Minería de Datos, 2008-2009
Transformación, Exploración y Selección con Weka

Instalación y visión preliminar

1. Ojear el contenido de *http://www.cs.waikato.ac.nz/ml/weka/*
2. Instalar Weka.
   (a) En la sección Download (*http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html*), seleccionar la última versión (weka-3-5-8jre.exe; 34,573,061 bytes) que incluye Java VM 5.0 en el enlace *http://prdownloads.sourceforge.net/weka/weka-3-5-8jre.exe*
   Nota: se puede intentar bajar e instalar sólo el ejecutable de Weka sin Java (weka-3-5-8.exe; 18,390,439 bytes) en el enlace *http://prdownloads.sourceforge.net/weka/weka-3-5-8.exe* por si funciona con la versión de Java que tenga el equipo, pero si no funciona, utilizar el procedimiento anterior.
   (b) Para instalar Weka, ejecutar el fichero bajado siguiendo las instrucciones.
   (c) Si se detecta algún problema, ver la página *http://weka.sourceforge.net/wekadoc/index.php/en:Troubleshooting*
3. Weka dispone de mucha documentación pero no hay ninguna que documente con detalle las técnicas estadísticas disponibles y, en su lugar, hay que acudir a los artículos científicos referenciados en la ayuda correspondiente. En la sección Documentation (*http://www.cs.waikato.ac.nz/ml/weka/index_documentation.html*) bajar la documentación que se estime más conveniente. En especial, son interesantes:
   (a) La presentación *http://prdownloads.sourceforge.net/weka/Weka_a_tool_for_exploratory_data_mining.ppt*
   (b) Tutoriales sobre los módulos de Weka (aportan sólo una visión demasiado general):
      1. Explorer: *http://prdownloads.sourceforge.net/weka/ExplorerGuide-3-5-8.pdf?download*
      2. Experimenter: *http://prdownloads.sourceforge.net/weka/ExperimenterTutorial-3-5-8.pdf?download*
      3. KnowledgeFlow: *http://prdownloads.sourceforge.net/weka/KnowledgeFlowTutorial-3-5-8.pdf?download*
   (c) Un tutorial en español (*http://www.metaemotion.com/diego.garcia.morate*) aunque de una versión anterior (3.4.2).
   (d) Las especificaciones del formato de fichero de datos *.arff* utilizado por Weka en *http://www.cs.waikato.ac.nz/~ml/weka/arff.html*

   Otra documentación de interés pero en linea es:

   (a) La documentación más reciente para la versión $3.5.8$ que está en *http://weka.sourceforge.net/wekadoc/index.php/en:Weka_3.5.8*

(b) Preguntas frecuentes: *http://weka.sourceforge.net/wiki/index.php/FAQ*

4. Una magnífica fuente de información sobre Weka son los archivos de la lista de correo de Weka: *https://list.scms.waikato.ac.nz/pipermail/wekalist/*

5. En el menú inicio de Windows, buscar la carpeta Weka 3.5.8 y ejecutar el acceso directo Weka 3.5. Como resultado se abre la ventana del programa. Explorar los diferentes menús.

   (a) *Program*: registro de mensajes (*LogWindow*) y *Exit*.

   (b) *Aplications*

      1. *Explorer*: este módulo permite utilizar cualquier técnica de minería de datos incluida en Weka, pero de manera individual, es decir, explorando sus capacidades y prestaciones de manera aislada, sin realizar tareas de comparación entre diferentes técnicas ni encandenar de manera automática tratamientos sucesivos.
      2. *Experimenter*: este módulo permite realizar el proceso de selección del modelo, es decir, comparar el comportamiento de diferentes técnicas en distintas configuraciones de test.
      3. *KnowledgeFlow*. este módulo permite diseñar un proceso de minería de datos completo, desde la conexión a una base de datos local o externa, la exploración, la selección, el preproceso de datos, la minería de datos con la selección del modelo, y la presentación de resultados.
      4. *Simple CLI* (*Command Line Interface*): es una consola para introducir comandos en línea. En principio, cualquier tarea realizable por los tres módulos anteriores puede ejecutarse desde aquí en modo de comandos. No lo utilizaremos durante este curso.

   (c) *Tools*:

      1. *ArffViewer*: visualizador y editor de ficheros de datos con formato *.arff* y otros.
      2. *SqlViewer*: interface para la conexión a bases de datos e interacción mediante comandos *SQL* (*Structured Query Language*), lenguaje estándar de consulta de bases de datos relacionales. Consultar p. ej. en *wikipedia*. No lo utilizaremos en el curso.
      3. *EnsembleLibrary*: módulo para la creación de modelos en formato *.xml* (*eXtensible Markup Language*). Consultar p. ej. en *wikipedia*. No lo utilizaremos en este curso.

   (d) *Visualization*: incluye un conjunto de programas para la visualización de información.

      1. *Plot*: visualización de ficheros de datos, en dos dimensiones eligiendo las variables a visualizar en cada dimensión y con la posibilidad de utilizar una tercera variable (categórica) para caracterizar los pares de puntos en el gráfico.
      2. *ROC*: visualización de curvas ROC, siglas de *Receiver Operating Characteristics*, herramienta de análisis de la eficacia de un clasificador.
      3. *TreeVisualizer*: visualizador de árboles de decisión.
      4. *GraphVisualizer*: visualizador de gráficos en diferentes formatos (*XML*, *BIF*, *DOT*), por ejemplo para redes bayesianas.
      5. *BoundaryVisualizer*: visualizador de fronteras discriminantes en un problema de clasificación.

   (e) *Help*: enlaces a diferentes páginas web con material de ayuda. La función SystemInfo muestra los valores de las diferentes variables de configuración del programa.

Ampliación de la memoria asignada a Weka

1. Si durante la utilización de Weka, se obtiene algún mensaje de memoria insuficiente puede ampliarse la cantidad de memoria asignada por defecto al programa durante el proceso de instalación. El procedimiento es el siguiente:

   (a) Buscar el fichero *RunWeka.ini* en el directorio *Archivos de programa/Weka-3-5*

   (b) Con el botón derecho pulsar sobre el fichero *RunWeka.ini* ubicado en el directorio *Archivos de programa/Weka-3-5* seleccionando la opción *Editar*.

   (c) Se mostrará el contenido del fichero. Por defecto, el proceso de instalación asigna 128 Mb a Weka mediante la instrucción *maxheap = 128m*. Sustituir dicha cantidad con el tamaño de memoria que se quiere asignar al programa teniendo en cuenta las posibilidades de la memoria RAM del equipo.

Weka ArffViewer

1. Este módulo permite abrir ficheros de datos en formato $.arff$ (formato original de Weka) y en formato $.csv$ delimitado por comas (por comas; no por punto y comas).

   Abrir un fichero en formato csv mediante los siguientes pasos:

   (a) Convertir el fichero *ordenadores* a formato $.csv$ delimitado por comas: abrir el fichero con Excel y guardarlo en formato *csv separado por comas* con nombre *ordenadores.csv*.

   (b) Abrir el fichero *ordenadores.csv* mediante *Tools-ArffViewer-File-Open* observando como no logra identificar los campos debido a que Excel utiliza punto y coma (en lugar de coma) como delimitador de campos. Cerrar este fichero mediante *File-Close* en el menú de $ArffViewer$.

   (c) El fichero *ordenadores.csv* está delimitado por punto y comas, en lugar de comas. Abrirlo con un procesador de texto y cambiar todas los punto y comas, por comas, guardando el resultado. Abrir de nuevo el fichero desde *Tools-ArffViewer-File-Open* observando que se visualiza correctamente.

2. En el menú de *ArffViewer* mediante *File-Save as...* guardar el fichero en formato $.arff$ con el nombre $ordenadores.arff$ (eliminar la extensión $.csv$ antes de pulsar *Guardar*).

3. Observar como al pulsar sobre la cabecera de cualquier variable, los datos se ordenan según ella, al hacer doble click sobre cualquier campo, puede modificarse su contenido y al pulsar con el botón derecho sobre cualquier variable se obtiene un menú desplegable (cuya primera función es obtener la media cuando la variable es numérica).

4. La utilización de cualquier técnica de aprendizaje de tipo supervisado (con una variable respuesta) requiere definir en el fichero qué variable va actuar como tal. Utilizaremos la variable $proc$ como variable respuesta (Weka la denomina $class$). Para ello, con el botón derecho sobre *proc* seleccionamos *Attribute as class*.

   Obsérvese como la variable se coloca en último lugar de la tabla (es un convenio del formato *arff*) y como en *File-Properties* van apareciendo los comandos correspondientes a los cambios realizados (*-weka.filters.unsupervised.attribute.Reorder-R1,3,4,5,6,7,8,2*).

5. En el menú de *ArffViewer* mediante *File-Save as...* guardar el fichero en formato $.arff$ con el nombre $ordenadores\_proc.arff$ (eliminar la extensión $.csv$ antes de pulsar *Guardar*).

6. (Opcional) **Formato *.arff***. Con un editor de texto (p. ej. WordPad o Word) abrir el fichero recién creado *ordenadores_proc.arff* y observad su contenido comparándolo con las especificaciones del formato $.arff$ descritas en el documento bajado en el punto 3d de la pág. 1.

   No borrar el fichero *ordenadores.csv* por si se necesita posteriormente.

Weka Explorer

Preprocess

1. Mediante el menú de Weka *Applications-Explorer* abrir el módulo *Explorer* y situarse en la pestaña *Preprocess*.

2. Pulsando el botón *Open file...* abrir el fichero $ordenadores\_proc.arff$ creado anteriormente y explorar superficialmente las funciones de los botones de la primera fila:

   - Observar como las funciones *Edit* y *Save* son análogas a las que realiza el módulo *ArffViewer* visto anteriormente.
   - La función *Generate...* sirve para generar un conjunto aleatorio de datos. Si se pulsa en el botón, se obtiene una ventana con dos botones: *Choose* para elegir el tipo de datos a generar y *Generate* para ejecutar la generación de datos. Posponer la prueba de estas funciones hasta el final de esta sección.

3. Explorar gráficamente el fichero de datos $ordenadores\_proc.arff$.

   (a) Seleccionar la variable $origen$ (y sucesivamente las demás) en la zona izquierda del gráfico y observar un resumen descriptivo en la zona derecha y el histograma que aparece debajo.

   (b) Cambiar la variable que actúa como *Class* encima del histograma observando el efecto en el gráfico cuando se elige una variable numérica y cuando se elige una variable nominal.

   (c) Por ejemplo, con la variable $origen$ como $Class$, pulsar el botón *Visualize All* entendiendo los gráficos resultantes.

4. Seleccionar la pestaña *Visualize* y observad la matriz de diagramas de dispersión dos a dos. Modificar algunas características pulsando *Update* después de cada modificación. Observar los gráficos tratando de encontrar posibles relaciones entre pares de variables.

5. Hacer lo mismo mediante mediante *Visualization-Plot* del menú de Weka: abrir el fichero *ordenadores_proc.arff* y explorar los gráficos que se obtienen cambiando las variables de los ejes. En los pequeños gráficos de dispersión de la zona derecha pulsar en diferentes filas para cambiar la variable que actúa como independiente.

6. (Opcional). Volviendo al *Explorer*, explorar la funcionalidad del botón *Generate* generando diferentes conjuntos de datos con diversos generadores y explorando gráficamente las características de los datos generados.

7. Además de las funciones propias de la fase de Exploración, la pestaña *Preprocess* incluye también funciones propias de las fases de Limpieza, Selección y Preproceso de un proyecto de Minería de Datos. En este sentido, contiene diversos filtros aplicables a los datos con anterioridad a la minería propiamente dicha.

Estos filtros se obtienen pulsando en el botón *Choose*. Al pulsar en dicho botón se obtiene un árbol desplegable con dichos filtros clasificados según se trate de:

(a) Filtros generales: *Allfilter* (sin utilidad para el preproceso) y *Multifilter* (para encadenar diversos filtros.

(b) Filtros de tipo supervisado o no supervisado. Los primeros utilizan la información de alguna variable en el preproceso al contrario que los segundos.

(c) Filtros aplicables a atributos o a instancias (individuos). Los primeros producen nuevas variables o modifican las existentes. Los segundos actúan sobre los individuos de la muestra.

8. Seleccionar uno cualquiera de estos filtros. Pulsar a continuación en el campo en el que aparece detallado. Se obtiene una ventana descriptiva de sus funciones que, a su vez, permite configurar sus opciones. Para obtener más información pulsar *More* y *Capabilities*.

9. La ventana que se obtiene al pulsar el botón *Choose* permite también aplicar filtros a los filtros mediante el botón inferior *Filter* que produce una nueva ventana en la que se pueden seleccionar la tipología de atributos o de datos con la que se quiere trabajar con objeto de que el árbol de filtros muestre en color negro sólo los compatibles con dichos requierimientos (y en color rojo los que no lo son). Weka establece por defecto un filtro asociado al tipo de datos cargados.

   Pulsar en *Remove filter* para quitar ese filtro y pulsar después en *Filter* seleccionando únicamente *Nominal attributes*. Observar el efecto en el árbol de filtros: aparecen en rojo los filtros que sólo utilizan atributos numéricos.

10. La variable $proc$ es para nosotros una variable categórica (identifica el fabricante del procesador del ordenador) pero está codificada numéricamente. Muchas funciones de Weka sólo pueden tratar variables tipo $Class$ de tipo nominal, así que cambiaremos la codificación de esta variable:

    (a) Elegir el filtro no supervisado para atributos *NumericToNominal* y aplicarlo sobre el atributo 9 dejando el resto de las opciones por defecto. Después de pulsar *Apply* observar los resultados en los histogramas eligiendo como variable $Class$ la nueva variable $proc$ y graficando las demás variables.

    (b) Guardar el fichero mediante el botón $Save$ de la pestaña *Preprocess* (Asegurarse de que el fichero no está abierto en *ArffViewer* o en otro módulo pues de lo contrario, no se guarda).

11. *Muestra aleatoria*. Elegir el filtro supervisado sobre instancias *Resample* y pinchar en la linea de especificaciones del filtro, pulsar en el botón *More* para ver el significado de los argumentos del filtro.

    Realizar una submuestra aleatoria simple sin reemplazamiento del fichero *ordenadores_proc.arff* que tenga el 70% de los datos con aproximadamente un equilibrio en el número de ordenadores de cada procesador (notar que esta posibilidad es la razón de que el filtro esté clasificado como supervisado). Pulsar *Apply* para aplicar el filtro. Observad el resultado mediante histograma y resumen descriptivo de la variable $tipo$. Repetir $Apply$ para comprobar si la representación de los tipos de procesador se equilibra en la muestra.

12. Pulsar el botón *Log* de la parte inferior derecha de la ventana. Aparece la ventana de mensajes de Weka que se utiliza también para mostrar los posibles errores que puedan obtenerse.

13. (Opcional) Revisar la funcionalidad de la mayoría de los filtros aplicándolos al fichero $ordenadores\_proc.arff$. Para volver a la situación original, volver a cargar el fichero mediante *Open file...*

Esto también puede consultarse en la sección 5.1.1 del tutorial en español (punto 3c, pág. 1) teniendo en cuenta posibles discrepancias debido que dicho tutorial se refiere a la versión 3.4.2[1].

14. *Outliers detection*: Utilizando todos los datos del fichero, realizar un análisis de datos atípicos y extremos utilizando la definición dada en la página 8 de los apuntes de estadística descriptiva, de la siguiente forma:

    (a) Seleccionar el filtro no supervisado sobre atributos *InterquartileRange*. Pulsar sobre la línea de especificaciones del filtro, introducir los necesarios *extremevaluesFactor* y *outlierFactor* y aceptar las demás opciones por defecto del método. Consultar las opciones del método pulsando en el botón *More*.

    (b) Como resultado aparecen dos nuevas variables (*Outlier* y *ExtremeValue*) que identifican los valores atípicos y los valores extremos, respectivamente. En la pestaña *Visualize* tratar de identificar los casos atípicos y extremos en los diferentes gráficos. (Elegir como *Colour* la variable *Outlier* para identificar en color rojo dónde están los atípicos en los gráficos). Si es necesario, consultar los datos mediante el módulo *ArffViewer*.

    (c) Con objeto de identificar los atributos responsables del carácter de atípico o extremo de los casos afectados, repetir el análisis pero esta vez eligiendo *detectionPerAttribute*. Estudiar el significado y contenido de las nuevas variables, concluyendo al respecto.

    (d) Reflexionar sobre las razones que en general pueden mover a eliminar los casos atípicos y, en particular, sobre la conveniencia de eliminarlos en nuestro problema (caracaterización del tipo de procesador –*proc*–en términos de las demás variables) o, por ejemplo, en el problema de investigar la relación entre ram y $tiempo2$.

15. *Tipificación de variables*. Mediante el filtro no supervisado para atributos *Standardize* tipificar las variables numéricas. Comprobar el resultado observando sus valores descriptivos. Realizar el análisis anterior de atípicos comparando los resultados.

16. *Selección de variables*

    (a) Con el fichero *ordenadores_proc.arff*, mediante el filtro *AttributeSelection* realizar un ranking de los atributos en función de su importancia sobre la variable $proc$ utilizando el test de la Chi cuadrado visto en los apuntes de Estadística Descriptiva (elegir como *evaluator* el método *ChisquaredAttributeEval* (una vez seleccionado, pulsar en la línea donde aparece su configuración para consultar su funcionalidad). Dejar el método de búsqueda (*search*) que aparece por defecto (*BestFirst*) y pulsar OK y después *Apply.*

    Se obtiene un error porque el método de búsqueda *BestFirst* utilizado sólo produce una selección de atributos y no un ranking de los mismos: sustituir el método *BestFirst* por el método *Ranker* y repetir la operación pulsando al final *Apply*.

    Como resultado, se obtiene una ordenación diferente de las variables en la que las primeras son las más influyentes en la variable $proc$ según el método utilizado.

    (b) Repetir el procedimiento pero ahora con intención de quedarse con las 2 variables más influyentes en $proc$ (según el criterio Chi cuadrado). Para ello, cambiar la opción *numToSelect* del método *Ranker*.

    Como resultado, se obtienen las variables $origen$ y $tipo$ como las más influyentes y el fichero estaría preparado para ulteriores análisis.

[1] De entrada, en la versión 3.4.2 todos los filtros se encuadran bajo el grupo no supervisado (véase nota al pie de la pág. 10 de dicho tutorial).

(c) Cargando de nuevo el fichero original, probar el mismo filtro *AttributeSelection* con el método *CfsSubsetEval* dejando la opción de búsqueda *BestFirst*.

(Opcional) Identificar en el diálogo del método *CfsSubsetEval* pulsando *More* identificar una publicación de referencia y buscarla en internet.

17. *Pestaña Select attributes*. La mayor parte de las funciones de la pestaña *Preprocess* realizan transformaciones (*preproceso*) a los datos y a las variables sin que se pueda evaluar la calidad y oportunidad de esas transformaciones pues dichas funciones no aportan información. Por ejemplo, la tarea anterior se ha realizado a ciegas pues no obtuvimos el valor del estadístico χ^2 para cada uno de los atributos con objeto de conocer la fuerza de su asociación con la variable de interés.

    Esta información puede obtenerse en la pestaña *Select attributes* del módulo *Explorer* y lo normal es realizar el análisis primero en este módulo y después aplicar sus resultados mediante los filtros del módulo *Preprocess.*

    (a) Por ejemplo, en dicha pestaña en el botón *Choose* volver a elegir el método *ChisquaredAttributeEval* con todos los datos (*Use full training set*), seleccionar $proc$ como variable respuesta y pulsar $Start.$ En la zona derecha de la ventana se presentan los resultados del análisis. Obsérvese el valor de la χ^2 de todas las variables es nulo excepto para las variables $origen$ y $tipo$. Nótese que para la aplicación de este método, las variables continuas han sido discretizadas con anterioridad al cálculo del estadístico χ^2.

    (b) No sabemos si los resultados obtenidos se obtendrían consistentemente con varias muestras de datos. Repetir el proceso anterior pero ahora seleccionando *Cross-validation* utilizando 10 *Folds* (Grupos) y dejando la $Seed$ (semilla aleatoria utilizada en el proceso de selección aleatoria de los 10 grupos en los que se divide la muestra) en el valor propuesto. Pulsar finalmente en *Start*.

    El método 10*-fold Cross Validation* divide la muestra en 10 grupos y evalúa el estadístico χ^2 10 veces dejando cada vez un grupo fuera. Como resultado proporciona la media del valor χ^2 y su error estándar en esas 10 observaciones, así como el orden medio (y error estándar) que cada atributo ocupa en esas 10 clasificaciones.

    (c) Cambiar el número de grupos utilizando valores entre 2 y 21 analizando los resultados.

    (d) Comparar los resultados obtenidos anteriormente con *ChisquaredAttributeEval* con los que se obtienen mediante el método *CfsSubsetEval* con *GreedyStepwise*, con todos los datos y con 10-fold cv.

18. *Filtros Múltiples*. Weka permite encadenar diferentes filtros mediante el filtro *Multifilter*. Encadenar el filtro *Resample* y el filtro *AttributeSelection* con *CfsSubsetEval* realizados en el ejercicio 16c, aplicar el filtro compuesto resultante y comprobar que los resultados son los esperados.

19. *Discretización de variables*. Existen técnicas de machine learning que sólo trabajan con variables discretas. Por ello, a veces se requiere discretizar variables a pesar de perder con ello información. La discretización puede ser no supervisada o supervisada. En este último caso, se pretende que la pérdida de información sea mínima en cuanto a su relación con la variable dependiente.

    (a) Discretización no supervisada.

        1. Volver a cargar el fichero $ordenadores\_proc.arff$ y mediante el filtro no supervisado para atributos *Discretize* construir variables discretas que sustituyan a las continuas, mediante intervalos de igual longitud determinados automáticamente (mediante *leave-one-out*). Comprobar los resultados observando los valores descriptivos de cada variable.

2. Mediante la pestaña *Select attributes* evaluar la asociación entre las variables así discretizadas y la variable $proc$ al modo del ejercicio 17a, analizando sus resultados.
3. Repetir los dos pasos anteriores ahora con una discretización basada en 10 intervalos con igual frecuencia (deshacer antes la transformación anterior pulsando *Undo* en la pestaña *Preprocess*).

(b) Discretización supervisada. Deshacer la discretización anterior mediante *Undo* en la pestaña *Preprocess* y mediante el filtro supervisado *Discretize*, discretizar las variables utilizando las opciones por defecto.

Mediante la pestaña *Select attributes* evaluar la asociación entre las variables así discretizadas y la variable $proc$ al modo del ejercicio 17a, comparando los resultados con los obtenidos con el método de discretización no supervisado y con los obtenidos utilizando las variables sin discretizar.

20. *Asociación según Modelo*.

(a) Deshacer la transformación anterior mediante *Undo* de la pestaña *Preprocess*. En esa misma pestaña, mediante *Edit* abrir el fichero y definir como variable respuesta la variable $tipo$ pulsando en ella con el botón derecho y seleccionando *Attribute as class*. Pulsar *OK* para cerrar y mediante *Save...* guardar el fichero con nombre *ordenadores_tipo.arff*.

(b) (Opcional) Repetir el ejercicio 17a con la variable $tipo$ como variable respuesta.

(c) Mediante la pestaña *Select attributes* seleccionar los atributos que resultarían relevantes para la variable $tipo$ según una regresión logística utilizando el método *WrapperSubsetEval* (consultar su funcionalidad en el botón *More* de su ventana y buscar la referencia Kohavi and John, en Internet) y eligiendo como *classifier* en la carpeta *functions* el método *Logistic* y como *folds* (número de grupos para validación cruzada) el número propuesto.

Para comprender el significado del análisis anterior, téngase en cuenta que el coeficiente de correlación lineal determina los atributos relevantes para una variable respuesta según un modelo de regresión lineal, y, por tanto, si se utiliza un modelo de regresión logística podríamos hablar de un *coeficiente de correlación logística*. Igualmente, si utilizásemos otras técnicas (lineales, no lineales o no paramétricas), podríamos hablar de coeficientes de correlación de tipo lineal, no lineal o no paramétrico (asociados a cada técnica, respectivamente).

21. *Clasificación*. Mediante la pestaña *Classify* utilizar los siguientes métodos de clasificación para la variable *tipo* utilizando 10-*fold-cv* y con las opciones propuestas por defecto: De la carpeta *functions*: *logistic*, *LibSVM* [2]y *Multilayer Perceptron.* De la carpeta *trees*: *Simple CART*. De la carpeta *Bayes*: *BayesNet*. De la carpeta *Rules*: *DecisionTable*, *PART.*

Para cada método, analizar la salida del programa entendiendo su contenido (de la sección *Detailed Accuracy by Class*, por ahora sólo las columnas TP y FP) y comparar los resultados de los diferentes clasificadores.

[2]Es posible que la utilización de *LibSVM* produzca el siguiente error "*Problem evaluating classiffier: libsvm not in CLASSPATH!*". En ese caso, para añadir *libsvm* al CLASSPATH:

(a) Bajarse el fichero *wlsvm.zip* de Faitic y extraer el contenido de su carpeta *lib* a la carpeta *C:\Program Files\Weka-3-5* donde está instalado el programa.

(b) Si se utiliza XP, pulsar con el botón derecho en Mi PC, y si se utiliza Vista, pulsar con el botón derecho en Equipo y seleccionar, Configuración avanzada del sistema.

(c) Para ambos sistemas: elegir Opciones avanzadas – Variables de entorno y en Variables de usuario, pulsar en Nueva y escribir como Variable CLASSPATH y como valor: *C:\Program Files\Weka-3-5\wlsvm.jar;libsvm.jar*. Salir con Aceptar.

22. Una estrategia interesante en el preproceso de datos es utilizar las componentes principales de las variables explicativas pues se elimina posible ruido y se evita asociación entre covariables (colinealidad).

    (a) Realizar un análisis de componentes principales de las covariables mediante *Select attributes* utilizando *PrincipalComponents* como evaluador, entendiendo la composición de cada componente principal y el porcentaje de varianza explicada por cada una.

    (b) Transformamos el fichero para tener como covariables las componentes principales:

        1. En la pestaña *Preprocess*, elegir como filtro *PrincipalComponents* eligiendo las que cubren el 95% de varianza. Pulsar *Apply* y observad las nuevas variables. Revisar los histogramas de la derecha para cada componente (comparar con los resultados de varianza explicada obtenidos en el punto anterior) y determinar qué componentes pueden ser relevantes para explicar o predecir la variable $tipo$.
        2. Pulsar *Undo* y repetir el proceso pero ahora eligiendo sólo el 90% de varianza explicada.

    (c) Veamos qué componentes podrían resultar relevantes para un modelo de regresión logística que tratase de predecir la variable $tipo$: en la pestaña *Select attributes* utilizar como evaluador *Wrapper* con regresión logística y 10-*fold-cv*, observando las componentes principales seleccionadas.

    (d) Repetir los métodos de clasificación del ejercicio 21 con las componentes principales obtenidas anteriormente al requerir un mínimo del 90% de varianza explicada. Reflexionar sobre los resultados obtenidos.

23. (Opcional) Realizar el ejemplo de la sección 5.4.3 de Hernández-Orallo.

    Sugerencias:

    (a) Para construir la tabla de datos pasar primero los datos a fichero txt y después importar con Excel con formato de ancho fijo. Copiar y pegar entre hojas de Excel a conveniencia hasta construir la tabla, colocando las variables en columna y las instancias en fila.

    (b) Crear una nueva variable dicotómica que idenfique la muestra de entrenamiento y de test.

    (c) Crear una variable respuesta con el resultado bad/good de la negociación.

    (d) Revisar los datos y corregir los errores producidos por un ancho inadecuado en alguna columna.

    Otra forma de cargar los datos es utilizando el formato C4.5 (directorio C4.5 en la web de UCI) y abriéndolo desde *ArffViewer*. Pero el fichero de entrenamiento tiene ahora 40 instancias en lugar de 27 de antes. Para abrir el fichero de test desde *ArffViewer*, cambiar de nombre el de entrenamiento una vez cargado y dar al de test el de entrenamiento para que el *ArffViewer* pueda asociarlo con el fichero .names.

24. Analizar el fichero de datos EncuestaPeriodico. Se trata de los datos de una encuesta destinada a analizar el grado de aceptación que tendría el lanzamiento de un nuevo periódico en Vigo (véase el cuestionario correspondiente).

    El objetivo principal del estudio es tratar de delimitar el perfil de las personas que están a favor de que haya un nuevo periódico en Vigo.

    Para ello, mediante la adecuada selección de variables y explorando diferentes técnicas de clasificación como las vistas anteriormente, construir un modelo de clasificación con el menor número de varialbles y la máxima capacidad predictiva (menor tasa de error en validación cruzada).