ISSN 1870-4069

Algoritmos de aprendizaje supervisado para la clasificación de géneros musicales caracterizados mediante modelos estadísticos

Arturo Tepepa Cantero, Héctor Manuel Pérez Meana, Mariko Nakano Miyatake

Instituto Politécnico Nacional,
Escuela Superior de Ingeniería Mecánica y Eléctrica Unidad Culhuacán,
Sección de Estudios de Posgrado e Investigación,
México

sas_19_93@hotmail.com, {mnakano, hmperezm}@ipn.mx

Resumen. En nuestros días es común tener música almacenada en formato digital. Sin embargo debido a la gran cantidad de información que se tiene, es imposible realizar una clasificación adecuada de toda la música existente sin algún tipo software. Tomando en cuenta esta problemática, en el presente trabajo se desarrolló un algoritmo para realizar la clasificación automática de géneros musicales usando un proceso de segmentación empleando características espectrales tales como *centroide* (SC), *flatness* (SF) y *spread* (SS) así como una temporal, tal como la tasa de cruces por cero (ZCR); obteniendo vectores característicos de las pistas de audio. En la etapa de clasificación se utilizaron 4 clasificadores KNN, SVM, LDA y árboles de decisión, observando que la mejor clasificación para nuestros vectores se obtuvo usando la máquina de soporte vectorial (SVM). Finalmente se utilizó el proceso de *voting* para mejorar la clasificación obtenida usando segmentos individuales; dando como resultado un grado de clasificación mayor al 90%. Finalmente se clasificaron canciones en las cuales se utilizaba un solo instrumento musical por lo cual se obtuvieron mejores resultados muy próximos al 100% de clasificación.

Palabras clave: Clasificación géneros musicales, segmentación, clasificador, voting.

Supervised Learning Algorithms for the Classification of Musical Genres Characterized by Statistical Models

Abstract. Nowadays is common to have the music stored using some digital format. However, because the large amount of data, it is impossible to make an accurate classification of all existing music without some kind of software. Attending to this requirement, in this work we developed a musical genres using a segmentation process together with some spectral characteristics such as centroid (SC), flatness (SF) and spread (SS) as well as temporary characteristic

such as zero crossing rate (ZCR) getting the characteristic vectors of the music. In the step of classification 4 classifiers were used KNN, SVM, LDA and decision trees. From these results we noticed that the best classification using the estimated characteristics vectors was using support vector machine (SVM). To further improve the classification obtained using each individual segment, we use a voting method which provides a classification performance higher than 90%. Finally we classify several songs played with only one kind of musical instrument, obtaining classification results closely to 100%.

Keywords: Music genre classification, segmentation, classifier, voting.

1. Introducción

Durante los últimos años se ha tendido a almacenar las pistas de audio para su posterior uso ya sea en discos CD-DVD, Discos Duros HDD-SSD así como en internet, lo cual implica un desafío poder clasificar la información ya sea en línea o fuera de línea. Para realizar lo anterior se debe hacer un etiquetado de las pistas de audio se dice que las etiquetas son textos basados en la información semántica del sonido [1]. Así el análisis de la música se puede hacer de varias formas [2] donde se identifica la música por su género, artista, instrumentos y estructura, mediante el etiquetado, el cual puede ser manual o automático. El etiquetado manual permite una visualización del comportamiento de una pista de audio ya sea en dominio del tiempo o en dominio de la frecuencia tal como en el espectrograma, haciendo posible clasificar las canciones sin necesidad de escucharlas. Sin embargo la realización de este proceso conlleva mucho tiempo y esfuerzo, incluso problemas en la salud [3] en donde se muestra que "el volumen, la sensibilidad acústica, el tiempo y costo requeridos para un proceso manual de etiquetado es en general prohibitivo. Por su parte, para la realización de un etiquetado automático se necesitan 3 pasos fundamentales: el pre-procesamiento, la extracción de características y la clasificación [4]. En el presente trabajo se desarrolló un esquema de clasificación en el cual, inicialmente se procesa la señal de entrada para reducir el ruido, seguidamente se segmenta la señal, cuyos segmentos se procesan mediante dos esquemas de caracterización, una en el dominio de frecuencia y otra en el dominio del tiempo [5].

Se han hecho trabajos en la segmentación de audio [6] utilizando características básicas como la tasa de cruces por cero por siglas en inglés "ZCR" además del cálculo de energía en un periodo de tiempo muy corto "centroide", utilizando ventanas de 2.4s, donde se reportó una precisión de 98% en la clasificación. Existen además desarrollos en el procesamiento digital de imágenes [7] enfocándose en el espectrograma cuyo objetivo es la clasificación multiclase, donde el clasificador empleado fue la máquina de soporte vectorial (SVM) obteniendo resultados de 85% de clasificación multi-clase, en donde el sistema determina a cual clase pertenece la señal de audio bajo análisis. Aquí el clasificador SVM a pesar de ser un clasificador creado para clasificación binaria obtiene muy buenos resultados debido a la previa extracción de características en la pista de audio.

En este trabajo se utilizaron algunas características propuestas por Tzanetakis y Cook [8] cómo lo es el centroide espectral (punto donde el espectro se encuentra en equilibrio) y *ZCR* (valor promedio de las veces que la señal cruza cero en el eje x estando dominio temporal). Además se usaron otras características como *Spectral Flatness* (Valor de la cantidad de cambios en la frecuencia por trama) y *Spectral Spread* (Potencia alrededor de cada *Spectral Centroid* y la relación de un centroide con los demás).

2. Método propuesto

El sistema propuesto se muestra en la Fig. 1, el cual clasifica un conjunto de pistas de audio divididas entre 5 géneros. En particular durante el entrenamiento se emplearon 10 pistas de audio de cada género musical, las cuales fueron Cumbia, Pop, Rap, Rock y Salsa con frecuencia de muestro fs = 44100Hz, 16 bits de profundidad, con una duración de 5 – 10 minutos y formato wav. Se extrajeron 3 características espectrales *Spectral Centroid, Spectral Spread y Spectral Flatness*, así como una característica temporal *Zero Crossing Rate*. Finalmente se utilizaron 4 clasificadores: Decision Trees, Discriminant Analysis, Support Vector Machine (Gaussian), Nearest Neighbor Classifier.

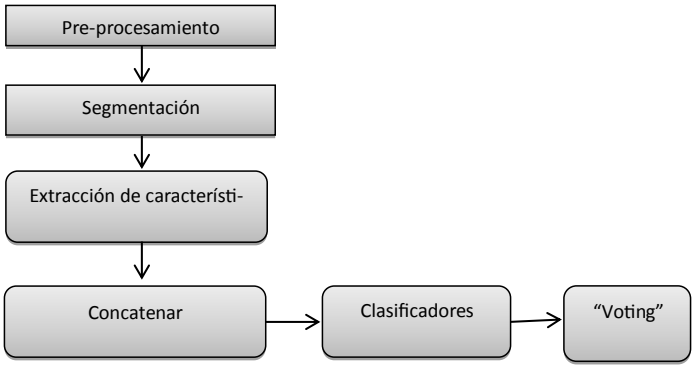

Fig. 1. Método de clasificación señales de audio.

2.1. Pre-procesamiento

En la etapa del análisis de datos, se observa que una parte importante de la música conserva información relevante a partir de una tercera parte de la canción hasta un poco más de las dos terceras partes, debido al silencio que usualmente se tiene en las pistas de audio al inicio y al final así que se toma como un factor importante la cantidad de muestras para enfocarse en esa región y observando que las canciones de mayor duración tienen el doble de duración entonces el número de muestras mínimas es la mitad de la duración máxima de las canciones que se tienen en la base de datos. Así la forma el número de muestras se calculad como se muestra en la ecuación (1):

$$\text{Numero de Muestras} = fs * \text{Tiempo}. \quad (1)$$

Las muestras para la canción más corta fueron alrededor de $13x10^6$, por lo que el máximo de muestras para la canción de mayor duración serán $26x10^6$. Estos valores son muy importantes para la segmentación que se hará posteriormente debido a que se tratará de eliminar partes de las canciones que no aportan información importante para su fututa clasificación.

2.2. Segmentación

En este proceso se toma la parte deseable de la canción "D" y se descartan las partes

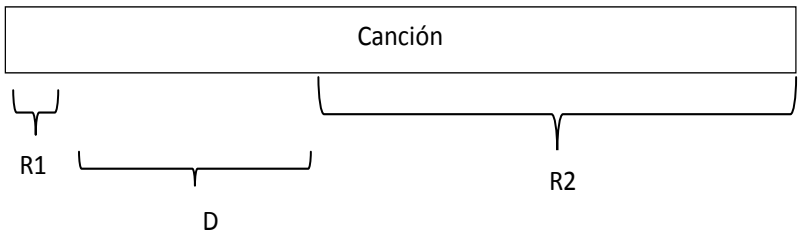

Fig. 2. Obtención de sección de análisis.

"R1y R2" correspondiente a posibles silencios como se muestra en la Fig. 2.

El valor R1 contiene de $.5x10^6$ muestras, para calcular el valor en segundo se usa la ecuación (2):

$$\text{Calculo de muestras en tiempo} = \text{número de muestras} / fs. \quad (2)$$

Son 11.2 segundos descartados (R1). La parte deseada "D" consistirá de 15 "*segmentos"* que sumados no deberán superar $13x10^6$ muestras, valor de muestras máximo de la canción con menos duración, así que se escoge el valor de $.332800x10^6$ el cual simboliza el número de muestras para cada "*segmento*". El tamaño de "D" esta expresado por la ecuación (3). Es importante señalar que los 15 "*segmentos"* deben ser etiquetas con el mismo género de la canción:

$$D = .3328x10^6 * 15. \quad (3)$$

D resulta de un tamaño de $4.992x10^6$ muestras, aplicando la ecuación (2) nos da un valor de 113 segundos de música para analizar sin el problema de silencio, aumentando la base de datos 15 veces y obteniendo segmentos del mismo tamaño sin importar la duración de la canción.

Finalmente cada uno de los 15 segmentos de "D" se divide en 650 "*sub-segmentos*" obteniendo 512 muestras por "*sub-segmento*", a las cuales se les aplicará la extracción de características. El número de vectores que serán utilizados para el entrenamiento está dado por la ecuación (4):

#_Vectores = #_Géneros * #_pistas_género*#_segmentos (4)
#_ Vectores = 5 * 10* 15 = 750.

2.3. Extracción de características

Como se mencionó en el inicio de este capítulo se utilizaran 4 modelos matemáticos que obtienen diversas características de la pista de audio, 1 de ellos se encuentra en un dominio temporal y 3 son espectrales así que se tendrá que calcular la FFT de las ultimas 3.

Zero Crossing Rate Este caracterizador es del tipo temporal, a cada instante de tiempo se le asigna un valor obtenido por un micrófono llamándose "muestra" que tiene valores positivos y negativos, que serán utilizados para calcular el número de cruces por cero con la ecuación (5):

$$\text{ZCR} = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} |sgn[x(m+1)] - sgn[x(m)]|, \quad (5)$$

donde x es el conjunto de muestras, m es la posición de la muestra y N el total de muestras. El objetivo del algoritmo es sumar las veces que cambia de signo una muestra con respecto a la anterior, significando que la señal de audio atravesó de valores positivos a negativos o viceversa en el eje x, se suman los valores obtenidos y se normaliza dividendo entre 2(N-1).

Como se observó en la sección de segmentación se obtuvieron 650 sub-segmentos con una longitud de 512 muestras a las cuales aplicando el algoritmo ZCR se obtienen 650 componentes normalizadas con 2 veces el total de muestras es decir 1024 formando un vector característico con su respectiva etiqueta.

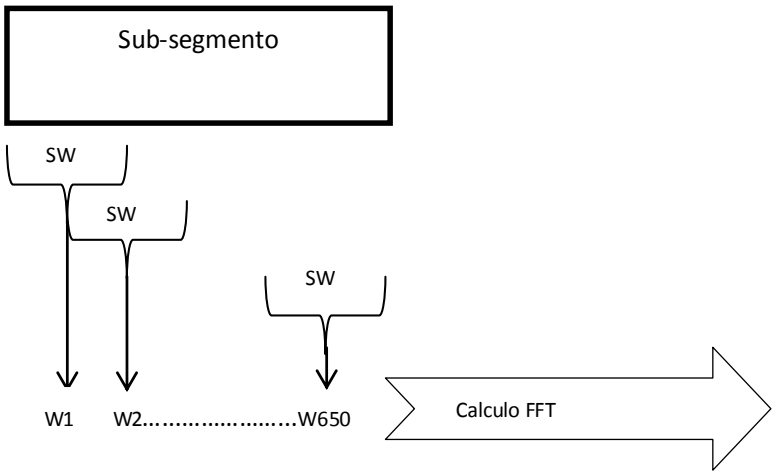

Fig. 3. Conversión dominio temporal a frecuencia.

Spectral Centroid Antes de calcular la Transformada Rápida de Fourier (FFT) y transformar los valores del dominio temporal al espectral es necesario un "ventaneo" de la

señal con un traslape, esto para disminuir el denominado efecto de "Gibbs" producido al cortar de forma abrupta la señal. Con esta finalidad se ocupó una venta Hamming de tamaño de 1024 (SW) que es el doble de tamaño de la cantidad de muestras para así traslapar al 50%, el esquema se muestra en la Fig. 3. Así después de "ventanear" la señal y obtener valores con un tamaño de 1024 se aplica FFT para posteriormente calcular *Spectral centroid* con la ecuación (6):

$$\text{SC} = \frac{\sum_{m=1}^{N-1} X(m) f(m)}{\sum_{m=1}^{N-1} X(m)}, \tag{6}$$

donde X representa los valores obtenidos de la FFT y f resulta de crear un vector de 1-1024 valores y dividir cada valor entre 1024, se crea una nueva escala que representa f.

Calculando los centroides en el "sub-segmento" se obtiene otro vector característico con un tamaño de 650 valores que se asemeja con el vector obtenido de *ZCR*, sin embargo esta característica representa la energía que se tiene en cada una de las ventanas.

Spectral Spread *Spectral Spread* representa la concentración de energía alrededor de cada *Spectral centroid*, Una característica importante es que entre mayor sea SS, significa que habrá una gran cambio en las frecuencias, se calcula enseguida de tener SC sobre la ventana de 1024 usando la ecuación (7):

$$\text{SS} = \sqrt{\frac{\sum_{m=1}^{N-1} \left(f(m) - SC(m)\right) * |X(m)|}{\sum_{m=1}^{N-1} |X(m)|}}. \tag{7}$$

La única diferencia en el cálculo de SS es que se realiza una sustracción a f con el SC que se obtuvo, además de calcular su raíz cuadrada en esa ventana.

Spectral Flatness Este rasgo pertenece al conjunto de características básicas [4] la cual indica que tan "plano" es el espectro con una serie de valores que expresan la energía del espectro dentro de una banda de frecuencia pre-definida, se ocupa la ecuación (8):

$$\text{SF} = \frac{\sqrt{\prod_{m=1}^{N-1} |X(m)|}}{\frac{1}{N} \sum_{m=1}^{N-1} |X(m)|}. \tag{8}$$

2.4. Concatenar vectores

Se tienen 4 vectores característicos cuyo tamaño es de 650, estos se concatenan y se obtiene 1 vector de tamaño de 2600, el orden será ZCR-SC-SS-SF-Etiqueta, este proceso se hará con los 750 vectores, obteniendo el descriptor final que se clasificará por 4 métodos.

2.5. Clasificadores

La clasificación se hace en cada sub-segmento cuya duración será de 11.2 ms, es decir habrá 750 clasificaciones y como se ha mencionado en el artículo se utilizaron 4 clasificadores; obteniéndose los valores de clasificación mostrados en la Figura 1. De estos resultados se observa que, como podría esperarse, empleando un solo sub-

Clasificadores	
Tipo	Precisión
Decision Tree	33.9 %
Discriminant Analysis	49.7 %
Support Vector Machine	58 %
K-Nearest Neighbor	49.3%

Fig. 1. Porcentaje de clasificación géneros musicales.

Gaussian SVM					
	Cumbia	Pop	Rap	Rock	Salsa
Cumbia	71	43	5	8	23
Pop	15	80	24	3	28
Rap	12	18	102	3	15
Rock	12	5	8	92	33
Salsa	20	8	17	15	90

Fig. 2. Matriz de confusión de géneros musicales.

segmento de 11.2 ms el porcentaje de acierto no es satisfactorio. Por otro lado se observa que el valor máximo de clasificación es obtenido usando SVM en una clasificación multiclase, donde el clasificador es requerido a determinar a cuál de las diferentes clases pertenece la señal de entrada. Así observando la matriz de confusión mostrada en la Figura 2. se tendrá una tendencia que aprovecharemos para utilizar el método denominado "Voting".

En la mayoría de los casos el valor obtenido en la diagonal principal de la matriz de confusión supera por más del doble y en el mejor de los casos (rap) la tendencia es 6 veces mayor que la mayor de las otras 4 opciones. Además observando los valores de la curva ROC del género cumbia (Representación gráfica de la comparación entre sensibilidad eje y contra la especificidad eje x y donde el valor máximo es 1) se aprecia una excelente clasificación Fig. 3. El valor de la curva ROC para todos los géneros se muestra en la Figura 4.

2.6. Voting

Una vez obtenida la clasificación de cada sub-segmento es decir 15 vectores clasificados de la misma canción se procede a escoger el valor que más se repite dentro de los 15 sub-segmentos como se observa en la figura 5.

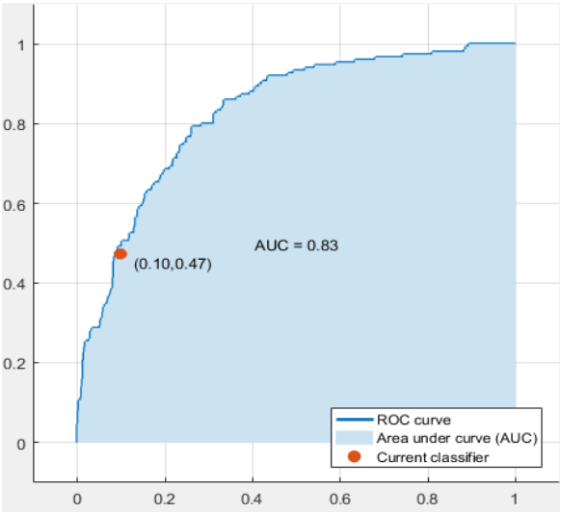

Fig. 3. Curva ROC clasificación géneros musicales.

Género	Valor curva ROC
Violín	0.99
Piano	0.96
Guitarra	0.97
Flauta T.	0.99

Fig. 4. Curva ROC clasificación géneros musicales.

C1	C1	C2	C1	C3	C1	C1	C1	C4	C5	C2	C1	C1	C1	C1

Fig. 5. Sub-segmentos clasificados.

En el caso de este ejemplo se tiene que hubo 10 valores para C1, 2 para C2, 1 para C3, 1 para C4 y 1 para C5, se escoge C1 debido a que es el que más se repite. Se aplicó esta clasificación para las 50 canciones de nuestra base de datos para obtener de resultado un 96% de clasificación, equivocándose en 2 canciones de cumbia, confundidas con pop.

3. Pruebas y resultados

Durante el proyecto se obtuvo una clasificación multiclase de 96% utilizando el Gaussian SVM y el método de voting con señales de audio que tienen componentes espectrales muy parecidas, se hizo el mismo procedimiento para clasificación de 10 canciones tocadas con 4 instrumentos musicales diferentes obteniendo excelentes

Clasificadores	
Tipo	Precisión
Decision Tree	71.2 %
Discriminant Analysis	80.8 %
Support Vector Machine	93 %
K-Nearest Neighbor	76%

Fig. 5. Porcentaje de clasificación instrumentos musicales.

Gaussian SVM				
	Violín	Piano	Guitarra	Flauta T.
Violín	146	1	2	1
Piano	0	137	10	3
Guitarra	12	0	135	3
Flauta T.	6	2	3	139

Fig. 6. Matriz de confusión instrumentos musicales.

Género	**Valor curva ROC**
Violín	0.99
Piano	0.96
Guitarra	0.97
Flauta T.	0.99

Fig. 7. Curva ROC clasificación instrumentos musicales.

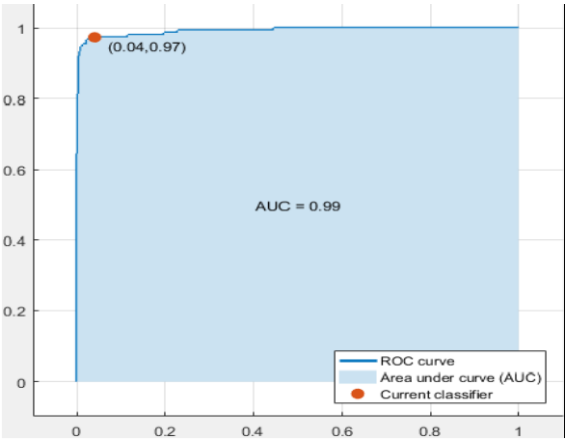

Fig. 8. Curva ROC clasificación instrumentos musicales.

resultados con los 5 clasificadores donde el SVM fue también el clasificador que proporción los mejores resultados, como se muestra en la Figura 5.

La matriz de confusión obtenida es bastante buena debido a la diagonal principal que contiene un porcentaje muy alto, como se muestra en la Tabla 5. Esto se observa también en la Tabla 6 y en la curva ROC Figura 6.

Aplicando voting la clasificación fue de 99% para instrumentos musicales obteniendo resultados satisfactorios y observando que dependiendo la cantidad de frecuencias involucradas (instrumentos y voz) será menor el índice de clasificación, aun así con el método de voting mejora notablemente.

4. Conclusiones

Para realizar una correcta clasificación es necesario conocer los vectores de entrada y discriminar defectos que pueden tener, además se observa que la clasificación depende en gran medida de que tan parecidos sean los vectores (géneros o instrumentos musicales), en estos casos las dos pruebas comprobaron esta idea aunque fuera el mismo tamaño, segmentación y extracción de características, los valores cambiaron ampliamente. El proceso de voting es muy eficiente si se tiene una clasificación de regular a buena en varios segmentos a pesar del nivel bajo obtenido por SVM.

Finalmente se puede asegurar que la utilización de los modelos estadísticos como lo es el SC, SF, SC y ZCR a pesar de tener un coste computacional bajo sirvieron bastante bien, esto nos da la pauta para aplicar posteriormente sistemas más robustos que obtengan características más eficientes que provoquen una mayor clasificación, inclusive no tan general como géneros sino como la pista en cuestión.

Referencias

1. Panagakis, Y., Kotropoulos, C.: Automatic music tagging via PARAFAC2. (ICASSP), IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (2011)
2. Mitrovic, D., Zeppelzauer, M., Eidenberger, H.: Analysis of the Data Quality of Audio Features of Environmental Sounds. Knowledge Creation Diffusion Utilization, pp. 4–17 (2006)
3. Lau, A., Mason, R., Pham, B., Richards, M., Roe, P., Zhang, J.: Monitoring the environment through acoustics using smartphone-based sensors and 3G networking. IEEE international conference on distributed computing in sensor systems (2008)
4. Greece-Duan, S., Zhang, J., Roe, P.: A survey of tagging techniques for music, speech and environmental sound, pp. 637–661 (2014)
5. Stowell, D., Plumbley, M.: A survey of UK birdsong and machine recognition for music researchers. Tech. Rep., pp. 09–12 (2011)
6. Lu, L., Zhang, H.J., Li, S.: Digital Object Identifier Multimedia Systems Content-based audio classification and segmentation by using support vector machines. Multimedia Systems, pp. 482–492 (2003)
7. Faisal-Ahmed, P.P., Paul, M.G.: Music Genre Classification Using a Gradiente-Based Local Texture descriptor. Springer International Publishing Switzerland, pp. 99–110 (2016)
8. Tzanetakis, G.: Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, pp. 293–302 (2002)