Evaluación de Rankings de Atributos para Clasificación

Roberto Ruiz, Jesús S. Aguilar–Ruiz, and José C. Riquelme

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla, Sevilla, España
{rruiz,aguilar,riquelme}@lsi.us.es

Resumen En este trabajo se presentan distintas formas de comparar los rankings generados por algoritmos de selección de atributos, mostrando la diversidad de interpretaciones posibles en función del enfoque dado al estudio que se realice. Partimos de la premisa de la no existencia de un único subconjunto ideal para todos los casos. La finalidad de este tipo de algoritmos es reducir el conjunto de datos a los primeros atributos de cada ranking sin perder predicción frente a los conjuntos de datos originales. En este trabajo proponemos un método que mide el comportamiento de un ranking de atributos generado y a su vez es válido para comparar diferentes algoritmos de listas de atributos. Para ello, nos basamos en un nuevo concepto, el area bajo la curva de comportamiento al clasificar un ranking de atributos ($AURC$). Las conclusiones y tendencias extraídas de este documento pretenden dar soporte ante la realización de tareas de aprendizaje donde intervengan algunos de los algoritmos de ranking aquí estudiados.

1. Introducción

Es un hecho que el comportamiento de los clasificadores mejora cuando se eliminan los atributos no relevantes y redundantes. En la selección de características se intenta escoger el subconjunto mínimo de atributos de acuerdo con dos criterios: que la tasa de aciertos no descienda significativamente; y que la distribución de clase resultante, sea lo más semejante posible a la distribución de clase original, dados todos los atributos. En general, la aplicación de la selección de características ayuda en todas las fases del proceso de minería de datos para el descubrimiento de conocimiento.

Los algoritmos de selección de atributos los podemos agrupar en dos categorías desde el punto de vista de la salida del método: subconjunto de atributos o ranking de atributos. La primera categoría escoge un subconjunto mínimo de características que satisfaga un criterio de evaluación; la segunda, elabora una lista de atributos ordenados según alguna medida de evaluación. Idealmente, los métodos de selección de características buscan a través de los subconjuntos de atributos intentando encontrar el mejor entre los 2^m (m: número total de atributos) subconjuntos candidatos según una función de evaluación. Sin embargo este proceso exhaustivo sería demasiado costoso incluso para bases de datos pequeñas.

Se han definido estrategias que permiten obtener un subconjunto de atributos que no aseguran el óptimo pero que tienen un valor próximo con respecto a la función de evaluación utilizada. En los algoritmos que elaboran un ranking no existe el proceso de búsqueda al evaluar individualmente cada atributo.

Cuando los algoritmos de selección de atributos se aplican como técnica de preprocesado para la clasificación, estamos interesados en aquellos atributos que clasifican mejor los datos desconocidos hasta el momento. Si los algoritmos proporcionan un subconjunto de atributos, este subconjunto se utiliza para generar el modelo de conocimiento que clasificará los nuevos datos. Sin embargo, cuando la salida del algoritmo es un ranking, no es fácil determinar cuántos atributos son necesarios para obtener un buen resultado de clasificación.

En este trabajo presentamos distintas formas de comparar rankings de atributos, y mostramos la diversidad de interpretaciones posibles según el enfoque del estudio que se realice. Intentamos saber si existe alguna dependencia entre clasificadores y métodos de ranking, además de dar respuesta a dos preguntas fundamentales ¿Qué es un ranking de atributos? y ¿Cómo valorar/medir un ranking? Para ello, realizaremos las distintas comparaciones utilizando cuatro métodos de rankings de atributos: χ^2, Ganancia de Información, ReliefF y SOAP, comentados posteriormente. Y comprobaremos los resultados calculando la tasa de aciertos con tres clasificadores: C4.5, Naïve Bayes y el vecino más cercano.

El documento se organiza de la siguiente forma: en la sección 2 se define lo que entendemos por ranking de atributos y otros conceptos que ayudarán en la comprensión del documento; en la sección 3 se exponen trabajos relacionados con la comparación de técnicas de ranking de atributos; en la sección 4 se presentan las motivaciones que nos han llevado a este estudio; a continuación se analizan los experimentos realizados en la sección 6; y finalmente, en la sección 7, se recogen las conclusiones más interesantes.

2. Ranking de Atributos

Los algoritmos de la categoría de ranking de atributos proporcionan una lista de características ordenada según alguna medida de evaluación. En [1], Dash-Liu realizan una clasificación de las medidas en: consistencia, información, distancia, dependencia y exactitud de algún algoritmo de aprendizaje. Los métodos de ranking asignan pesos a los atributos individualmente y los ordenan basándose en su relevancia con respecto al concepto destino o atributo clase. Los k primeros atributos formarán el subconjunto final.

Comparar dos métodos de selección de subconjuntos de atributos es trivial. Se generan los modelos de clasificación con los subconjuntos de atributos obtenidos y se evalúa la predicción de exactitud de ambos. Sin embargo, no está clara la comparación entre dos métodos de ranking de atributos, dado que la predicción de exactitud de los modelos de clasificación dependen del número (k) y de la calidad de las características seleccionas del ranking. El éxito de clasificación puede variar sustancialmente dependiendo de k como veremos mas adelante.

La utilidad de los rankings de atributos, no siempre va orientada a la selección de un número de ellos para aplicar un algoritmo de aprendizaje al conjunto de datos reducido. A veces, nos interesa saber qué características son importantes y cuales no, y si existe un orden en la relevancia de ellas. Esto dificulta la comparativa entre los rankings. En el análisis de microarrays, se utilizan estos algoritmos para descubrir qué fármacos son los más eficaces. En [2], se identifican los cincuenta genes más correlados con la clase, para clasificar pacientes con leucemia. Algunos algoritmos de selección de subconjuntos de atributos más complejos, suelen generar rankings en sus fases iniciales.

Antes de comparar rankings de atributos tendríamos que aclarar cuándo se considera que un ranking es bueno. En general, un ranking será bueno cuando la posición que ocupa cada atributo en la lista se corresponda con su relevancia, existiendo distintos puntos de vista sobre que entender por relevancia. En nuestro caso, teniendo en cuenta la aplicación que se le va a dar, selección de atributos para clasificación, un buen ranking será aquel que coloca en las primeras posiciones los mejores atributos para clasificar el conjunto de datos.

De la anterior definición, deducimos que si dos atributos tienen igual importancia, ocupan posiciones contiguas en el ranking, aún conteniendo la misma información (redundantes). Además, si los dos atributos son muy relevantes entrarían a formar parte del subconjunto final. Esta situación, que es normal en la salida de los algoritmos de rankings, es incompatible con aquella a la que se pretende llegar en cualquier proceso general de selección, donde se intenta obtener aquel subconjunto mínimo de atributos que sean relevantes para la clasificación, dejando fuera los atributos redundantes e irrelevantes, por el efecto negativo generado en el algoritmo de aprendizaje aplicado con posterioridad. Para evitar la inclusión de atributos redundantes en el subconjunto generado por un algoritmo de ranking, es necesario una fase posterior de refinamiento.

2.1. Algoritmos

En este trabajo hemos escogido cuatro criterios para ordenar atributos. Son muy diferentes entre ellos y los describimos brevemente a continuación:

- χ^2 (CH) fue presentado por primera vez por Liu y Setiono [3] como método de discretización y más tarde se mostró que era capaz de eliminar atributos redundantes y/o no relevantes.
- Ganancia de información (IG), basada en el concepto de entropía de la teoría de la información, es una medida de la incertidumbre de una variable aleatoria.
- Relief (RL) basándose en la técnica del vecino más cercano asigna un peso a cada atributo. Sus creadores fueron Kira y Rendell [4] y posteriormente fue modificado por Kononenko [5]. El peso de cada atributo se va modificando en función de la habilidad para distinguir entre los valores de la variable clase.
- SOAP [6] (SP) (Selection of Attributes by Projections) Este criterio se basa en un único valor denominado NCE (Número de Cambios de Etiqueta),

que relaciona cada atributo con la etiqueta que sirve de clasificación. Este valor se calcula proyectando los ejemplos de la base de datos sobre el eje correspondiente a ese atributo (los ordenamos por él), para a continuación recorrer el eje desde el menor hasta el mayor valor del atributo contabilizando el número de cambios de etiqueta que se producen.

Una vez obtenidos los rankings de atributos, comprobaremos los resultados calculando la tasa de aciertos con tres clasificadores. Los algoritmos de aprendizaje empleados se han elegido por ser representativos de diferentes tipos de clasificadores, y se usan con frecuencia en los estudios comparativos: C4.5 [7] es una herramienta que crea un árbol de decisión con los datos de entrenamiento y es un algoritmo de clasificación rápido, robusto y fácil de utilizar y entender, que le lleva a ser de los métodos más populares; Naïve Bayes [8] (NB) utiliza una versión simplificada de la fórmula de Bayés, y aun siendo uno de los clasificadores más simples, sus resultados son competitivos; El vecino más cercano [9] (1-NN), simplemente asigna la clase del ejemplo más próximo utilizando una función de distancia.

2.2. Definiciones

Antes de proseguir con el estudio de los rankings de atributos se darán algunas definiciones para describir formalmente los conceptos utilizados en el documento: ranking de atributos, clasificador, exactitud de clasificación y exactitud de clasificación basada en ranking.

Definition 1 (Datos). *Sea* D *un conjunto de ejemplos* $e_i = (\overline{x_i}, y_i)$*, donde* $\overline{x_i} = (a_1, \ldots, a_m)$ *es un conjunto de atributos de entrada e* y_i *es el atributo de salida. Cada uno de los atributos de entrada pertenece al conjunto de los atributos (*$a_i \in$ A*, continuo o discreto) y cada ejemplo pertenece al conjunto de datos (*$e_i \in$ D*). Sea* C *el atributo de decisión (*$y_i \in$ C*), denominado clase, el cual se utilizará para clasificar los datos. Por simplicidad en este trabajo,* y_i *significa "la etiqueta de la clase del ejemplo* e_i*"*

Definition 2 (Ranking de Atributos). *Sea* A *el conjunto de m atributos* $\{a_1, a_2, \ldots, a_m\}$*. Sea r una función* $r : \mathrm{A_D} \rightarrow \mathbb{R}$ *que asigne un valor de evaluación a cada atributo* $a \in$ A *de* D*. Un ranking de atributos es una función* F *que asigna un valor de evaluación (relevancia) a cada atributo* $a \in$ A *y devuelve una lista de los atributos (*$a_i^* \in$ A*) ordenados por su relevancia, con* $i \in \{1, \ldots, m\}$*:*

$$F(\{a_1, a_2 \ldots, a_m\}) =< a_1^*, a_2^*, \ldots, a_m^* >$$

donde $r(a_1^*) \geq r(a_2^*) \geq \ldots \geq r(a_m^*)$*.*

Por convención, asumimos que un valor alto es indicativo de un atributo relevante y que los atributos están en orden decreciente según $r(a^*)$. Consideramos definido el criterio de ordenación para atributos individuales, independientes del contexto de los demás, y nos limitaremos a criterios de aprendizaje supervisado.

Definition 3 (Clasificación). *Un clasificador es una función* H *que asigna una etiqueta de una clase a un nuevo ejemplo:* $\mathrm{H} : \mathrm{A}^p \rightarrow \mathrm{C}$*, donde* p *es el número de atributos utilizado por el clasificador ,* $1 \leq p \leq m$*.*

La exactitud de clasificación (CA) es la media de la tasa de aciertos proporcionada por el clasificador H *dado un conjunto de ejemplos de test, es decir, el promedio del número de veces que* H *fue capaz de predecir la clase de los ejemplos de test.*

Sea $\mathbf{x}$ *una función que extrae los atributos de entrada del ejemplo* e*,* $\mathbf{x} : \mathrm{A}^m \times \mathrm{C} \rightarrow \mathrm{A}^m$*. Para un ejemplo de test* $e_i^* = (x_i, y_i)$*, si* $\mathrm{H}(\mathbf{x}(e_i^*)) = y_i$ *entonces* e_i^* *está clasificado correctamente; en otro caso se considerará no clasificado.*

En este documento, utilizaremos la validación cruzada dejando uno fuera o leaving–one–out (explicada brevemente en la sección 6) para medir el comportamiento de los clasificadores, y no depender de la aleatoriedad al formar los conjuntos de entrenamiento, como ocurre en la validación cruzada con k conjuntos (k–fold cross–validation). En la expresión siguiente, si $\mathrm{H}(e_i) = y_i$ entonces se cuenta 1, y 0 en cualquier otro caso.

$$CA = \frac{1}{N} \sum_{i=1}^{N} (\mathrm{H}(\mathbf{x}(e_i)) = y_i)$$

Al estar interesados en rankings, la exactitud de clasificación se medirá con respecto a muchos subconjuntos diferentes del ranking proporcionado por alguno de los métodos de ranking de atributos.

Definition 4 (Clasificación Basada en Ranking). *Sea* S_k^{F} *una función que devuelve el subconjunto de los* k *primeros atributos proporcionados por el método de ranking* F *(*$S_k^{\mathrm{F}} : \mathrm{A}^m \rightarrow \mathrm{A}^k$*). La exactitud de la clasificación basada en ranking de* H *será la siguiente:*

$$CA_k(\mathrm{F}, \mathrm{H}) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathrm{H}(S_k^{\mathrm{F}}(\mathbf{x}(e_i))) = y_i\right)$$

Tener en cuenta que S_1^{F} es el primer (mejor) atributo del ranking elaborado por F; S_2^{F} son los dos primeros atributos, y así hasta m.

3. Trabajos relacionados

Existe poca bibliografía específica donde se defina cómo comparar rankings de atributos. En [10] se comenta brevemente el uso de la curva de aprendizaje para mostrar el efecto de añadir atributos cuando se dispone de una lista de atributos ordenada. Se empieza con el primer atributo (el más relevante) y gradualmente se añade el siguiente más relevante de la lista de atributos, uno a uno hasta el final del ranking, obteniendo la tasa de aciertos con cada subconjunto. Uniendo cada valor obtenido se forma la curva.

Nos encontramos trabajos [11] en los que se comparan rankings de atributos mediante un sólo subconjunto, el que mejor clasificación obtiene de todos los subconjuntos formados para obtener la curva de aprendizaje. Este estudio, nos puede servir para comparar rankings con algoritmos de selección de subconjuntos de atributos, pero no refleja si el comportamiento de una lista de atributos es mejor que otra.

Pero lo más simple y difundido es seleccionar los atributos cuyo peso de relevancia sea mayor que un valor umbral [6,12], normalmente establecido por el usuario. O si se quiere escoger un conjunto con k atributos, simplemente seleccionamos los k primeros de la lista ordenada. Es un método rápido de elegir atributos y comparar resultados, pero con la dificultad de establecer un parámetro que difícilmente maximiza resultados en todas las bases de datos, teniendo que estudiarlo particularmente para cada caso.

En [13], se inserta en la base de datos un atributo irrelevante (cuyos valores son aleatorios) que se utiliza de umbral al aplicar el algoritmo de ranking.

En todos los casos mencionados anteriormente, la medida utilizada para cuantificar el comportamiento de un ranking es la exactitud obtenida por un clasificador con los k primeros atributos de la lista, diferenciándose en el modo de fijar el umbral. Esto nos lleva a plantear las siguientes cuestiones: ¿Qué se está evaluando exactamente, el ranking o el método para seleccionar los atributos? ¿Es correcto?

El valor que se utiliza en la comparación depende de tres factores: ranking generado, método de fijar el umbral y algoritmo de aprendizaje. Es un hecho que la exactitud de los modelos de clasificación puede variar sustancialmente según los atributos que intervengan; por consiguiente, la forma de elegir los atributos influirá considerablemente en el resultado final. Obviamente, se está evaluando ranking y método de selección, pero parece que cobra más importancia la forma de elegir los atributos que el orden en el que se encuentran. Por lo tanto, podemos afirmar que las comparaciones serán justas, pero no completas. Nuestra propuesta es evaluar el ranking directamente, sin esa dependencia con el método de selección.

4. Motivación

En primer lugar, observemos la calidad de los cuatro métodos de ranking de atributos con respecto a los tres clasificadores, utilizando la base de datos Glass2 (214 ejemplos, 9 atributos, 2 clases) como caso representativo motivador de nuestro análisis. Para ello, en el apartado anterior, se comentó que el poder predictivo de un subconjunto de variables se puede medir en términos de tasa de error o de aciertos de algún clasificador. Sin embargo, en nuestro caso, queremos evaluar una lista de atributos ordenada, no un subconjunto en particular.

La Tabla 1 muestra los ranking para χ^2, Ganancia de información, Relief y SOAP. Para cada método de ranking, la fila rk presenta el ranking de atributos generado con uno de los métodos, y bajo esta fila, el resultado de clasificación con C4.5, Naïve Bayes y la técnica del vecino más cercano utilizando el número

Tabla 1. Ranking de Atributos para G2. FR: método de Feature–Ranking (CH: χ^2; IG: Information Gain; RL: Relief; SP: Soap); Cl: Clasificador (c4: C4.5; nb: Naïve Bayes; nn: 1–Nearest Neighbour); y rk: ranking de atributos.

		Subconjunto								
FR	Cl	1	2	3	4	5	6	7	8	9
CH	rk	7	1	4	6	3	2	9	8	5
	c4	**73.6**	77.9	**82.2**	78.5	75.5	74.8	73.6	76.1	75.5
	nb	57.1	57.1	66.9	69.9	63.8	63.8	63.2	62.0	62.0
	nn	66.9	**79.7**	75.5	**82.8**	**88.3**	**81.0**	**77.9**	**77.9**	**77.3**
IG	rk	7	1	4	3	6	2	9	8	5
	c4	**73.6**	77.9	**82.2**	82.2	75.5	74.8	73.6	76.1	75.5
	nb	57.1	57.1	66.9	63.8	63.8	63.8	63.2	62.0	62.0
	nn	66.9	**79.7**	75.5	**84.7**	**88.3**	**81.0**	**77.9**	**77.9**	**77.3**
RL	rk	3	6	4	7	1	5	2	8	9
	c4	57.7	67.5	80.4	76.7	75.5	75.5	74.8	77.9	75.5
	nb	**62.0**	62.6	65.0	64.4	63.8	63.8	63.8	62.6	62.0
	nn	58.9	**75.5**	**81.0**	**83.4**	**88.3**	**83.4**	**81.6**	**81.6**	**77.3**
SP	rk	1	7	4	5	2	3	6	9	8
	c4	**77.3**	77.9	**82.8**	**81.6**	**81.6**	**84.1**	74.9	73.0	75.5
	nb	52.2	57.1	66.9	65.6	62.6	63.2	63.8	62.0	62.0
	nn	72.4	**79.7**	75.5	79.8	80.4	82.2	**81.6**	**77.3**	**77.3**

de atributos del ranking indicado en la primera fila, bajo el título "Subconjunto". Resaltamos, la diferencia existente en los resultados de los distintos clasificadores con el mismo conjunto de datos (con el conjunto de datos completo: 75.5, 62.0 y 77.3, respectivamente).

Se puede observar que el atributo más relevante para χ^2 e IG fue 7, para RL 3 y para SP 1. Utilizando sólo el atributo 7 (CH e IG), C4.5 obtiene una tasa de aciertos de 73.6, siendo 57.7 en el caso de utilizar sólo el atributo 3 (RL) y 77.3 en el caso del 1 (SP). El segundo atributo seleccionado por χ^2 e IG fue el 1, el 6 por RL y el 7 por SP, siendo el mismo subconjunto (y por tanto el mismo resultado obtenido) para χ^2, IG y SP. Al ser escogido el atributo 4 en los cuatro casos como tercer atributo relevante, se mantiene el mismo subconjunto para χ^2, IG y SP, pero a partir de este punto, SP ya no coincidirá con los otros dos, que si mantendrán el mismo ranking a partir del sexto atributo seleccionado.

Analizando la Tabla 1 se puede inferir varias conclusiones interesantes: (a) Los cuatro métodos de ranking proporcionan diferentes listas de atributos, que obviamente conducen a diferentes comportamientos en el algoritmo de aprendizaje. (b) El par SP+C4.5 es el único que obtiene una tasa de aciertos (77.3) utilizando sólo un atributo (atributo 1) que mejora a la tasa obtenida con el conjunto completo de atributos (75.5). (c) En principio, la secuencias que mejores resultado de clasificación obtiene son arbitrarias (SP+C4,77.3), (SP+NN,79.8),

(SP+C4,82.8), (IG+NN,84.7), ({CH,IG,RL}+NN,88.3), (SP+NN,84.1), ({RL + SP}+NN,81.6), (RL+NN,81.6) y el último mejor valor 77.3 con 1NN. (d) Parece que 1NN ofrece buenos resultados cuando el número de atributos es mayor que $m/2$. Un hecho significante es que los mejores cinco atributos con 1NN son 1,3,4,6,7, pero los mejores seis atributos son 1,2,3,4,5,7. El atributo 6 no es relevante cuando los atributos 2 y 5 son tenidos en cuenta. En general, un atributo que es completamente irrelevante por si mismo, puede proporcionar una mejora significativa en el comportamiento del modelo cuando se toma junto a otros atributos.

La Figura 1 es ilustrativa de las situaciones en las que nos podemos encontrar al comparar diferentes rankings para un conjunto de datos. La pregunta que se plantea es ¿Qué ranking es mejor para clasificar? La respuesta estaría condicionada a lo que el usuario esté buscando. Es decir, si lo que interesa es identificar el método de ranking que obtenga el subconjunto con mejor clasificación para un algoritmo de aprendizaje dado, escogeríamos Metodo1 sabiendo que para ello necesita el ochenta por ciento de los atributos. Sin embargo, se observa que en el resto de la curva, los resultados de clasificación casi siempre están por debajo de los otros dos métodos. Si escogemos un número de atributos inferior al setenta por ciento, el resultado del Metodo1 será el peor de los tres. Si lo que buscamos es el método con un comportamiento mejor a lo largo de toda la curva, comparamos punto a punto la evolución de las tres curvas. El Metodo2 pierde al principio (hasta el treinta por ciento de los atributos) frente al Metodo3, posteriormente siempre mejora a los otros dos, salvo puntualmente en el caso comentado anteriormente (con el ochenta por ciento de los atributos). Y por último, si se quiere seleccionar menos del treinta por ciento de los atributos, el mejor es el Metodo3.

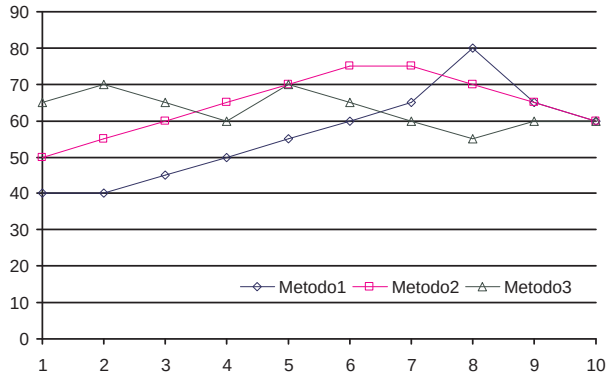

Figura 1. Ejemplo ficticio de tres tipos diferentes de curvas de aprendizaje.

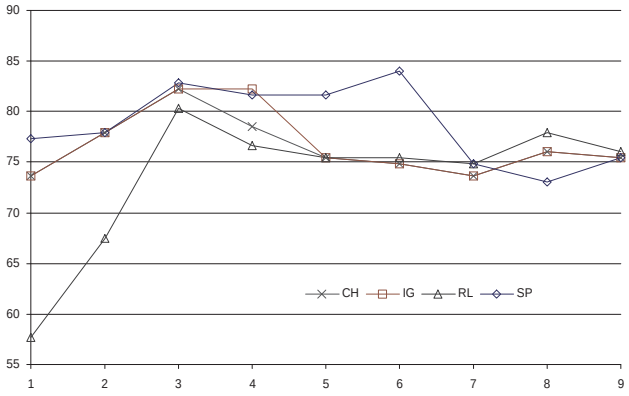

Figura 2. La exactitud obtenida con C4.5 para la base de datos Glass2 (datos de la Tabla 1). En el eje de abscisas el nº de atributos utilizados en la clasificación y en el de ordenada la tasa de aciertos

La figura 2 muestra la exactitud en la clasificación de C4.5 con los cuatro métodos de ranking de atributos para la base de datos Glass2. Aunque la exactitud del mejor subconjunto es parecida, el comportamiento de SOAP es excelente para cualquier número de atributos y es el único que en casi todos los subconjuntos aparece por encima de la media. Por lo tanto podríamos afirmar que es mejor ranking que los demás.

El análisis basado en el mejor subconjunto, no refleja exactamente la bondad del ranking de atributos, puesto que antes o después de ese subconjunto los resultados podrían ser pésimos. La Figura 3 muestra un comportamiento en la clasificación con C4.5 muy diferente para los rankings de los distintos métodos, sin embargo, si extraemos el resultado del mejor subconjunto de cada uno, llegamos a la conclusión de que son rankings muy parecidos (CH:80.77, IG:80.29, RL:80.77 y SP:81.25).

En la Figura 4 observamos las curvas generadas con los cuatro métodos de ranking para la base de datos *Segment*, utilizando el clasificado NB. Si evaluamos los rankings por el mejor subconjunto de cada método, al comparar obtendríamos el mismo resultado, al coincidir ese valor en las cuatro listas de atributos. Sin embargo, la evolución de las curvas es diferente, observamos que el comportamiento de los métodos CH e IG es inferior en casi todos los subconjuntos a RL y SP.

Teniendo en cuenta estas conclusiones, queremos considerar la posibilidad de llegar a comprender mejor cuando un ranking de atributos es mejor que otros para un clasificador dado. Además, sería interesante analizar el comportamiento del método de ranking a lo largo de la curva de aprendizaje descrita, y sacar conclusiones en función de la proporción de atributos utilizada.

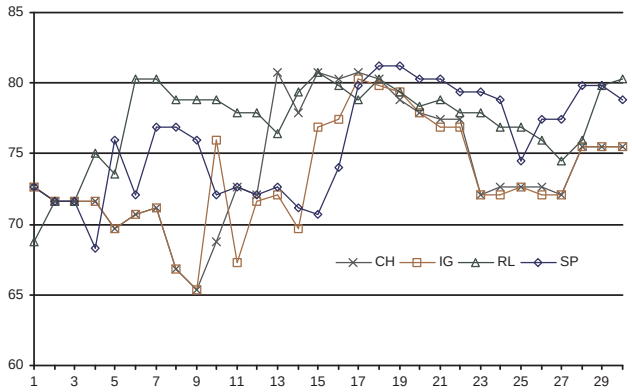

Figura 3. La exactitud obtenida con C4.5 para la base de datos Sonar. En el eje de abscisas el nº de atributos utilizados en la clasificación y en el de ordenada la tasa de aciertos

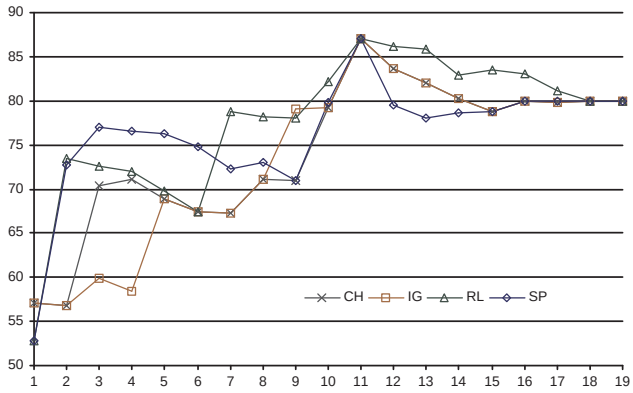

Figura 4. La exactitud obtenida con NB para la base de datos Segment. En el eje de abscisas el nº de atributos utilizados en la clasificación y en el de ordenada la tasa de aciertos

5. Metodología propuesta

Una comparación más completa entre dos rankings de atributos sería comparar subconjunto a subconjunto, es decir, comparar los resultados de clasificación obtenidos con el primer (mejor) atributo de las dos listas, con los dos mejores, y así sucesivamente hasta los m atributos del ranking. Calculando el promedio de los resultados obtenidos con cada lista, podríamos utilizarlo para comparar los rankings. Un estudio muy parecido sería el calcular el área bajo la curva que describen los resultados anteriores.

El area que queda bajo una curva (AUC-Area Under the Curve) se calcula aplicando la fórmula del trapecio. En nuestro caso hacemos referencia a la curva de aprendizaje obtenida al ir añadiendo atributos según el orden asignado por el método de ranking, pudiéndose cuantificar su evolución y así comparar los resultados de los distintos rankings según su área.

$$\sum_{i=1}^{m-1} (x_{i+1} - x_i) * \frac{(y_{i+1} + y_i)}{2}$$

Definition 5 (AURC). *Dado un método de ranking de atributos* F *y un clasificador* H*, podemos obtener el comportamiento del método de clasificación con respecto a la lista de atributos suministrada por el método de ranking, por medio del valor del área bajo la curva* $AURC(\mathrm{F},\mathrm{H})$*. La curva se dibuja al unir cada dos puntos* $(CA_k(\mathrm{F},\mathrm{H}), CA_{k+1}(\mathrm{F},\mathrm{H}))$*, donde* $k \in \{1,\ldots,m-1\}$ *y* m *es el número de atributos. El área bajo la curva de comportamiento de clasificación basado en ranking (AURC–The Area Under Ranking Classification performance Curve)* $AURC(\mathrm{F},\mathrm{H})$ *se calculará:*

$$AURC(\mathrm{F},\mathrm{H}) = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (CA_i(\mathrm{F},\mathrm{H}) + CA_{i+1}(\mathrm{F},\mathrm{H}))$$

Con esta expresión, para cada par (F,H), el área bajo la curva $AURC(\mathrm{F},\mathrm{H}) \in [0,1]$ (en las tablas aparece multiplicado por cien para mayor claridad), lo que nos proporciona un excelente método para comparar la calidad de los rankings de atributos con respecto a los algoritmos de clasificación. Una propiedad interesante observada en la curva, es que no es monótona creciente, es decir, para algunos i, sería posible que $CA_i(\mathrm{F},\mathrm{H}) > CA_{i+1}(\mathrm{F},\mathrm{H})$.

En la Figura 4, tomando los valores de AURC como valor representativo de los ranking generados por los distintos métodos, CH:74.63, IG:73.79, RL:78.26 y SP:76.77, se comprueba que el resultado de la comparación es más fiel a la realidad que el obtenido mediante la comparación por el mejor subconjunto. Lo mismo ocurre en la Figura 3, donde los valores de AURC son: CH:73.92, IG:73.71, RL:76.06 y SP:75.15, reflejando la diferencia de calidad entre ellos.

En la Figura 4, donde el mejor AURC lo tiene el ranking elaborado por el método RL, vemos que el comportamiento de las dos mejores curvas (RL y SP) cambia a partir del séptimo atributo, siendo hasta ese instante SP el mejor método. Lo que nos lleva a realizar un estudio por porcentajes de atributos.

Definition 6 (Comportamiento de un ranking de atributos). *El comportamiento de un ranking de atributos se mide como la evolución del AURC a lo largo del ranking de características, con paso δ %. Los puntos de la curva, para la que se calcula el AURC, se calculan para cada δ % de los atributos de la siguiente manera:*

$$AURC_{\delta}(\mathrm{F},\mathrm{H}) = \frac{1}{2\delta(m-1)} \sum_{i=1}^{\delta(m-1)} (CA_i(\mathrm{F},\mathrm{H}) + CA_{i+1}(\mathrm{F},\mathrm{H}))$$

Debemos tener en cuenta que la idea de todo método de selección de atributos, entre ellos los de ranking, es la de quedarse con el menor número de atributos posible, esto unido a la posibilidad de que dentro de la curva de aprendizaje se compensen exactitudes altas y bajas para el cálculo del AURC, nos lleva a realizar un estudio del comportamiento de los método con los primeros atributos y con porcentajes fijos.

6. Resultados

6.1. Experimentos

La implementación de los algoritmos de inducción y de los selectores fue desarrollada en Java, utilizando la librería Weka [14]. Para llevar a cabo la comparativa se utilizaron dieciseis bases de datos de la Universidad de California Irvine [15], cuyas características se muestran en la Tabla 2. Las bases de datos tienen en común la ausencia de valores perdidos entre sus atributos, por lo que los resultados no se verán afectados por la forma en la que se resuelva este inconveniente.

El proceso seguido en este trabajo para medir la capacidad de un clasificador en predecir las nuevas instancias correctamente, es la validación cruzada dejando uno fuera o "leaving-one-out", es decir, para N instancias, se usa como conjunto de entrenamiento N-1 instancias y se comprueba la clasificación de la instancia que queda fuera. Repitiendo el proceso con cada instancia, y dividiendo el número de aciertos por N obtendríamos la tasa de exactitud. Este proceso es computacionalmente muy costoso, al tener que construir el clasificador N veces, produce una varianza muy alta y los modelos generados están sobreajustados, pero se escoge esta validación cruzada por su ausencia de aleatoriedad, pues el objetivo de estudio no es la exactitud de los modelos, sino realizar una comparación entre los rankings obtenidos por los distintos métodos.

A cada base de datos se le aplica los cuatro métodos de ranking de atributos, y para cada ranking se calcula la curva de aprendizaje con cada uno de los tres clasificadores. Al aplicar validación cruzada leaving-one-out a cada subconjunto ordenado de atributos del ranking, la cantidad de modelos generados para cada selector-clasificador asciende a 286.009 (16 bd x nºatt's de cada una), como tenemos doce combinaciones selector-clasificador (4 selectores x 3 clasificadores), el total de modelos generados es de 3.432.108.

Tabla 2. Bases de datos utilizadas en los experimentos.

Dataset	Id	Instances	Attributes	Classes
anneal	AN	898	38	6
balance	BA	625	4	3
g_credit	GC	1000	20	2
diabetes	DI	768	8	2
glass	GL	214	9	7
glass2	G2	163	9	2
heart-s	HS	270	13	2
ionosphere	IO	351	34	2
iris	IR	150	4	3
kr-vs-kp	KR	3196	36	6
lymphography	LY	148	18	4
segment	SE	2310	19	7
sonar	SO	208	60	2
vehicle	VE	846	18	4
vowel	VW	990	13	11
zoo	ZO	101	16	7

6.2. Área Bajo la Curva

La Tabla 3 muestra, para cada conjunto de datos, el área bajo la curva de clasificación basada en ranking obtenida con cada combinación ranking-clasificador. Los valores en **negrita** son los mejores para los tres clasificadores, y los que aparecen subrayados son los mejores para el clasificador correspondiente. No se pueden extraer conclusiones claras, pero sí ciertas tendencias:

- Atendiendo a los promedios que aparecen en la última fila, los resultados son muy parecidos bajo cada clasificador, pero sí existen diferencias entre cada uno de ellos. El clasificador que ofrece mejor comportamiento con los cuatro métodos de ranking de atributos es 1NN, seguido muy de cerca por C4.5, y en último lugar con una diferencia significativa NB.
- Si tenemos en cuenta el mejor AURC para cada base de datos, observamos que en la mitad de las bases de datos, el mejor AURC se alcanza con el clasificador 1NN, lo que viene a reforzar la conclusión anterior.
- Por clasificadores, comprobamos que en los tres casos RL es el que mas veces gana, por lo que podríamos concluir que es mejor método de ranking.

6.3. Porciones de Área Bajo la Curva

En la Figura 5, siguiendo la definición 6, se usa un paso de 5 % (δ=0.05). Aunque la evolución del AURC es más apropiado analizarlo independientemente para cada base de datos, esta figura muestra el promedio del comportamiento del método de ranking de atributos para las dieciseis bases de datos. Se puede distinguir tres grupos de curvas, cada uno con cuatro curvas (una por método de ranking).

Tabla 3. Valor AURC para cada combinación ranking-clasificador.

	C4.5				NB				1NN			
BD	CHI2	IG	RLF	SOAP	CHI2	IG	RLF	SOAP	CHI2	IG	RLF	SOAP
an	97.30	97.12	96.90	97.11	85.82	86.30	86.50	86.47	**98.20**	98.09	97.54	97.71
bs	68.61	68.61	68.83	68.61	**75.55**	**75.55**	72.56	**75.55**	72.77	72.77	69.79	72.77
gc	72.39	72.39	71.71	72.31	**74.74**	**74.74**	73.89	74.22	70.16	70.16	70.38	66.83
di	72.85	72.89	73.30	72.52	75.36	**75.73**	75.68	75.15	68.87	69.52	68.09	67.87
gl	64.57	66.09	67.09	67.32	49.15	49.85	47.34	51.37	63.49	67.67	68.17	**71.12**
g2	76.65	77.11	74.35	79.03	63.27	62.50	63.50	62.27	79.41	79.64	**80.37**	78.91
hs	78.23	78.23	77.04	76.54	**83.09**	**83.09**	81.53	80.80	78.43	78.43	75.94	74.34
io	88.69	89.18	**90.03**	86.94	84.52	85.11	85.66	80.23	88.57	88.36	88.57	86.92
ir	95.11	95.11	95.22	95.22	95.56	95.56	95.56	95.56	95.00	95.00	95.56	95.56
kr	95.22	95.13	**96.48**	95.56	87.47	87.47	89.81	86.99	93.97	93.89	95.79	94.63
ly	74.84	75.97	75.66	75.42	78.76	80.25	80.37	80.09	76.61	81.28	**82.31**	80.56
se	92.23	92.15	93.37	92.96	74.63	73.79	78.26	76.77	92.78	93.11	**93.87**	93.26
so	73.92	73.71	76.06	75.15	67.62	67.44	69.57	68.83	84.15	83.94	**84.41**	83.87
ve	64.59	65.79	**68.10**	67.43	41.65	41.54	41.72	41.04	66.09	65.83	65.79	65.69
vw	73.98	74.20	73.96	74.59	61.96	62.46	61.65	62.17	**90.67**	90.66	89.63	90.52
zo	88.18	87.56	86.88	88.27	88.95	88.21	86.42	89.36	**91.34**	90.84	87.69	90.87
Pr	79.83	80.08	80.31	80.31	74.25	74.35	74.37	74.18	81.91	82.45	**82.12**	81.96

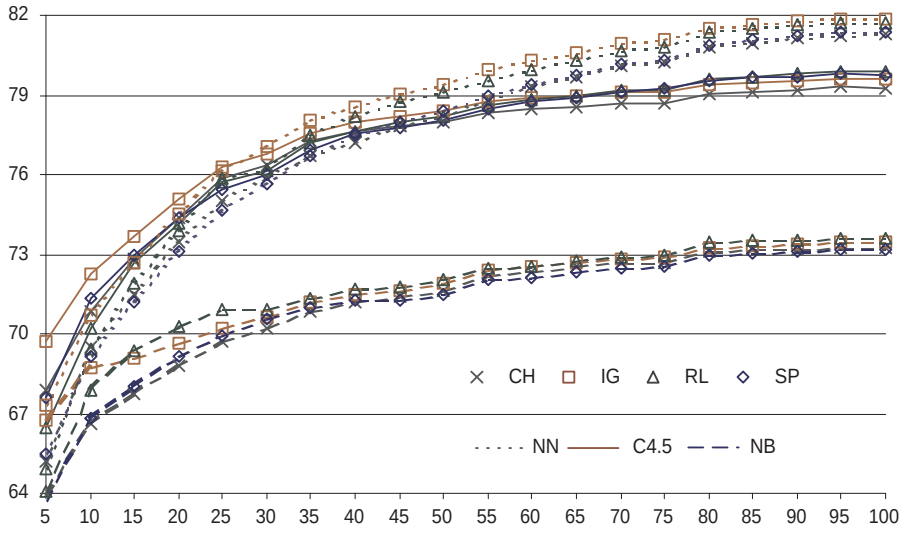

Figura 5. AURC utilizando porcentajes de atributos (con paso 5 %) . En el eje de abscisas el porcentaje de atributos utilizados en la clasificación y en el de ordenadas el promedio de AURC para todas las bases de datos.

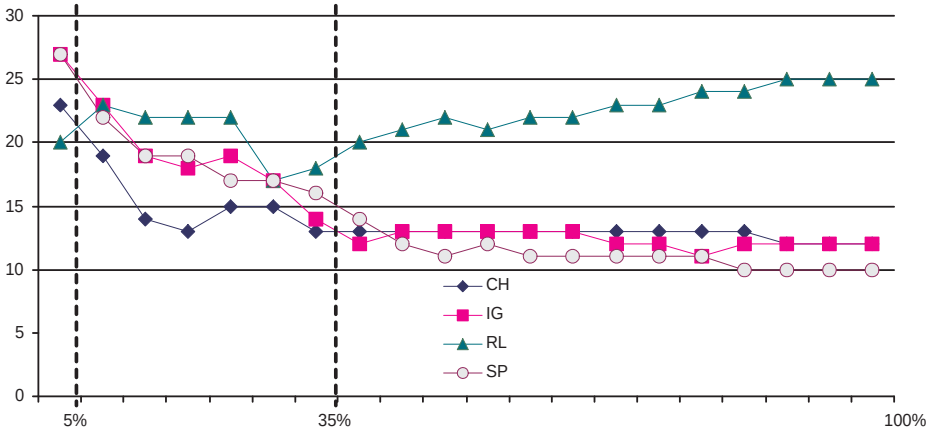

Figura 6. Relación entre el porcentaje de atributos seleccionado del ranking y la mejor clasificación obtenida. En el de ordenadas, el número de veces que un método de ranking fue mejor que los demás.

En promedio, IG es el mejor método de ranking de atributos. Referente a los clasificadores, cuando seleccionamos un número inferior al 15 % de los atributos de un ranking, C4.5 es el mejor clasificador. Sin embargo, por encima del 35 % 1–NN es el mejor, especialmente con RL e IG (ver en la figura que las curvas de C4.5 interseccionan con las de 1–NN aproximadamente en el 25 %). Excepto en muy pocos casos, especialmente cuando están involucrados muy pocos atributos relevantes, el clasificador Naïve Bayes no obtiene buenos resultados de comportamiento.

Si en vez de trabajar con los promedios de todas las bases de datos, donde se pueden solapar unos resultados buenos con otros, lo hacemos contando el número de veces que un método es mejor que los demás, obtenemos la gráfica que aparece en la Figura 6. Para menos del 5 % de los atributos, las mejores técnicas de ranking son IG y SP, las cuales mantienen una tendencia similar. También es obvio que RL es el más estable cuando el porcentaje de atributos está por encima del 35 % del total, quedando un intervalo entre ambas zonas (5 %–35 %) de cambios en las tendencias de las curvas.

Además, queremos saber las exactitudes obtenidas con cada clasificador para los conjuntos de datos reducidos al primer atributo de las listas generadas por cada método de selección. Además la AURC de las curvas obtenidas después de reducir los datos a dos, tres, cuatro y cinco atributos, puesto que entendemos que este será el objetivo de cualquier selector de atributos. Para terminar con este juego de comparaciones, estudiaremos la AURC para los conjuntos obtenidos con el 25 %, 50 % y con el 100 % de los atributos de los rankings de cada método.

Se utilizan los resultados de clasificación con los algoritmos C4.5, NB y NN mediante validación cruzada leaving-one-out, después de reducir cada base de datos con los cuatro métodos de selección con las opciones antes descritas, se

calcula el AURC según lo indicado en el párrafo anterior, disponiendo finalmente de mil quinientos treinta y seis valores (16bd.x3cl.x4ra.x8pruebas=1536) para extraer conclusiones. Serán cuatro las posiciones posibles en cada comparación, interesándonos sólo la primera (en caso de empate cuenta para todos los implicados).

Tabla 4. Resumen de las veces que cada método de ranking ocupa la 1ª posición. Resultados agrupados por: los primeros atributos, porcentajes y clasificadores

Resultados por	CH	IG	rl	SP
Exactitud-1at:	26	28	20	**30**
AURC-2at:	18	22	16	**24**
AURC-3at:	15	17	15	**21**
AURC-4at:	14	15	16	**20**
AURC-5at:	15	**20**	18	17
AURC-25 %:	17	**22**	21	17
AURC-50 %at:	13	12	**21**	13
AURC-all at:	14	12	**25**	12
C4.5:	39	46	50	**51**
NB:	41	**59**	58	46
NN:	50	43	44	**55**
Total:	130	148	**152**	**152**

En la Tabla 4 disponemos de un resumen de las veces que cada método de ranking ocupa la 1ª posición. Se establecen distintos grupos de comparaciones: En el primer bloque se encuentran los resultados obtenidos con los primeros atributos (con el primer atributo se compara la tasa de aciertos y el AURC con dos, tres, cuatro y cinco atributos); El segundo bloque muestra los resultados de las comparaciones por porcentajes (25, 50 y con todos los atributos); Y en el último grupo se desglosa por clasificadores.

Si atendemos a las pruebas realizadas con los primeros atributos, destaca el método de ranking SOAP, sobre todo con los clasificadores C4.5 y 1NN, siendo IG el que mejor resultado ofrece con NB utilizando sólo los primeros atributos del ranking.

Al 25 % del ranking IG y RL obtienen mejores resultados (IG: 22, rl: 21 y CH,SP: 17). Parcialmente por clasificador, se mantiene esta posición con C4.5 y NB pero con NN la primera posición al 25 % es para relief. Y a partir de aquí hasta el total del conjunto de atributos, RL es el que más veces ocupa la primera posición. Al 100 % del ranking, relief gana con diferencia 25 veces frente a CH y IG con 12 y SP con 10 e igualmente al 50 % de los atributos del ranking. Los resultados se mantienen con estos porcentajes (50 y 100) para los tres clasificadores.

Si el estudio de todas las ocho pruebas se hace por clasificadores, no existen grandes diferencias. Destacamos con el clasificador C4.5 los métodos SP y RL, y

con muy poca diferencia sobre IG. Con NB, los que ocupan los primeros puestos son IG y RL, mientras que con 1NN, es SP.

Los que más veces han quedado en primera posición en el conjunto de todas la pruebas (480) fueron SOAP y Relief con 152, seguido de IG con 148 y chi2 con 130.

Con los resultados obtenidos en las tres pruebas anteriores (AURC, porcentajes de AURC y AURC con los primeros atributos de las listas ordenadas) podemos realizar las siguientes recomendaciones:

- El AURC da una idea más completa de la bondad del ranking que la exactitud obtenida con un subconjunto de atributos.
- La lista completa mejor valorada es la generado mediante el algoritmo *RL*, sin embargo, si se va a trabajar con los primeros atributos o con menos del 25 % de los atributos, los métodos *SP* e *IG* ofrecen mejores resultados en menos tiempo.
- En general, los mejores resultados de clasificación se obtienen con *1NN*, aunque cuando el número de atributos seleccionado es pequeño (inferior al 25 %) C4.5 se comporta mejor en los cuatro casos que los demás clasificadores.

7. Conclusiones

Los trabajos tradicionales donde se realizan comparaciones de algoritmos de rankings de atributos, principalmente evalúan y comparan el método de selección de atributos, en vez del ranking. En este trabajo presentamos una metodología para la evaluación del ranking, partiendo de la premisa de la no existencia de un único subconjunto ideal para todos los casos, y de que el mejor ranking va a depender de lo que el usuario esté buscando.

Podemos concluir que el área bajo la curva del ranking (*AURC*) refleja el comportamiento completo de la lista de atributos ordenada, indicando su poder predictivo en global. En base al análisis de la evolución del *AURC*, se ha recomendado el uso de los algoritmos *SP* e *IG* y del clasificador *C4.5* cuando queremos clasificar con pocos atributos, y *RL* y el clasificador *1NN* en los demás casos.

Nuestros trabajos irán a partir de ahora a confirmar si se pueden extrapolar estos resultados a otras bases de datos con mucho mayor tamaño, así como en profundizar si existe alguna relación entre el método de ranking y el clasificador elegido. Además, se pretende ampliar el estudio con otras medidas de evaluación de atributos.

8. Agradecimientos

Este trabajo ha sido elaborado en el marco del proyecto de investigación oficial TIC2001–1143–C03–02, financiado por la Comisión Interministerial de Ciencia y Tecnología (CICYT).

Referencias

1. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analisys **1** (1997)
2. Golub, T.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–537
3. Liu, H., Setiono, R.: Chi2: Feature selection and discretization of numeric attributes. In: 7th IEEE International Conference on Tools with Artificial Intelligence. (1995)
4. Kira, K., Rendell, L.: A practical approach to feature selection. In: 9th International Conference on Machine Learning, Aberdeen, Scotland, Morgan Kaufmann (1992) 249–256
5. Kononenko, I.: Estimating attributes: Analysis and estensions of relief. In: European Conference on Machine Learning, Vienna, Springer Verlag (1994) 171–182
6. Ruiz, R., Riquelme, J., Aguilar-Ruiz, J.: Projection-based measure for efficient feature selection. Journal of Intelligent and Fuzzy System **12** (2002) 175–183
7. Quinlan, J.: C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, California (1993)
8. Mitchell, T.: Machine Learning. McGraw Hill (1997)
9. Aha, D., Kibler, D., Albert, M.: Instance-based learning algorithms. Machine Learning **6** (1991) 37–66
10. Liu, H., Motoda, H.: Feature Selection for Knowlegde Discovery and Data Mining. Kluwer Academic Publishers, London, UK (1998)
11. Hall, M., Holmes, G.: Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering **15** (2003)
12. Hall, M.: Correlation-based feature selection for discrete and numeric class machine learning. In: 17th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2000) 359–366
13. Stoppiglia, H., Dreyfus, G., Dubois, R., Oussar, Y.: Ranking a random feature for variable and feature selection. Journal of Machine Learning Research **3** (2003) 1399–1414
14. Witten, I., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco (2000)
15. Blake, C., Merz, E.K.: Uci repository of machine learning databases (1998)