Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 9, Mes Septiembre, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 179-187
http://publicaciones.uci.cu

Tipo de artículo: Artículo de revisión
Temática: Ingeniería y gestión de software
Recibido: 10/03/2020 | Aceptado: 04/08/2020 | Publicado: 01/09/2020

Extensión de índices de validación de grupo en la herramienta WEKA para la evaluación de algoritmos de agrupamiento

Extension of group validation indices in the WEKA tool for the evaluation of clustering algorithms

Yohandra Echeverria Castillo [1*], Barbara Laborí de la Nuez [2], Roberto Soriano Sifonte [3]

[1] Departamento Sistemas Digitales, Facultad 2, Universidad de Ciencias Informática, Carretera a San Antonio de los Baños, Km 2 ½ La Lisa, La Habana, Cuba. yoha@uci.cu
[2] Departamento Sistemas Digitales, Facultad 2, Universidad de Ciencias Informática, Carretera a San Antonio de los Baños, Km 2 ½ La Lisa, La Habana, Cuba. barbaral@uci.cu
[3] Centro CIGED, Facultad 2, Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, Km 2 ½ La Lisa, La Habana, Cuba. rsoriano@uci.cu

[*] Autor para correspondencia: yoha@uci.cu

Resumen
En el ámbito del creciente desarrollo de los algoritmos de agrupamiento en innumerables áreas de la sociedad, y debido a la imprecisión de la herramienta WEKA para la validación de la calidad de resultados de estos algoritmos, surge el Trabajo de Diploma: "Extensión de índices de validación de grupo en la herramienta WEKA para la evaluación de algoritmos de agrupamiento". Esta investigación tiene como objetivo medir la calidad de las particiones resultantes de los algoritmos de agrupamiento y así contribuir en el diseño de experimentos, para recopilar información sobre diferentes tipos de datos. Es por ello que se integra un conjunto de métricas o índices de validación externas e internas a la herramienta. La solución ofrecerá que la herramienta WEKA permita la validación de calidad para los algoritmos de agrupamiento, lo que permitirá llevar acabo comparación entre algoritmos.

Palabras clave: agrupamiento, validación, calidad, índice de validación de grupo, validación interna, validación externa, WEKA.

Abstract
In the space of increasing development of the clustering's algorithms in innumerable areas of the society, and due to the imprecision of the tool WEKA for the validation of quality of aftermath of these algorithms, Diploma's Work appears: "Extending validity index WEKA cluster in the tool for evaluating clustering algorithms". This research aims to measure the quality of partitions resulting from the clustering algorithms and thus contribute to the design of experiments, to collect information about different types of data. That is why a set of metrics or indices of external

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 9, Mes Septiembre, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 179-187
http://publicaciones.uci.cu

and internal validation tool is integrated. The solution will allow the validation tool WEKA quality for clustering algorithms, which will carry out comparison between algorithms. .

Keywords: *clustering, validation, quality, cluster validity index, internal validity, external validity, WEKA.*

Introducción

El aprendizaje automático es una técnica que utiliza la minería de datos en su proceso de conversión de datos en conocimiento, Para agilizar el mismo se maneja la extracción de modelos, utilizando los árboles de clasificación como herramientas para eliminar los resultados innecesarios. Esto lo convierte en un motor de consultas que permite realizar ordenamientos y selección de datos. Entre los diferentes algoritmos de aprendizaje automático, se encuentran los de clasificación supervisada y los de clasificación no supervisada (Caparrini, 2013).

Entre las tareas más utilizadas del aprendizaje supervisado se encuentra la predicción y la clasificación, mientras que en el aprendizaje no supervisado están agrupamiento y asociación (Weiss, 1998). La clasificación supervisada es conocida también como clasificación con aprendizaje, se conoce a priori la clase a la que pertenecen cada uno de los objetos. La clasificación no supervisada es conocida como clasificación sin aprendizaje, no se toma en cuenta la información de las clases, debido a que aprende a partir de la naturaleza intrínseca de los datos.

Entre los algoritmos que utiliza la clasificación no supervisada se encuentran: Los algoritmos de agrupamiento. Estos algoritmos tienen como objetivo aglomerar un conjunto de objetos, en dependencia de la naturaleza de los rasgos que caracterizan dichos objetos, basándose en la similitud entre estos. Para medir la similitud entre objetos se utilizan diferentes funciones de distancia: distancia Euclídea, de Manhattan, de Mahalanobis, etc.

Existen varios enfoques de algoritmos de agrupamiento, debido a que, para un mismo conjunto de datos, aplicando diferentes algoritmos de agrupamiento se pueden obtener resultados diferentes. Es por esto que surge la necesidad de evaluar las particiones obtenidas y poder determinar la calidad de los resultados alcanzados. Existen medidas que permiten realizar una evaluación de la estructura resultante de cada algoritmo de agrupamiento, obteniendo de manera cuantitativa un valor de calidad de las mismas. Estas medidas de calidad se conocen bajo el nombre de índices de validación de grupo.

Un índice de validación de grupo es una función, que mide la bondad o calidad de una partición. Se define que a menor valor de la función mejor será la partición. Un índice de validación proporciona una medida objetivo del resultado de un agrupamiento, y su valor óptimo, se usa frecuentemente para indicar la mejor selección posible ya que permiten:

- Cuantificar a través de una medida la calidad del agrupamiento para una determinada base de datos.

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 9, Mes Septiembre, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 179-187
http://publicaciones.uci.cu

- Determinar una configuración adecuada de los parámetros de entrada para cierto algoritmo de agrupamiento. (Ej. Número óptimo de grupos para un conjunto de datos)

La validación de grupo se clasifica en tres categorías: externa, interna y relativa. La investigación se centrará en las categorías externa e interna, debido a que la categoría relativa algunos autores la tratan como un caso particular de la validación interna.

Existen un conjunto de herramientas con técnicas y algoritmos de aprendizaje automático, que su uso se ha extendido por la comunidad internacional. Se destacan las herramientas Matlab y R, las cuales contienen índices de validación de agrupamiento. Otra de las herramientas más utilizada en el aprendizaje automático es WEKA. Desarrollada en el departamento de Ciencias de la Computación de la Universidad de Waikato en Nueva Zelanda. Es una herramienta de código abierto, funciona en plataforma Windows y Linux, contiene una gran colección de algoritmos de aprendizaje automático, implementado en Java. Entre los principales tipos de problemas de aprendizaje que pueda hacer frente se encuentran la clasificación, regresión, agrupamiento y existe cierto apoyo a la minería de reglas de asociación (Jímenez, 2003). Esta herramienta crea un entorno favorable para el diseño de experimentos y la comparación de algoritmos para el desarrollo de investigaciones científicas.

En cuanto a los algoritmos de aprendizaje automático que soporta WEKA esencialmente cuenta con un conjunto de herramientas de pre-procesamiento de datos, alrededor de 76 algoritmos de clasificación/regresión para resolver problemas de clasificación supervisadas, 8 algoritmos de agrupamiento y 3 algoritmos de reglas de asociación para problemas no supervisados. Este balance entre los algoritmos supervisados y no supervisados en WEKA, denota que el mayor esfuerzo en el desarrollo de esta herramienta ha estado dirigido en los algoritmos de clasificación supervisados. Esta herramienta no tiene una implementación de los diferentes índices de validación de agrupamiento, que pueden ser utilizados para evaluar cuál de los algoritmos de agrupamiento funciona mejor sobre un conjunto de datos.

Metodología

Para la determinación del problema real de la investigación se utilizó como caso de estudio el algoritmo SimpleKMeans, realizando experimentos a diferentes bases de datos. El histórico lógico se refleja a partir del estudio de la evolución de los métodos de validación de grupo, para así lograr un mejor entendimiento de estos enfoques y trabajar en su mejoramiento o actualización. El analítico sintético es reflejado a través del análisis de la bibliografía disponible para realizar la investigación, y a su vez realizar una extracción de las características y elementos más

importantes sobre los métodos de validación de grupo. Este método permite definir los principales conceptos, definiciones y otras soluciones ya existentes.

La evaluación estadística de resultados experimentales, es una parte esencial para la validación de los métodos de aprendizaje automático. Para el diseño de la propuesta presentada se tuvo en cuanta lo expresado por (Vázquez, 2001) y (Pizarro, 2002) pues los autores consideran que las técnicas utilizadas son las más idóneas para la comparación entre múltiples modelos en solo una colección de datos. De ahí que se utilice el test de Friedman para determinar si existen diferencias entre los algoritmos propuestos.

En el caso de la validación externa se identifican cuatro índices fundamentales, como son entropía, pureza, Rand y Jaccard coefficient, a partir de las métricas analizadas. Por otra parte, en la validación interna se utiliza la distancia euclidiana para medir la similitud entre objetos, teniendo en cuenta los siguientes índices: Calinski-Harabasz, C-Index, Dunn, Davies Bouldin y Xie-Beni para la implementación de los índices internos. Se propone un paquete de métricas, utilizando como medidas de validación los índices internos y externos expresados anteriormente. Este paquete se integrará a la herramienta WEKA, debido a que es una herramienta de experimentación y de fácil integración con aplicaciones empresariales.

Resultados y discusión

Para evaluar la partición obtenida como resultado de un algoritmo de agrupamiento se tiene en cuenta los métodos de validación externa e interna. Estos métodos conformarán un paquete de métricas con el objetivo de validar a través de la herramienta WEKA los algoritmos de agrupamiento. Esta herramienta servirá de apoyo para poder determinar si el algoritmo de agrupamiento realizado a un conjunto de datos es apropiado.

Índices de validación externa

La validación externa se basa en determinar cuán cercanas se encuentra la partición obtenida producto de un algoritmo de agrupamiento y la partición ideal. Como parte de la solución para dar respuesta a un problema de agrupamiento, se proponen los índices basados en teoría de la información y los índices basados en el recuento de pares, para la validación de grupo en la implementación de este enfoque.

Índices de validación interna

Los índices de validación internos no utilizan la información de la clase para evaluar la calidad del agrupamiento, dado que se concentran en el trabajo sobre una sola estructura de agrupamiento. Estos se dividen en dos grandes grupos los basados en suma y los basados en radio (centroide) como se mencionaba anteriormente. Se propone para la implementación de este enfoque a los índices: Davies-Bouldin, Dunn, Calinski-Harabasz, Xie-Beni y C-Index

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 9, Mes Septiembre, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 179-187
http://publicaciones.uci.cu

(Calinski, 1974), (Bouldin, 1979), (Rousseeuw, 1987), (Goikoetxea, 2010). La anexión de nuevos índices de validación de grupos resulta muy sencilla, pues existen funcionalidades en la implementación que se han definido, que pueden ser reutilizadas para la construcción de los mismos. A continuación, se mencionan estas funcionalidades:

Tabla 1 Funcionalidades definidas para la implementación de nuevos índices (Elaboración propia)

Funcionalidades	**Expresión**
Mínima distancia entre dos objetos de un mismo grupo	$d = min_x i,xj \in Cd(x_i, x_j)$
Mínima distancia entre dos objetos de grupos diferentes	$d = min_x i \in Ci,xj \in Ckd(x_i, x_j)$
Máxima distancia entre dos objetos de un mismo grupo	$d = max_x i,xj \in Cd(x_i, x_j)$
Máxima distancia entre dos objetos de grupos diferentes	$d = max_x i \in Ci,xj \in Ckd(x_i, x_j)$
Mínima distancia entre dos centroides	$d = mind(\overline{C_i}, \overline{C_j})$
Máxima distancia entre dos centroides	$d = maxd(\overline{C_i}, \overline{C_j})$
Distancia de un objeto a su centroide	$d = dxi \in Ci(xi, \overline{C})$

Diagrama de paquete

El siguiente diagrama representa el diseño de paquete propuesto para la integración de las métricas de evaluación de los algoritmos de agrupamiento a la herramienta Weka.

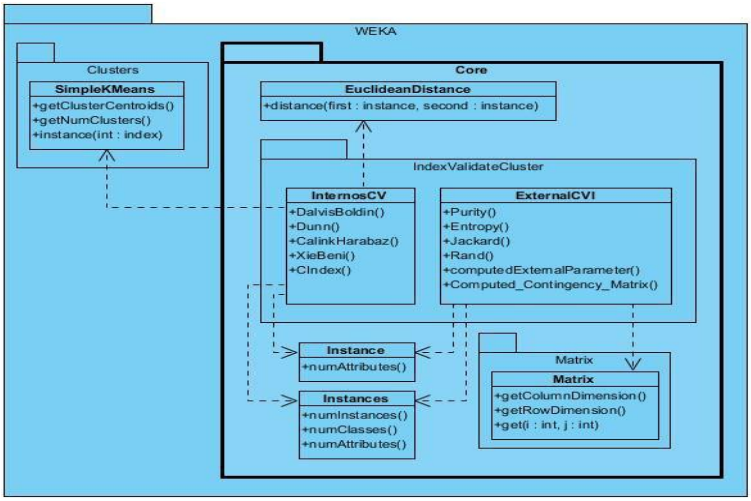

Figura 1 Integración del paquete de metricas a Weka (Elaboración propia)

Se representa el paquete propuesto para dar solución al problema planteado, así como la relación con algunos paquetes de Weka. El paquete Index_Validate_Cluster está compuesto por las clases InternosCV y la clase ExternalCVI, estas contienen los índices de validación internos y externos respectivamente. Mientras que en el

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 9, Mes Septiembre, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 179-187
http://publicaciones.uci.cu

paquete Weka se encuentran los paquetes Core y Clusteres, entre otros. El paquete Core está compuesto por las clases e interfaces que conforman la infraestructura de WEKA, define las estructuras que contienen los datos a manejar por los algoritmos de aprendizaje automático, por lo que la propuesta de solución está integrada a este paquete. Contiene además las clases Instance, Instances, EuclideanDistance y el paquete matrix, entre otros.

La clase Instances y la clase Instance permitieron acceder a los datos de las particiones y la base de datos. La clase Instances contiene las bases de datos junto con los métodos para su manejo. Mientras que la clase Instance encapsula cada uno de los ejemplos individuales que forman una base de datos, almacenando los valores de los respectivos atributos. La clase Matrix del paquete Matrix, se utilizó para la implementación del método Computed_Contingency_Matrix, el cual es esencial en la implementación de los índices externos. En la clase EuclideanDistance se encuentra el método distance, que calcula la distancia euclidiana al cuadrado entre dos casos de un agrupamiento. La clase SimpleKMeans está contenida dentro del paquete clusteres. Esta clase contiene los atributos necesarios para la implementación de los índices internos.

Validación de los índices de validación de agrupamiento

Descripción de las bases de datos

Para validar el funcionamiento de los índices de validación de agrupamiento, se propone la realización de experimentos, utilizando como caso de estudio al algoritmo SimpleKmeans existente en Weka. Se emplearon 13 bases de datos estándares extraídas del UCI Repository (Iris, breast, diabetes, ecoli, banknote, glass, ionosphere, transfusión, vehicle, vertebral_columm, user_modeling, segment, messidor_features). A continuación, se realiza una descripción de estas bases de datos, las cuales se relacionan en la siguiente tabla, donde la primera columna indica el nombre de la base de datos, la segunda se refiere a la cantidad de objetos y la tercera se refiere a la cantidad de rasgos:

Tabla 2. Descripción de las bases de datos utilizadas en los experimentos(descripción propia)

Bases de datos	Cantidad de objetos	Cantidad de rasgos
banknote	1372	5
breast	106	10
diabetes	768	9
ecoli	336	8
glass	214	10
ionosphere	351	35
segment	2002	4
iris	150	5
vehicle	846	19
transfusion	748	5
user_modeling	258	6

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 9, Mes Septiembre, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 179-187
http://publicaciones.uci.cu

vertebral_columm	310	7
messidor_features	1151	20

Test de Friedman.

La prueba Friedman es un equivalente no paramétrico de ANOVA[1] (Friedman, 1937), su objetivo fundamental es determinar si existen o no diferencia entre los algoritmos analizados. Ordena los algoritmos para cada conjunto de datos separadamente, donde el algoritmo con mejor desempeño obtiene la jerarquía de 1, el segundo mejor la jerarquía 2 y así sucesivamente. La prueba Friedman compara las jerarquías comunes entre los algoritmos a través de la siguiente fórmula:

$$R_j = \frac{1}{N}\sum_i r_i^j$$

Ecuación 1 Ecuación para calcular la jerarquía entre los datos

Dónde:

r_i^j : Es el valor del algoritmo j en la base de dato i.

N: El número de bases de datos.

Friedman declara como hipótesis nula que los algoritmos son equivalentes y sus rankings iguales. Si esta hipótesis nula es negada, es necesario realizar una prueba post-hoc (Nemenyi) que se basa en el cálculo de la diferencia crítica, comparando los algoritmos entre sí. A partir de esta comparación establece que los algoritmos son significativamente diferentes, si el ranking resultante de su comparación es mayor o igual a esta diferencia, calculada de la siguiente forma (Demsar, 2006):

$$DC = q_\sigma\sqrt{\frac{K(K+1)}{6N}}$$

Ecuación 2 Ecuación para calcular la diferencia entre jerarquía

Siendo K el número de algoritmos.

Mientras que q_σ es un constante, dada por K y el valor de σ , en la siguiente tabla se muestra esta relación.

Tabla 3 Valores de q_σ

σ \ K	2	3	4	5	6	7	8	9	10
0.05	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
0.10	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

1 Análisis de varianza

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 13, No. 9, Mes Septiembre, 2020
ISSN: 2306-2495 | RNPS: 2343
Pág. 179-187
http://publicaciones.uci.cu

Resultados del Test de Friedman.

En la siguiente tabla se muestra el ranking obtenido por el test de Friedman para K=3.

Tabla 4.Valores obtenidos para los índices de validación externos para K=3(Elaboración propia)

Algoritmo	Ranking
Pureza	2,80
Entropía	3,61
Jaccard	1,07
Rand	2,50

Tabla 5. Valores obtenidos para los índices de validación internos para K=3(Elaboración propia)

Algoritmo	Ranking
CIndex	1,69
Calinski-H	3,23
Davies-B	4,03
Dunn	2,26
Xie-B	3,76
Dunn	2,26
Xie-B	3,76

Al analizar los resultados se concluye que existen diferencias entre los algoritmos. Se determinó que el mejor algoritmo para K = 3, K=5, K=7 y K=9 en los índices externos es Jaccard. Por otra parte, para K = 3, K=5 y K=9 en los índices internos es CIndex, mientras que para K = 7 es Calinski-Harabaz (Ver Anexos). A raíz de que no se cumple la hipótesis nula se hace necesario aplicar la prueba post-hoc de Nemenyi, se calcula el valor de la distancia crítica para los algoritmos analizados anteriormente. El valor obtenido es de 2.35 para $q_{\sigma} = 0.05$ en los índices externos, mientras que para los internos obtiene un valor de 5.47. A continuación se muestran los resultados obtenidos por el test de Friedman de las comparaciones entre los algoritmos de la presente investigación, para los valores de $q_{\sigma} = 0.05$. mientras que para los internos obtiene un valor de 5.47. (Ver anexos)

Como se evidencia en los índices Pureza y Entropía presentan diferencias significativas con el resto de los índices externos. Por otra parte, los índices internos no presentan diferencias significativas entre ellos para K = 3 y $q_{\sigma} = 0.05$, esto sucede también para K = 5 y $q_{\sigma} = 0.05$.

Conclusiones

El estudio de los métodos de validación de grupo permitió identificar dos enfoques: índices de validación externa e índices de validación interna, en el caso de la validación externa se identificaron los índices a partir de las métricas analizadas y en el caso de la validación interna se basa en el cálculo de distancias utilizando la distancia euclidiana.

Con el test de Friedman aplicado se determinó que existen diferencias entre los algoritmos analizados, por lo que se hizo necesario aplicar una prueba post-hoc de Nemenyi, obteniendo como resultado que los índices Pureza y Entropía presentan diferencias significativas respecto a los otros algoritmos externos, mientras que los internos no presentan diferencias significativas para K = 3.

Con la integración del paquete de métricas implementado a la herramienta WEKA permite la evaluación de la calidad de las particiones resultantes de evaluar los algoritmos de agrupamiento.

Referencias

A. González, C. (s.f.). Comparación de dos métodos de aprendizaje sobre el mismo problema.

Bouldin. 1979. A clustering separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979.

Calinski. 1974. A dendrite method for cluster analysis. Communications in Statistics. 1974.

Caparrini, Fernando Sancho. 2013. Introducción al Aprendizaje Automático. s.l. :Dpt. Ciencias de la Computación e Inteligencia Artificial. Univ. de Sevilla , 2013.

Demsar, J. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. Jurnal of Machine Learning Research 7,. 2006. 1-30.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American StatisticalAssociation. 1937.

Jímenez, M. G. 2003. Análisis de Datos en WEKA– Pruebas de Selectividad. 2003.

Pizarro, J., Guerrero, E., & Galindo, P. 2002. Multiple comparison procedures applied to model selection. 2002. Neurocomputing 48 155–173.

Amigo, Enrique. 2009. A comparison of ExtrinsicClusteringEvaluation. Madrid, Spain: UNED.: Departamento de Lenguajes y Sistemas Informáticos, 2009.

Rand. 1971. Objective criteria for the evaluation of clustering methods. s.l.: Journal of American Statistical Association, 1971. 66:846-850.

Goikoetxea, IbaiGurrutxaga. 2010. Aportaciones a la clasicación no supervisada y a su validación. Aplicación a la seguridadinformática. Donostia :s.n., 2010.

Rousseeuw. 1987. Silhouttes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. s.l.: Journal of Computational and Applied Mathematics, 1987. 20, 53 – 65.