Temas Selectos de Inteligencia Artificial

Reporte Clasificación de Textos

Edgar Adrián Nava Romo

Profesor: Dr. Hiram Calvo

24 de Abril de 2018

Introducción

Existen dos tipos de clasificaciones, cuando el experto indica las clases en las que debe dividirse el dominio se llama clasificación supervisada; cuando el procedimiento de clasificación genera automáticamente las clases, sin intervención del experto, se llama clasificación no supervisada.
Se tiene una base de datos de una escuela primaria donde se pidió a 129 alumnos que proporcionaran una breve auto-biografía, se sabe que 48 son alumnos primogénitos y 81 no lo son, por lo tanto el método que se usará para clasificarlos será de manera supervisada

Objetivo

Clasificar por medio de un software clasificador de datos llamado Weka en su versión 3.2.8 y diferentes algoritmos de clasificación de texto de forma supervisada, esto es, indicando las clase en la que pertenece cada biografía, en este caso si es primogénito o no lo es, se intentará encontrar algún patrón entre los alumnos que ayude a facilitar la clasificación de alumnos de esta manera en un futuro, de igual forma se verificará si el algoritmo Naive Bayes visto en clase pertenece a uno de los mejores clasificadores de datos dentro de su campo.

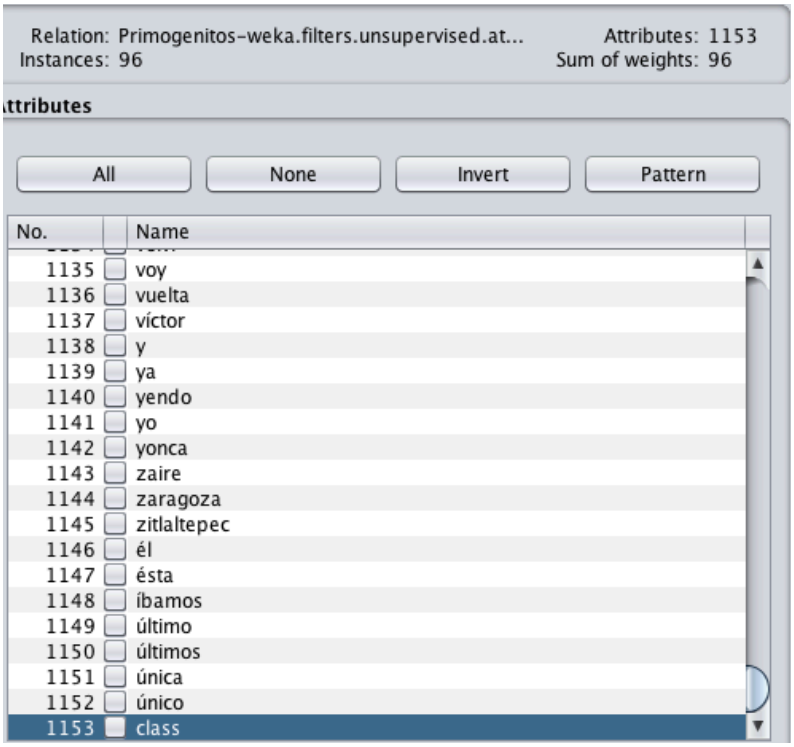
Desarrollo experimental

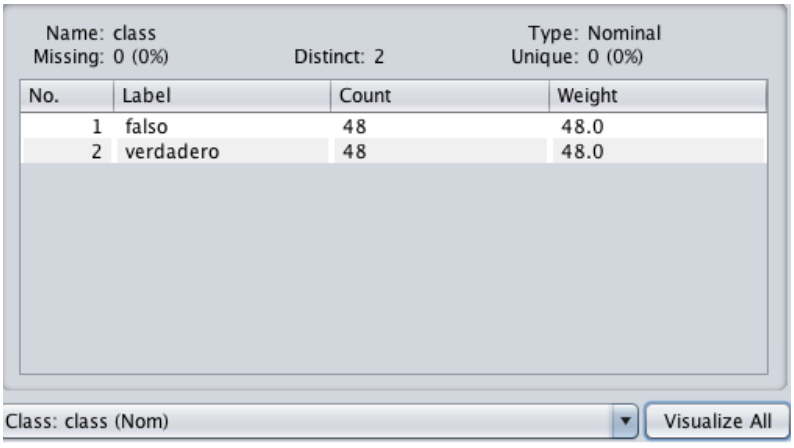
Fue necesario convertir el archivo .txt a formato ARFF para que el software utilizado pudiera leer los datos adecuadamente. Para esto, se agregó una relación al principio del documento (@RELATION) y dos atributos (@ATTRIBUTES) llamados Class y Prim, el atributo class es usado para clasificar el texto desde un principio si el resultado es verdadero o falso, el segundo atributo fue llamado Prim declarada como variable *string*, en el que se encuentra la biografía del estudiante para que se pudiera leer el texto en @DATA, se separó cada biografía de todos los alumnos entre comillas y agregando la clase en la que se encuentra al principio del párrafo.

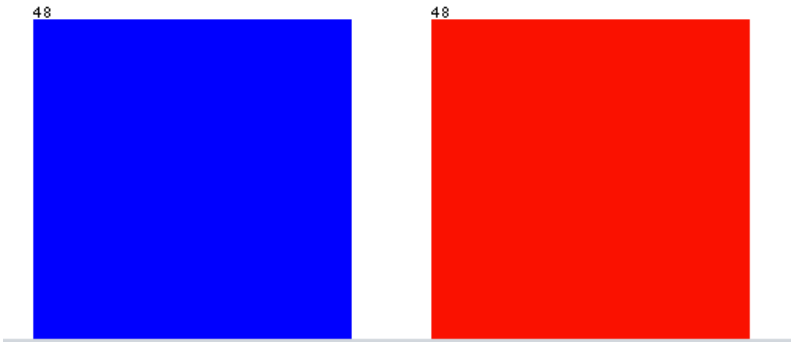
Este mismo proceso se repitió cuatro veces para crear cuatro archivos .ARFF diferentes a partir del DataSet proporcionado, en los primeros dos archivos se experimentó con 129 instancias, se tenían 48 verdaderas y 81 falsas, en los siguientes 2 archivos se tuvieron 96 instancias de forma balanceada, 48 verdaderas y 48 falsas.

Preprocesamiento

Como preprocesamiento en los cuatro casos se convirtió cada palabra en un vector con el filtro “StringtoWordVector” que dejaba el atributo class hasta el principio, por lo que se tuvo que editar y cambiarlo de atributo a clase principal, una vez hecho esto se podían ver el número de atributos así como las siguientes gráficas:







Clasificación

Al finalizar este preprocesamiento se probaron todas las diferentes clasificaciones posibles, sin embargo la mayoría de estás dieron un número bastante bajo de clasificación correcta, luego de obtener estos resultados se usó un segundo filtro llamado “AttributeSelection” el cuál selecciona los atributos más relevantes dentro del texto para poder obtener mejores resultados y una búsqueda mucho más rápida, se compararon los resultados sin este filtro y con el a 5 pliegos, a continuación se muestran los mejores resultados de clasificación dentro de dos tablas:

Desbalanceado 48 Verdaderos + 81 Falsos							
Clasificador	**Precisión**	**Recuperación**	**Área bajo la Curva**	**Exactitud**			
				Correctos		**Incorrectos**	
SIN ATTRIBUTE SELECTION (1382 Atributos)							
HoeffdingTree	?	0.628	0.483	81	62.79%	48	37.21%
IBK	0.592	0.628	0.496	81	62.79%	48	37.21%
Naive Bayes	0.581	0.589	0.608	76	58.91%	53	41.09%
CON ATTRIBUTE SELECTION (43 Atributos)							
HoeffdingTree	0.970	0.969	0.969	125	96.90%	4	3.10%
Logistic	0.969	0.969	0.969	125	96.90%	4	3.10%
SDG	0.925	0.915	0.885	118	91.47%	11	8.53%
Naive Bayes Multinomial	0.913	0.899	0.978	116	89.92%	13	10.08%

Como se puede observar, el número de atributos con el filtro “Atributte Selection” redujo un 97% el número total de atributos, sin embargo la clasificación resultante incrementó en precisión, recuperación, área bajo la curva, exactitud y de igual forma el tiempo en realizar la misma tarea.

Balanceado 48 Verdaderos + 48 Falsos							
Clasificador	**Precisión**	**Recuperación**	**Área bajo la Curva**	**Exactitud**			
				Correctos		**Incorrectos**	
SIN ATTRIBUTE SELECTION (1153 Atributos)							
Naive Bayes	0.573	0.573	0.573	55	57.29%	41	42.71%
Naive Bayes Multinomial	0.531	0.531	0.605	51	53.13%	45	46.88%
RandomForest	0.555	0.552	0.546	53	55.21%	43	44.79%
CON ATTRIBUTE SELECTION (49 Atributos)							
Naive Bayes	0.938	0.938	0.984	90	93.75%	6	6.25%
Naive Bayes Multinomial	0.917	0.917	0.978	88	91.67%	8	8.33%
HoeffdingTree	0.950	0.948	0.982	91	94.79%	5	5.21%

En un intento de obtener mejores resultados, se balancearon los datos, es interesante mencionar que en ambas tablas se obtuvo que por el algoritmo de Naive Bayes estos resultados salen bastante altos, aunque en ninguno llegó a destacar.

De igual forma que en la tabla pasada, sin el filtro de Attribute Selection, los resultados salieron demasiado bajos, sin embargo aplicando este filtro los resultados suben en gran manera, aunque los resultados no fueron igual de altos que en la tabla anterior, se acercó mucho el mismo algoritmo de HoeffdingTree, teniendo 91 aciertos y 5 errores.

Conclusión

En conclusión, el mejor clasificador que se tuvo fue el método HoeffdingTree y Logistic que obtuvieron el mayor porcentaje de aciertos en toda la prueba con un total de 125 aciertos y solo 4 incorrectas con una precisión de 0.970, recuperación de 0.969 y un área bajo la curva de 0.969, aunque sin el filtro de Attribute Selection el mismo algoritmo no obtuvo ningún grado de precisión, marcándolo con un signo de interrogación.
Naive Bayes puede ser de gran utilidad en muchos casos y, aunque no obtuvo el mayor porcentaje de acierto fue de los más altos en las 2 tablas, quedando como una opción bastante confiable en muchos casos

Referencias

Weka Tutorial 01: ARFF 101 https://www.youtube.com/watch?v=gd5HwYYOz2U

Ocw.uc3m.es. [online] Available at: http://ocw.uc3m.es/ingenieria-informatica/herramientas-de-la-inteligencia-artificial/contenidos/transparencias/TutorialWeka.pdf.

Brownlee, J. *How to Run Your First Classifier in Weka*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/how-to-run-your-first-classifier-in-weka/