Comparación de medidas de evaluación de reglas de asociación

César Hervás-Martínez, Cristóbal Romero Morales, Sebastián Ventura Soto
Departamento de Informática y Análisis Numérico. Universidad de Córdoba.
chervas@uco.es, cromero@uco.es, 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 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 basada en gramáticas. 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.**

***Palabras clave*- tests de hipótesis, medidas de interés, reglas de asociación, componentes principales.**

I. Introducción

En los últimos años esta apareciendo en la comunidad científica que trabaja en minería de datos una gran dedicación por determinar el interés 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. Así, 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 aseguran cuanto conocimiento se gana en la distribución de probabilidad 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 de valores de soporte que se encuentran de forma habitual en la práctica.

En las publicaciones recientes sobre este tipo de medidas es habitual realizar comparaciones de las propiedades de las medidas estudiadas [11] [20], pero no lo es comparar, si las características de localización y dispersión de los valores que estas medidas asocian a un conjunto de reglas, difieren significativamente. Para poder analizar si sus distribuciones son normales, o si están correladas linealmente, y en este caso reducir el número de medidas o realizar combinaciones lineales de ellas, es necesario realizar contrastes de hipótesis paramétricos y no paramétricos.

En este trabajo consideramos realizar contrastes de hipótesis para determinar si algunas de las distribuciones, de las puntuaciones que estas medidas asocian a las reglas extraídas de una base de datos en entorno educativo, son iguales, o en otro caso, si tienen características de localización iguales, o en el supuesto de que estén correladas linealmente, poder reducir la dimensión del espacio de características mediante un análisis en componentes principales

Un problema común en Investigación Operativa y en Computación Evolutiva es la comparación de diferentes algoritmos o heurísticas para determinar cuál es el que funciona mejor y que factores afectan al rendimiento. Existen básicamente dos aproximaciones para entender el rendimiento de un algoritmo: i) El análisis teórico centrado en probar las características fundamentales del mismo. Análisis asintóticos del caso-peor o del caso-medio (resultados de orden "O") que describan la conducta en el límite del tiempo de ejecución como una función de los parámetros que describen el tamaño del problema. ii) Análisis empíricos que conllevan la implementación del algoritmo en un código de una computadora y evalúan el tiempo de ejecución, el número de operaciones o la calidad de la solución obtenida por las heurísticas sobre determinadas partes del problema.

Estas dos aproximaciones son complementarias. Cada análisis puede dar luz para analizar el otro, y ambos pueden dar luz sobre las características de los algoritmos que pueden ayudar a un decidor a la hora de elegir la mejor herramienta para atacar un problema particular. A menudo los estudios en computación evolutiva se analizan mediante tablas donde se incluye el número de ejecuciones, el número de evaluaciones, el tiempo de cpu, 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.

Hipótesis asociadas a la independencia entre las variables objeto de estudio, o la normalidad de sus distribuciones, o de los resultados obtenidos por el algoritmo de clasificación, modelado, o en general, de optimización, 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 para que las afirmaciones basadas en gráficas, o estadísticos (casi siempre la media muestral) asociados a resultados obtenidos en unas pocas ejecuciones, sean determinantes a la hora de afirmar que un algoritmo es mejor que otro para un determinado problema, porque para una clase de problemas esta afirmación aparece como fantasiosa y carente de toda verosimilitud, sobre todo si tenemos en cuenta el teorema de "not free lunch".

Un algoritmo es una descripción de un procedimiento mecánico para llevar a cabo una tarea computacional. Esta descripción puede ser más o menos detallada, por lo que diferentes niveles de implementación pueden ser apropiados para diferentes clases de análisis. Cuando el nivel de implementación crece muchas más características deben de especificarse con mucho más detalle.

En problemas de optimización utilizando soluciones mediante computador, el término exacto o algoritmo óptimo se refiere al procedimiento que computa una probablemente solución óptima global (el término algoritmo es a menudo sobredimensionado por la palabra algoritmo exacto). Un algoritmo heurístico es un algoritmo que no garantiza una solución óptima. Un algoritmo aproximado tiene un nivel de rendimiento garantizado (en términos de calidad de la solución o de la convergencia), pero una heurística no necesita, en general, tal garantía. Una metaheurística es una forma de trabajo para producir heurísticas, tales como enfriamiento o recocido simulado y búsqueda tabú. Para ello se deben de especificar las características del problema, tales como definición de las soluciones factibles, el entorno de una solución, reglas para la transformación de soluciones, y reglas para el cambio de ciertos parámetros a lo largo del proceso de ejecución.

II. 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] [20] [17] [9]. Estos trabajos se centran en primer lugar, en que es 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 y donde se mencionan criterios de diseño de experimentos, pero no muestran claridad acerca de un análisis 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.

Un número de trabajos tratan de aplicar análisis estadístico para cuantificar la relación entre el rendimiento de los algoritmos y varios factores que describen las características del problema. 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 de distribución libre (esto es, las variables aleatorias subyacentes no tienen una distribución conocida) [22] [6]. Por último citamos algunos libros sobre técnicas de validación cruzada, remuestreo y diseño de experimentos muy utilizados en computación [7][4].

III. 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, la regresión L_1, etc [6].

IV. 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. De esta manera hemos considerado la necesidad de realizar los contrastes de comparaciones utilizando una base de datos de la cual se pudieran extraer un número suficientemente grande de reglas para poder sacar algunas conclusiones acerca de las medidas de interés que habitualmente se aplican a las reglas extraídas. Así, 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 ha implementado en Java utilizando la biblioteca de clases Java para Computación Evolutiva desarrollada por el grupo "Aprendizaje y Redes Neuronales Artificiales" de la Universidad de Córdoba [23]. 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, así como 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. Además, la localización de nodos es muy eficiente, reduciéndose el tiempo de cómputo en las operaciones de cruce y mutación selectivos. 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.

V. MÉTRICAS PARA LA VALORACIÓN Y ORDENACIÓN DE 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 $Conf(A \rightarrow B) = \max(P(B/A), P(A/B)) = \max(\frac{P(A\cap B)}{P(A)}, \frac{P(B\cap A)}{P(B)})$ siempre que $P(A)\neq 0$ y $P(B)\neq 0$, 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 y es simétrica.

Interés (Int). Definida por Silverstein et al. en 1998 [18] y denominada también medida de independencia representa un test para medir la dependencia estadística de la regla y se mide como $Int(A \rightarrow B) = \frac{P(A\cap B)}{P(A)P(B)}$ siempre que P(A) y $P(B)\neq 0$. 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. El rango de valores entre 0 e ∞, al no estar limitado, propicia que no sea 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.

Factor de Certeza (FC). Es una medida definida por Shortliffe and Buchanan en 1975 que sirve para representar la incertidumbre en reglas de un sistema experto y que se está aplicando en la actualidad en minería de datos. $FC(A \rightarrow B) =$

$$\max\left(\frac{P(B/A)-P(B)}{1-P(B)}, \frac{P(A/B)-P(A)}{1-P(A)}\right)$$

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 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:

$$\chi^2(A\rightarrow B)=\sum_j\sum_k \frac{(n(A_j\cap B_k)-n(A_j)n(B_k))^2}{n(A_j)n(B_k)} = \frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}$$

. 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.

Medida de interés (MI). Definida por Tan y Kumar [20], 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:

$$MI(A\rightarrow B)=\frac{P(A\cap B)}{\sqrt{P(A)P(B)}}$$

siempre que P(A) y P(B) sean distintos de cero. Es simétrica y toma valores entre 0 e ∞.

Entropía (S). Es una medida de asociación derivada de la entropía de Shanon por Tan and Kumar [20], 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 et al en 1999 y por Pietetsky-Shapiro está relacionada con la generabilidad 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 en nuestra base de datos entre = -0.1 y 0.22.

Coeficiente de correlación lineal (ϕ). Este coeficiente mide el grado de correlación lineal entre dos variables aleatorias, y en el caso de una regla de asociación se define en la forma

$$\rho(A\rightarrow B)=\frac{Cov(A,B)}{\sqrt{V(A)}\sqrt{V(B)}}$$

. Es un coeficiente simétrico, adimensional y que toma valores entre -1 y 1, cuyo estimador es $r(A\rightarrow B)=\frac{S_{A,B}}{S_A S_B}$ siendo $S_{A,B}$ la covarianza muestral, y S_A y S_B las desviaciones típicas maestrales. Este valor para variables dicotómicas es

$$\phi(A\rightarrow B)=\frac{P(A\cap B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$$

, cuyo estimador muestral es $\frac{n_{11}n_{22}-n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}$.

En los trabajos de Tan and Kumar [21] y Jaroszewiccz and Simovici [11] se muestra que estas medidas y otras análogas como Laplace, Ganancia de Entropía etc. son muy similares en naturaleza al coeficiente de correlación lineal o unas son casos particulares de medidas más generales. Pero experimentalmente estas afirmaciones no están soportadas por ningún estudio estadístico, y es lo que queremos contrastar en términos de test estadísticos de igualdad de medias o medianas poblacionales, dependiendo de test previos de normalidad de las nueve variables objeto de estudio.

VI. TEST DE COMPARACIONES DE ESTADÍSTICOS 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, basada en el interés, 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..

A. Diseño experimental

El método de obtención de información consiste en aplicar las nueve medidas a una base de datos de 265

reglas descrita en Romero **[26]**. Los valores obtenidos con estas medidas usando tres cifras decimales significativas se muestran en la Tabla 2 para las tres primeras reglas:

Acierto,Testf_Unix-Media(5)=NO→
Nivel,Automatizar_Unix-Alta = EXPERTO

Tiempo,Testf_Unix-Baja=ALTO→
Tiempo,Demonios_Unix-Alta(2) = ALTO

Acierto,Testf_Unix-Baja(2)=SI→
Acierto,Testf_Unix-Baja(0)= NO

Sop	Conf	Int	FC	χ^2	MI	S	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. Valores de las 9 medidas propuestas

B. Análisis de normalidad de los datos de aptitud

Para poder realizar contrastes de hipótesis acerca de si los parámetros de localización 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 nueve medidas propuestas. El test no-paramétrico de Kolmogorov-Smirnov (K-S) cuyos resultados se muestran en la Tabla 3, indica que para todas medidas excepto para MI se rechaza la hipótesis nula de normalidad para un α= 0.05, puesto que los niveles críticos, o valores p, son respectivamente 0.00 o 0.01 a excepción de MI cuyo valor es 0.08.

Métrica	Sop	Conf	Int	FC	χ^2	IS	E	PRP	ϕ
Media	0.29	0.61	1.17	0.17	23.03	0.57	1.46	0.03	0.13
Des	0.10	0.16	0.27	0.28	16.96	0.13	0.89	0.06	0.26
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. Estadísticos y Test de Kolmogorov-Smirnov

En las filas segunda y tercera de la tabla 3 mostramos los estadísticos media y desviación típica, mientras que en la cuarta fila situamos el valor del estadístico Z de K-S; mientras que en la quinta 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-paramétrico de Friedman.

C. 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 que se muestran en la tabla IV y

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

La región de aceptación unilateral del contraste es C_0 = (0; F_α), donde F_α se obtiene a partir de unas tablas construidas por Friedman para muestras de tamaño pequeño. La regla de decisión es por tanto que "Si $F \in C_0$ Se acepta la hipótesis nula para un nivel de confianza α, prefijado".

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. Esto es, para k≥ 5 y n≥ 10 el valor del estadistico de contraste es

$$F = \frac{12\times 3793153.4}{265\times 9\times 10} = 1908.5,$$ puesto que en este caso

$$S= \sum_{i=1}^{k} (R_i - \frac{n(k+1)}{2})^2 = \sum_{i=1}^{9} (R_i - 1325)^2$$

S= 3793153.44. Este estadístico converge a una distribución $\chi^2_{(8)}$.

A partir de los valores de los rangos promedio podemos obtener los valores de rango total R_i multiplicando dichos valores por 265. Estos valores se muestran en la tabla 4.

Mét	Sop	Con	Int	FC	χ^2	MI	S	PRP	ϕ
R.	3.48	5.58	7.46	2.62	9.00	5.37	7.51	1.60	2.38
R_i	922.2	147	197	694.3	238	142	199	424	630.7

Tabla 4. Rango promedio y Suma de los rangos de las métricas, R_i, para todas las reglas.

Con los resultados anteriores C_0 = (0; $\chi^2_8(0.05)$), donde $\chi^2_8(0.05)$ = 15.51 se obtiene a partir de la tabla de la $\chi^2_{(8)}$ y por tanto F= 1908.5$\notin$ C_0, pues 1908.5 > 15.51. Se rechaza la hipótesis nula de igualdad de medianas en los valores de aptitud para las 9 métricas propuestas, para un nivel de confianza del 5%. Estos resultados corroboran los obtenidos mediante el software SPSS, donde la significación asintótica es 0.00 y para un nivel de significación α= 0.05 al ser mayor que la Sig. asintótica = 0.00. 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 9 métricas. La cuestión es que habría que realizar 36 contrastes.

D. Test de Wilcoxon

En concreto utilizaremos la mediana M de la diferencia de 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 se plantea en la forma:

$$\text{Hipótesis} \begin{matrix} H_0: M_X - M_Y = 0 \\ H_1: M_X - M_Y \neq 0 \end{matrix} \Big\}$$

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.

En primer lugar se construyen los rangos de los n valores de z_i, de forma tal que r_i= rang.(z_i) y con estos valores se definen dos estadísticos equivalentes.

$$W^+ = \sum_{s_i=1} s_i r_i \quad \text{y } W^- = \sum_{s_i=-1} s_i r_i = -|W^-|$$

La región de aceptación para aceptar la hipótesis nula es C_0= ($W_{1-\alpha/2}$, $W_{\alpha/2}$) y la distribución de W^- para muestras de tamaño mayor de 30, como es nuestro caso, se demuestra que converge a una normal, esto es

$$W^- \xrightarrow{C.L.} N\left(-\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$$

siendo por tanto

$$W_{\alpha/2} = -\frac{n(n+1)}{4} + z_{\alpha/2}\sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$\text{y } W_{1-\alpha/2} = \frac{2W_{\alpha/2}}{n(n+1)}$$

Las salidas de SPSS de la Tabla 5 muestran los valores de W^- y de p de las comparaciones de las medianas de cada métrica con todas las demás métricas, donde se observa que existen diferencias significativas entre cada par individual de medianas para α= 0.05, dado que el nivel crítico es 0.00 o 0.02.

	Conf	Int	FC	χ^2	MI	S	PRP
Sop	-14.12	-14.11	-7.43	-14.11	-14.11	-14.11	-14.12
	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Conf		-14.11	-13.95	-14.11	-5.38	-14.10	-14.11
		0.00	0.00	0.00	0.00	0.00	0.00
Int			-14.16	-14.11	-14.11	-3.03	-14.11
			0.00	0.00	0.00	0.02	0.00
FC				-14.11	-13.77	-14.11	-8.83
				0.00	0.00	0.00	0.00
χ^2					-14.11	-14.11	-14.11
					0.00	0.00	0.00
MI						-14.08	-14.11
						0.00	0.00
S							-14.11
							0.00
ϕ	-14.02	-14.11	-3.35	-14.11	-14.11	-14.11	-7.53
	0.00	0.00	0.001	0.00	0.00	0.00	0.00

Tabla 5. Salida de 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 ocho métricas para cualquier valor de α.

VII. MÉTODO DE ANÁLISIS EN COMPONENTES PRINCIPALES

El análisis factorial y en particular 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 se pueden obviar los supuestos de normalidad, homocedasticidad, y linealidad de las variables originales, aunque 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.

Con el fin de analizar las correlaciones parciales entre variables, el contraste de Kaiser-Meyer-Olkin esta asociado a medir la relación entre las variables a través de sus coeficientes de correlaciones parciales. El índice KMO se obtiene a partir del valor:

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} a_{ij,kl\ldots p}^2}$$

Donde r_{ij} es el coeficiente de correlación simple entre las variables i y j, $a_{ij,\ kl\ldots p}$ es el coeficiente de correlación parcial entre esas mismas variables. Si la suma de los coeficientes de correlación parciales al cuadrado entre todos los pares de variables es pequeña comparada con la suma de los coeficientes de correlación al cuadrado, la medida KMO se aproxima a la unidad. Por ello, valores pequeños de KMO cuestionan la aplicación del análisis factorial, dado que, en ese caso, las correlaciones entre pares de variables no pueden explicarse por otras variables. Este índice toma valores entre 0 y 1, tomando un valor de 1 cuando cada variable se puede predecir sin error a partir de las demás; valores inferiores a 0.5 hacen el análisis factorial "inaceptable" y valores de 0.9 indican una medida "maravillosa". En nuestro caso obtenemos un valor de 0.629 por lo que el análisis en CP es adecuado. Otra forma de determinar la conveniencia del citado análisis 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=0.00 muestra que se

rechaza la hipótesis nula por lo que existen correlaciones significativas entre las nueve medidas propuestas y es interesante realizar un análisis en CP.

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. Se basa en la idea de que si un factor es significativo, en cuanto a explicar la correlación lineal entre variables, entonces debería absorber por lo menos tanta varianza como una variable promedio de las de partida, 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 nueve 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, esto es, proporciona una mayor capacidad explicativa a los factores y un mejor panorama de interpretación.

Si cualquier variable tiene una carga alta (correlación alta) sólo en una componente, es fácil dar a cada componente una interpretación que surgirá de las variables con cargas más altas con respecto a la CP. 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 adjunta.

	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
S	-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 6. 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, así como Soporte y Medida de Interés y explica el 56.1% de 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 Entropía 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.

Una vez que hemos conseguido reducir el número de medidas, (en nuestro caso de nueve pasamos a dos), quizás nos interese conocer la puntuación que tendrá cada regla en los factores o componentes, puesto que estas nos permiten saber en qué medida los factores se dan en los individuos, por lo que es necesario calcular las puntuaciones factoriales. Para calcular estas puntuaciones es necesario utilizar los coeficientes obtenidos en la matriz de componentes sin rotar y multiplicarlos por el valor de la medida tipificada correspondiente. Esto es, para obtener la puntuación de la regla i-ésima de la base de datos de alumnos con las dos nuevas medidas asociadas a las componentes principales tenemos:

Puntua en 1ª CPi = $0.654\times Zi_{SOP}+0.712\times Zi_{CONF}+ ...+ 0.889\times Zi_{PRP}+ 0.892\times Zi\phi$
Puntua en 2ª CPi= $0.619 \times Zi_{SOP}+ 0.449\times Zi_{CONF} + ...- 0.431\times Zi_{PRP}- 0.419\times Zi\phi$

VIII. Conclusiones y Trabajo Futuro

En este trabajo se ha presentado una metodología de comparación de medidas del grado de asociación en reglas extraídas mediante GBGP en una base de datos en entorno educativo. De estas comparaciones se desprende que las distribuciones de las medidas no son normales salvo para MI y que al aplicarles los contrastes de igualdad de medianas se observa que estas son diferentes entre si para $\alpha= 0.05$ y que si sólo consideramos contrastes dos a dos se pueden considerar también diferentes, por lo que en general no toman valores igualmente distribuidos. Por otra parte existe entre ellas correlación lineal y un análisis en CP muestra que es posible situarlas en dos CP que extraen el 90% de

la varianza total aportada por las 9 métricas, esto nos lleva a concluir que algunas medidas miden características similares de las reglas y que se pueden definir otras métricas como combinación lineal de varias de las iniciales. Como futuro trabajo se va a investigar sobre la utilización de un algoritmo evolutivo multiobjetivo basado en el Frente Pareto, pero que en lugar de utilizar en el vector de objetivos directamente una serie de medidas seleccionadas, cada una para medir una cualidad deseada, se va a utilizar estas dos componentes principales.

Agradecimientos

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

IX. 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 JJJournal 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 on 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 Knowledge 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] 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.

[19] Srinivas N., Deb K., "Multiobjetive optimization using nondominated sorting in genetic algorithms". Tech. Rep. Department of Mechanical Engineering. 1993.

[20] Tan P., Kumar V., "Interesting Measures for Association Patterns". Technical Report TR00-036. Department of Computer Science. University of Minnnesota. 2000.

[21] Tan P., Kumar V., Srivastava J., "Selecting the right Interestingness measures for association patterns". SIGKDD´02 Edmonton, Alberta. 2002.

[22] Tukey J. W., Exploratory Data Analysis. Addison-Wiley. Reading. MA. 1977.

[23] 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.

[24] Whigham P.A., "Gramatically-based Genetic Programing". Proceeding of the Workshop on Genetic Programming. pp. 33-41. 1995.

[25] Zytkow J., Klosgen W., Handbook of Data Mining and Knowledg Discovery. Oxford University Press. 2001.