erróneamente entre el número total de unidades en la muestra.

3.4 Comparativo entre árboles de clasificación

Para conocer el árbol de clasificación que selecciona mejor a los individuos de *Potimirim mexicana* se necesita comparar todos los árboles de clasificación descritos a lo largo del capitulo. La Tabla 27 describe la precisión con respecto a la clasificación errónea de cada árbol.

Árbol de clasificación (sección)	3.1	3.2	3.3.1	3.3.2	3.3.3	3.3.4
Número de nodos	5	9	13	7	9	11
Número de nodos terminales	3	5	7	4	5	6
Número de individuos clasificados erróneamente	44	34	33	38	34	32
Porcentaje de individuos clasificados erróneamente	23.03	17.8	17.27	19.89	17.8	16.75

Tabla 3.27. Comparativo entre árboles de clasificación.

Eliminando los 2 árboles con mayor porcentaje de clasificación errónea, los árboles de las secciones 3.2, 3.3.1 y 3.3.3 son muy parecidos en cuanto al porcentaje de clasificación errónea. De hecho, sólo difieren en un elemento erróneamente clasificado.

El árbol de clasificación de la sección 3.3.2 sólo difiere de los anteriores en 4 unidades y tiene dos nodos menos y un nodo terminal menos. Sin embargo, su porcentaje se eleva 2 puntos.

El árbol que produce menor porcentaje con respecto a la clasificación errónea es el de la sección 3.3.4.
El inconveniente es debido al número de nodos que contiene, 11 en total y 6 terminales. La diferencia con el número de individuos clasificados erróneamente en los árboles de clasificación que se encuentran en la parte media de la Tabla 3.27, es de 2 unidades.

Conclusiones

Se desarrollaron seis árboles de clasificación y se vuelve complicado decidir tajantemente cuál es el mejor, ya que depende del interés particular que se tenga, así, se procederá a elegir uno de ellos.

Si se estuviera interesado en el número de nodos, los nodos terminales o en el tamaño del árbol, la elección idónea sería el árbol de la sección 3.3.2, con 7 nodos, de los cuales 4 son terminales y 38 individuos erróneamente clasificados, conteniendo un nodo más que el árbol mínimo; seguido por los árboles de las secciones 3.2 y 3.3.3 que son esencialmente iguales con 9 nodos, 5 nodos terminales y 34 individuos erróneamente clasificados.

Comparando los árboles del párrafo anterior, sólo difieren por 2 nodos y por cuatro individuos clasificados erróneamente; que es el precio que se paga. Por lo tanto se elige cualesquiera de los árboles que son iguales.

Si el criterio de decisión fuera el número de individuos erróneamente clasificados, se elegiría el árbol de la sección 3.3.4.

Este árbol de clasificación contiene 11 nodos, 6 terminales y 32 individuos clasificados erróneamente que, con respecto a los árboles de las secciones 3.2 y 3.3.3 sólo es de dos individuos erróneamente clasificados y de un individuo erróneamente clasificado con respecto al árbol de clasificación de la sección 3.3.1.

Debido a lo anterior se concluye que los mejores árboles de clasificación son los de las secciones 3.2 y 3.3.3 debido a la diferencia de sólo 2 individuos clasificados erróneamente y además de que contiene una cantidad menor de nodos.

Cambiando el criterio de decisión por el menor número de variables, entonces solamente resta elegir al árbol de clasificación de la sección 3.3.3, debido a que se genera el mismo árbol con el mismo número de individuos clasificados erróneamente, pero con 3 variables.

Con lo anterior, las variables que clasifican el sexo de *Potimirim mexicana* son la longitud total, la longitud del segundo pleópodo y la longitud del carpo, debido a que con estas tres variables se desarrolló el mejor árbol de clasificación con respecto a los criterios de decisión de los párrafos anteriores y es igual al desarrollado con la totalidad de las variables (árbol de clasificación de la sección 3.2).

Equivalentemente, si se desea clasificar o predecir el sexo de un nuevo individuo de *Potimirim mexicana* o una nueva muestra y si no ha evolucionado o mutado, entonces es posible clasificarla con el árbol de clasificación de la sección 3.3.3 de la siguiente forma:

Si la longitud del segundo pleópodo del individuo o individuos es menor o igual a 1.186 Mm., entonces se clasificará como indeterminado. En otro caso, si la longitud del segundo pleópodo es mayor a 1.186 Mm. y la longitud total del individuo es mayor a 17.628 Mm., es clasificado como hembra.

Si la longitud del segundo pleópodo es mayor a 1.186 Mm. y la longitud total del individuo es menor o igual a 17.628 Mm. y la longitud del carpo es mayor a 0.81395 Mm., entonces es clasificado como macho.

Si la longitud del segundo pleópodo es mayor a 1.186 Mm. y la longitud total del individuo es menor o igual a 17.628 Mm. y la longitud del carpo es menor o igual a 0.81395 Mm. y la longitud total del individuo mayor a 14.767 Mm., entonces es clasificado como hembra.

Si la longitud del segundo pleópodo es mayor a 1.186 Mm. y la longitud total del individuo es menor o igual a 17.628 Mm. y la longitud del carpo es menor o igual a .81395 Mm. y la longitud total del individuo menor o igual a 14.767 Mm., entonces es clasificado como macho.

Se observa que los individuos indeterminados son los más precisamente clasificados, debido a que 41 de 43 de los mismos miden, en la longitud de su segundo pleópodo igual o menor a la constante de división.

En las subsecuentes divisiones, se produce mayor confusión, que posiblemente es generada por el tamaño de longitud total y la madurez sexual. Si se ordenan los datos observados de hembras y machos con respecto a la longitud total, existe una intersección entre hembras y machos. El rango de machos es de 12.51 a 20.37 Mm. y el de las hembras es de 11.40 a 29.25 Mm. Entonces, 55 hembras están en el rango de los machos.

Lo discutido en el párrafo anterior, provoca la elevación de los porcentajes, en las hembras y los machos, de clasificación errónea. Es decir, que sea más inexacta. Por lo tanto, se originan más particiones o se necesitarán en más ocasiones la misma variable al realizar las divisiones.

No es la única técnica de clasificación que se puede utilizar para seleccionar, pero tiene la ventaja de ser más sencilla y práctica, por lo que puede ser usada con confianza y certeza de que hará una buena clasificación.

Literatura consultada

Barros, M. P. y N. F. Fontoura. 1996. Biología reproductiva de *Potimirim glabra* (Kingsley, 1878) (Crustacea, Decapoda, Atyidae), na Praia de Vigia, Garopaba, Santa Catarina, Brasil. Nauplius, Rio Grande, 4: 1-10.

Bauer, R. T. 2000. Simultaneous hermaphroditism in caridean shrimps: an unique and puzzling sexual system in the Decapada. Journal of Crustacean Biology 20 (Special issue 2): 116-128.

Breiman L., J. H. Friedman, R. A. Olshen, C. J. Stone. 1984. Classification and Regression Trees. Chapman and Hall/CRC. 341 p.

Charnov, E. L. 1979. Natural selection and sex change in pandalid shrimp: test of life-history theory. Am. Nat. 113: 715-734.

Cheng, C. S., Chen, L., 1990. Growth characteristics and relationships among body length, body weight and tail weight of *Penaeus monodon* for a culture environment in Taiwan. Aquaculture 91, 253-263.

Dall, W., B. J. Hill, P. C. Rothlisberg, D. J. Staples, 1990. The Biology of the Penaeidae. Adv. Mar. Biol. 27, Academic Press, London. 1-489 p.

Ghiselin, M. T. 1969. The evolution of hermaphroditism among animals. Q. Rev. Biol. 44: 189-208.

Luna, M. M. L. (1989) Aspectos biológicos de *Potimirim mexicana* bajo la influencia estuarina del Río La Antigua, Veracruz. Tesis de Licenciatura. Universidad Veracruzana. 23pp.

Martínez, M. M. 2003. Contribución al conocimiento de la Biología y Ecología de *Potimirim glabra* Kingsel (Decapada, Atyidae) en el Río Coyuca, Guerrero. Tesis de Maestría. Universidad Nacional Autónoma de México. 64 pp.

Peixoto, S., R. Soares, W. Wasielesky, R. O. Cavalli and Luciano Jensen. 2004. Morphometric relationship of weight and length of cultured *Farfantepenaeus paulensis* during nursery, grow out, and broodstock production phases. Aquaculture. 241: 291-299.

Primavera, J. H., F. D. Parado-Estepa and J. L. Lebata. 1998. Morphometric relationship of length and weight of giant tiger prawn *Penaeus monodon* according to life stage, sex and source. Aquaculture. 164: 67-75.

Villalobos, F. A. 1959. Contribución al conocimiento de los Atyidae de México. II (Crustacea, Decapoda). Estudio de algunas especies del género *Potimirim* (=*Ortamannia*), con descripción de una especie nueva en Brasil. An. Inst. Biol. Méx. 30: 269-330.

Anexo 1

En este anexo se presentan el complemento de tablas y gráficas para los árboles de clasificación de las secciones 3.3.1 a la 3.3.4. Los árboles de clasificación se pueden consultar en la sección respectiva.

Árbol de clasificación de la sección 3.3.1, medida Ji-cuadrada probabilidades *a priori* iguales y con la totalidad de variables

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2	Clase predicha.	Constante de división	Variable de división
1	2	3	43	64	84	0	-1.1860	L2DOPLEO
2			41	5	6	0		
3	4	5	2	59	78	2	-2.6047	LC
4			1	0	0	0		
5	6	7	1	59	78	2	-17.6279	LTO
6	8	9	1	51	23	1	-1.1395	LCA
7			0	8	55	2		
8	10	11	0	49	23	1	-0.8140	LCA
9			1	2	0	1		
10	12	13	0	8	14	2	-14.7674	LTO
11			0	41	9	1		
12			0	6	2	1		
13			0	2	12	2		

Tabla A.1. Estructura de árbol. Nodos descendientes. Número de individuos en clase observada, clase predicha y condiciones de división para cada nodo.
Medida Ji-cuadrada y probabilidades *a priori* iguales.

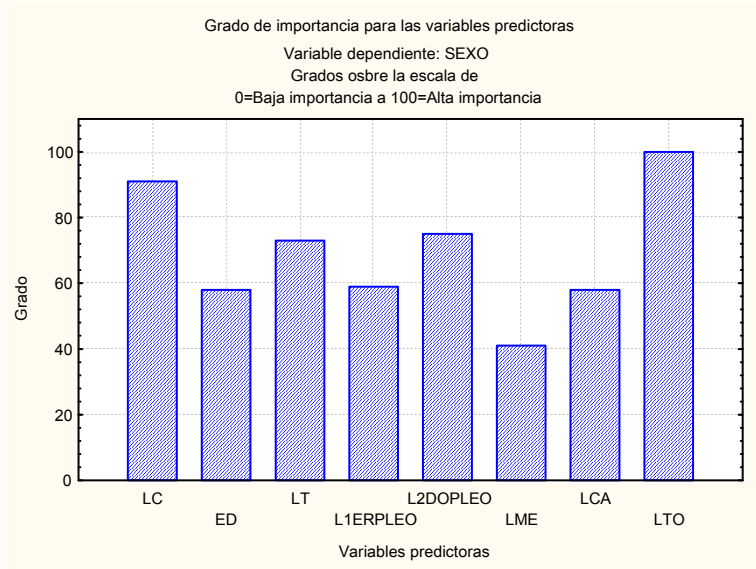


Figura A.1. Medida Ji-cuadrada y probabilidades *a priori* iguales.

Totalidad de las variables.

Variable	Grado
L_C_	91
D_O_	58
L_T_	73
L_1ERPL	59
L_2DOPL	75
L__ME_	41
L__CA_	58
L_TO_	100

Tabla A.2. Grado de Importancia para las variables predictoras

Basado en divisiones univariadas. 0 = Baja importancia; 100 =Alta importancia

Medida Ji-cuadrada y probabilidades *a priori* iguales. Totalidad de variables.

Figura A.2. Medida Ji-cuadrada y probabilidades *a priori* iguales.

Número de árbol	Nodos terminales	Validación cruzada	Error estándar	Resustitución	Complejidad del nodo
1	30	0.291309	0.033605	0.045946	0.000000
2	28	0.291309	0.033605	0.048427	0.001240
3	26	0.273140	0.032820	0.052395	0.001984
4	19	0.273140	0.032820	0.068020	0.002232
5	18	0.273140	0.032820	0.071988	0.003968
6	16	0.265389	0.032261	0.082405	0.005208
7	15	0.260180	0.032097	0.089101	0.006696
8	13	0.260180	0.032097	0.103734	0.007316
9	11	0.265389	0.032261	0.119607	0.007937
***10**	7	0.262660	0.032203	0.153337	0.008433
11	4	0.305377	0.031484	0.190540	0.012401
12	2	0.360436	0.032035	0.372647	0.091054
13	1	0.666667	0.000000	0.666667	0.294020

Tabla A.3. Sucesión de árboles. Estadísticas para árboles sucesivos.

Árbol seleccionado denotado por * Medida Ji-cuadrada y probabilidades *a priori* iguales.

Nodo	Clase 0	Clase 1	Clase 2
2	41	5	6
4	1	0	0
7	0	8	55
9	1	2	0
11	0	41	9
12	0	6	2
13	0	2	12

Tabla A.4. Número de individuos en clase observada por nodo terminal.
Medida Ji-cuadrada y probabilidades *a priori* iguales.

Árbol de clasificación de la sección 3.3.2, medida de Gini, probabilidades *a priori* estimadas y variables importantes

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2	Clase predicha.	Constante de división	Variable de división
1	2	3	43	64	84	2	-1.1860	L2DOPLEO
2			41	5	6	0		
3	4	5	2	59	78	2	-17.6279	LTO
4	6	7	2	51	23	1	-0.8140	LCA
5			0	8	55	2		
6			0	8	14	2		
7			2	43	9	1		

Tabla A.5. Estructura de árbol. Nodos descendientes. Número de individuos en clase observada, clase predicha y condiciones de división para cada nodo.
Medida de Gini, probabilidades *a priori* estimadas y variables importantes.

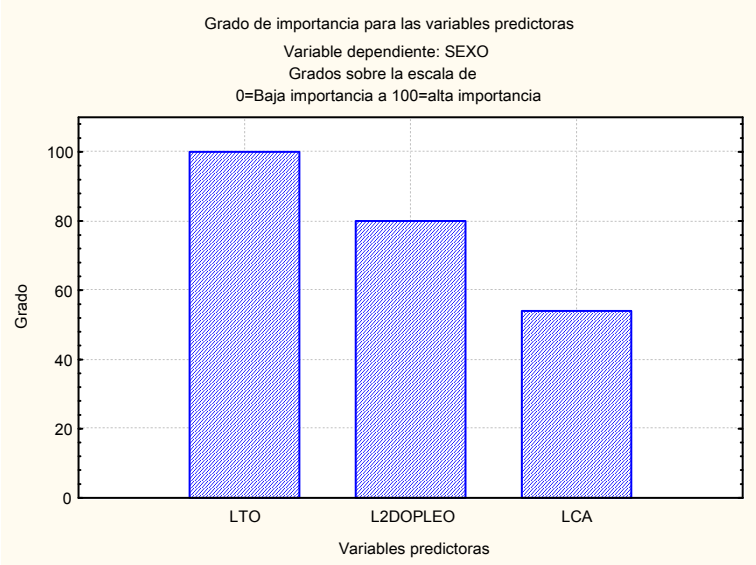


Figura A.3. Medida de Gini, probabilidades *a priori* estimadas y variables importantes.

Variables	Grado
LTO	100
L2DOPLEO	80
LCA	54

Tabla A.6. Grado de Importancia para las variables predictoras

Basado en divisiones univariadas. 0 = Baja importancia; 100 =Alta importancia

Medida de Gini, probabilidades *a priori* i estimadas y variables importantes.

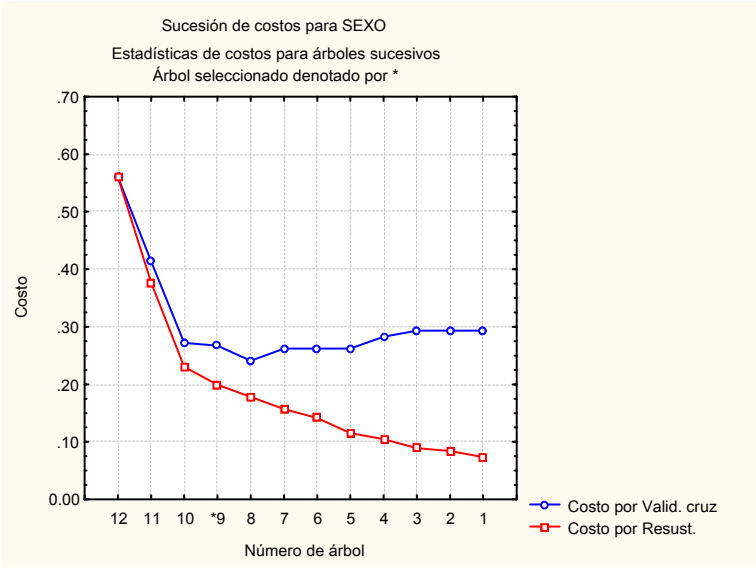


Figura A.4. Medida de Gini, probabilidades *a priori* estimadas y variables importantes.

Número de árbol	Nodos terminales	Validación cruzada	Error estándar	Resustitución	Complejidad del nodo
1	27	0.293194	0.032939	0.073298	0.000000
2	21	0.293194	0.032939	0.083770	0.001745
3	19	0.293194	0.032939	0.089005	0.002618
4	15	0.282723	0.032584	0.104712	0.003927
5	13	0.261780	0.031809	0.115183	0.005236
6	9	0.261780	0.031809	0.141361	0.006545
7	7	0.261780	0.031809	0.157068	0.007853
8	5	0.240838	0.030939	0.178010	0.010471
***9**	4	0.267016	0.032011	0.198953	0.020942
10	3	0.272251	0.032208	0.230366	0.031414
11	2	0.413613	0.035635	0.376963	0.146597
12	1	0.560209	0.035915	0.560209	0.183246

Tabla A.7. Sucesión de árboles. Estadísticas para árboles sucesivos.

Árbol seleccionado denotado por *

Medida de Gini, probabilidades *a priori* estimadas y variables importantes.

Nodo	Clase 0	Clase 1	Clase 2
2	41	5	6
5	0	8	55
6	0	8	14
7	2	43	9

Tabla A.8. Número de individuos en clase observada por nodo terminal.

Medida de Gini, probabilidades *a priori* estimadas y variables importantes.

Árbol de clasificación de la sección 3.3.3, medida de Gini probabilidades *a priori* iguales y variables importantes

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2	Clase predicha.	Constante de división	Variable de división
1	2	3	43	64	84	0	-1.1860	L2DOPLEO
2			41	5	6	0		
3	4	5	2	59	78	2	-17.6279	LTO
4	6	7	2	51	23	1	-0.8140	LCA
5			0	8	55	2		
6	8	9	0	8	14	2	-14.7674	LTO
7			2	43	9	1		
8			0	6	2	1		
9			0	2	12	2		

Tabla A.9. Estructura de árbol. Nodos descendientes. Número de individuos en clase observada, clase predicha y condiciones de división para cada nodo.
Medida de Gini, probabilidades *a priori* iguales y variables importantes.

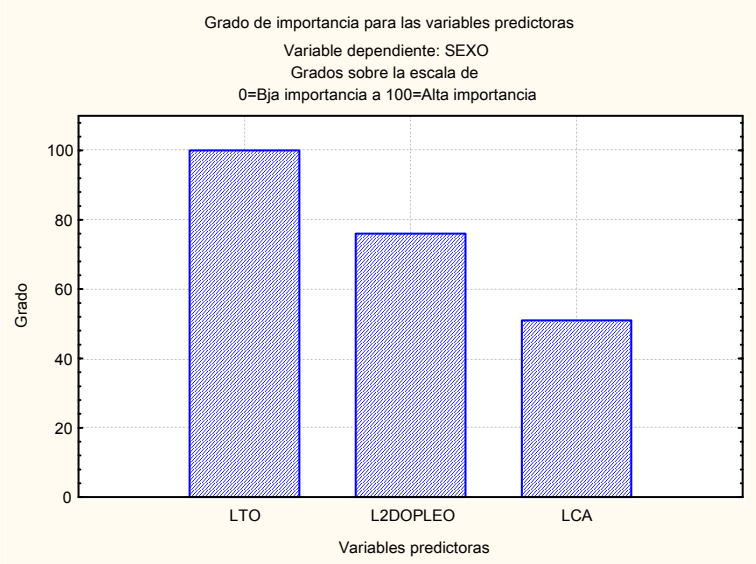


Figura A.5. Grado de importancia para las variables predictoras.
Medida de Gini, probabilidades *a priori* iguales y variables importantes.

Variables	Grado
LTO	100
L2DOPLEO	76
LCA	51

Tabla A.10. Grado de Importancia para las variables predictoras

Basado en divisiones univariadas. 0 = Baja importancia; 100 =Alta importancia

Medida de Gini, probabilidades *a priori* iguales y variables importantes.

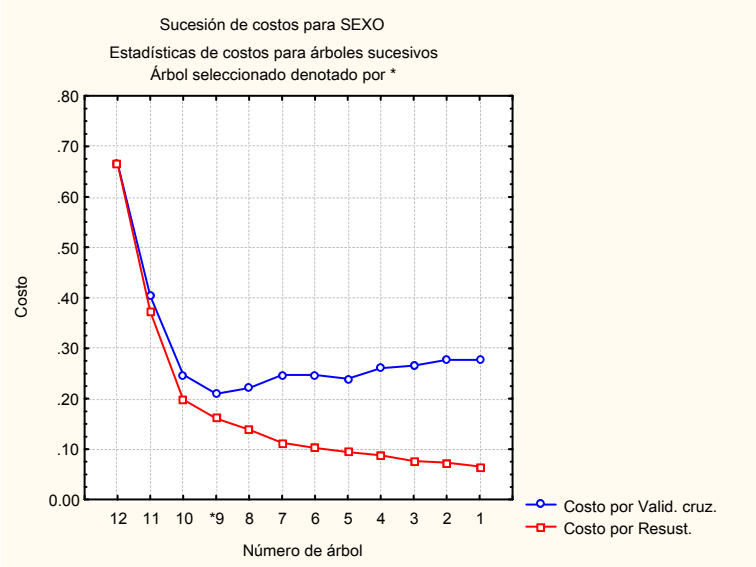


Figura A.6. Medida de Gini, probabilidades *a priori* iguales y variables importantes.

Nodo	Clase 0	Clase 1	Clase 2
2	41	5	6
5	0	8	55
7	2	43	9
8	0	6	2
9	0	2	12

Tabla A.11. Número de individuos en clase observada por nodo terminal.

Medida de Gini, probabilidades *a priori* iguales y variables importantes.

Número de árbol	Nodos terminales	Validación cruzada	Error estándar	Resustitución	Complejidad del nodo
1	28	0.277109	0.032914	0.065540	0.000000
2	22	0.277109	0.032914	0.073413	0.001312
3	21	0.265389	0.032261	0.076141	0.002728
4	17	0.260180	0.032097	0.088046	0.002976
5	15	0.239347	0.031294	0.094494	0.003224
6	13	0.247099	0.031870	0.102431	0.003968
7	11	0.247099	0.031870	0.112847	0.005208
8	7	0.220930	0.030425	0.138889	0.006510
***9**	5	0.209395	0.029099	0.161089	0.011100
10	3	0.247589	0.030257	0.198292	0.018601
11	2	0.404087	0.030715	0.372647	0.174355
12	1	0.666667	0.000000	0.666667	0.294020

Tabla A.12. Sucesión de árboles. Estadísticas para árboles sucesivos.

Árbol seleccionado denotado por *

Medida de Gini, probabilidades *a priori* iguales y variables importantes.

Árbol de clasificación de la sección 3.3.4, medida Ji cuadrada, probabilidades *a priori* iguales y variables importantes

Nodo	Rama izquierda	Rama derecha	Individuos en clase 0	Individuos en clase 1	Individuos en clase 2	Clase predicha.	Constante de división	Variable de división
1	2	3	43	64	84	0	-1.1860	L2DOPLEO
2			41	5	6	0		
3	4	5	2	59	78	2	-13.8372	LTO
4			2	2	4	2		
5	6	7	0	57	74	1	-17.6279	LTO
6	8	9	0	49	19	1	-0.8140	LCA
7			0	8	55	2		
8	10	11	0	7	13	2	-14.7674	LTO
9			0	42	6	1		
10			0	5	1	1		
11			0	2	12	2		

Tabla A.13. Estructura de árbol. Nodos descendientes. Número de individuos en clase observada, clase predicha y condiciones de división para cada nodo.
Medida Ji-cuadrada, probabilidades *a priori* iguales y variables importantes.

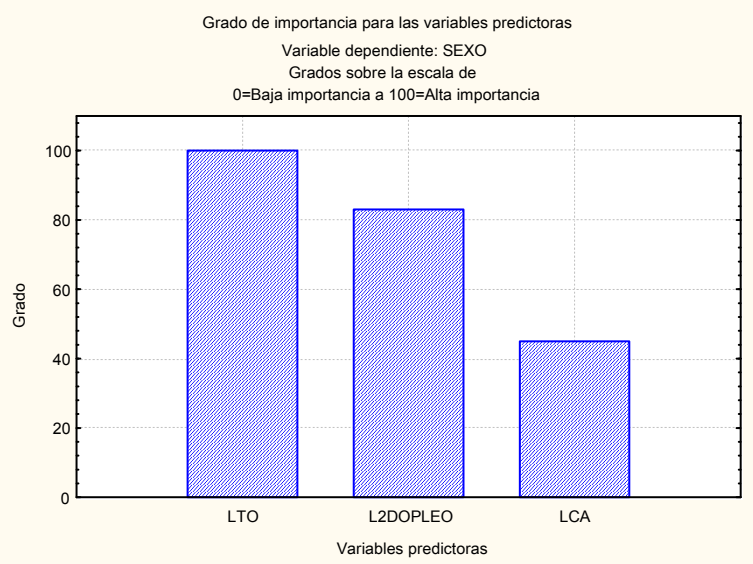


Figura A.7. Medida Ji-cuadrada, probabilidades *a priori* iguales y variables importantes.

Variables	Grado
LTO	100
L2DOPLEO	83
LCA	45

Tabla A.14. Grado de Importancia para las variables predictoras

Basado en divisiones univariadas. 0 = Baja importancia; 100 =Alta importancia

Medida Ji-cuadrada, probabilidades *a priori* iguales y variables importantes.

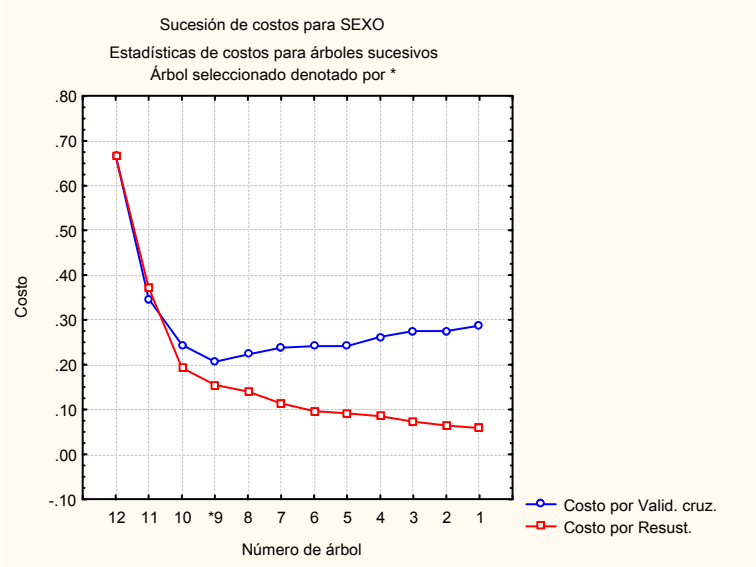


Figura A.8. Medida Ji-cuadrada, probabilidades *a priori* iguales y variables importantes.

Nodo	Clase 0	Clase 1	Clase 2
2	41	5	6
4	2	2	4
7	0	8	55
9	0	42	6
10	0	5	1
11	0	2	12

Tabla A.15. Número de individuos en clase observada por nodo terminal.

Medida Ji-cuadrada, probabilidades *a priori* iguales y variables importantes.

Número de árbol	Nodos terminales	Validación cruzada	Error estándar	Resustitución	Complejidad del nodo
1	28	0.287341	0.033508	0.059028	0.000000
2	25	0.275621	0.032898	0.064236	0.001736
3	21	0.275621	0.032898	0.073413	0.002294
4	17	0.262660	0.032203	0.085317	0.002976
5	15	0.241827	0.031511	0.091766	0.003224
6	14	0.241827	0.031511	0.096974	0.005208
7	11	0.238043	0.030996	0.114087	0.005704
8	7	0.223410	0.030572	0.140129	0.006510
***9**	6	0.206666	0.029054	0.155633	0.015504
10	4	0.244861	0.030273	0.192835	0.018601
11	2	0.347291	0.031989	0.372647	0.089906
12	1	0.666667	0.000000	0.666667	0.294020

Tabla A.16. Sucesión de árboles. Estadísticas para árboles sucesivos.

Árbol seleccionado denotado por *

Medida Ji-cuadrada, probabilidades *a priori* iguales y variables importantes.