Selección de medidas de evaluación de reglas obtenidas mediante programación genética basada en gramática

César Hervás-Martínez, Cristóbal Romero, Sebastián Ventura

Universidad de Córdoba, Campus Universitario de Rabanales, 14071, Córdoba, España
{chervas,cromero,sventura}@uco.es

Resumen. La mayoría de las técnicas de asociación en minería de datos necesitan una métrica adecuada para poder extraer el grado de dependencia que existe entre las variables asociadas a un conjunto de datos. En este trabajo analizamos, evaluamos y comparamos diferentes medidas de reglas de asociación, unas que provienen del campo de la estadística y otras definidas de forma especifica en minería de datos para la evaluación de reglas obtenidas en un entorno educativo mediante algoritmos de programación genética. La observación de los resultados muestra que algunas de estas medidas aportan información redundante y abren la posibilidad de reducir el número de estas reglas mediante un análisis en componentes principales, donde con sólo dos componentes se explica el 90% de la varianza aportada por nueve de las principales medidas propuestas en la literatura. Concluimos analizando el significado de cada componente y planteamos la necesidad de definir nuevas medidas como combinación lineal o no lineal de las utilizadas a priori.

1 Introducción

En los últimos años esta apareciendo en la comunidad científica que trabaja en minería de datos una gran importancia por determinar la cantidad de información de una regla puesto que muchos algoritmos de análisis de datos por computadora producen una gran cantidad de reglas, dando lugar a que el usuario no sea capaz de analizarlas. De esta forma es necesario establecer alguna medida que sea capaz de dar de forma numérica el grado de interés que tiene una regla determinada. De esta forma se han propuesto diferentes medidas algunas de ellas surgidas del campo de la estadística, otras del aprendizaje de máquinas y otras definidas de forma expresa para la minería de datos. Algunas de estas medidas muestran cuanto conocimiento se gana en la distribución de un atributo a partir del conocimiento de la distribución de otro, ejemplos de estas medidas son la ganancia de entropía, la información mutua, la ganancia de Gini y la medida χ^2 [11]. Por otra parte existen medidas (como el interés, la confianza, la ganancia en entropía, etc.) que son muy similares al coeficiente de correlación en la región donde los valores de soporte se encuentran de forma habitual.

En las publicaciones recientes sobre este tipo de medidas es habitual realizar comparaciones de las propiedades de las medidas estudiadas [11] [21], pero no lo es

comparar si las características de localización y dispersión de los valores que se asocian a un conjunto de reglas difieren significativamente. Este será por tanto uno de los objetivos del presente trabajo.

A menudo los estudios sobre rendimiento de ciertas heurísticas se analizan mediante tablas donde se incluye el número de ejecuciones, el número de evaluaciones, el tiempo de cpu, etc; mientras que por otra parte se analiza la calidad de las soluciones en clasificación (basada en el porcentaje de patrones mal clasificados), la calidad de las soluciones en modelado (basada en el error cuadrático medio o en el error estándar de predicción SEP), y en general, en optimización, basada en la cercanía al optimo cuando este se conoce. Las publicaciones estándar, de cierto nivel, siguieren la conveniencia de análisis estadísticos, pero existen pocas directrices de cómo deben de hacerse estos estudios.

En estos análisis es fundamental realizar test de hipótesis asociados a la independencia entre las variables objeto de estudio, o asociados a la normalidad de sus distribuciones, o de los resultados obtenidos por el algoritmo de clasificación, modelado, o en general, de optimización. Pero estos contrastes no son fáciles de mantener a lo largo del proceso computacional y mucho menos probar. Es por ello necesario plantear alguna metodología de diseño de experimentos y de contraste de hipótesis sobre la comparación de las distribuciones de los resultados.

2 Análisis Estadístico de Algoritmos

La literatura de Investigación Operativa dedicada a la metodología de comparación de algoritmos es difusa y no está ampliamente distribuida. Los trabajos que en la actualidad comparan algoritmos son de una calidad muy variable. Revistas que aceptan artículos en computación, tales como Operational Research, INFORMS Journal on Cumputing, Mathematical Software, Journal on Heuristics y Data Mining and Knowledge Discovery, han publicado trabajos estándar y guías sobre tests en computación, [3] [21] [17], [9]. Estos trabajos se centran en lo que hay que medir y comparar en un artículo que describe investigación computacional. En ellos se muestran criterios de comparación y evaluación de medidas acordes con el criterio planteado, pero no aclaran cuál es el un análisis más apropiado de dichos diseños.

Una revisión de los primeros trabajos que muestran resultados de test computacionales se puede ver en Jackson and Mulvey [10]. Una excelente revisión de simulación y análisis estadístico de algoritmos ha sido hecha por McGeoch [13], que aparece como un artículo característico en INFORMS Journal on Computing. McGeoch incluye un conjunto comprensivo de referencias básicas sobre las técnicas estadísticas más apropiadas además de dar una visión general de cómo diseñar y conducir los experimentos.

Existen trabajos con directrices para implementar diseño de experimentos y análisis de la varianza [2] [12] y estudios metodológicos donde se discuten test de

hipótesis libres de distribución (esto es las variables aleatorias subyacentes no tienen una distribución conocida) [23] [6]. Por último citamos algunos libros sobre técnicas de validación cruzada, remuestreo y diseño de experimentos muy utilizados en computación [7][4].

3 Análisis basados en la media versus en la mediana

La literatura estadística contiene muchos métodos tradicionales de contraste de hipótesis, de determinación de modelos lineales y no lineales, de construcción de intervalos de confianza etc basados en la decisión de tomar como estimador de localización o de centralización la media de la población. Frente a estos métodos, en el caso de que las distribuciones de las variables aleatorias sean significativamente no normales o contengan valores claramente espúreos, es conveniente utilizar como estimador del parámetro de localización la mediana de la distribución dando lugar a métodos robustos basados en la mediana, por ejemplo el test de Friedman.

4 Diseño de experimentos

Antes de aplicar los algoritmos de minería de datos sobre la información disponible, es necesario llevar a cabo una recopilación de la información generada y un preprocesado de ésta, que la ponga en un formato apto para su utilización. Se ha desarrollado un curso de Sistema Operativo Linux sobre un sistema adaptativo para la educación basada en web [15]. Este curso ha sido realizado por 50 alumnos de primer año de ciclo formativo de grado superior en Informática de Sistema, pertenecientes al I.E.S. “Gran Capitán” de Córdoba, que lo realizaron en horas de prácticas de la asignatura “Sistemas Operativos”. Estos datos se han tenido que preprocesar para adaptarlos a la tarea de descubrimiento de conocimiento que se desea realizar.

A partir de los datos preprocesados de los aciertos y fallos cometidos por los alumnos del curso a las preguntas propuestas en test iniciales y finales para cada tema y en las actividades de cada concepto del curso, se han aplicado algoritmos evolutivos para el descubrimiento de reglas [16]. En concreto el paradigma utilizado es la Programación Genética Basada en Gramáticas, Grammar Based Genetic Programming, GBGP, que se han implementado en Java utilizando la biblioteca de clases Java JCLEC [24]. La implementación que presenta esta biblioteca de clases para el paradigma de la GBGP codifica los árboles sintácticos como vectores de enteros ordenados según el recorrido del árbol en preorden. El valor almacenado en el vector codifica, en forma de campos de bits, el símbolo contenido en el nodo, toda la información necesaria para su manipulación y posterior conversión a una consulta SQL. Esta implementación presenta la ventaja de permitir la reutilización de los individuos generados, reduciendo sensiblemente los requisitos de memoria de la aplicación y, consiguientemente, aumentando su eficiencia. La valoración de individuos consiste en la conversión de la cadena de enteros en una serie de consultas

SQL mediante las cuáles se determinan los valores necesarios para el cálculo de la métrica o métricas empleadas como objetivos a optimizar.

5. Métricas para la valoración y ordenación de las reglas

Las reglas de asociación son patrones evaluables puesto que ofrecen información sobre el tipo de dependencias que existen entre los atributos de una base de datos. Debido a la naturaleza de completitud de los algoritmos para análisis de patrones de tipo regla de asociación, el número de patrones extraídos es a menudo muy grande. De esta forma es necesario ordenar o acotar los patrones descubiertos de acuerdo a algunas medidas de interés. El objetivo de nuestro trabajo es analizar como estas medidas reflejan la noción estadística de correlación lineal y como a través de ellas se pueden ordenar las reglas de asociación de una base de datos pudiendo cuantificar de forma objetiva el interés de las mismas. Todas estas medidas se pueden calcular en base a la tabla de contingencia de la regla que se define como:

	B	B^c	Total
A	$n(A \cap B)=n_{11}$	$n(A \cap B^c)=n_{12}$	$n(A)=n_1$
A^c	$n(A^c \cap B)=n_{21}$	$n(A^c \cap B^c)=n_{22}$	$n(A^c)=n_2$
Total	$n(B)= n_{.1}$	$n(B^c)= n_{.2}$	n

Tabla 1. Tabla de contingencia de la regla ($A \rightarrow B$).

Las medidas asociadas a la regla $A \rightarrow B$ consideradas aquí son:

Soporte (Sop). El soporte o frecuencia definido por Agrawal et al en 1993 [1] indica el porcentaje de instancias que contienen tanto A como B, y se define como $Sop(A \rightarrow B) = P(A \cap B)$ y para la muestra se estima mediante $\frac{n(A \cap B)}{n}$. Indica el porcentaje, en tanto por uno, de instancias a las que se les puede aplicar la regla. Toma valores entre 0 y 1 y es simétrica.

Confianza (Conf). También definida como exactitud o precisión en [1] indica el máximo en tanto por uno de instancias que conteniendo a A contienen también a B o que conteniendo a B contienen a A y se define en la forma $Conf(A \rightarrow B) = \max(P(B/A), P(A/B))$, y se estima a partir de las frecuencias relativas que estiman a los valores de probabilidad de los sucesos. Esta medida por tanto, mide la probabilidad condicionada de los sucesos asociados con una regla particular. Por ejemplo si la regla tiene una confianza c_1, esto significa que el c_1% de todas las transacciones que contienen a A contendrán también a B. Toma valores entre 0 y 1.

Interés (Int). Definida por Silverstein et al. en 1998 [19] y denominada también medida de independencia representa un test para medir la dependencia estadística de la regla y se mide como Int(A $\rightarrow$ B) = $\frac{P(A \cap B)}{P(A)P(B)}$. Es por tanto el cociente entre la distribución de probabilidad conjunta de dos variables con respecto a sus probabilidades esperadas bajo la hipótesis de independencia de ambas variables. De nuevo las probabilidades se estiman a partir de las frecuencias. Esta medida verifica las propiedades de RI, pero su rango no está limitado y por ello no es fácil comparar reglas con esta medida y es difícil definir para ella un umbral. Además el interés es simétrico Int(A $\rightarrow$ B)= Int(B $\rightarrow$ A), por lo que sólo mide el grado de independencia y no la implicación en ambas direcciones. Toma valore entre 0 e ∞.

Medida de interés (MI). Definida por Tan y Kumar [21], medida altamente lineal con respecto al coeficiente de correlación para muchas reglas interesantes. Esta definida a partir de la medida de interés I, y presenta según los autores, una alta correlación estadística en la región de bajo soporte y alto interés, y se define como:

MI(A $\rightarrow$ B) = $\frac{P(A \cap B)}{\sqrt{P(A)P(B)}}$. Es simétrica y toma valores entre 0 e ∞.

Factor de Certeza (FC). Es una medida definida por Shortliffe y Buchanan en 1975 [18] que sirve para representar la incertidumbre en reglas de un sistema experto y que se está aplicando en la actualidad en minería de datos.

FC(A $\rightarrow$ B) = max ($\frac{P(B/A)-P(B)}{1-P(B)}$, $\frac{P(A/B)-P(A)}{1-P(A)}$)

El factor de certeza toma valores entre -1 y 1 y se interpreta como una medida de variación de la probabilidad de que el consecuente B esté en una transacción cuando se consideran las transacciones en las que está A. Es simétrica.

Chicuadrado (χ^2). Es una medida estadística aplicada a reglas en [21] y esta asociada al contraste de independencia de dos variables dicotómicas donde la primera tiene los sucesos A y A^c y la segunda los sucesos B y B^c. Los datos muestrales se estructuran en una tabla de doble entrada que se muestra en la tabla adjunta y el estadístico de contraste es el estadístico P de Pearson cuya distribución asintótica es una distribución χ^2 con 1 grado de libertad. Se define en la forma:

χ^2(A$\rightarrow$B)= $\sum_j \sum_k \frac{(n(A_j \cap B_k) - n(A_j)n(B_k))^2}{n(A_j)n(B_k)}$. Cuanto más grande es el valor de la χ^2 más probabilidad existe de rechazar la hipótesis de independencia, pero no nos da la fuerza de la correlación entre el antecedente y el consecuente. Toma valores entre 0 e ∞ y es simétrica.

Entropía (E). Es una medida de asociación derivada de la entropía de Shanon, por Tan y Kumar [21], es una medida de incertidumbre. Esta medida del grado de asociación para una variable y extendida a dos variables es de la forma:

$$H(A)= -\sum_{k=1}^{m} P(A_k)\log P(A_k)\,;\; H(A,B)= -\sum_{k=1}^{m}\sum_{j=1}^{l} P(A_k \cap B_j)\log \frac{P(A_k \cap B_j)}{P(A_k)P(B_j)}$$

La medida completa de asociación entre A y B se puede expresar en términos del cociente

$$S(A\rightarrow B)= \frac{H(A)+H(B)-H(A\cap B)}{\min[H(A),H(B)]}$$

Estimándose de nuevo las probabilidades a partir de las frecuencias maestrales. La información mutua especifica el aumento de reducción en incertidumbre de una variable B cuando se conoce una variable A. Esta medida es simétrica para A y B. Toma valores entre 0 e ∞, y es simétrica.

Precisión Relativa Ponderada *(PRP)*. Esta medida introducida por Lavrac y otros en 1999 [25] y por Pietetsky-Shapiro está relacionada con la generalidad y exactitud de la regla. Se define como: $PRP(A\rightarrow B) = P(A)(P(B/A)-P(B))$ o también $PRP(A\rightarrow B) = P(A\cap B)- P(A)\,P(B)$. Se puede utilizar como una medida filtro y es simétrica para A y B. Es una de las medidas más utilizadas en la evaluación de reglas. Las probabilidades se estiman mediante las frecuencias. Toma valores entre -0.1 y 0.22. En el trabajo de Tan and Kumar [22] se muestra que estas medidas y otras análogas como Laplace, Ganancia de Entropía etc. son muy similares en naturaleza al coeficiente de correlación lineal $\varnothing(A\rightarrow B)= \frac{Cov(A,B)}{\sqrt{V(A)}\sqrt{V(B)}}$, coeficiente simétrico, adimensional y que toma valores entre -1 y 1, cuyo estimador es el coeficiente de regresión lineal $r(A\rightarrow B)= \frac{S_{A,B}}{S_A S_B}$. Este valor para variables dicotómicas tiene un estimador muestral de $\phi(A\rightarrow B)= \frac{n_{11}n_{22}-n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}$.

Esta afirmación no está soportada por ningún estudio estadístico, y es la que queremos contrastar en términos de test estadísticos de igualdad de medias o medianas poblacionales, dependiendo de test previos de normalidad de las ocho variables objeto de estudio.

5. Test de comparaciones estadísticas de las aptitudes de reglas en función de las diferentes métricas utilizadas

En esta sección consideraremos los elementos necesarios para plantear test de hipótesis parametricos y no parametricos para comparar las distribuciones de aptitud

de las reglas de asociación en función de las métricas consideradas. Estas aptitudes se obtienen para poder extraer reglas de predicción mediante algoritmos evolutivos, en este ejemplo mediante Programación Genética Basada en Gramáticas (PGBG). La valoración de los individuos consiste en obtener una regla correcta a partir del árbol de derivación que contiene y a continuación aplicar una función que produzca una medida de calidad de la regla. En [16] se consideran diferentes funciones de agregación o de algoritmos multiobjetivo [20].

5.1 Diseño experimental

El método de obtención de información consiste en aplicar las ocho medidas a una base de datos de 265 reglas descrita en [16]. Estas medidas usando tres cifras decimales significativas se muestran en la Tabla 2 para las tres primeras reglas, que exponemos a continuación:

Acierto,Testf_Unix-Media(5)=NO $\rightarrow$ *Nivel,Automatizar_Unix-Alta = EXPERTO*

Tiempo,Testf_Unix-Baja=ALTO $\rightarrow$ *Tiempo,Demonios_Unix-Alta(2) = ALTO*

Acierto,Testf_Unix-Baja(2)=SI $\rightarrow$ *Acierto,Testf_Unix-Baja(0)= NO*

Sop	**Conf**	**Int**	**FC**	χ^2	**IS**	**E**	**PRP**	ϕ
0.370	1.000	1.227	1.000	41.727	0.674	2.020	0.069	0.366
0.259	0.540	1.211	0.169	16.154	0.560	0.984	0.045	0.182
0.296	0.800	1.964	0.662	19.236	0.763	0.718	0.145	0.613

Tabla 2. Tabla de contingencia de la regla (A $\rightarrow$ B).

5.2 Análisis de normalidad de los datos de aptitud

Para poder realizar contrastes de hipótesis acerca de si las medias de las distribuciones son iguales o no es necesario hacer previamente un test de normalidad de los valores de aptitud de las reglas para las ocho medidas propuestas. El test no-paramétrico de Kolmogorov-Smirnov (K-S) cuyos resultados se muestran en la Tabla 3, indica que para todas las medidas excepto para IS se rechaza la hipótesis nula de normalidad para un α= 0.05, puesto que los niveles críticos, o valores p, son respectivamente 0.000 o 0.010 a excepción de IS cuyo valor es 0.080.

Métrica	**Sop**	**Conf**	**Int**	**FC**	χ^2	**IS**	**E**	**PRP**	ϕ
Z K-S	2.58	2.37	2.57	2.27	2.84	1.26	3.82	2.12	1.62
p	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.01

Tabla 3. Test de Kolmogorov-Smirnov (K-S).

En la segunda fila de la tabla situamos el valor del estadístico Z de K-S; mientras que en la tercera mostramos los valores de p. Con estos resultados el test de comparaciones más adecuado es el de igualdad de medianas de valores de aptitud dados por las ocho medidas para las 265 reglas propuestas; por lo que hacemos un test no-parametrico de Friedman.

5.3 Test de Friedman

El estadístico F de Friedman es de la forma:

$$F= \frac{12}{nk(k+1)}\left[\left(\sum_{i=1}^{k} R_i^2\right)-3n(k+1)\right]= \frac{12S}{nk(k+1)}$$

Siendo n el tamaño muestral, 265 en nuestro caso, k el número de poblaciones a comparar, 9 en nuestro caso, R_i la suma de los rangos de todos los individuos de la población i-ésima y siendo:

$$S= \sum_{i=1}^{k}(R_i-\hat{R}_i)^2 = \sum_{i=1}^{k}(R_i-\frac{n(k+1)}{2})^2.$$

Cuando el tamaño muestral es suficientemente grande, como es nuestro caso n= 265, se demuestra que la distribución del estadístico F converge en distribución a una distribución χ^2 de Pearson con k-1 grados de libertad, 8 en nuestro caso. A partir de los valores de los rangos promedio podemos obtener los valores de rango total R_i multiplicando dichos valores por 265.

Con los resultados anteriores C_0 = (0; $\chi_8^2(0.05)$), donde $\chi_8^2(0.05)$ = se obtiene a partir de la tabla de la $\chi^2_{(8)}$ y por tanto F= 1613.47ɾ C_0, pues 1613.47> 14.06. Se rechaza la hipótesis nula de igualdad de medianas en los valores de aptitud para las 8 métricas propuestas, para un nivel de confianza del 5%. Estos resultados corroboran los obtenidos mediante el software SPSS, donde la significación asintótica es 0.000 y para un nivel de significación α= 0.05 al ser mayor que la Sig. asintótica = 0.000. Se rechaza la hipótesis nula, por lo que los valores de aptitud mediana difieren significativamente para al menos una de las nueve medidas.

Para analizar cual de las aptitudes medianas difiere de las demás seria conveniente realizar test no parametricos de comparaciones múltiples de medianas, no existentes en nuestro conocimiento y por ello deberían de realizar test de Wilcoxon de pares de variables dependientes, puesto que ya hemos visto que existen relaciones de dependencia lineal entre las 8 métricas. La cuestión es que habría que realizar 36 contrastes. Presentamos a continuación alguno de ellos.

5.4 Test de Wilcoxon

En concreto utilizaremos la diferencia de medianas M de las aptitudes proporcionadas por cada una de las dos métricas como parámetro de localización dado que las distribuciones de las variables X e Y son desconocidas y las hipótesis de normalidad no son apropiadas. El contraste bilateral es $H_0: M_X - M_Y = 0$. El estadístico de contraste se construye a través de dos variables auxiliares, transformaciones de X e Y. Z= |X-Y| y S= sig.(X-Y), de forma tal que los valores muestrales de las citadas transformaciones z_i y s_i son los que utilizaremos.

La salida del software SPSS de la Tabla 4 muestra las comparaciones de las medianas de la métrica Soporte con todas las demás métricas (Confianza, Interés, Factor de Certeza y Chi Cuadrado, etc.) donde se observa que existen diferencias significativas entre cada par de medianas para α= 0.05, dado que el nivel crítico es 0.000.

		Conf	Int	FC	χ^2	λ	S	PRP
Sop	W^-	-14.45	-14.45	-6.87	-14.53	-14.54	-12.85	-14.53
	p	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Conf	W^-		-2.19	-14.34	-14.53	-14.35	-2.35	-14.53
	p		0.02	0.00	0.00	0.00	0.00	0.00
Int	W^-			-14.27	-14.53	-14.53	-3.08	-14.53
	p			0.00	0.00	0.00	0.00	0.00
FC	W^-				-5.58	-14.34	-12.42	-5.65
	p				0.00	0.00	0.00	0.00
χ^2	W^-					-14.53	-14.52	-4.91
	p					0.00	0.00	0.00
λ	W^-						-14.23	-14.53
	p						0.00	0.00
S	W^-							-14.52
	p							0.00

Tabla 4. Salida del SPSS para el test de Wilcoxon.

De esta forma podemos concluir que la distribución de las medidas de las reglas obtenida por una métrica cualquiera es diferente de las distribuciones de las medidas de las reglas para las otras siete métricas para cualquier valor de α.

6. Método de Análisis de Componentes Principales

El análisis en Componentes Principales, CP, es una técnica de análisis multivariante que consiste en tratar de reducir la matriz de datos inicial de un conjunto de variables o características que identifican a los elementos de la población objeto de análisis. El análisis CP esta relacionado con la identificación de la

estructura dentro de un conjunto de variables observadas. Establece dimensiones dentro del conjunto de datos y sirve como una técnica de reducción de variables. Para aplicar esta metodología es necesario que exista un cierto grado de multicolinealidad entre las variables para que sea conveniente realizar combinaciones lineales de ellas y que estas combinaciones lineales sustituyan a las variables originales.

Una forma de determinar la conveniencia del análisis factorial es considerar la matriz de correlación R, en su conjunto. El contraste de esfericidad de Bartlett determina la presencia de correlaciones lineales entre las variables originales. Este contraste mide la probabilidad de que la matriz de correlaciones de las variables sea una matriz identidad, esto es, que las variables asociadas a toda la población están incorreladas linealmente y que las correlaciones no difieren significativamente de cero. En nuestro caso el nivel crítico p=0.000 muestra que se rechaza la hipótesis nula por lo que existen correlaciones significativas entre las nueve medidas propuestas y es interesante realizar un análisis factorial.

Otra decisión no menos importante a tomar es el número de factores o componentes a extraer, existen varias metodologías de extracción aunque una de las utilizadas expresa que los factores incluidos deben de explicar tanta varianza como la variable promedio. En nuestro caso como tenemos 9 variables, entendemos que debería de explicar al menos 1/9%= 0.11% de varianza. Siguiendo este criterio debemos de elegir dos componentes principales, puesto que la primera explica un 56.1% de la varianza total de las mueve medidas, la segunda componente principal explica un 32.3%, y ya la tercera, tan sólo un 5.6%. De esta forma elegimos dos componentes principales que explican entre las dos el 88.4% de la varianza total.

Otros elementos a tener en cuenta en el análisis en CP son los valores de las saturaciones de los dos factores “sin rotar” y “rotados”, estas saturaciones recogen el grado de correlación entre los factores seleccionados y las medidas consideradas. La rotación de los factores sirve para poder, en su caso, interpretar mejor el significado de los factores, puesto que algunas de estas saturaciones se acercan a valores de +1 o –1, y las demás se acercan a 0. En nuestro caso hemos elegido una rotación Varimax por Kaiser. El método consiste en aumentar la varianza de cada factor consiguiendo que algunos números-peso tiendan a acercarse a 1 mientras que otros tiendan a hacerse 0, con lo que obtendremos una pertenencia más clara de cada variable a ese factor.

Estas saturaciones pesos o cargas factoriales nos indican el grado de correlación entre la variable y la componente correspondiente. Elevando al cuadrado el peso factorial obtenemos la proporción de varianza compartida por la variable y la componente, por lo que valores de peso inferiores a 0.3 no comparten ni un 10% de varianza, por lo que puede no considerarse esta variable como elemento de la componente. Los valores de las componentes de la matriz rotados y sin rotar se muestran en la tabla 5.

	Componentes sin rotar		Componentes rotadas	
medidas	1	2	1	2
Sop	0.654	0.619	0.313	0.844
Conf	0.712	0.499	0.418	0.762
Int	0.835	-0.479	0.961	-6.07e-02
FC	0.897	-0.132	0.863	0.278
χ^2	0.382	0.886	-4.9e-02	0.964
MI	0.938	0.196	0.755	0.590
E	-5.8e-03	0.918	-0.411	0.820
PRP	0.889	-0.431	0.988	5.98e-03
ϕ	0.892	-0.419	0.986	1.80e-02

Tabla 5. Componentes de la matriz rotados y sin rotar.

Las componentes rotadas no aportan en este caso una mejor interpretabilidad de los factores por lo que elegimos las componentes sin rotar. De esta forma la componente principal primera está formada por las medidas de Confianza, Interés, Factor de Certeza, Precisión Relativa Ponderada, Coeficiente de correlación lineal asi como Soporte y Medida de Interés y explica el 56.1% fe la varianza total. Las cinco primeras son medidas de la exactitud o precisión de la regla, como de exacta es la regla, desde un punto de vista de clasificación sería el porcentaje de clasificación correcta, mientras que las dos últimas son medidas del interés de las reglas, en el sentido de porcentaje de posible aplicación de la regla sobre los datos. La componente principal segunda esta asociada a las medidas Chi-cuadrado y Entropia y explica el 32.3% de la varianza total. Ambas son medidas de dependencia estadística que indica el mayor o menor grado de independencia de las condiciones que forman la regla.

Concluimos que es posible construir nuevas medidas de calidad de reglas de aprendizaje mediante combinación lineal de otras que están correladas linealmente.

Agradecimientos

Este trabajo ha sido financiado por el MCYT a través del proyecto TIC2002-04036-C05-02 y de fondos FEDER.

Referencias

1. Agrawal R., Imielinski T., Swami A., “Mining association rules between sets og ítems in large databases”. Conference on Management of Data. Washington, 1993.
2. Amini M. M., Barr R. S., “Network Reoptimization Algorithms: A statistically designed comparison”. INFORMS JJournal on Computing:4, pp. 395-409. 1993.
3. Barr R. S., Golden B. L., Kelly J. P., Resende M. G. C., Stewart W.R., “Designing and Reporting on Computational Experiment with Heuristic Methods”. Journal of Heuristics, 1, pp. 1-32. 1995.
4. Good P. I., Resampling methods. A practical guide to data analysis. Birkhauser. Boston. 1999.

5. Greenberg H. J., “Computational testing :Why, how and how much”. ORSA Journal on Computing 2:1, pp. 94-97. 1990.
6. Hettmansperger T. P., McKean J. W.. Robust Nonparametric Statistical Methods. Kendall´s Library of Statistics 5. John Wiley and Sons. 1998.
7. Hinkelmann K., Kempthorne O. “Design and analysis of experiments”. Vol I. John Wiley & Sons, Inc. N Y. 1994.
8. Hollander M., Wolfe D. A., “Nonparametric statistical methods”. 2nd ed. Wiley, New York, pp. 56-59. 1999.
9. Jackson R. H. F., Boggs P. T., Nash S. G., Powell S., “Guidelines for reporting results of computational experiments: Reports od the ad hoc committee”. Mathematical Programming. 49, pp. 413-425. 1991.
10. Jackson R. H. F., Mulvey J. M., “A critical rewiev of comparisons of mathematical programming algorithms and software”. Journal of research of the national bureau of standards 83:6, pp. 563-584. 1978.
11. Jaroszewiccz S., Simovici D.. “A general measure of rule interestingness”. Conference on Principles and Practice of Knowledge Discovery in Databases. pp. 253-265. 2002.
12. Lim T., Loh W., Shih Y., “A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms”. Kluwer Publisher, pp. 1-27. 2000.
13. McGeoch C., “Towards and Experimental method for algorithms simulation”. INFORMS Journal on Computing 8:1, pp. 1-15. 1996.
14. Montgomery D. C., Design and analysis of experiments, Wiley. New York. 1991.
15. Romero C., De Bra P., Ventura S., De Castro C., “Using Knoledge Level with AHA! For Discovering Interesting Relationship”. World Congress ELEARN. Montreal. 2002.
16. Romero C., Ventura S., de Castro C., De Bra P., “Discovering Prediction Rules in AHA! Courses”. LNCS User Modeling’03. 2003.
17. Salzberg S.L., “On comparing classifiers:A critique of current research and methods”. Data mining and knowledge discovery, 1, pp. 1-12. 1999.
18. Shortliffe E., Buchanan B. “A model of inexact reasoning in medicine”. Mathematical Biosciences, 23, pp. 351-379. 1975
19. Silverstein A., Brin S., Motwani R., “Beyond market baskets: Generalizing association rules to dependence rules”. Data Mining and Knowledge Discovery, 2, pp. 39-68. 1998.
20. Srinivas N., Deb K., “Multiobjetive optimization using nondominated sorting in genetic algorithms”. Tech. Rep. Department of Mechanical Engineering. 1993.
21. Tan P., Kumar V., “Interesting Measures for Association Patterns”. Technical Report TR00-036. Department of Computer Science. University of Minnnesota. 2000.
22. Tan P., Kumar V., Srivastava J., “Selecting the right Interestingness measures for association patterns”. SIGKDD´02 Edmonton, Alberta. 2002.
23. Tukey J. W., Exploratory Data Analysis. Addison-Wiley. Reading. MA. 1977.
24. Ventura S., Ortiz D., Hervás C., “JCLEC: Una biblioteca de clases java para computación evolutiva”. I Congreso Español de Algoritmos Evolutivos y Bioinspirados. pp 23-30. 2001.
25. Zupan B., Lavrac N., Flach P.. “Rule Evaluation Measures: A Unifying View”. Ninth International Workshop on Inductive Logic Programming. pp. 174--185. 1999.