Proyecto Final

Aprendizaje de Máquina - Otoño 2017

Juan Pablo Fonseca Correa
138263
15/12/2017

Introducción

En América del Sur existen distintos tipos de especies de ranas. La gran mayoría emiten llamados (croídos y cantos) distintivos a la familia taxonómica a la que pertenecen. Es de utilidad para los biólogos y zoólogos lograr identificar la familia de las ranas solamente por escuchar los llamados. Por ejemplo, para detectar las familias de ranas que habitan una zona, es mucho más rápido y efectivo realizar grabaciones de audio y luego analizarlas, que tener que buscarlas y verlas individualmente. Además, escuchar a las ranas genera una perturbación mucho menor en el hábitat que la anteriormente obligada visualización de ellas.

Se utilizó una base de datos generada a partir de llamados (croídos y cantos) de anuros (ranas), con tres columnas de etiquetas: la familia, el género y la especie a la que pertenecen. Este conjunto de datos se ha utilizado en varias tareas de clasificación o reconocimiento de especies de ranas a través de sus llamados. Se creó para clasificar 60 registros de audio pertenecientes a 4 familias diferentes, 8 géneros y 10 especies. Cada uno de los audios corresponde a una rana individual y la identificación del registro también se incluye como una columna adicional. Cada registro se capturó en condiciones de ruido real. Después del registro se identificaron 7195 sílabas distintas. Algunas especies son del campus de la Universidad Federal de Amazonas, Manaus, otras de Mata Atlántica, Brasil, y una de ellas de Córdoba, Argentina. Las grabaciones se almacenaron en formato wav con 44,1 kHz de frecuencia de muestreo y 32 bits de resolución, lo que permite analizar señales de hasta 22 kHz. De cada sílaba extraída, 21 MFCC (Coeficientes Cepstrales en las Frecuencias de Mel) se calcularon mediante el uso de 44 filtros triangulares. Estos coeficientes se normalizaron entre -1 y 1.

MFCC

Los Coeficientes Cepstrales en las Frecuencias de Mel (MFCC) son coeficientes que forman colectivamente un cepstrum de las frecuencias de mel (MFC). El cepstrum de una señal se obtiene tomando la transformada de Fourier inversa del logaritmo del espectro estimado de una señal. Los coeficientes se derivan de un tipo de

representación cepstral del clip de audio (un "espectro-de-espectro" no lineal). El MFC deforma la frecuencia para que se parezca más a la gama de frecuencia sonora humana que con otros tipos de cepstrums.

Metodología

La base de datos fue tomada del archivo de bases de datos de la UCI Machine Learning Repository. Contiene 7195 observaciones y 21 atributos sin contar la etiqueta de clasificación. Lo primero que se realizó con ella fue limpiarla. Se revisó que no tuviera valores nulos o inconsistentes y estos se removieron. Afortunadamente la base está prácticamente completa, por lo que la labor de limpieza fue mínima.

Lo siguiente que se hizo fue seleccionar los atributos relevantes o de interés. Dicha base tiene tres columnas de etiquetas: la familia, el género y la especie de las ranas. Para simplificar el problema, se descartaron las últimas dos columnas y se retuvo únicamente la de la familia taxonómica. Se observan cuatro familias registradas: Leptodactylidae, Hylidae, Dendrobatidae y Bufonidae.

No fue necesaria una transformación de los datos numéricos, puesto que estos ya están escalados al intervalo del -1 al 1. Es decir, la transformación ya ha sido previamente realizada.

Después de realizar este trabajo de preprocesamiento, se procedió a aplicar diversas técnicas de aprendizaje de máquina a la base de datos. En la sección de resultados se evalúan algunas de estas técnicas con las medidas de *precision*, *recall* y *accuracy*. Dichas medidas son relevantes en el contexto del problema ya que indican para cada clase, respectivamente, el porcentaje de ranas correctamente clasificadas en relación al número total de ranas seleccionadas para la clase, el porcentaje de ranas correctamente clasificadas en relación al número total de ranas realmente pertenecientes a esa clase, y el porcentaje de ranas clasificadas correctamente en general. Altos porcentajes de estas medidas en un clasificador indican que éste es en general uno bueno.

Análisis de Componentes Principales

La técnica de análisis de componentes principales o PCA por sus siglas en inglés nos permite distinguir las direcciones en las cuales existe una mayor variabilidad. Es decir,

PCA puede reducir la dimensionalidad de los datos proyectando los datos originales a un hiperplano de dimensión posiblemente mucho menor al original, conservando quizá el 80% o 90% de la información de la base. Dicho hiperplano consiste precisamente de las direcciones de mayor variabilidad.

El análisis de componentes principales puede ser considerado también una técnica de preprocesamiento, pues a partir de los resultados se pueden aplicar otros métodos de aprendizaje de máquina como K vecinos más cercanos, por ejemplo.

La base de datos estudiada contiene 21 atributos, por lo cual es deseable revisar si es posible reducir dimensionalidad por medio de componentes principales. Así, se aplicó PCA probando con distintos valores de componentes (dimensiones) que rescatamos. Se probó con distintos números de dimensiones. A continuación se muestra la varianza explicada de los primeros 7 componentes con mayor variabilidad:

Número de componente principal	**Variabilidad**
1	0.20178777
2	0.11151925
3	0.0528705
4	0.03357853
5	0.02834302
6	0.02183891
7	0.0170243

De la tabla mostrada arriba podemos observar que los primeros 4 componentes explican alrededor de un 40% de la variabilidad. Para la utilidad de PCA normalmente vista con otras bases de datos, esta cifra es de hecho bastante baja. Existen bases de datos para los cuales solamente los primeros 2 componentes principales logran explicar arriba de un 85% de la variabilidad total de los datos. En esos casos resulta muy conveniente, pues permite graficar los datos en dos dimensiones sin perder mucha información.

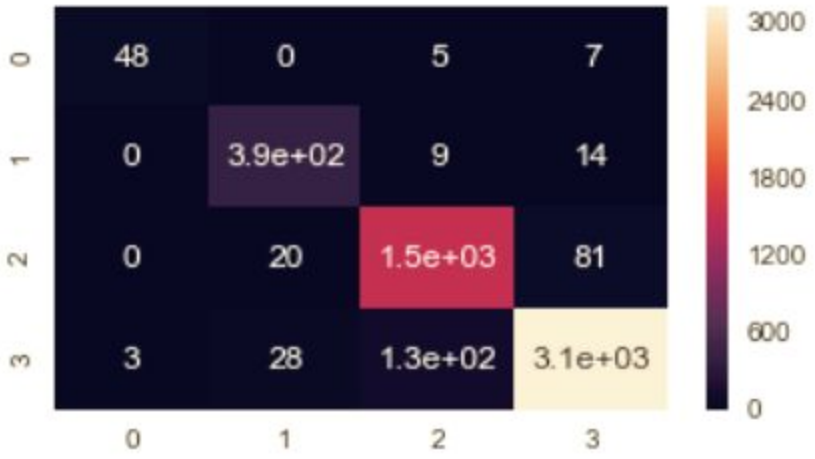
En la sección de resultados se analizará un poco más a fondo acerca de las implicaciones que los valores de la variabilidad obtenida tienen.

K vecinos más cercanos

A la base de datos original se le aplicó la técnica de los K vecinos más cercanos. Dicha técnica genera regiones en el hiperplano, cada una de las cuales corresponde a una categoría de la base estudiada. Así, un nuevo elemento sin etiqueta puede ser analizado desde el punto de vista de la región en la que caería bajo el modelo construido con K-vecinos cercanos, y así poder clasificarlo. Como se mencionó anteriormente, las clases con base en las cuales los datos serán entrenados y luego clasificados, son las familias taxonómicas de las ranas.

Para este modelo de aprendizaje supervisado, es necesario dividir los datos en un conjunto de entrenamiento y un conjunto de prueba. Se realizó esto con la base de datos de las ranas, dejando un 75% de las observaciones para el conjunto de entrenamiento y el resto para el conjunto de prueba. El hiperparámetro k fue seleccionado mediante la técnica de validación cruzada, la cual nos arrojó un valor de cinco.

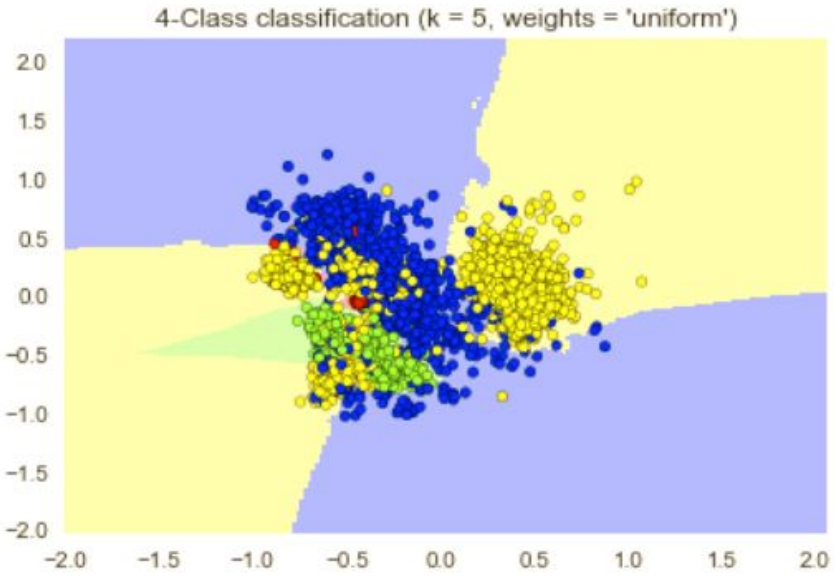
La clasificación obtenida para los datos de prueba es relativamente buena. Es importante considerar que la base de datos no está uniformemente distribuida en las familias taxonómicas de las ranas. Es decir, para algunas familias se tienen varias veces más datos que para otras. Incluso así, la matriz de confusión obtenida es la siguiente:



Podemos observar que la clase 3, Leptodactylidae, es claramente la más abundante de todas. Notamos que las observaciones se concentran principalmente en la diagonal de la matriz, lo que indica que el clasificador es generalmente bueno. En la sección de resultados se profundizará más al respecto.

Las regiones generadas por el clasificador de K vecinos más cercanos viven en un hiperespacio de dimensión 21, por lo que no es posible mostrar una gráfica de ello. Se pueden hacer, sin embargo, proyecciones bidimensionales en varias direcciones para quizá tener una idea mejor sobre el tipo de regiones.

Para el presente trabajo esto no se llevó a cabo. Sin embargo, se hizo algo parecido, aprovechando el trabajo de PCA que se describió en la sección anterior. Más concretamente, se tomaron las dos primeras componentes principales (las dos que más explican a los datos), y se proyectaron los datos al plano resultante de dos dimensiones. Ya teniendo datos bidimensionales en lugar de 21-dimensionales, se aplicó nuevamente K-vecinos más cercanos. Es esperable que el desempeño no vaya a ser tan bueno, ya que se está perdiendo información al hacer PCA. Sin embargo, ya se puede mostrar una gráfica de las regiones generadas por K-vecinos en el plano con mayor variabilidad definido por PCA. La gráfica obtenida es la siguiente:



Es interesante observar en la gráfica que la segunda clase más grande (círculos amarillos) está aparentemente dividida en dos regiones. Esto probablemente no sea así en dimensiones mayores. Otro tipo de análisis haría ver si en verdad está dividida la

clase de esta manera o no. Las dos clases más pequeñas también se pierden casi completamente entre las dos más grandes. Esto hace ver que una gran desventaja del método de K vecinos más cercanos es que si se tienen clases distribuidas no uniformemente, quizá unas sean más difíciles de identificar. Se probó decreciendo el valor de la k, pero esto ocasionó un problema de sobreajuste.

Naïve Bayes

Uno de los clasificadores más intuitivos de entender es Naïve Bayes, o Bayes Ingenuo. Se basa en la regla o teorema de Bayes, que habla acerca de cálculo de probabilidades a priori y a posteriori. El clasificador aproxima la probabilidad de una observación de pertenecer a cada una de las clases, y asigna la clase con la mayor probabilidad obtenida. Las probabilidades a priori se obtienen del conjunto de entrenamiento.

Después de aplicar esta técnica, la matriz de confusión obtenida es la siguiente:

De primera impresión, parece que la clasificación realizada es buena. La mayoría de los datos se encuentran distribuidos en la diagonal de la matriz. Sin embargo, la clase más pequeña, Bufonidae, es difícilmente clasificada bien. El detalle del desempeño se muestra en la sección de resultados.

Agrupación jerárquica

Finalmente se realizó una agrupación jerárquica. La descripción del proceso se describe a continuación:

Primero, cada una de las observaciones vistas como vectores se dividieron entre el

tamaño que tienen como vector, lo que se conoce como normalización. Al realizar esto, todas las observaciones transformadas viven ahora en una hiperesfera de dimensión 21. Después, se construyó una matriz de distancias utilizando la distancia coseno. La distancia coseno permite identificar cercanía entre dos observaciones/vectores si el ángulo formado entre ellos es pequeño. A partir de esta matriz de distancias, se aplicó clusterización jerárquica utilizando *complete linkage*. El parámetro de altura de corte del dendograma se dejó libre para poder comparar distintos niveles de generalidad/particularidad de los clusters formados.

Los resultados obtenidos fueron interesantes; dependen de la especificidad pedida para los clusters (la altura de corte del dendograma). Como se sabe, en la agrupación jerárquica se obtienen clusters de distintos tamaños dependiendo de un parámetro de tolerancia relacionado con la distancia entre los clusters. Se hablará más de esto en la sección de resultados.

Resultados

En las secciones de arriba ya se han comentado algunos de los resultados obtenidos. En esta sección se detalla más al respecto. Los resultados obtenidos fueron los siguientes:

Análisis de Componentes Principales

Como se mencionó anteriormente, después de realizar PCA se observó que la variabilidad explicada por los primeros cuatro componentes era apenas del 40%. De todos modos, este análisis nos da información importante. Por un lado, nos dice que no es viable reducir dimensionalidad drásticamente, pues se perdería demasiada información. Por otro lado, nos informa que los datos no están predominantemente dispersos en una o pocas direcciones/dimensiones; al contrario, estos se dispersan a lo largo de distintas dimensiones.

Obsérvese de todas maneras que los primeros dos componentes nos brindan un 30% de información/variabilidad de los datos. Si bien es muy poca, considérese que el total de componentes son 21. Dicho esto, quizá podemos quedarnos con estos dos componentes como los más representativos de la base de datos.

K vecinos más cercanos

El detalle del desempeño de K vecinos más cercanos se dará en relación a tres medidas: *precision*, *recall* y *accuracy*. Considérese que dichas medidas fueron después de una evaluación del método al conjunto de prueba de los datos. A continuación se muestra la información obtenida:

	Bufonidae	Dendrobatidae	Hylidae	Leptodactylidae
Precision	0.8	0.944	0.938	0.952
Recall	0.941	0.891	0.916	0.968

Accuracy para el método: 0.945

A excepción del valor de *precision* para la clase Bufonidae (la más pequeña), observamos valores arriba de 0.9 en las demás medidas de evaluación. Esto nos indica que el método en general, si bien tuvo un buen desempeño, no fue excelente. Un error del 10% puede ser importante. Esperablemente, la clase con mejor evaluación de desempeño fue Leptodactylidae, la más grande. En ella observamos errores de clasificación por debajo de 5%. El valor alto de *accuracy* obtenido, 94.5%, no debe de engañarnos, pues está fuertemente influenciado por el alto desempeño de la clase más grande.

Naïve Bayes

El detalle del desempeño de Naïve Bayes se dará en relación a tres medidas: *precision*, *recall* y *accuracy*. Nuevamente es importante considerar que dichas medidas fueron obtenidas a partir de una evaluación del método al conjunto de prueba de los datos. A continuación se muestra dicha información:

	Bufonidae	Dendrobatidae	Hylidae	Leptodactylidae
Precision	0.75	0.945	0.865	0.865
Recall	0.2	0.553	0.857	0.965

Accuracy para el método: 0.87

En comparación con K vecinos más cercanos, vemos un desempeño menos deseable. Sí observamos un par de evaluaciones alrededor del 95%, pero las demás se encuentran alrededor del 80% o incluso más bajas. El *recall* de la clase más pequeña, Bufonidae, es de solamente 20%, lo cual indica que la técnica no logra categorizar correctamente al 80% de los integrantes reales de esta clase. Se puede decir algo similar del *recall* de la clase Dendrobatidae, cuyo valor es de 55.3%. Nuevamente, el mejor desempeño se nota en la clase más grande, Leptodactylidae.

Agrupación jerárquica

Como se comentó anteriormente, los resultados de la agrupación jerárquica dependen de la generalidad pedida de los clusters. Si bien se tienen cuatro clases en realidad, el método permite refinar en más clases o generalizar en menos.

Si se establece un corte de dendograma muy arriba, los clusters obtenidos son muy generales. Esto quiere decir que la clase más grande, Leptodactylidae, es exitosamente agrupada en un solo cluster. También ocurre con la segunda clase más grande. Sin embargo, las otras dos clases en ocasiones se mezclan entre sí o con las otras dos grandes; hay tan pocas observaciones que el algoritmo no logra detectar una agrupación significativa de datos.

Por otro lado, si se establece un corte de dendograma muy abajo, ocurre aproximadamente lo inverso. Es decir, las clases pequeñas son correctamente identificadas como clusters aislados. Sin embargo, las clases grandes se dividen en varios clusters pequeños, lo cual tampoco es conveniente.

Finalmente, se buscó una altura de corte tal que se generaran cuatro agrupaciones, y con ayuda de la similitud de Jaccard, se asoció cada agrupación con cada una de las etiquetas originales. Para esta configuración, se calcularon nuevamente *precision*, *recall* y *accuracy*. Nótese que como el método es no supervisado, no hay división en un conjunto de entrenamiento y uno de validación, por lo que las mediciones fueron calculadas sobre todos los datos. Considerando esto, a continuación se muestra la información:

	Bufonidae	Dendrobatidae	Hylidae	Leptodactylidae
Precision	0.191	0.243	0.42	0.941
Recall	0.953	0.912	0.831	0.652

Accuracy para el método: 0.74

Definitivamente el método que peor desempeño tuvo es éste. Esto se puede deber a varias razones. Primero, el método es no supervisado, por lo que la información de las etiquetas no es utilizada en absoluto para construir el modelo. Segundo, como las clases no están distribuidas uniformemente, los valores de *precision* son muy bajos para las clases más pequeñas, lo que indica que en sus clusters contienen demasiado sobre otras clases. Sin embargo, el *recall* más alto se observa precisamente en las clases más pequeñas. Esto significa que ambos clusters contienen a la mayoría de sus integrantes originales. De cualquier forma, clusterización jerárquica no es la manera óptima de abordar el problema en cuestión.

Conclusiones

Después de haber realizado un análisis sobre la base de datos mediante diversas técnicas de aprendizaje de máquina, observamos que el mejor clasificador fue K vecinos más cercanos. Éste tuvo métricas arriba del 90% en general, lo cual indica que es bueno. En el contexto del problema de reconocer familias de ranas con base en sus croídos y llamados, notamos que al menos para las especies estudiadas de América del Sur, un clasificador que ayudaría con esta tarea es K vecinos más cercanos. Esto permitiría con un alto grado de exactitud identificar ranas con sólo escucharlas y sin necesidad de verlas o atraparlas. Así, el hábitat en donde se encuentran se vería menos perturbado, sin esto afectar los estudios zoológicos y biológicos que se quieran realizar.

Existe una gran cantidad de técnicas de aprendizaje de máquina. En este trabajo solamente se exploraron unas pocas; un siguiente paso deseable sería explorar una mayor cantidad y diversidad de técnicas sobre la misma base de datos para intentar encontrar un aún mejor desempeño que el de K vecinos más cercanos.

Bibliografía

1. “Anuran Calls (MFCCs) Data Set” en *UCI Machine Learning Repository*, Homepage, 5 de diciembre de 2017, <https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29>
2. “List of Anuran families” en *Wikipedia*, Homepage, 8 de diciembre de 2017, <https://en.wikipedia.org/wiki/List_of_Anuran_families>
3. “Mel frequency cepstrum” en *Wikipedia*, Homepage, 8 de diciembre de 2017, <https://en.wikipedia.org/wiki/Mel-frequency_cepstrum>
4. “Jaccard index” en *Wikipedia*, Homepage, 10 de diciembre de 2017, <https://en.wikipedia.org/wiki/Jaccard_index>