Reconocimiento de patrones para el análisis poblacional

Alexis Herrera Saucedo
Héctor Mauricio Zamudio Domínguez

***Abstract*—El propósito de esta práctica se hizo con el fin de hacer un data set de las diferentes pirámides de población, dicho data set se utilizará para comparar la eficacia de dos métodos de clasificación los cuales son: mínima distancia y k_{nn} vecinos, con el fin de determinar las mejores características que arrojen un mayor porcentaje de eficacia en los clasificadores supervisados.**

***Keywords*—** Reconocimiento de patrones, Clasificación supervisada, Pirámides Poblacionales

I. Introducción

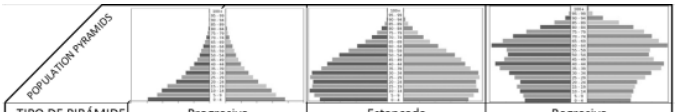
El reconocimiento de patrones es una actividad que se realiza cada día. Todos los días extraemos características de objetos para poder reconocer a que son, características como el color y sabor que nos podrían ayudar a clasificar si dicho objeto inspeccionado es una manzana o un plátano, incluso alguien puede tomar la característica de sabor para realizar la misma clasificación. El objeto el cual está siendo reconocido es un patrón. Un patrón es el conjunto de características que describen a un objeto y una vez realizada dicho reconocimiento se determina a que clase pertenece mediante un proceso de clasificación. Lo pasos para clasificar los objetos son:

- Extracción de características: Es el proceso de generar características que puedan ser usadas en el proceso de clasificación de los datos.
- Selección de variables: Aquí se debe seleccionar cuál es la característica más adecuada para describir los objetos.
- Obtener los datos: En base a las características seleccionadas, obtener los datos de dichas características que representa el Objeto.
- Clasificar: Determinar en base al Patrón a que clase pertenece.

En esta práctica los clasificadores utilizados son "Clasificadores Supervisados" son aquellos donde para cada elemento dentro de nuestro conjunto de entrenamiento conocemos su valor objetivo, es decir, conocemos la clase a la cual realmente pertenece. Esto permite proporcionar nuestro conocimiento al clasificador con el propósito de optimizar su criterio. Lo cual se expondrá su resultado al final de este reporte.

II. Creación del training test

Para comprobar el funcionamiento de distintos clasificadores y su eficacia, se tomó como objeto a clasificar pirámides de población. La recuperación de datos se extrajo de la página populationpyrmid debido a que era una de las páginas con más información sobre las poblaciones de todo el mundo. En las pirámides de población existen tres gráficas representativas las cuales se muestran a continuación:



TIPO DE PIRÁMIDE	Progresiva	Estancada	Regresiva
DESARROLLO	Subdesarrollado	En vías de desarrollo	Desarrollado
Tasa de natalidad	Muy alta	Reduciéndose	Baja
Tasa de mortalidad	Muy alta	Alta, pero reduciéndose	Baja
Esperanza de vida	Baja	Creciente	Alta
Grupos de edad	Joven - Adulta - Anciana	Adulta - Joven - Anciana	Adulta - Anciana - Joven

Fig. 1. Tipos de pirámides de población

La imagen anterior se tomó como referencia para clasificar las distintas pirámides de población. Ahora bien, las características que se tomaron para formar un Patrón para las pirámides poblaciones son el porcentaje de la población respecto a la población total del país o ciudad, este porcentaje se toma por edades con incrementos de 5 años y separados por sexo. En el eje X se utiliza para los porcentajes de población y en el eje Y los rangos de edades, en el lado derecho se muestran los datos de los Hombres y al izquierdo los datos de las Mujeres.

Una vez elegido las características se procedió a rellenar el data set de la siguiente manera:

País	0 a 4 años -> H	0 a 4 años -> M	5 a 9 años -> H	5 a 9 años -> M	
Alemania 2018	2.2	2.1	2.2	2	
Japón	2.1	2	2.2	2.1	
Italia	2.1	2	2.3	2.2	
Austria	2.5	2.3	2.4	2.3	...
Bulgaria	2.4	2.2	2.6	2.5	
Canadá	2.7	2.6	2.8	2.7	
Suiza	2.7	2.5	2.5	2.4	
España	2.2	2.1	2.6	2.4	

Fig. 2. Formato de captura de datos

Por último, todos los datos se copiaron a un archivo tipo csv para que el clasificador los pudiera interpretar, quedando de la siguiente manera:

9.8	8.1	7.8	6.5	6.3	5.5	5.3	estancada
9.2	7.7	7.5	6.5	6.4	5.1	5.1	estancada
9.8	8.1	7.6	6.7	6.2	5.5	5	estancada
8	7.1	6.9	6.3	6.2	5.5	5.4	Regresiva
7.8	7	7	6.2	6.2	5.4	5.3	Regresiva
8	7.2	7.1	6.1	6	5.1	5.1	Regresiva
7.3	6.7	6.4	5.9	5.7	5.6	5.3	Regresiva

Fig. 3. Estructura del archivo csv

Es importante mencionar el proceso de normalización de los datos para la creación de los datos de entrenamiento y prueba, debido a que inicialmente se pensó en tomar los atributos en base al número de personas por edad y no el porcentaje. Esto presentaría un problema importante para el correcto funcionamiento de los clasificadoras ya que, aunque dos ciudades presentaran un comportamiento similar los datos son radicalmente diferentes, por ejemplo, para un país como Estados Unidos donde un 3.0% son más de 9 millones de personas, en Guyana este mismo 3.0% son poco más de 23 mil personas. Utilizando el porcentaje logramos comparar el comportamiento de distintas ciudades o países sin importar el volumen de población que posean.

III. Clasificador mínima distancia

En el clasificador de Mínima Distancia se llevan a cabo dos etapas una de entramiento y otra de clasificación. En la etapa de entrenamiento se lee el data training generando una colección de Patrones, debido a que estamos realizando una clasificación supervisada, esto es, conocemos a que clase pertenecen estos Patrones; Mínima Distancia crea un Patrón representativo por cada una de las clases contenidas en el data training. En la etapa de clasificación,

se lee (data test) la colección de patrones a clasificar en base al entrenamiento antes realizado por el clasificador, para cada patrón p se estima que pertenezca a una clase en base a la distancia de p hacia los Patrones representativos y se asigna una clase resultante a p al cual se encuentra a la mínima distancia. Dicha distancia es calculada usando la Distancia Euclidiana entre el patrón p y cada uno de los Patrones Representativos.

$$d = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

donde
d: Es la distancia Euclidiana
n: Número de características
x: Son las características del Patrón x
y: Son las características del Patrón y

IV. Clasificador de vecinos cercanos

Este clasificador parte de la idea de que una nueva muestra será clasificada a la clase a la cual pertenezca la mayor cantidad de vecinos más cercanos del conjunto de entrenamiento. Para llevar a cabo el proceso de entrenamiento se recibe un conjunto de patrones de los cuales se conocen las clases involucradas. Por otro lado, los pasos para el proceso de clasificación son:

1) Se recibe el patrón a clasificar.
2) Se calculan la distancia para cada patrón con respecto a cada una las instancias restantes a clasificar, para calcular las distancias se utilizó la distancia euclidiana.
3) Se ordena el arreglo de instancias basándose en los patrones que se encuentran a menor distancia del patrón a clasificar.
4) Por último la clase a la que pertenece el patrón a clasificar se determina en base a la primera clase que logre tener k- vecinos más cercanos a dicho patrón.

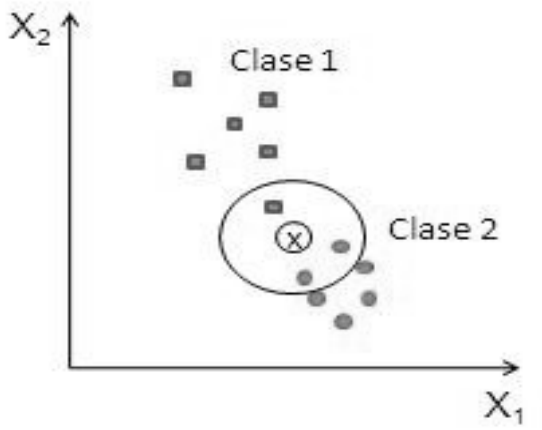


Fig. 4. Clase resultante: Clase 2. Para k=2.

V. Análisis de resultados

En el clasificador de Mínima Distancia en su etapa de entrenamiento generó los siguientes patrones representativos para cada uno de los tipos de pirámides poblaciones. Como se puede observar los patrones representativos se asemejan de manera correcta a los tipos de pirámides de la Fig. 1.

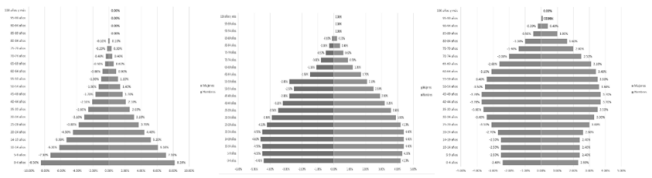


Fig. 5. Patrones Representativos de Mínima Distancia

Para medir la eficacia del clasificador, se creó una matriz de confusión la cual es una matriz que sirve como herramienta estadística para la evaluación de la eficacia de un clasificador o para el análisis de observaciones emparejadas.
Se trata de una matriz cuadrada de dimensión MxM donde M denota el número de clases en consideración. Las celdas de la diagonal de la matriz de confusión contienen las cantidades correspondientes a los ítems bien clasificados (coincide una con su correspondiente). Estas celdas las denominamos CC (celdas coincidencia). Las celdas de fuera de la diagonal contienen las cantidades correspondientes a las confusiones. La siguiente Matriz de confusión surge de clasificar el mismo data set con el cual fue entrenado mínima distancia.

		Referencia		
		Progresiva	**Estancada**	**Regresiva**
Conjunto de datos	**Progresiva**	20	0	0
	Estancada	0	20	0
	Regresiva	0	0	20

Fig. 6. Matriz de Confusión de Mínima Distancia

La manera de leer esta matriz es que para $M_{1,1}$ Son los datos de clase progresiva clasificados como clase progresiva, para $M_{1,2}$ Son los datos de clase progresiva clasificados como Estancada y así sucesivamente. Dado que Mínima distancia tiene una eficacia del 100% esto es, que no clasifica mal ningún patrón, podemos determinar que se trabaja con un proceso de clasificación lineal en el cual las clases son suficientemente distintas entre sí logrando así que el clasificador no cometa errores.

Por otra parte, el clasificador de K_{nn} vecinos también emplea la matriz de confusión para obtener la eficiencia de las diferentes combinaciones de k vecinos posibles. Todas las combinaciones posibles obtuvieron una eficacia del 100%, esto se debe a que las características seleccionadas son las más representativas lo que implica que ningún patrón está mal clasificado, a continuación, se muestra el resultado para k=3:

		Referencia		
		Progresiva	**Estancada**	**Regresiva**
Conjunto de datos	**Progresiva**	20	0	0
	Estancada	0	20	0
	Regresiva	0	0	20

Fig. 7. Matriz de Confusión de K_{nn} vecinos más próximos

Una vez entrenados los clasificadores, se procedio a generar un nuevo data set para testear dichos clasificadores dicho data set

contiene piramides poblacionales con fenomoenos de migracion y emigración para probar su funcionamiento en base a lo aprendido.

Para el clasificador de mínima distancia la mejor combinación posible arrojo una eficiencia del 55%, si se seleccionaban todas las características la eficiencia era del 30% menor que el clasificador K_{nn} vecinos. No hubo eficacia mayor al 55%.

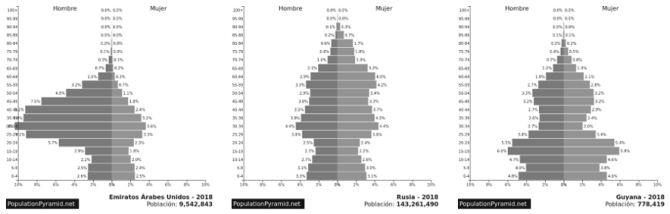


Fig. 8. Pirámides poblaciones con fenómenos de migración o emigración.

Para el clasificador K_{nn} vecinos, su eficacia máxima fue de 33% para k=3, esto se debe a que la única pirámide de población progresiva se clasifica mal lo cual reduce la eficiencia, además se probaron todas las combinaciones posibles y no hubo eficacia mayor al 33%.

En este caso el clasificador de mínima distancia resulto mejor que el clasificador K_{nn} vecinos.

VI. Conclusiones

La eficacia que pueden tener los clasificadores depende de distintos factores, entre ellos destaca la correcta selección de características de los cuales se almacenar los datos y que es más útil para el proceso de reconocimiento de patrones. Aunado a esto una correcta normalización de los datos logrando una mejor atracción de como percibimos los patrones.

Los clasificadores supervisados logran aprender lo enseñado se manera sobresaliente para esta prueba que fue clasificar Pirámides Poblacionales, reduciendo su eficacia solamente en casos muy particulares de poblaciones que presentan anomalías en su comportamiento.

VII. Bibliografía

1. Ariza, F. J., Rodríguez, J., y Fernández, V. (2018, 5 julio). CONTROL ESTRICTO DE MATRICES DE CONFUSIÓN POR MEDIO DE DISTRIBUCIONES MULTINOMIALES. Recuperado 16 septiembre, 2019, de http://www.geofocus.org/index.php/geofocus/article/view/591/460

2. Hidalgo, P. V. (2019). Clasificadores supervisados para el análisis predictivo de muerte y sobrevida materna. Recuperado 18 septiembre, 2019, de http://www.informatica2013.sld.cu/index.php/informaticasalud/2013/paper/viewFile/292/212

3. Friedman, M., y Kandel, A. (1999). Introduction to Pattern Recognition: Statistical, Structural, Neural, and Fuzzy Logic Approaches. Covent Garden, Lodon: World Scientific

4. Sancho, F. (2018, 26 diciembre). Clasificación Supervisada y No Supervisada. Recuperado 18 septiembre, 2019, de http://www.cs.us.es/%7Efsancho/?e=77

5. Duda, R. O., Hart, P. E., y Stork, D. G. (2001). Pattern classification (2ª ed.). New York, U.S.A: Wiley.

6. Ruiz, S. (2017, 20 julio). El algoritmo K-NN y su importancia en el modelado de datos. Recuperado 19 septiembre, 2019, de https://www.analiticaweb.es/algoritmo-knn-modelado-datos/