



Alumno:

Ricardo Jara.

Materia:

SE

Ciclo:

9no

Fecha:

24/07/2020

Waikato Environment for Knowledge Analysis - Entorno para Análisis del Conocimiento de la Universidad de Waikato)

Entorno para análisis del conocimiento de la Universidad de Waikato, se denomina a sí mismo un conjunto de Librerías para tareas de minería de datos. Las librerías pueden ser llamadas desde la interfaz de weka o desde tus propias clases Java. Weka contiene herramientas para diferentes tareas básicas:

Preprocess: Multitud de herramientas para el preprocesamiento de los datos (como por ejemplo discretización de variables).

Classify: Algoritmos de clasificación, distribuidos por paquetes, como por ejemplo ID3 o C4.5

Cluster: Diferentes algoritmos de segmentación como el simple k-means.

Associate: Algoritmos para encontrar relaciones de asociación entre variables (Apriori entre otros).

Select attributes: Aquí, una vez cargados los datos, Weka es capaz de buscar por nosotros las mejores variables del modelo.

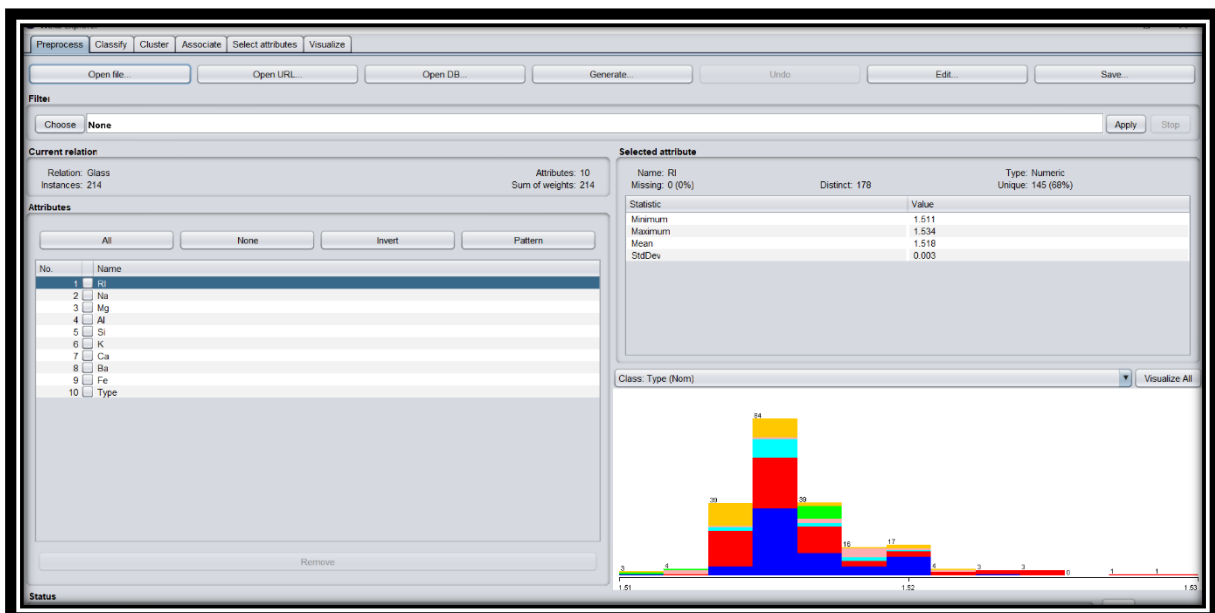
Visualize: Herramienta de visualización de datos en los ejes cartesianos, con muchas posibilidades

A. Realizamos la carga de datos, en weka, para esto se eligió el dataset, “glass.arff”.

Vina realizó una prueba de comparación de su sistema basado en reglas, BEAGLE, el algoritmo vecino más cercano y análisis discriminante. BEAGLE es de un producto disponible a través de VRS Consulting, Inc Al determinar si el vidrio era un tipo de vidrio "flotante" o no, se obtuvieron los siguientes resultados (# respuestas incorrectas): Tipo de muestra Beagle NN DA De ventanas que fueron procesadas flotantemente (87) 10 12 21 De ventanas que no eran: (76) 19 16 22 El estudio de clasificación de tipos de vidrio fue motivado por investigación criminológica. En la escena del crimen, el cristal se fue El% se puede usar como evidencia ... ¡si se identifica correctamente! 5.

Número de instancias: 214

Título: Base de datos de identificación de vidrio



1. Analizamos los datos mediante reglas, existen distintas opciones para poder realizar, la clasificación.

The screenshot shows the Weka Explorer interface with the ZeroR classifier selected. The 'Test options' panel on the left shows 'Use training set' selected. The 'Classifier output' panel on the right displays the following results:

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	76	35.514 %
Incorrectly Classified Instances	138	64.486 %
Kappa statistic	0	
Mean absolute error	0.2116	
Root mean squared error	0.3244	
Relative absolute error	100	%
Root relative squared error	100	%
Total Number of Instances	214	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,000	0,000	?	0,000	?	?	0,500	0,327	build wind float
1,000	1,000	0,355	1,000	0,524	?	0,500	0,355	build wind non-float
0,000	0,000	?	0,000	?	?	0,500	0,079	vehic wind float
?	0,000	?	?	?	?	?	?	vehic wind non-float
0,000	0,000	?	0,000	?	?	0,500	0,061	containers
0,000	0,000	?	0,000	?	?	0,500	0,042	tableware
0,000	0,000	?	0,000	?	?	0,500	0,136	headlamps
Weighted Avg.	0,355	0,355	?	0,355	?	0,500	0,263	

=== Confusion Matrix ===

a	b	c	d	e	f	g	<-- classified as
0	70	0	0	0	0	0	a = build wind float
0	76	0	0	0	0	0	b = build wind non-float
0	17	0	0	0	0	0	c = vehic wind float
0	0	0	0	0	0	0	d = vehic wind non-float
0	13	0	0	0	0	0	e = containers
0	9	0	0	0	0	0	f = tableware

Se puede generar un test aplicando la regla de Bayes, clasificatoria donde nos dice los porcentajes de datos analizados, nos ubica en un clúster y nos arroja la cantidad de predicciones correctas e incorrectas, de un de 214.

The screenshot shows the Weka Explorer interface with the BayesNet classifier selected. The 'Test options' panel on the left shows 'Use training set' selected. The 'Classifier output' panel on the right displays the following results:

Time taken to test model on training data: 0.04 seconds

=== Evaluation on training set ===

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	163	76.1682 %
Incorrectly Classified Instances	51	23.8318 %
Kappa statistic	0.6718	
Mean absolute error	0.0843	
Root mean squared error	0.2183	
Relative absolute error	39.8495	%
Root relative squared error	67.279	%
Total Number of Instances	214	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,914	0,236	0,653	0,914	0,762	0,639	0,912	0,790	build wind float
0,671	0,065	0,850	0,671	0,750	0,645	0,902	0,821	build wind non-float
0,000	0,000	?	0,000	?	?	0,918	0,452	vehic wind float
?	0,000	?	?	?	?	?	?	vehic wind non-float
0,923	0,015	0,800	0,923	0,857	0,850	0,986	0,710	containers
1,000	0,020	0,692	1,000	0,818	0,824	0,998	0,972	tableware
0,931	0,005	0,964	0,931	0,947	0,939	0,986	0,969	headlamps
Weighted Avg.	0,762	0,103	?	0,762	?	0,927	0,801	

=== Confusion Matrix ===

- B. Aquí nos indica que realizó un total de 9 iteraciones, donde nos clasifica el vendedor, los porcentajes de venta y las marcas que han sido vendidas, este clasifica en dos clústeres.

```

Number of iterations: 9
Within cluster sum of squared errors: 118.20374073549189

Initial starting points (random):

Cluster 0: 1.52152,13.05,3.65,0.87,72.32,0.19,9.85,0,0.17,'build wind float'
Cluster 1: 1.51618,13.53,3.55,1.54,72.99,0.39,7.78,0,0,'build wind float'

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute                                Full Data                                Cluster#                                1
                                           (214.0)                                (88.0)                                (126.0)

```

- C. Se puede ver en el grafico a continuación como clasifica en dos cluster lo que es valido para construir vidrio y el material que posiblemente no fuese útil para su uso.
- D. Genera un entrenamiento previo con los datos: Donde nos dice que tiene una predicción de 214 datos logrando clasificar en el primer cluster 88 que equivale al 44% del total del corpus, y el 126% que equivale al 56%.

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer

Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka core EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

☒ Use training set

☐ Supplied test set Set...

☐ Percentage split % 66

☐ Classes to clusters evaluation (Nom) Type

☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

02:42:16 - Cobweb

02:42:38 - SimpleKMeans

Clusterer output

Cluster 1: 1.51618,13.53,3.55,1.54,72.99,0.39,7.78,0,0,'build wind float'

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute	Full Data (214.0)	Cluster# 0 (88.0)	1 (126.0)
RI	1.5184	1.5186	1.5182
Na	13.4079	13.2822	13.4956
Mg	2.6845	3.5483	2.0813
Al	1.4449	1.1718	1.6356
Si	72.6509	72.575	72.704
K	0.4971	0.4412	0.536
Ca	8.957	8.7949	9.0702
Ba	0.175	0.0118	0.289
Fe	0.057	0.0564	0.0575
Type	build wind non-float	build wind float	build wind non-float

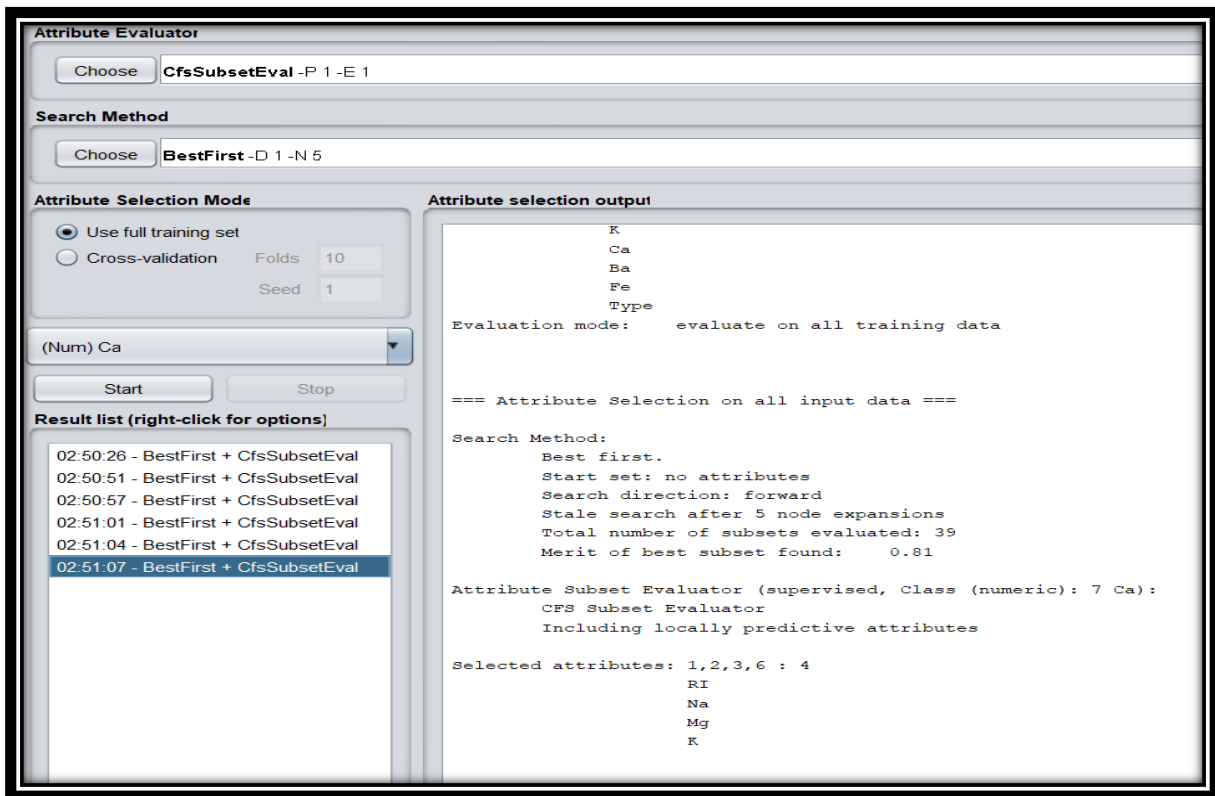
Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

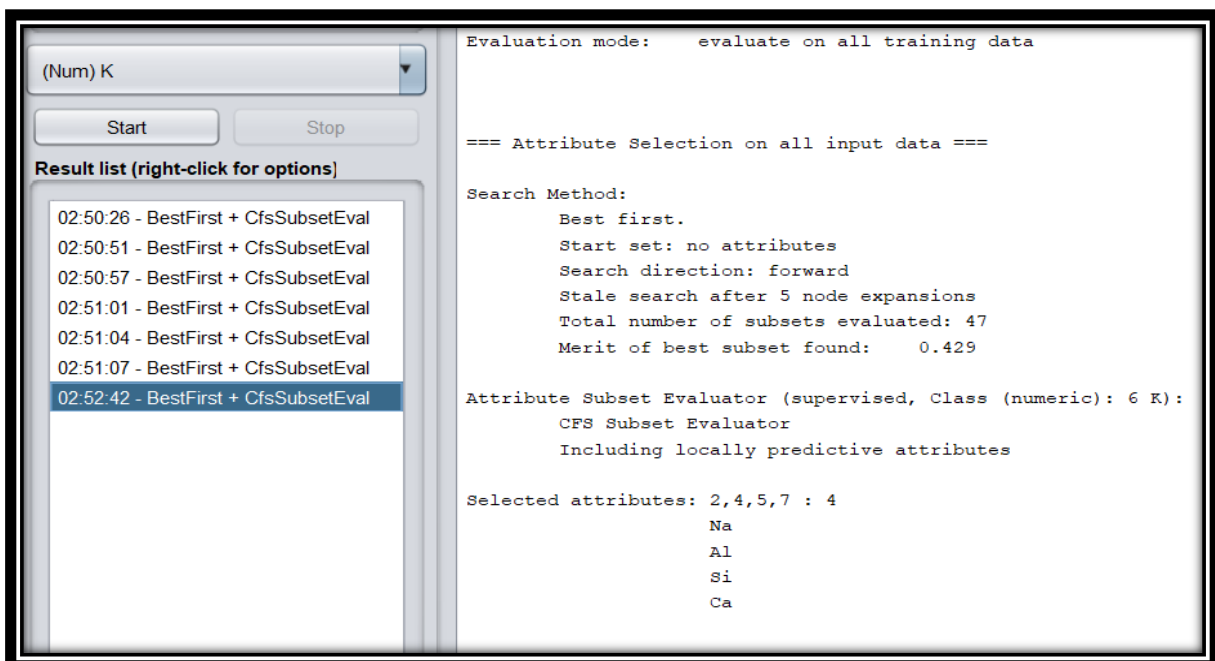
Clustered Instances

0	88 (41%)
1	126 (59%)

Aquí se realiza la selección de atributos, donde nos indica que el uno , dos, tres y el 6 son los datos que realmente tienen importancia, donde el 4 es el más importante,



Volvemos a seleccionar el segundo atributo, K y nos tendría que dar el resultado de 4, ya que nuestro interés ha sido analizar la importancia que tienen cada uno de los elementos en la producción del vidrio.



Se puede ver en la gráfica, los datos del corpus con la grafica, la grafica nos indica que elementos son los mas importantes. Gráficas 2D que relacionan pares de atributos.

