Comparador del Clasificador de Centroides vs Clasificador Gaussiano

Jordy Joaquin Cuan Robledo [1], Eduardo Cantorán Flores [1],
Rafael Pérez Aguirre [1]
[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, México

Resumen

En este reporte se presentan los detalles de dos clasificadores, un clasificador de centroides y un clasificador gaussiano. Se ha estudiado a detalle los clasificadores con sus fórmulas matemáticas que les describen y además se creó con un pequeño programa para poder clasificar nuevos valores. Más adelante se muestra el proceso utilizado para comparar estos dos clasificadores y por último se exponen los resultados conseguidos en esta investigación.

1. Introducción

Clasificar significa asignar un objeto en una clase. El clasificar lo que percibimos con los sentidos es algo natural en los seres humanos, tomamos ciertos patrones en común como referencia para hacer más fácil esta tarea, esto nos permite abstraer información la cual podemos representarla para la toma de decisiones.

Clasificar un objeto consiste en asignarlo a una de las clases disponibles. Los objetos se pueden definir por una serie de características, como pueden ser el color de sus píxeles, su textura o su tamaño.[2]

Para poder clasificar objetos es necesario definir las fronteras entre las diferentes clases. Normalmente estas fronteras se calculan mediante un proceso de entrenamiento en el que se usan las características de una serie de prototipos de ejemplo de las clases. Hablamos de fronteras por claridad, en general el clasificador infiere unas reglas de decisión durante el entrenamiento. [2]

Algunos ejemplos para poder clasificar eficientemente son:

- Segmentación de imágenes (por color, textura, etc.)
- Reconocimiento de objetos
- Control de calidad
- Detección de novedad (novelty detection), para detectar cambios o defectos en los objetos.
- Reconocimiento óptico de caracteres (OCR, Optical Character Recognition)

1.1. Recolección de datos

Para poder realizar técnicas de reconocimiento de patrones, en necesario contar con un subsistema de adquisición de datos, ya sea con ayuda de transductores que recolecten variables o, recolectando información manualmente referente al tema a analizar.

1.2. Extracción de características

La Extracción de Características es la etapa que se encarga, a partir del patrón de representación, de extraer la información discriminatoria eliminando la información redundante e irrelevante.

La cantidad de características de un patrón deben ser suficientes para poder determinar las diferencias de los objetos que se desean reconocer. Un buen clasificador debería utilizar el menor número de características que le permitan diferenciar eficazmente.

1.3. Selección del clasificador

Los clasificadores pueden ser representados mediante un modelo, el cual incluye las características que los patrones utilizaran y la manera en la que los datos de entrenamiento se utilizarán para poder reconocer correctamente.

El clasificador se encarga de la toma de decisiones en el sistema. Su rol es asignar los patrones de clase desconocida a la categoría apropiada.

1.4. Entrenamiento

Un sistema de reconocimiento de patrones debe aprender a clasificar objetos, este proceso se conoce como entrenamiento. En el entrenamiento supervisado los datos de entrenamiento consisten de parejas de entradas y sus salidas correspondientes. Aunque el entrenamiento supervisado es el más común, algunas veces no es posible tener etiquetados los datos de entrenamiento por lo que es necesario realizar un entrenamiento no supervisado, a este se le conoce también como agrupamiento pues eso es lo que realmente se tiene que hacer para asignarles una etiqueta sintética.

2. Conjunto de datos

Los datos empleados para los experimentos fueron tomados de un dataset que creamos, el cual contiene hojas de diferentes clases; se recolectaron 250 hojas por cada tipo, de las cuales medimos su ancho y su largo y de cada uno determinamos la media y la desviación estándar. Mostrado en las siguientes tablas.

Conjunto de Datos para Training de 225 elementos

CLASE	Media x	DsvStand x	Media y	DsvStand y
Níspero	4.9644	1.39315	16.4526	3.89438
Dolar	5.7196	1.48783	5.6544	1.55336

capulin	2.61486	1.20133	9.07189	1.37513
Bambú	2.5816	0.476329	13.7788	2.19698
ficus	3.9352	0.339325	8.3436	0.797171
Eucalipto	1.4976	0.302457	12.2568	3.22683
Manzana	3.5388	0.589308	6.6184	1.00803
Granada	2.15694	0.656764	5.72306	1.16845
Nogal	5.388	1.48619	11.042	3.04891
Aguacate	5.8996	1.62346	12.7552	2.85788
Nanche	2.63831	0.588282	7.9	1.9697
Fresno	4.052	0.904438	9.5904	2.21121
Naranja	3.8412	0.861226	8.1832	1.62173
Rosa	3.2944	0.679629	5.5756	1.0913

Conjunto de Datos para Test de 25 elementos

CLASE	Media x	DsvStand x	Media y	DsvStand y
Níspero	4.87511	1.33677	16.1193	3.7254
Dolar	5.62089	1.41631	5.58889	1.53029
capulin	2.575	1.23891	8.96161	1.34786
Bambú	2.568	0.469563	13.7467	2.2239
ficus	3.928	0.341697	8.31956	0.794102
Eucalipto	1.48844	0.305675	12.1733	3.29568
Manzana	3.52	0.592588	6.56844	1.01222
Granada	2.17748	0.672271	5.70252	1.12146
Nogal	5.37156	1.53375	10.9422	3.08864

Aguacate	5.844	1.65742	12.6591	2.90279
Nanche	2.66307	0.596322	8.075	1.9506
Fresno	4.02889	0.886864	9.304	1.992
Naranja	3.77467	0.846225	8.05911	1.57617
Rosa	3.25378	0.671084	5.50356	1.0682

3. Clasificadores

El término clasificador se utiliza en referencia al algoritmo utilizado para asignar un elemento entrante no etiquetado en una categoría concreta conocida. Dicho algoritmo, permite pues, ordenar o disponer por clases elementos entrantes, a partir de cierta información característica de éstos. Una manera de implementar un clasificador es seleccionar un conjunto de ejemplos etiquetados y tratar de definir una regla que permita asignar una etiqueta a cualquier otro dato de entrada.

En ocasiones, el término clasificador también es utilizado para referirse a la función matemática que implementa el algoritmo de clasificación.

Para la implementación de un clasificador es necesario tener en cuenta una serie de características concretas. La selección de éstas, sin embargo, no es una tarea sencilla.
La adición de muchos parámetros irrelevantes hace más difícil la clasificación para todos los métodos. Además, a medida que vamos añadiendo más información se van incrementando las dimensiones del espacio, hecho que supone que la optimización sea cada vez más difícil.

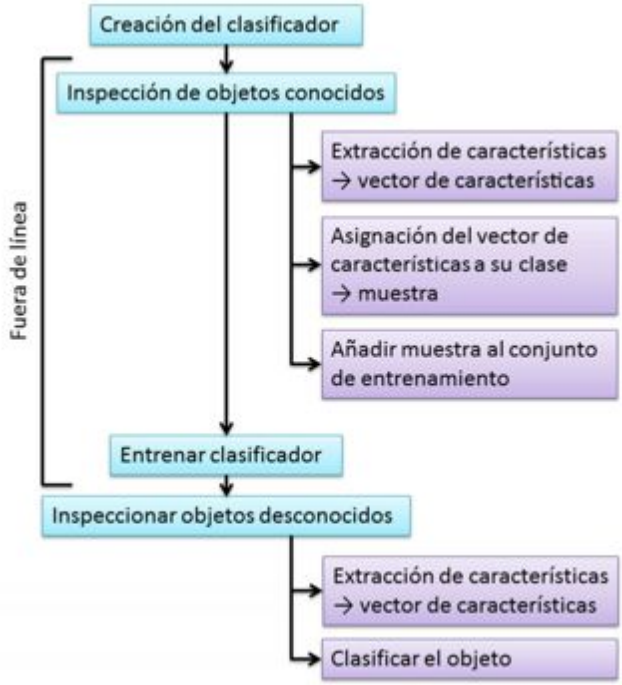

Figura 1. Pasos del proceso de entrenamiento y uso de un clasificador.

Fase de Entrenamiento

La finalidad de esta fase es construir un conjunto/grupo para que los objetos que se han clasificado como ciertos, sean conocidos/identificados. Los parámetros característicos que describen a los objetos deben ser discriminatorios para la clasificación.

Se emplea un conjunto de entrenamiento el cual debe contener una lista de objetos con tipos conocidos. Idealmente este conjunto de entrenamiento debería contener muchos ejemplos, de este modo, se incluirían objetos comunes y no comunes.

Para crear el conjunto de entrenamiento se requiere una fuente de objetos clasificados de forma cierta.

Fase de Prueba

Una vez construido el clasificador se debe medir la precisión. Este paso es necesario tanto en la aplicación del clasificador como también para poder compararlo con otros clasificadores diferentes.

La precisión se puede determinar aplicando al clasificador un entrenamiento independiente de un conjunto de objetos de los que se conoce la clasificación. Se hace uso de conjuntos conocidos ya que a veces no se tienen las fuentes necesarias para construir un nuevo modelo que se utilice puramente para su testeo. Se debe evitar el entrenamiento y el testeo con el mismo conjunto.

3.1. Clasificador de Centroides

La distancia permite cuantificar el concepto de similitud entre dos elementos respecto a un conjunto de características que ambos exhiben. Establece lejanía y proximidad entre dos objetos.

En el conjunto R^n, definido por vectores reales de dimensión n, la función de distancia más común es la Distancia Euclidiana. La siguiente expresión corresponde con dicha función de distancia.

$$d(x,\ y)\ = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} = \sqrt{(x-y)(x-y)^T}$$

En el espacio R^2 corresponde con la longitud del segmento de recta que une dos puntos. En este mismo espacio, el conjunto de los puntos equidistantes a un punto $x,\ C\ =\ \{c_i\ :\ d(x,c_i)\ =\ r\}$, forma una circunferencia, de radio r con centro en x .

3.2. Clasificador Gaussiana

El clasificador basado en gaussiana es óptimo cuando las diferencias entre los componentes de los patrones de la clase correspondiente siguen una distribución normal de media cero y una determinada varianza. No obstante, basta que aparezcan desviaciones con respecto a la hipótesis de normalidad, relativamente pequeñas, o que las distribuciones de las diferencias presenten heterocedasticidad para que este clasificador deje de ser óptimo. Por otra parte, no todas las situaciones en que debe de realizarse una clasificación de patrones responden al esquema en que las fluctuaciones con respecto al patrón de referencia de la clase, vienen determinadas por la adicción de un término estocástico, con distribución normal.

La fórmula que describe este clasificador se muestra a continuación.

$$\frac{1}{\sigma\sqrt{2\pi}}\ .\ e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

$\Theta = \{\mu\ ,\ \sigma\}$

Gaussiana $(x^*)\ =\ arg_c\ max\ N(x^*;\ \Theta)$

Con esta fórmula comparamos las distancias obtenidas contra la distancia global, logrando así terminar con la clase ganadora.

4. Experimentos

Para los experimentos se realizó un 10-Fold Cross Validation en la que conseguimos los siguientes resultados.

RESULTADOS	**% Gaussiano**	**% Centroides**

modelo 1	48.2857	42
modelo 2	48.8571	43.1429
modelo 3	52.2857	48
modelo 4	54.2857	45.4286
modelo 5	57.1429	47.1429
modelo 6	56	44.5714
modelo 7	48.5714	38.2857
modelo 8	45.7143	41.4286
modelo 9	52.9052	40.367
modelo 10	43.2099	40.4321
PROMEDIO	50.7258	43.0799

5. Conclusiones

Se han propuesto dos clasificadores los cuales hemos probado con diferentes conjuntos de entrenamiento y ejecutado diferentes pruebas, donde se consiguió una comparación más sólida del clasificador de centroides con el clasificador gaussiano. Se ha conseguido demostrar que con aproximadamente un 7%, el clasificador gaussiano es ligeramente mejor que un clasificador por centroides.

Hemos empleado un nuevo método llamado 10-Fold Cross Validation el cual nos ha devuelto diferentes porcentajes de precisión, de los cuales conseguimos el promedio de cada uno y podemos observar una precisión del clasificador más detallada.

Como trabajo a futuro, nos gustaría estudiar otros tipos de clasificadores y realizar una comparación de presición para mejorar la viavilidad y experiencia en el la clasificación de datos y poder tener mejores resultados en el tratamiento de información.

6. Referencias

[1] Raúl Benítez, Gerard Escudero, Samir Kanaan (2014). Inteligencia artificial avanzada. Editorial UOC

[2] http://bibing.us.es/proyectos/abreproy/70448/fichero/05_Capitulo4.pdf

[3]

[4]