Alternativa de herramienta libre para la implementación de aprendizaje automático.

Corso, Cynthia Lorena
Alfaro, Sofía Lorena
Universidad Tecnológica Nacional, Facultad Regional Córdoba

Abstract

En una organización existe gran cantidad de datos, lo que en muchas ocasiones representa dificultades en el proceso de toma de decisiones; es por ello que es necesario disponer de información justa y precisa.
La Minería de Datos consiste en un proceso de extracción de información útil y patrones ocultos en los datos que surgen de la aplicación de algoritmos estadísticos y computacionales.
Todos estos mecanismos de la minería de datos aportan información útil, lo que facilita y optimiza la toma de decisiones.
Existen diversas herramientas computacionales libres como comerciales que permiten implementar técnicas del aprendizaje automático.
El objetivo del trabajo es dar a conocer las técnicas más usadas de minería de datos que ofrece el Software Libre "Weka".

Palabras Clave

Minería de datos, Weka, Software libre, patrones ocultos, algoritmos de minería de datos.

Introducción

La toma de decisiones consiste en un proceso que implica elegir entre diversas alternativas para lograr un objetivo. Para lo cual es necesario tener información significativa que permita una toma de decisiones objetiva.
En la gran mayoría de los casos es necesario cumplimentar una etapa previa, que es la de preparación de los datos para aplicar alguna técnica de minería de datos. Alguna de las tareas más frecuentes en la etapa de pre procesamiento es filtrar, es decir seleccionar los datos para que realmente sean útiles. Una vez finalizada la etapa de preprocesado se implementa alguna de las técnicas de minería de datos más adecuadas.

La minería de Datos puede definirse como "extracción no trivial de información implícita, previamente no conocida y potencialmente útil desde los datos". [William J. Frawley, Gregory Platestky-Shapiro and Christopher Matheus]
El objetivo de la misma consiste en la generación de información, descubrimiento de relaciones entre variables y de patrones ocultos.
Para llevar a cabo este propósito utiliza un conjunto de algoritmos, los cuales se basan en distintas áreas como matemática, estadística, inteligencia artificial y redes neuronales.

La Minería de Datos produce cinco tipos de información: asociaciones, secuencias, clasificaciones, agrupamientos, pronósticos.
Actualmente existe variedad en cuanto a herramientas computacionales que implementan algoritmos de Minería de Datos.
Podemos encontrar herramientas libres, gratuitas y comerciales.
"Weka" (Waikato Environment for Knowledge Analysis) es una herramienta que permite el análisis de datos mediante aplicación de técnicas de minería de datos. Se distribuye como software libre y gratuito, escrita en lenguaje Java [kk]. Fue desarrollado en la Universidad de Waikato, Nueva Zelanda.

WEKA ofrece cuatro interfaces, a las que pueden acceder desde su pantalla inicial. Cada una de estas interfaces, permite trabajar en un entorno diferente.

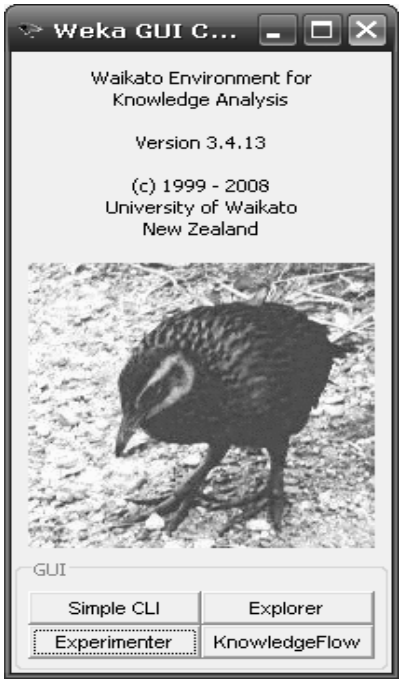

Figura 1. Ventana Inicial de WEKA

Simple CLI: la interfaz "Command-Line Interfaz" es una ventana de comandos java para ejecutar las clases de WEKA. Es poco usada.

Explorer: permite llevar a cabo la ejecución de los algoritmos de análisis sobre un solo archivo de datos.

Experimenter: esta opción permite definir experimentos más complejos, con objeto de ejecutar uno o varios algoritmos sobre uno o varios conjuntos de datos de entrada, y comparar estadísticamente los resultados.

KnowledgeFlow: permite llevar a cabo las mismas acciones del "Explorer", con una configuración totalmente gráfica, permite seleccionar componentes y conectarlos en un proyecto de minería de datos.

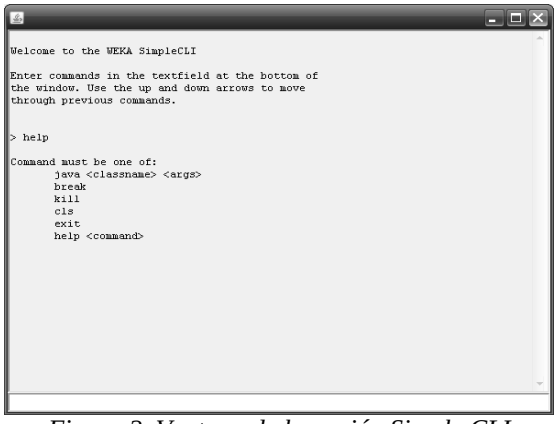

Figura 2. Ventana de la opción Simple CLI

Figura 3. Ventana de la opción Explorer

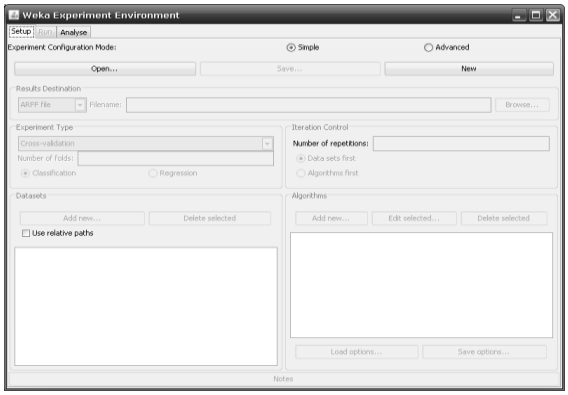

Figura 4. Ventana de la opción Experimenter

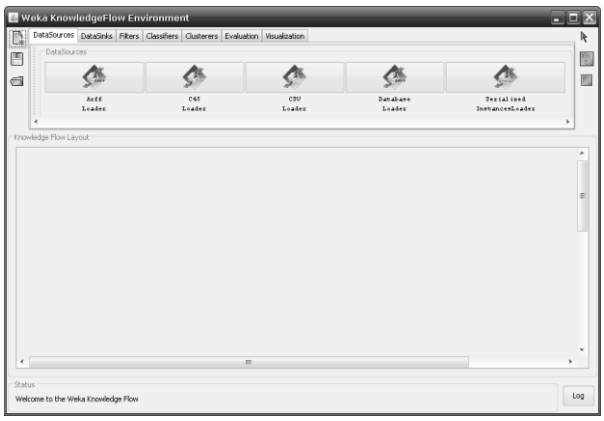

Figura 5. Ventana de la opción KnowledgFlow

WEKA trabaja con datos provenientes bases de datos, archivos y datos que residen en servidores de Internet. Una de las características principales de WEKA, es que trabaja con un archivo de formato arff (Attribute- Relation- File- Format). Su estructura está conformada por tres secciones:

Cabecera. Se define el nombre de la relación. Su formato:

@relation <*nombre-de-la-relación*>

Donde <*nombre-de-la-relación*> es de tipo String.

Declaraciones de atributos. Se declaran los atributos que compondrán el archivo, junto a su tipo. La sintaxis es la siguiente:

@attribute <*nombre-del-atributo*> <tipo>
Donde <*nombre-del-atributo*> es de tipo String.

Sección de datos. Se declaran los datos que componen la relación separando entre comas los atributos y con saltos de línea las instancias o registros.

@data
valor-atributo1,valor-atributo2,..
valor-atributon,valor-atributon+1
[l]

Con respecto a la sintaxis que utiliza éste formato podemos comentar lo siguiente:

"@": indica una definición. Sea de nombre de relación, atributo o sección de datos.
"%": indica comentario hasta final de la línea.
"?": se emplea para expresar un dato desconocido.

El objetivo de este trabajo es dar a conocer las técnicas más usadas de minería de datos como la clasificación y el agrupamiento, del Software WEKA y la importancia de la aplicación de esta ciencia en diversos ámbitos como el de las pequeñas y medianas empresas y otros.

Elementos del Trabajo y metodología.

En esta sección se detallarán las funciones principales de la sección Explorer de Weka. Explorer se utiliza para ejecutar y comparar resultados sobre un único conjunto de datos.
Para ejemplificar cada una de las tareas que se puede desempeñar hemos seleccionado un archivo "labor.arff", que es lo que tomaremos como base para mostrar los resultados.

Este archivo cuenta con 57 instancias y 17 atributos que caracterizan a la relación. Los datos incluyen todos los convenios colectivos alcanzados en el sector de servicios de negocio y personal para vecinos con al menos 500 miembros (profesores, enfermeras, el personal de universidad, la policía, etc.) en Canadá en 87 y el primer trimestre de 88.
Los datos fueron usados para obtener conocimiento la descripción de un contrato aceptable e inaceptable.

A continuación se visualiza la pantalla principal de la opción Explorer de Weka.

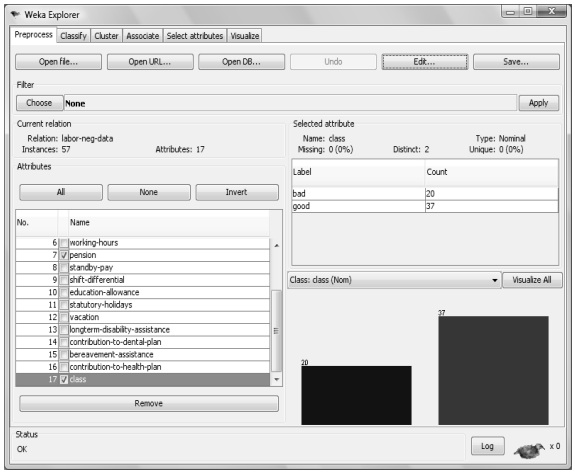

Figura 6. Ventana de la opción Explorer.

En esta primera pestaña de Preprocesado nos permite carga la fuente de datos sobre la cual se aplicará algunas de las técnicas de minería de datos.

La fuente de datos puede ser una base de datos, un archivo Excel transformado a un formato .csv o un archivo con formato .arff que la herramienta Weka es capaz de cargar y procesar.
Si queremos visualizar o actualizar el contenido de los datos disponemos de la opción Edit. Nos aparecerá una pantalla como se muestra en la figura 7.

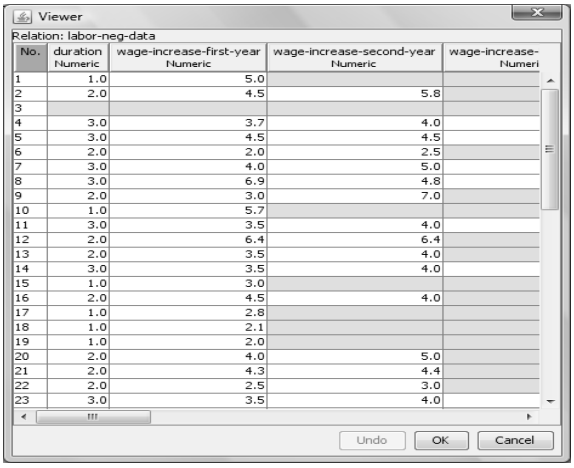

Figura 7. Ventana de Edit de Explorer.

Además también en esta pantalla nos visualiza los atributos de la fuente de datos cargada.
En la parte derecha aparecen las propiedades del atributo seleccionado. Detalla si es un atributo simbólico, se presenta la distribución de valores de ese atributo (número de instancias que tienen cada uno de los valores).
Si es numérico aparece los valores máximo, mínimo, valor medio y desviación estándar. Otras características que se destacan del atributo seleccionado son el tipo, número de valores distintos, número y porcentaje de instancias con valor desconocido para el atributo, y valores de atributo que solamente se dan en una instancia (Unique).

En la parte inferior se puede visualizar un histograma que representa gráficamente los valores asumidos por el atributo.

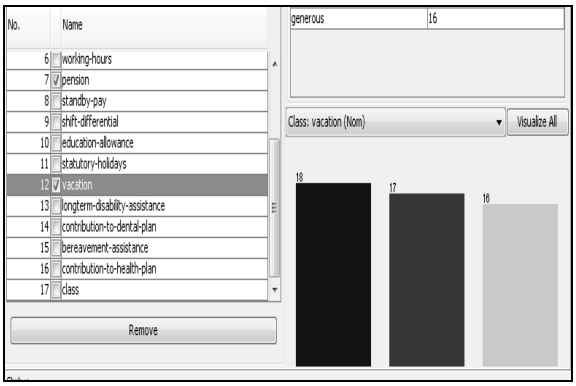

Figura 8. Visualización del histograma del atributo vacation.

Es probable que todos los atributos no sean lo suficientemente relevante y significativos para obtener conclusiones adecuadas mediante las técnicas de minería de datos.
Weka permite aplicar una gran diversidad de filtros sobre los datos, permitiendo realizar transformaciones sobre ellos de todo tipo.
Los filtros pueden ser de tipo no supervisado: son operaciones independientes del algoritmo análisis posterior, a diferencia de los filtros supervisados de "selección de atributos", que operan en conjunción con algoritmos de clasificación para analizar su efecto. Están agrupados según modifiquen los atributos resultantes o seleccionen un subconjunto de instancias (los filtros de atributos pueden verse como filtros "verticales" sobre la tabla de datos, y los filtros de instancias como filtros "horizontales").

Los filtros implementados por Weka pueden clasificarse en filtros de atributos y de instancias.
Los filtros de atributos se pueden utilizar para transformar los datos (por ejemplo convirtiendo datos numéricos en valores discretos).
Mientras que los filtros a nivel de instancias se aplican para eliminar registros o atributos según ciertos criterios previamente especificados.
Dentro de los atributos de instancias disponemos:

[2] **Filtros de selección**: Uno de los más utilizados es Remove, este tipo de filtro es útil en el caso de que no se quiera considerar uno o más atributos para el estudio.
Filtros de discretización: Estos filtros son muy útiles cuando se trabaja con atributos numéricos, puesto que muchas herramientas de análisis requieren datos simbólicos, y por tanto se necesita aplicar esta transformación antes.
Filtros para añadir expresiones: Muchas veces es interesante incluir nuevos atributos resultantes de aplicar expresiones a los existentes, lo que puede traer información de interés o formular cuestiones interesantes sobre los datos.

A continuación examinaremos la opción Classify, que dispone de una gran variedad de herramientas de clasificación. En general las técnicas de clasificación nos permiten predicir futuros comportamientos ante la ocurrencia de nuevas instancias u ocurrencias.

Weka nos permite seleccionar el algoritmo, los mismos están agrupados en las siguientes categorías:

- ✓ Bayes
- ✓ Functions
- ✓ Lazy
- ✓ Meta
- ✓ Misc
- ✓ Trees
- ✓ Rules

Se ha seleccionado uno de los algoritmos clásicos de clasificación que es JRip y es una de las técnicas que se encuentra en la categoría de Rules.
A continuación se visualiza en la Figura 9, los resultados de aplicar esta técnica de clasificación.

```
            vacation
            longterm-disability-assistance
            contribution-to-dental-plan
            bereavement-assistance
            contribution-to-health-plan
            class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

JRIP rules:
===========

(wage-increase-first-year <= 2.5) => class=bad (15.0/2.0)
(statutory-holidays <= 10) and (wage-increase-first-year <= 4) => class=bad
(longterm-disability-assistance = no) => class=bad (2.0/0.0)
 => class=good (35.0/0.0)

Number of Rules : 4


Time taken to build model: 0.05 seconds
```

Figura 9. Ventana de clasificador que visualiza las reglas generadas.

La confiabilidad del modelo aplicado sobre los datos fuente, se logra visualizar en la Figura 10. En la misma se visualiza la proporción de instancias bien y mal clasificadas; que nos permite dar una primera aproximación de cuan bueno es el modelo. Además existe otra herramienta que se utiliza para verificar la bondad del modelo, que es la matriz de confusión.
[3b] La matriz de confusión es una herramienta de visualización que se emplea en aprendizaje supervisado. Cada columna de la matriz representa las el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

Classifier output

```
Correctly Classified Instances          44               77.193  %
Incorrectly Classified Instances        13               22.807  %
Kappa statistic                          0.4935
Mean absolute error                      0.2256
Root mean squared error                  0.456
Relative absolute error                 49.3094 %
Root relative squared error             95.4949 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.65      0.162      0.684     0.65      0.667    bad
  0.838     0.35       0.816     0.838     0.827    good

=== Confusion Matrix ===

  a  b   <-- classified as
 13  7 |  a = bad
  6 31 |  b = good
```

Figura 10. Ventana de clasificador que visualiza los resultados arrojados por la aplicación del modelo.

Además Weka nos ofrece otra herramienta para visualizar los errores cometidos por el modelo.
Esta opción se denomina "Error de clasificación" es una herramienta gráfica que se detalla en la Figura 11.

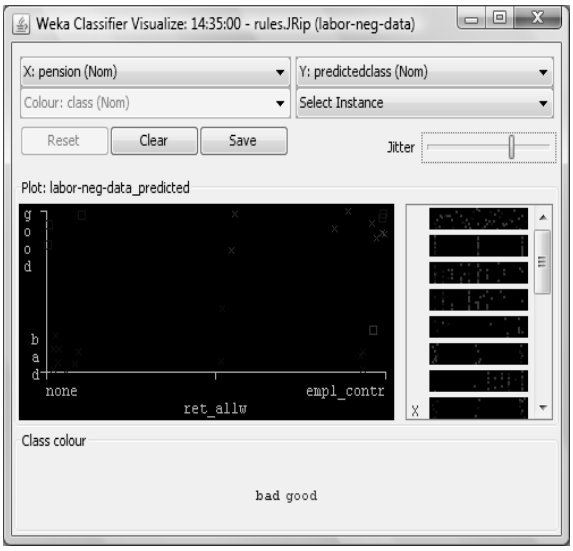

Figura 11. Ventana que muestra los errores de clasificación.

Este recurso clasificador de errores, nos brinda la posibilidad de seleccionar un atributo y lo representa en el eje de las X en el gráfico contra otro atributo que estará representado en el ejes de la Y.
En nuestro trabajo hemos seleccionado como primer atributo si el empleado recibe algún aporte de su jefe en concepto de pensión y como segundo atributo si él las condiciones laborales pactadas en el contrato laboral son buenas o malas.
En este caso los valores representados por un cuadrado representan instancias correctamente clasificadas y la cruz instancias mal clasificadas. En nuestro caso al estar los errores muy cercas no se podía visualizar adecuadamente. Para lograr una mejor visualización de los resultados se le incorpora una especie de ruido aleatorio, para obtener una mejor interpretación de los resultados arrojados.

En cuanto a la opción Cluster, ésta nos permite seleccionar algoritmos de agrupamiento.

Los algoritmos de Clustering buscan fragmentar un conjunto de instancias o registros en diferentes grupos. Las instancias que pertenecen a un grupo poseen características similares. Esta fragmentación se realiza de acuerdo a algún criterio o métrica tomada sobre valores de atributos de las instancias que usa el algoritmo para diferenciar entre las mismas y separarlas.
A continuación se visualiza la pantalla correspondiente a la opción Cluster de Explorer.

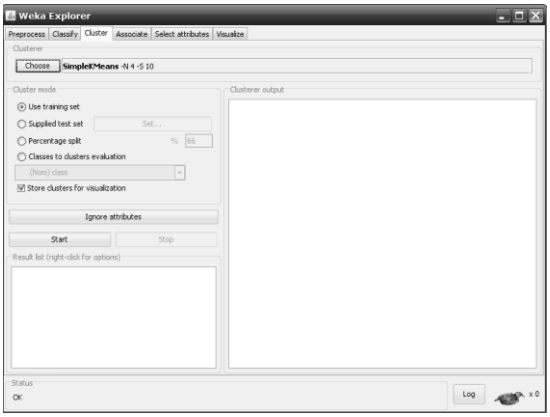

Figura 12. Ventana de Cluster de Explorer.

De igual modo que para la opción Classify, Cluster posee distintos algoritmos que podemos seleccionar. Para ello empleamos la opción Choose. Y para ejecutar el algoritmo seleccionado, la opción Start.
Los resultados obtenidos se visualizarán en la parte derecha de la pantalla, en el marco Clusterer Output.
Podemos seleccionar también el modo de evaluar los resultados. Esto, mediante la opción Cluster Mode. Así por ejemplo el modo "Use Tranining set" (que será el que emplearemos) indica qué porcentaje de instancias se van a cada grupo.

La opción "Ignore Atributtes" permite ignorar atributos que no se tendrán en cuenta en la ejecución del algoritmo.
En la figura se muestra la pantalla en donde podemos seleccionar los atributos que se ignorarán.

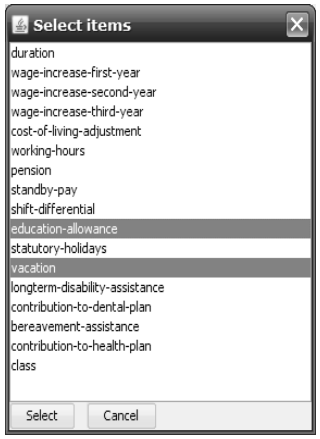

Figura 13. Ventana de selección de atributos ignorados.

Finalmente, podemos ver todos los algoritmos empleados en la sesión actual en la sección "Result list".

El algoritmo seleccionado para ejemplificar la aplicación de Cluster es "SimpleKMeans". Éste algoritmo se aplica sobre atributos numéricos. Por cada grupo se obtiene la media de los valores de los atributos tenidos en cuenta en la ejecución del algoritmo. Vale decir también que éste algoritmo requiere el número de categorías en las que queremos fragmentar las instancias.
En la Figura 14 se muestra la opción para ingresar el número de categorías.

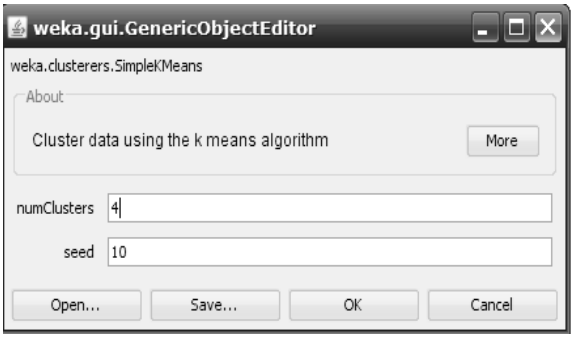

Figura 14. Ventana de selección de número de categorías de grupos

A continuación se visualiza los resultados de aplicar esta técnica de clustering.

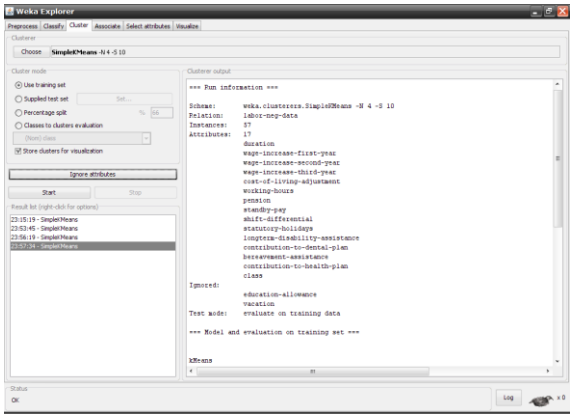

Figura 15. Resultados de algoritmo SimpleKmeans -1

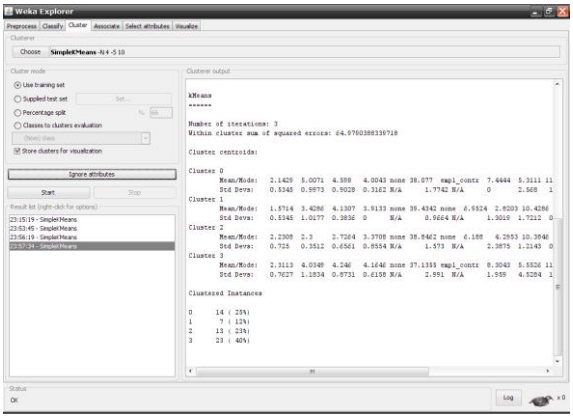

Figura 16. Resultados de algoritmo SimpleKmeans -2

En la primer sección de la pantalla de resultados la información que nos muestra: nombre del algoritmo empleado, nombre del archivo de datos arff., cantidad de instancias, los atributos empleados, los atributos ignorados y el modo de fragmentación seleccionado.

En la segunda sección la información que nos muestra: las categorías obtenidas, con sus medias y desviaciones estándar, y el porcentaje de instancias que pertenece a cada grupo generado.

Resultados

En esta sección se detallarán los resultados de las técnicas de minería de datos implementadas en la sección anterior, tomando como base el archivo "labor.arff".

De las técnicas de clasificación el resultado de aplicar el algoritmo **JRip** es que se han obtenido 4 reglas que se detallan a continuación y el nivel de confiabilidad obtenido por el modelo es del 77%:

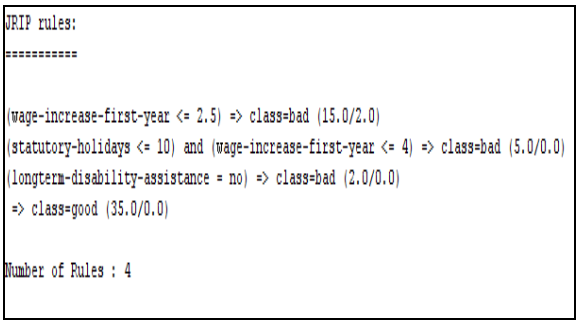

Figura 17. Ventana que muestra los errores de clasificación.

Estas reglas que se detallan a continuación nos permiten predecir ante nuevas instancias si las condiciones de contratación laboral son buenas o no.
Una de la regla generada establece que:

statutory_holidays<=10 and
wage_increase_first year<=4 → class=good

Esta regla se lee de la siguiente manera: Si el número de días de vacaciones estatutarias es menor igual a 10 y el aumento salarial del primer año de contrato del empleado es menor a 400 se predice que el empleado está clasificado en la clase good (buenas condiciones laborales).
Es decir que este modelo obtenido conformado por un conjunto de reglas nos permite predecir que ante la presencia de determinados atributos que asuman determinado rango de valores, para la regla ejemplificada podemos decir que el empleado ha obtenido una serie de beneficios que aseguran una buena condición laboral.

Como se visualiza en la Figura17, al lado de cada regla generada por el modelo aparece una valoración, en el caso de la regla que hemos considerado para la simplificación la valoración es (5.0/0.0) esto se denomina exactitud de la regla. Esto se obtiene verificando cuantas instancias cumplen con las características establecidas en la regla, representada por el numerador y en el denominador establece el número de instancias que no cumple con la regla.

Con respecto a los resultados arrojados por las técnicas de agrupamiento resultaron cuatro grupos:

Grupo 0 con 14 instancias representando un 25% del total de instancias

Grupo 1 con 7 instancias representando un 12% del total de instancias

Grupo 2 con 13 instancias representando un 23% del total de instancias

Grupo 3 con 23 instancias representando un 40% del total de instancias

Discusión.

En la actualidad, existen diversas paquetes de software que implementan dichas técnicas, tanto comerciales (Oracle Data Mining, Clementine, SQL Server etc.) y otras alternativas de software libre como Weka, Yale.

Muchas algunas son muy costosas o no se ajustan a las necesidades de las pequeñas y medianas empresas donde ubicamos el problema que nos concierne.

Las PyMEs, se caracterizan porque cuentan generalmente con un número reducido de trabajadores y tienen una facturación baja o media. Como consecuencia ninguna de ellas tiene el soporte financiero suficiente, como para adquirir herramientas o soluciones costosas.

Es por ello la importancia de este trabajo en resaltar una herramienta para el aprendizaje automático y libre, que aunque su interfaz es muy sencilla es una alternativa muy potente para facilitar el proceso de toma de

decisiones en las medianas empresas y en otros ámbitos como educación, industria, medicina etc.

Conclusión.

En este trabajo se ha expuesto una alternativa de software libre para la aplicación de aprendizaje automático.
Weka proporciona una amplia variedad de algoritmos de aprendizaje y minería de datos, además al disponer de un entorno estadístico, proporciona facilidades que normalmente no se encuentran en una herramienta de minería de datos.

Además esta alternativa se considera que tiene un interfaz más amigable e intuitiva, permitiendo una fácil navegabilidad, es multiplaforma, dispone de una serie de algoritmos de regresión y clasificación, incorpora herramientas para la visualización de los datos y resultados.

Esta herramienta ha sido una de las primeras dentro de la filosofía de software libre, de hecho existen números trabajos en los que se ha implementado minería de datos con esta herramienta. El hecho de que esta herramienta ha tenido una importante difusión, es que la universidad ha ocupado un lugar importante, ya que la mayoría de los proyectos han sido incubados en este ámbito.

Referencias

[1] Diego García Morate, "Weka Tutorial", 2006 http://www.metaemotion.com/diego.garcia.morate/download/weka.pdf

[2] "Página oficial de Weka", http://www.cs.waikato.ac.nz/~ml/weka/index.html

[3] Jordi Porta Zamorano ,"Técnicas cuantitativas para la extracción de términos en un corpus"; Escuela Politécnica Superior – Universidad Autónoma de Madrid, Julio de 2006.

[4] Dapozo, Gladys; Porcel, Eduardo; López, María V.; Bogado, Verónica; Bargiela, Roberto, "Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la FACENA-UNNE"

[5] López Molina José Manuel, Herrero José García, "Técnicas de Análisis de Datos. Aplicaciones prácticas utilizando Microsoft Excel y Weka", 2006

[6] María García Jiménez, Aránzazu Álvarez Sierra, "Análisis de Datos en Weka. Pruebas de selectividad.", http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf

Datos de Contacto:
Cynthia Lorena Corso.
Universidad Tecnológica Nacional.5009.
E-mail: cynthia@bbs.frc.utn.edu.ar