Representación de clases jerárquicas mediante una Red Bayesiana

Rosales Pérez Alejandro Ruiz Blanco Jaime Rafael

Instituto Nacional de Astrofísica, Óptica y Electrónica
{arosales,jrruiz}@ccc.inaoep.mx

Resumen Existen muchos dominios en los cuales la manipulación de la información es indispensable para la solución de un problema determinado. Tales son los casos de la clasificación de textos y el reconocimiento de objetos. Estos son dos procesos especiales en donde una instancia puede ser clasificada como múltiples clases que están organizadas en una jerarquía.La clasificación jerárquica proporciona una solución a este problema de aislamiento y permite combinar la información mutua entre clases para poder clasificar a una instancia de manera correcta y consistente, para evitar problemas ambigüedad. Para este problema, el siguiente trabajo propone un enfoque basado en una red bayesiana para combinar la información de las clases y disminuir la incertidumbre en la clasificación de una instancia. Para ello, se diseña un modelo en el cual cada nodo en la red corresponde a un clasificador bayesiano simple de una determinada clase. La red resultante sirve para propagar la información generada por cada clasificador. Además, para evaluar la red se compara los resultados obtenidos de la propagación de la misma con los obtenidos usando clasificadores bayesianos simples de forma independiente. Entre los resultados más importantes, se obtiene un porcentaje de clasificación correcta del 59.17 % en las clases más generales y de 55.52 % para las clases más específicas, contra un 25.74 % y 56.10 %, respectivamente, del clasificador bayesiano simple.

Palabras claves: Clases jerárquicas, redes Bayesianas, Naive Bayes.

1. Introducción

La clasificación se refiere a la tarea de aprender, a partir de un conjunto de instancias clasificadas, un modelo que pueda predecir las clases de instancias previamente no consideradas [1]. Asimismo, un clasificador jerárquico es un clasificador que mapea datos de entrada a un subconjunto de clases de salida. La clasificación ocurre por primera vez en un bajo nivel con una alta especificación de los datos de entrada. Las clasificaciones de las instancias individuales de los datos se combinan de forma sistemática y es clasificada en un nivel más alto iterativamente hasta que se obtiene una salida. Esta salida es la clasificación general de los datos. La clasificación jerárquica tiene dos características principales:

1. Un ejemplo simple puede pertenecer a múltiples clases simultáneamente.
2. Las clases están organizadas en una jerarquía, esto es, si un ejemplo pertenece a una clase automáticamente pertenece a todas sus superclases.

Existen muchas aplicaciones que pueden ser representadas más fácilmente con clasificadores jerárquicos, quizás, entre los ejemplos más claros están las que involucran en área de visión por computadora, clasificación de textos [2] y reconocimiento de objetos [3]. Un ejemplo de lo anterior es el reconocimiento de objetos dentro de imágenes, donde un objeto puede ser visto como una colección de objetos, es decir, una imagen de un tigre puede ser vista como un objeto dentro de un conjunto de imágenes de felinos que, a su vez, puede ser vista como un conjunto de mamíferos y así sucesivamente.

Hoy en día, existen diversas técnicas que realizan la tarea de clasificación, entre las más conocidas se encuentran los árboles de decisión, reglas de clasificación y clasificadores basados en probabilidad como el clasificador bayesiano simple. Además, entre las técnicas más modernas involucran el uso de máquinas de soporte vectorial como se describe en [2]. La clasificación jerárquica proporciona un nuevo enfoque para atacar el problema de clasifición ya que ésta considera la información de la relación existente entre las clases en los diferentes niveles de jerarquía. Esto es motivo del presente trabajo que presenta el uso combinado de una red bayesiana con clasificadores bayesianos simples. Cada clasificador bayesiano representa un nodo en la red bayesiana, mientras que esta última representa la jerarquía.

El trabajo está organizado de la siguiente manera, en la sección 3 se describe el desarrollo del proyecto en sus diferentes etapas como es el procesamiento de los datos y la discretización, así como el diseño de los clasificadores bayesianos simples y la red bayesiana. Seguidamente, en la sección 4 se detallan las pruebas realizadas y los resultados alcanzados con el proyecto. Finalmente, en la sección 5 se presenta las conclusiones y el trabajo futuro.

2. Trabajo relacionado

Gran parte del trabajo en la clasificación jerárquica de múltiple etiquetado ha sido motivada por la clasificación de texto. Koller and Sahami [4], consideran el problema de clasificación jerárquica de texto en la que

cada documento de texto llega a ser exactamente una clase en el último nivel de la jerarquía.

En el trabajo de Cesa-Bianchi et al.[5], cada instancia de datos es etiquetada con un conjunto de etiquetas de clase, las cuales pueden llegar a tener más de una ruta en la jerarquía.

Barutcouglu et al. [6], presentó una aproximación a dos pasos donde una SVM es aprendida para cada clase por separado. Y luego combinándolas usando un modelo de red bayesiana entonces las predicciones son consistentes con la jerarquía.

Vens et al. [1], utilizan árboles de decisión para la clasificación jerarquíca múlti-etiqueta.

3. Metodología y Desarrollo

3.1. Procesamiento de los datos

Para el presente trabajo se usó el conjunto de datos "The SAIAPT 12 Collection" [7] disponible en el sitio web del TIA[1]. El conjunto de datos cuenta con un total de $99,535$ ejemplos, los cuales representan las características extraídas de un conjunto de imágenes segmentadas manualmente. Los datos están organizados de la siguiente manera, la primera columna representa el ID de la imagen segmentada, la segunda columna indica la region cuyas características son extraídas, de las columnas 3 a la 29 representan los artibutos (características) propios de la región tales como área de la región, convexidad , entre otros. Por último, en la columna 30 se indica la etiqueta a la clase que pertenece. Es importante recalcar que las clases ya se encuentran en una jerarquía definida previamente.

Para el desarrollo del trabajo no se usó el conjunto completo sino sólo una rama de la jerarquía, la rama de "Animal". Esta rama cuenta con un total de $1,999$ instancias

Al conjunto de datos se le introdujo una nueva columna que describe el nivel de cada instancia en la jerarquía (veáse la figura 1). Este proceso puede ser manual o automatizado, en este caso se desarrolló una función en MatLab para realizar este proceso.

Una vez establecido los niveles en el conjunto de datos estos fueron ordenados y, posteriormente, agrupados en subconjuntos de datos por niveles, teniendo un total de 27 conjuntos.

Seguidamente, cada uno de los subconjuntos de datos obtenidos se proceso para obtener los datos característicos de cada clase en ese ni-

[1] http://ccc.inaoep.mx/~tia/TIA/

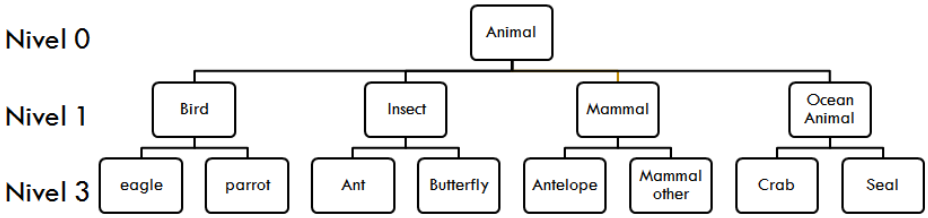


Figura 1. Ejemplo de la representación de niveles dentro de la jerarquía.

vel de la jerarquía y de sus hijos. Posteriormente, se procede al paso de discretización de los datos.

3.2. Discretización de los datos

Para facilitar el manejo de la información debido a la gran cantidad de datos con los que se cuenta y a que estos siguen una distribución continua, los datos fueron discretizados.

Básicamente, se usó un enfoque no supervisado aplicando la relación descrita en la ecuación 1. El $Vmin$ corresponde al menor valor de un atributo dentro del conjunto completo de datos, por el contrario, el $Vmax$ corresponde al valor mayor dentro del mismo atributo y el $beam$ es el número valores discretizados que se desean tener, en este proyecto se usó un $beam = 5$.

$$lenghtInterval = \frac{Vmax. - Vmin.}{beam} \tag{1}$$

3.3. Desarrollo del clasificador bayesiano simple

Después del proceso de creación de los subconjuntos y discretización de los mismo se continua con el diseño del clasificador bayesiano simple.

Un clasificador bayesiano simple asume que todos los atributos son condicionalmente independientes entre sí dada la clase[2], esto puede expresarse como sigue:

$$P(A_1, A_2, ..., A_n|V_j) = \prod_i P(A_i|v_j) \tag{2}$$

Como ya se mencionó previamente, sólo se trabajó con la rama de "Animal", sin embargo, esta rama cuenta con 67 clases por lo cual sólo se tomó en cuenta 27 clases, éstas fueron elegidas de manera arbitraria.

[2] http://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/node66.html

Aunado a lo anterior, y por cuestiones de tiempo de procesamiento de la información, sólo se consideraron 3 niveles dentro de la jerarquía de la rama "Animal" -Animal, hijos de Animal e hijos de los hijos de Animal. Por tanto, se tienen 27 clasificadores bayesianos simples para cada una de las clases.

Los clasificadores bayesianos fueron diseñados en MatLab, para facilitar el proceso se implementó una función destinada a esta tarea -el diseño del clasificador.

Finalmente, cada uno de estos clasificadores representa un nodo dentro de la red bayesiana, cuya descripción de su desarrollo se presenta enseguida.

3.4. Construcción de la red bayesiana

El primer paso fue establecer el nodo raíz de la red, en nuestro caso, este se representa como el nodo en el nivel más alto en la jerarquía.

Ya que se cuenta con el nodo raíz, los siguientes nodos a agregar son las clases que se encuentran en el nivel inmediato después de la clase raíz. Este proceso de conexión entre nodos en la red se repite hasta alcanzar los 4 niveles que se cubren el presente reporte -Entidad más los tres niveles de animal descritos en la sección 3.3.

Una vez terminada la estructura básica de la red, el siguiente paso es incorporar la información resultante del proceso de los datos y la creación de los clasificadores bayesianos simples.

La construcción fue realizada con el programa de "Elvira", mismo programa que se usa para la propagación de la red -que se describe en la sección 3.4.3.

2.4.1. Obtención de los parámetros de la red

En una red bayesiana, cada nodo posee una probabilidad asociada a su información a priori o condicional. Para el nodo raíz, su probabilidad es asignada a priori, en el caso del conjunto de datos con el que se trabaja, el nodo "Entity" tiene una probabilidad del 100 %, ya que esta clase engloba a las demás clases y cualquier vector de entrada para la red es clasificada como "Entity".

Para nodos internos, el proceso que se sigue es establecer sus probabilidades de acuerdo a la probabilidad condicional de que el nodo tome cierto valor dado que el padre ha tomado un valor específico. En nuestro modelo, asumimos que todas las clases son binarias, por lo tanto las clases pueden tomar valores únicamente de "si" o "no".

Por ejemplo, para el nodo "Animal", las probabilidades asociadas al nodo, son calculadas de acuerdo a la frecuencia de las instancias del nodo dentro del conjunto de datos del padre, en este caso "animal" (veáse figura 2).

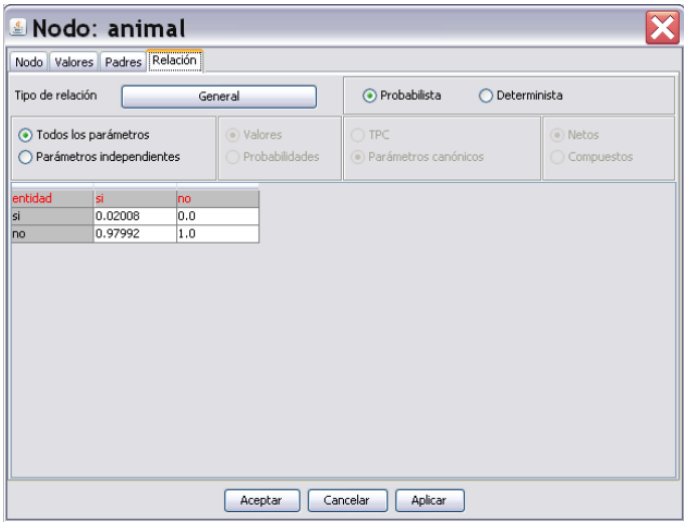


Figura 2. Ejemplo de las probabilidades a priori para la clase Animal.

2.4.2. Atributos como nodos en la red

Después de establecer la estructura de la red y las probabilidades asociadas a cada uno de las clases dentro de ella, se procede a establecer el criterio para procesar un vector de entrada.

La manera directa para representar la influencia de este vector en la red es crear nuevos nodos en la red, los cuales representan los 27 atributos que caracterizan a una región en el conjunto de datos. El proceso conlleva a incorporar las relaciones de los atributos con sus respectivas clases, como consecuencia el tamaño de la red crece rápidamente en función del número de nodos que se agreguen a la misma.

Poniendo estos datos a números; se cuenta con 27 nodos que representan las clases, entonces, se añaden 27 nodos por cada una de las clases para representar los atributos característicos de la región. En suma, se cuenta con un total de 629 nodos en la red.

En seguida, las probabilidades de cada nodo agregado son obtenida de los clasificadores bayesianos simples. Como resultado, el conocimiento de cada clasificador es incorporado a la red a través de estos nuevos nodos.

2.4.3. Proceso de propagación en la red

En este punto del proceso en la creación del modelo, se cuenta con una red bayesiana que combina la información de clasificadores bayesianos simples.

A partir de esto, se procede a obtener las probabilidades posteriores de la red cuando un vector de entrada es introducido. Esto se logra a través de los siguientes pasos:

1. El vector de entrada es leído y discretizado de acuerdo a los criterios establecidos para esta fase.
2. Con el vector discretizado, se genera la evidencia para que sea incluida en la red bayesiana.
3. La evidencia es instanciada en los nodos atributos agregados en la red para todas las clases.
4. Se realiza la propagación sobre la red.
5. Se guardan las probabilidades posteriores de la red para futuro análisis.

El método de propagación utilizado para la red fue el algoritmo de eliminación de variable y, a través de un script se codificaba la evidencia. Los resultados de la propagación eran guardados en un archivo de texto, para posteriormente analizarlos y obtener las estadísticas de clasificación.

4. Pruebas y resultados

En esta sección se describen las pruebas y resultados alcanzados con el método propuesto.

El desempeño de la red es evaluado mediante un promedio de instancias clasificadas correctamente. Para realizar lo anterior, se usó un conjunto de ejemplos para pruebas cuya clasificación es conocida. Posteriormente, los ejemplos se usan como evidencia en la red bayesiana y se obtienen las probabilidades generadas por la red después del proceso de propagación. Para determinar si el ejemplo pertenece a la clase o no se usa el criterio del argumento máximo.

El conjunto de pruebas consta de un total de $1,041$ ejemplos elegidos aleatoriamente. El tiempo de clasificación de los ejemplos fue de una hora

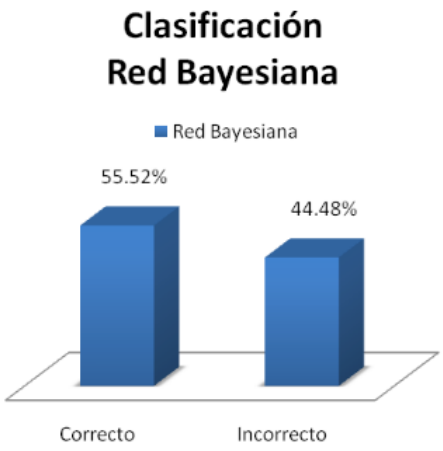


Figura 3. Resultado general de la clasificación con la red bayesiana

y treinta minutos[3] para 520 ejemplos y dos horas y treinta minutos[4] para los 521 ejemplos restantes.

Los resultados generales de la clasificación se muestran en la gráfica de la figura 3, estos resultados son un promedio de las clasificaciones de la clase general "Animal". Se tiene un 55.52 % de éxito en las clasificaciones.

Además, para saber que tan buen desempeño tiene la clasificación jerárquica, con una red bayesiana, se compararon los resultados obtenidos en la fase de pruebas con las clasificaciones obtenidas usando únicamente clasificadores bayesianos simples, cabe destacar que se usó el mismo conjunto de pruebas para los bayesianos simples.

En la figura 4 se muestra una gráfica comparativa de los resultados obtenidos de la clasificación usando el método propuesto contra un clasificador bayesiano simple (Naive Bayes), se tiene que para las clases más específicas el bayesiano simple brinda mejores resultados. Sin embargo al observar las gráficas de la figura 5 se observa un mejor desempeño de la clasificación con la red bayesiana.

[3] Con un procesador Intel Celeron a 1.72 GHz, 1.5 GByte en RAM y S.O. Fedora Linux 12.
[4] Con un procesador Intel Centrino Duo a 1.66 GHz, 2 GBytes en RAM y S.O. Windows XP

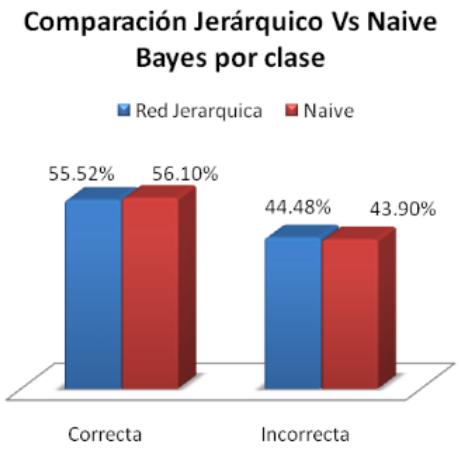


Figura 4. Comparación entre los resultados de clasificaciones usando un clasificador bayesiano simple y una red bayesiana

5. Discusiones y conclusiones

En el presente trabajo se presentó un enfoque de clasificación basado en una red bayesiana en combinación con clasificadores bayesianos simples. Asimismo, se mostraron los resultados obtenidos de la clasificación.

Es importante mencionar que durante la etapa de experimentación, primeramente, se probó con una red bayesiana con pocos nodos y de un sólo nivel, es decir, un padre y sus hijos. Posteriormente, la red bayesiana se fue extendiendo hasta llegar a tener tres niveles. Se observó que al ir incrementando el número de nodos en la red bayesiana, ir considerando más clases, e ir aumentando el número de niveles, la clasificación lograda por la red iba aumentando gradualmente.

Por otro lado, conforme la red crecía en número de nodos, el tiempo de procesamiento de la información así como el tiempo requerido para la propagación e inferencia se incrementaba considerablemente, razón por la cual se optó por trabajar sólo con una rama del conjunto de datos.

Como trabajo futuro se tiene la construcción de la red jerárquica bayesiana considerando todos los niveles de la jerarquía, de igual manera, trabajar para reducir la complejidad de la red bayesiana y con ello ayudar en el tiempo de procesamiento.

En resumen, la clasificación jerárquica es un nuevo enfoque de clasificación que promete brindar buenos resultados, en general, si se tiene en consideración todos los niveles de la jerarquía. Aún cuando los resultados

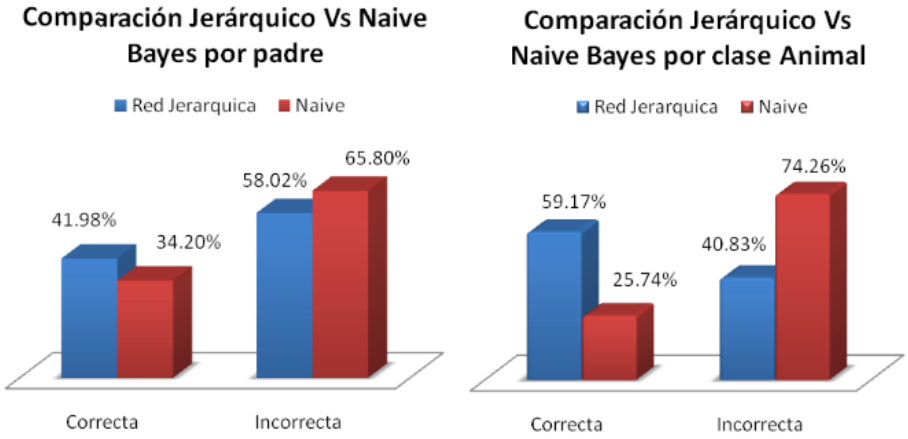


Figura 5. Gráfica comparativa de los resultados obtenidos con la red bayesiana y el bayesiano simple para las clases padres y animal.

fueron debajo del 60 %, en base a la experimentación, se tienen antecedentes de lograr una mejora en estos por lo cual aún queda trabajo por realizar.

Referencias

1. Vens, C., Struyf, J., Schietgat, L., Dzeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification
2. Rousu, J., Saunders, C., Szedmak, S., Shawe-Taylor, J.: Kernel-based learning of hierarchical multilabel classification models. Journal of machine learning **7** (2006) 1601–1626
3. Stenger, B., Thayananthan, A., Torr, P., Cipolla, R.: Estimating 3d hand pose using hierarchical multi-label classification. Image and Vision Computing **5** (2007) 1885–1894
4. Koller, D., Sahami, M.: classifying documents using very few words. In proc.of the 14th International Conf. on Machine Learning (1997) 170–178
5. Cesa-Bianchi, ., Gentile, C., Zaniboni, L.: Incremental algorithms for hierarchical classification. Journal of machine learning
6. Barutcuoglu, Z., Schapire, R., Troyanskaya, O.: Hierarchical multi-label prediction of gene function (2006)
7. Escalante, H.J., Hernández, C.A., Gonzalez, J.A., López-López, A., Montes, M., Morales, E.F., Sucar, L.E., Villaseñor, L., Grubinger, M.: The segmented and annotated iapr tc-12 benchmark. Computer Vision and Image Understanding **In press** (2009) doi: http://dx.doi.org/10.1016/j.cviu.2009.03.008.

Anexos

A. Diagrama de la jerarquía

En la figura 6 se muestra el diagrama jerárquico con las clases tomadas de las base de datos para la construcción de la red bayesiana.

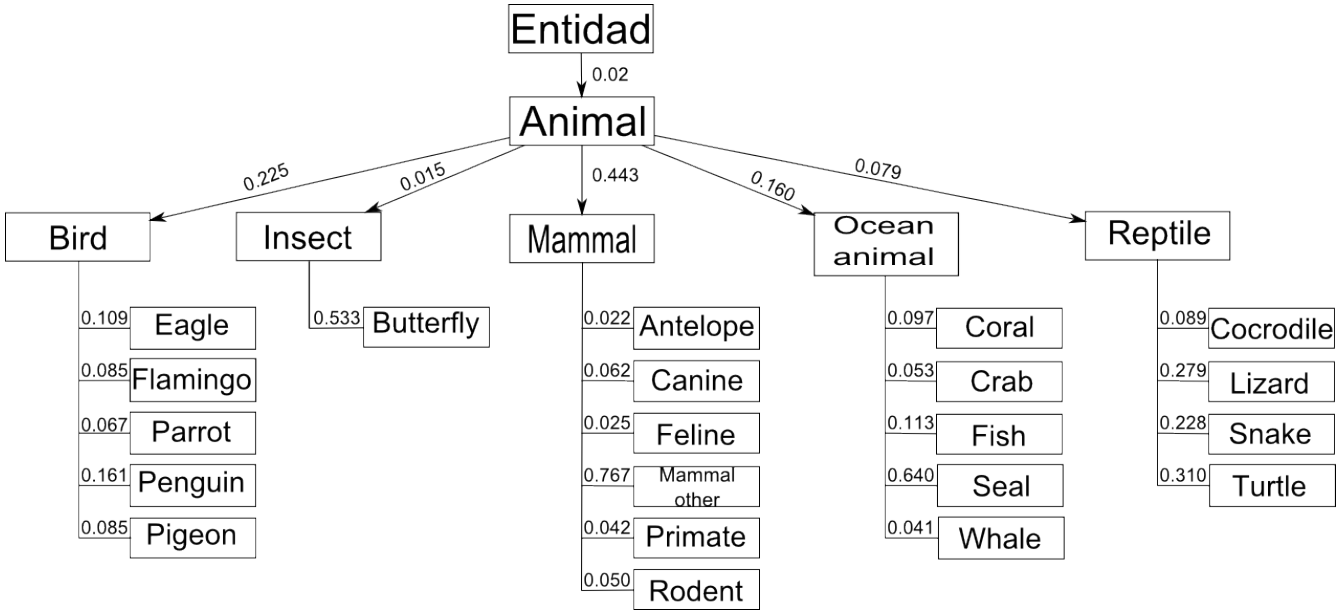


Figura 6. Diagrama de la jerarquía del conjunto de datos para la rama Animal.

B. Resultados

Los presentado en la gráfica mostrada en la figura 7, corresponde a los resultados obtenidos en la evaluación general del modelo en comparación con los resultados obtenidos del clasificador bayesiano simple, con un umbral de exactitud del 50 %.

Estos resultados corresponden a la asignación de probabilidad a priori de 50 % para el valor "si" y 50 % para el valor "no", del nivel superior de nuestro modelo (Nodo Animal).

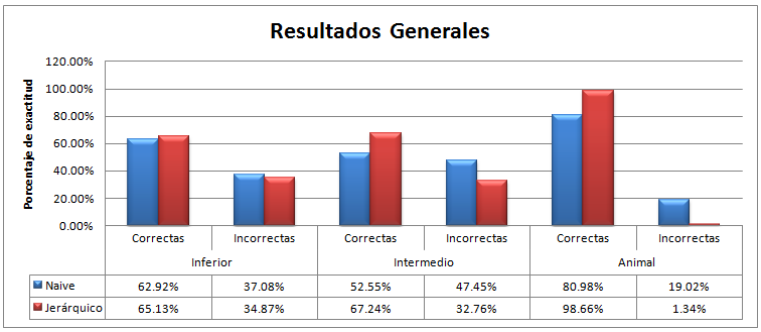
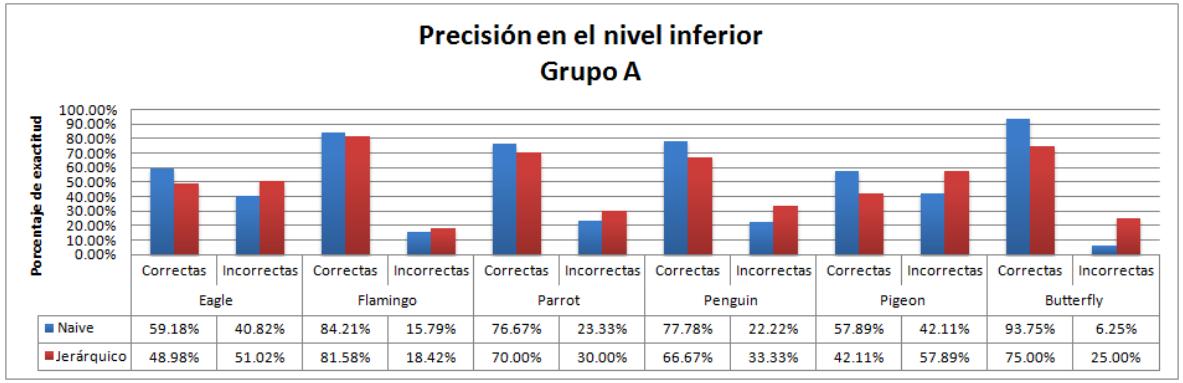


Figura 7. Gráfica de los resultados obtenidos considerando una probabilidad a priori de 0.5 para la clase Animal

Durante el proceso de evaluación, se observó que esta configuración para el nodo superior, brindaba mejores resultados que considerando la jerarquía completa. Esta suposición se hace valida debido a que al tomar unicamente la rama animal de la jerarquía, no se toma en cuenta su probabilidad con respecto a las demás ramas en su mismo nivel, ya que la red no propagará información de ninguna de las otras ramas.

El clasificador bayesiano simple que se utilizo para la comparación en nuestro modelo, fue diseñado con las probabilidades condicionales utilizadas para cada clasificador bayesiano de los atributos de la red, con la modificación de asignar como probabilidades a priori a cada clasificador, sus correspondientes probabilidades condicionales tomadas de los nodos en la red bayesiana.

En las gráficas mostradas en las figuras 8 a la 13 se presentan los resultados obtenidos de la clasificación por cada clase y por niveles inferior, intermedio y superior.



	Eagle		Flamingo		Parrot		Penguin		Pigeon		Butterfly	
	Correctas	Incorrectas	Correctas	Incorrectas	Correctas	Incorrectas	Correctas	Incorrectas	Correctas	Incorrectas	Correctas	Incorrectas
Naive	59.18%	40.82%	84.21%	15.79%	76.67%	23.33%	77.78%	22.22%	57.89%	42.11%	93.75%	6.25%
Jerárquico	48.98%	51.02%	81.58%	18.42%	70.00%	30.00%	66.67%	33.33%	42.11%	57.89%	75.00%	25.00%

Figura 8. Gráfica comparativa del porcentaje de clasificaciones correctas obtenidas con el Naive Bayes y la red Jerárquica

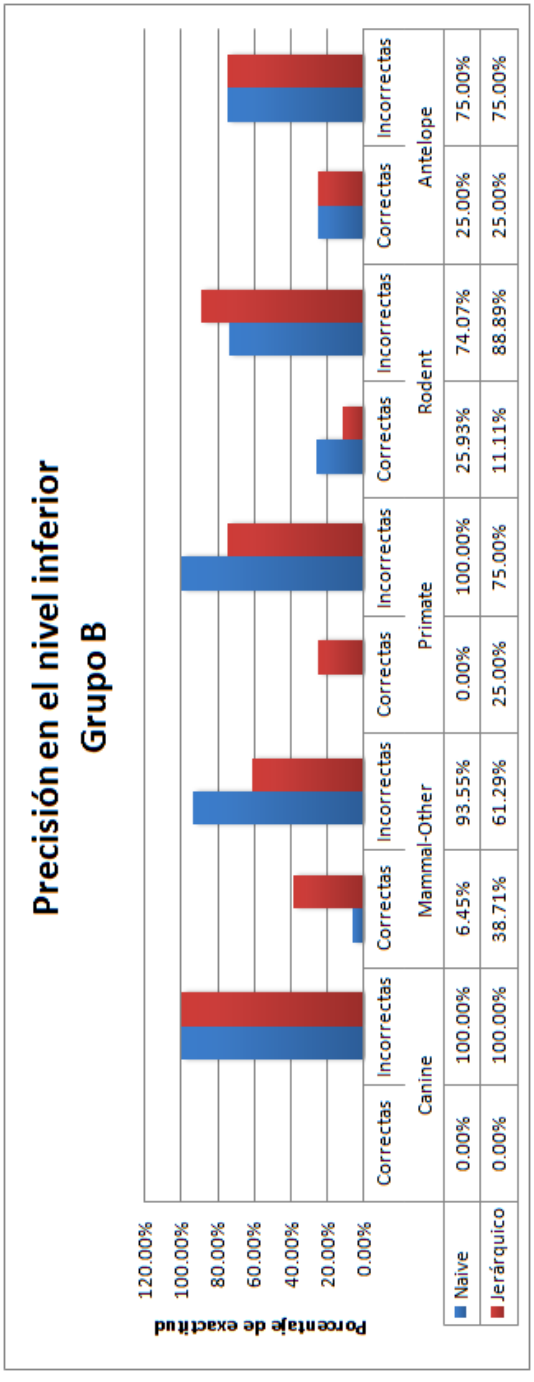


Figura 9. Gráfica comparativa del porcentaje de clasificaciones correctas obtenidas con el Naive Bayes y la red Jerárquica

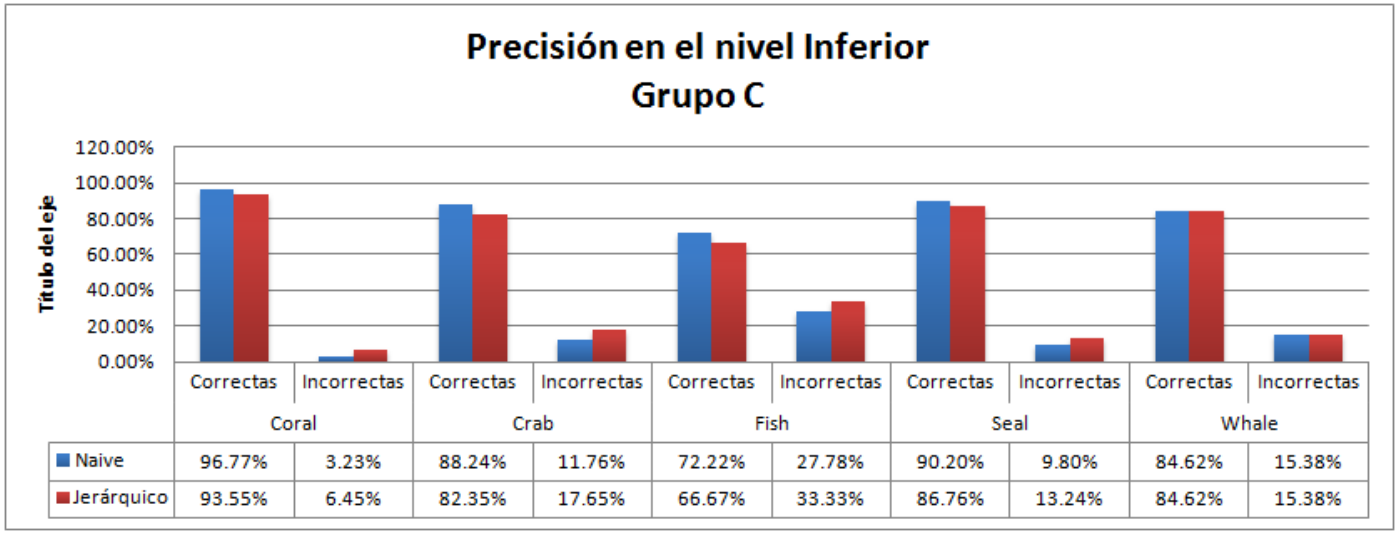


Figura 10. Gráfica comparativa del porcentaje de clasificaciones correctas obtenidas con el Naive Bayes y la red Jerárquica

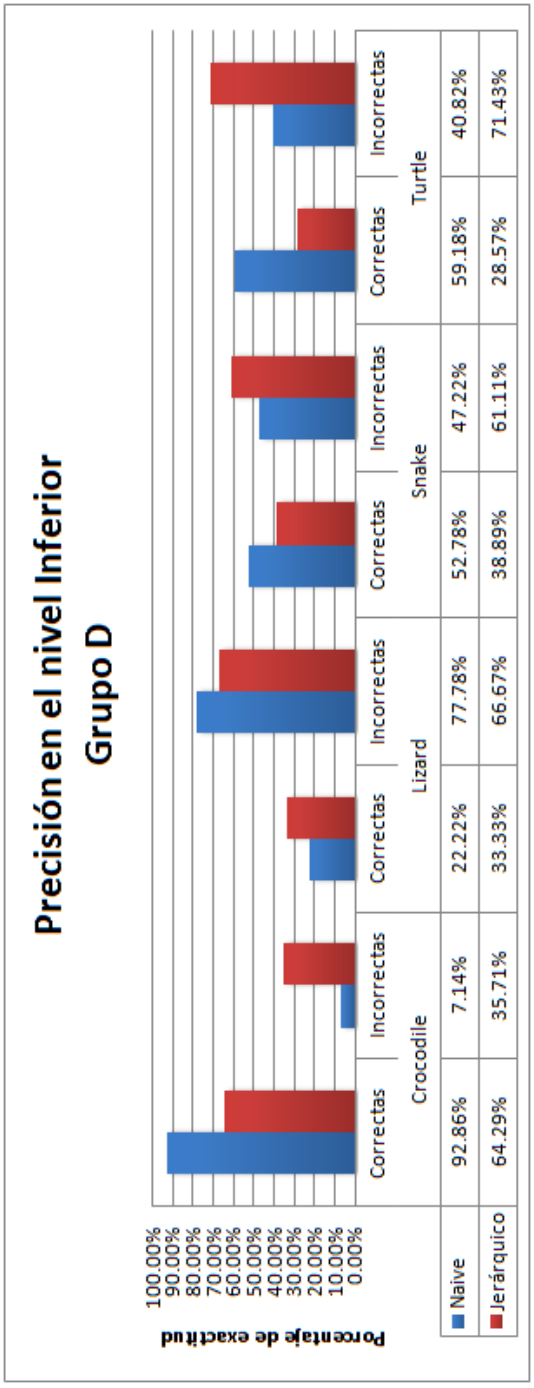


Figura 11. Gráfica comparativa del porcentaje de clasificaciones correctas obtenidas con el Naive Bayes y la red Jerárquica

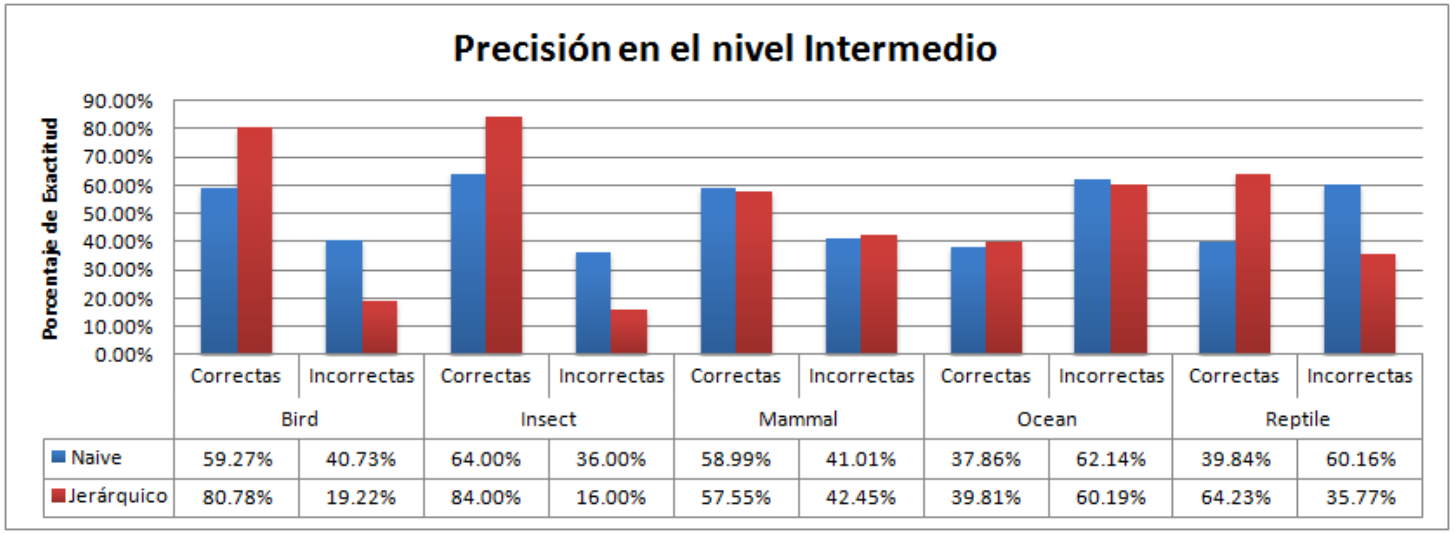


Figura 12. Gráfica comparativa del porcentaje de clasificaciones correctas obtenidas con el Naive Bayes y la red Jerárquica

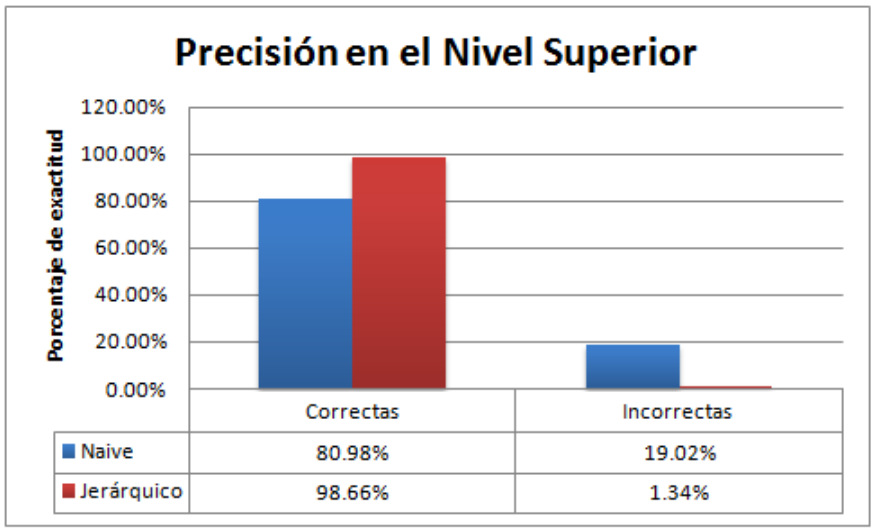


Figura 13. Gráfica comparativa del porcentaje de clasificaciones correctas obtenidas con el Naive Bayes y la red Jerárquica