Aproximación al índice externo de validación de clustering basado en chi cuadrado

José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros and José C. Riquelme Santos
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Sevilla, España

***Abstract*—El clustering es una de las técnicas más utilizadas en minería de datos. Tiene como objetivo principal agrupar datos en clusters de manera que los objetos que pertenecen al mismo clúster sean más similares que los que pertenecen a diferentes clusters. La validación de un clustering es una tarea que se realiza aplicando los llamados índices de validación. Estos índices miden la calidad de la solución del clustering y se podrían clasificar como índices internos, los cuales calculan la calidad del clustering en función de los propios clusters; e índices externos, que miden la calidad mediante información externa de los datos, como puede ser la clase. Los índices externos que nos encontramos en la literatura están sujetos a una interpretación que puede dar lugar a error, por ello, el objetivo de este artículo es presentar un nuevo índice de validación externa basado en el test estadístico de chi cuadrado que mide la calidad del clustering de forma exacta, sin necesidad de tener que ser interpretado. Se ha realizado una experimentación usando 6 datasets que podrían ser considerados big data y los resultados obtenidos son prometedores ya que mejoran la tasa de aciertos y porcentaje de error respecto a los índices de la literatura.**

***Index Terms*—Análisis de clustering, validación de clustering, índices de validación externa, Big Data**

I. Introducción

El clustering es una técnica de minería de datos que agrupa datos no supervisados en clusters de manera que las instancias que pertenecen al mismo clúster son similares. El clustering se ha usado en diferentes áreas de conocimiento como las ciencias sociales [1], la biología [2], la electricidad [3] o la agricultura [4].

Existen numerosos métodos de clustering en la literatura, y por lo general, cada uno genera una solución de clustering diferente. En algunos casos, se pueden obtener diferentes soluciones con el mismo método con tan solo cambiar alguno de parámetros de entrada. Una de las principales tareas del clustering es el análisis y evaluación de las distintas soluciones. Para medir la calidad de la solución de clustering, existen los llamados índices de validación de clustering (CVI).

Los CVI se podrían dividir en dos categorías: índices internos e índices externos. Los índices internos miden la calidad de la solución en función de la distribución de las instancias por los clusters, es decir, evalúan la separación que existe entre los clusters y la compacidad que hay entre las instancias que pertenecen al mismo clúster. Este tipo de índice es el único que se puede aplicar cuando el dataset no aporta ningún dato adicional. Por otra parte, los índices externos son aquellos que evalúan los clusters en función de algún atributo externo como puede ser la clase. Los índices de este tipo comparan el resultado del clustering con el de una solución global denominada *ground truth*. De esta forma los índices saben a priori la solución óptima así como el número óptimo de clusters del dataset ya que el *ground truth* contiene esta información. Por lo general, los índices de la literatura indican la solución óptima a través de una representación gráfica, y los resultados podrían ser interpretados de manera imprecisa. Los CVI se han usado en [5]–[8]. En este trabajo vamos a centrarnos en los índices de validación externos.

El objetivo de este artículo es presentar un nuevo CVI de clustering externo basado en el test estadístico de chi cuadrado cuyo resultado no necesite ser interpretado. La efectividad de este nuevo índice se ha comparado con 3 índices de la literatura en 6 datasets que podrían ser consideradas big data. La implementación del índice se ha realizado haciendo uso de las librerías MLlib de Spark [9] por lo que el índice estaría preparado para ejecutarse con datasets que podrían considerarse big data.

Este artículo se organiza de la siguiente forma: Sección II trata la literatura sobre los índices externos de validación. En la Sección III se introduce el nuevo índice propuesto en este trabajo. Sección IV presenta los experimentos, la metodología y los resultados obtenidos. Y por último, las conclusiones y trabajos futuros en la Sección V.

II. Trabajos relacionados

Los índices externos evalúan los resultados de un clustering comparándolo con el *ground truth*. Los índices externos se podrían clasificar dependiendo del criterio de comparación del clustering con el *ground truth* [10]. Los índices podrían clasificarse en: *set matching*, *pair-counting* y *information theory*.

- *Set matching* es la categoría que establece que la etiqueta de cada instancia se corresponde con un clúster. Alguno de los índices de la literatura son *purity* [11], *F-measure* [12] y *Goodman-Kruskal* [13].
- Los índices *pair-counting* se basan en la comparación entre el número de instancias con la misma etiqueta y el resultado del clúster. Esta categoría incluye índices como: *rand index* [14], *adjusted rand index* [15], *Jaccard* [16], *Fowlkes-Mallows* [17], *Hubert Statistic* [18] y *Minkowski score* [19].

- Los índices basados en *information theory* como la *entropy* [11], *variation of information* [20] and *mutual information* [21] también se han aplicado en la literatura.

En los últimos años se han publicado en la literatura numerosos estudios que proponen nuevos índices externos para la validación de clusters. En [7] encontramos un nuevo índice *pair-counting* basado en un enfoque probabilístico intuitivo, que se utiliza para comparar soluciones que pueden tener un cierto grado de solape. Este índice fue probado usando 4 datasets artificiales con 6 clases y 4 datasets reales del repositorio UCI [22].

También se presentó un nuevo CVI en [23], pero en este caso, el índice se basa en la distancia Max-Min usando lo que denominan información previa. Este índice externo podría clasificarse en la categoría de *Set matching*. El rendimiento se comparó con índices de tipo *Set matching* y *Counting pairs* utilizando 6 datasets artificiales y 2 datasets reales también del repositorio UCI.

Los autores del trabajo presentado en [24] propusieron un nuevo índice basado en un conjunto de clasificadores supervisados. Podemos clasificar este índice como índice *pair-counting*. Para los experimentos se utilizaron 50 datasets reales del repositorio de la UCI y los resultados se compararon con algunos índices internos.

En [25] se presentó un nuevo CVI *pair-counting* para comparaciones analíticas. Aplica una corrección por azar y normaliza para cada grupo por separado. Los experimentos se llevaron a cabo con datasets artificiales con 3 clases y 6000 instancias cada una. Este nuevo índice obtuvo mejores resultados que otros CVI externos, tales como *purity*, *adjusted rand index* o *mutual information*.

Otros autores sugirieron en [26] un nuevo CVI de concordancia de texto basado en una concepción del grado de libertad que mide el intervalo de decisión entre dos clases. Este índice mide la calidad del clustering comparándolo con índices internos y externos. Se utilizaron 14 datasets reales para probar el rendimiento del índice.

La mayoría de estos CVI se comprueban comparando los resultados de los clusters con algunos de los CVI de la literatura y utilizando datasets sintéticos. Sin embargo, ninguno de estos índices ha sido probado en entornos paralelos y distribuidos utilizando grandes datasets. Este trabajo pretende proporcionar un CVI que permita trabajar con datasets que podrían considerarse big data y basado el test estadístico de chi cuadrado.

A. Chi cuadrado

El test estadístico de chi cuadrado es un método que mide la diferencia entre los valores esperados y los valores observados en una distribución entre dos variables [27]. La siguiente ecuación se utiliza para verificar esta correlación:

$$\chi^2 = \sum_{i}^{r}\sum_{j}^{c}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \quad (1)$$

donde r es el número de filas, c es el número de columnas, n_{ij} es el valor observado y E_{ij} es el valor esperado. E_{ij} viene dado por

$$E_{ij} = \frac{n_{ij}}{n} \quad (2)$$

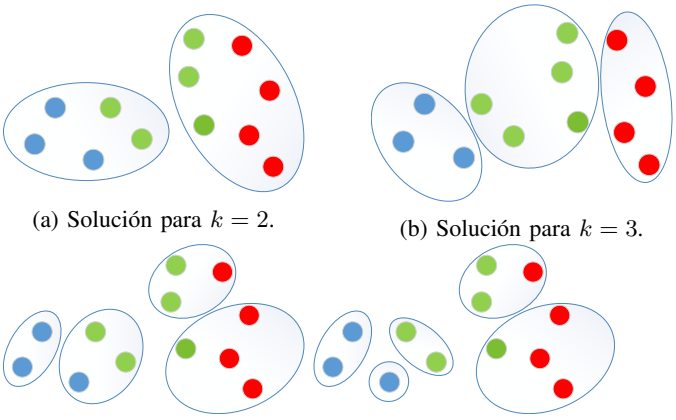
donde n es el número total de instancias. De manera que el valor de χ^2 estará más cerca de cero cuanto más se asemeje el valor observado al valor esperado.

III. Índice de validación de clustering externo basado en chi cuadrado

Supongamos una distribución de 12 instancias y 3 clases tal y como muestra la Figura 1, en la que cada punto representa una instancia y su color define la clase a la que pertenece.

Fig. 1: Representación de un dataset con 12 instancias y 3 clases donde los puntos representan a las instancias y los colores a las clases a las que pertenecen.

Antes de aplicar un método de clustering a este dataset tenemos que decidir el número de clusters (k) en el que lo vamos a dividir. Nuestro objetivo es encontrar el número óptimo de clusters de manera que las instancias que pertenecen a la misma clase queden agrupadas en los mismos clusters, y que los clusters tengan mayoritariamente instancias de una sola clase. La Figura 2 muestra las soluciones de clustering con $k = 2$ hasta $k = 5$, donde k es el número de clusters.



(a) Solución para $k = 2$.

(b) Solución para $k = 3$.

(c) Solución para $k = 4$.

(d) Solución para $k = 5$.

Fig. 2: Resultados de clustering para $k = 2$ hasta $k = 5$.

Para medir objetivamente la calidad de cada solución de clustering necesitaríamos un índice de validación externo. El

denominado *chi index* mide la calidad del clustering basandose en el test estadístico de chi cuadrado. Chi index aplica el test estadístico a la tabla de contingencia generada por la solución de clustering. Una tabla de contingencia muestra la distribución de frecuencias de las instancias, teniendo en cuenta las clases, en forma de matriz.

Siguiendo el ejemplo de la Figura 1, la Tabla I presenta la tabla de contingencia para de la solución de clustering para $k = 2$. Aquí podemos ver que el clúster 1 tiene 3 instancias de la clase azul y 2 instancias verdes, mientras que el clúster 2 tiene 4 instancias rojas y 3 verdes. Esta tabla puede ser analizada desde dos puntos de vista diferentes, teniendo en cuenta que queremos que cada clúster tenga instancias de la misma clase y que las instancias de una misma clase queden agrupadas en mismos clusters:

- Si la analizamos por filas, podemos decir que el clúster 1 está compuesto solo por instancias azules y verdes, sin instancias rojas. Y el clúster 2 está formado por instancias rojas y verdes.
- Si analizamos la tabla por columnas, podemos concluir que las instancias azules están sólo en el clúster 1, las instancias rojas están sólo en el clúster 2, y las verdes están repartidas en ambos clusters.

TABLE I: Tabla de contingencia para la solución de clustering con $k = 2$.

Clúster	Azul	Rojo	Verde	Total
1	3	0	2	5
2	0	4	3	7
Total	3	4	5	12

Estas lecturas se representan en las Tablas II y IIb, que son las tablas de contingencia expresadas con las frecuencias relativas al total de filas (Tabla IIa) y de columnas (Tabla IIb).

TABLE II: Tablas de contingencia relativas para $k = 2$.

(a) Frecuencias relativas al total de cada fila.

Clúster	Azul	Rojo	Verde	Total
1	60%	0%	40%	100%
2	0%	57%	43%	100%

(b) Frecuencias relativas al total de cada columna.

Clúster	Azul	Rojo	Verde
1	100%	0%	40%
2	0%	100%	60%
Total	100%	100%	100%

El siguiente paso sería obtener el valor de chi cuadrado para estas tablas, y realizar el mismo procedimiento en cada iteración.

En este ejemplo de juguete, hemos calculado el índice de chi para las soluciones de clustering desde $k = 2$ hasta $k = 6$. El objetivo es maximizar los valores del índice de chi en ambas tablas y minimizar la diferencia entre ellas. De esta manera, el resultado del índice de chi intentará que los valores observados y esperados sean lo más diferentes posible. Esto obligará a mantener la solución con el porcentaje más alto de clases en cada grupo.

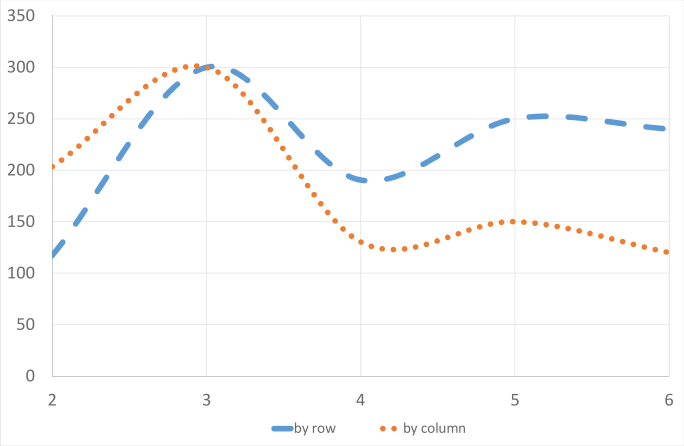


Fig. 3: Chi index result for $k = 2$ to 6.

La figura 3 muestra los resultados de nuestro ejemplo de juguete desde $k = 2$ hasta $k = 6$. Como se puede ver, en $k = 3$ las curvas alcanzaron un máximo, y también ambas curvas tienen el mismo valor. Por lo tanto, la mejor solución de clustering para este conjunto de datos es $k = 3$. Cabe destacar que este índice señala la solución cuando ambas curvas están en la distancia mínima o cruzadas. La interpretación de la solución es bastante simple porque sólo hay que mirar donde se cruzan ambas curvas.

Las tablas III muestran las tablas de contingencia relativas para la solución $k = 3$. Como se puede observar, cada clúster solo tiene instancias de una misma clase (Tabla IIIa), y cada clase está distribuida en un mismo clúster (Tabla IIIb).

TABLE III: Tablas de contingencia relativas para la solución de clustering $k = 3$.

(a) Frecuencias relativas al total de cada fila.

Clúster	Azul	Rojo	Verde	Total
1	**100%**	0%	0%	100%
2	0%	0%	**100%**	100%
3	0%	**100%**	0%	100%

(b) Frecuencias relativas al total de cada columna.

Clúster	Azul	Rojo	Verde
1	**100%**	0%	0%
2	0%	0%	**100%**
3	0%	**100%**	0%
Total	100%	100%	100%

Por lo tanto, *chi index* podría definirse como:

$$chi\,index = \min(\chi^2_{fila} - \chi^2_{columna}) \quad (3)$$

, donde

$$\chi^2_{fila} = \sum_i^r \sum_j^c \frac{(\frac{n_{ij}}{n_{i\cdot}} - \frac{n_{ij}}{n})}{\frac{n_{ij}}{n}} \quad (4)$$

$$\chi^2_{columna} = \sum_i^r \sum_j^c \frac{(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij}}{n})}{\frac{n_{ij}}{n}} \quad (5)$$

IV. Resultados

Esta sección describe los experimentos llevados a cabo para testear el CVI propuesto. Para realizar la comparativa se han usado 6 datasets públicos que podrían considerarse big data y se han comparado los resultados con 3 CVI de la literatura. Esta sección se compone de la Sección IV-A donde se detallan los datasets que se han utilizado en los experimentos. La Sección IV-B presenta el diseño de los experimentos seguidos. La Sección IV-D muestra los resultados de los experimentos realizados. La sección IV-D1 incluye un análisis estadístico para probar la efectividad de nuestro índice propuesto para los conjuntos de datos públicos. Finalmente, se incluye una discusión sobre los resultados en la Sección IV-D2.

A. Datasets

La tabla IV muestra los datasets utilizados para los experimentos. La tabla muestra las siguientes características: el nombre del dataset, el número de clases que va a ser usado como el número óptimo de clusters, el número de características y el número de instancias. Todos estos conjuntos de datos fueron descargados del repositorio de machine learning de UCI [22]. Todos los datasets incluyen la etiqueta de la clase, pero ésta no se ha utilizado para el proceso de clustering, solo se ha usado en la etapa de análisis.

TABLE IV: Descripción de los datasets.

Datasets	Clases	Atributos	Instancias
airlines	2	7	539,383
convtype	7	54	581,012
higgs	2	28	11,000,000
kddcup99	2	41	494,020
poker	10	10	829,202
susy	2	12	5,000,000

B. Diseño de experimentos

Para generar las soluciones de clustering, se han aplicado 3 métodos de clustering de Spark incluidos en la librería MLlib [9]: k-means, bisecting k-means y Gaussian mixture.

Estos métodos de clustering necesitan el número de clusters (k) en los que se va a particionar el dataset. Los experimentos se han realizado tomando los valores de k en el intervalo $[D_k-10, D_k+10]$ donde D_k es el número correcto de clusters del dataset siendo $k > 1$. Cada dataset se ha ejecutado con cada uno de estos 3 métodos de clustering con los que se han obtenido un total de 360 soluciones de clustering para probar los CVI. La comparativa se ha realizado entre 3 CVI de la literatura descritos en la Sección II y nuestro *chi index* propuesto.

C. Efectividad del CVI

La efectividad se mide en función de la cercanía a una solución ya dada (*ground-truth*) y realizar una comparativa de los resultados. El primer paso consiste en aplicar el algoritmo de clustering al dataset y obtener las múltiples soluciones. En el segundo paso se evalúan las soluciones de clustering aplicando los CVI. En el tercer y último paso se comparan los resultados del CVI y se selecciona el que mejor puntuación haya obtenido siguiendo.

Para hacer una comparativa entre los diferentes CVI se van a tener en cuenta los siguientes valores:

- Media de aciertos: este valor viene dado por la media de veces que el índice acierta el número óptimo de clusters por el total de datasets.
- Error medio al cuadrado: se calcula como la media de las distancias entre la predicción del índice I_i y el número correcto n_i por el total de datasets:

$$Error = \frac{\sum_{i \in n} d(I_i, n_i)^2}{n} \quad (6)$$

, donde n es el número total de datasets.

1) Test estadísticos: Por último, se ha aplicado un marco estadístico para probar el rendimiento de los CVI. Se ha seleccionado el test no paramétrico de Quade y el procedimiento post-hoc de Holm para validar estadísticamente las diferencias en los rangos medios de los *p-values* correspondientes alcanzados. Este análisis estadístico se realizó utilizando la plataforma de código abierto StatService [28].

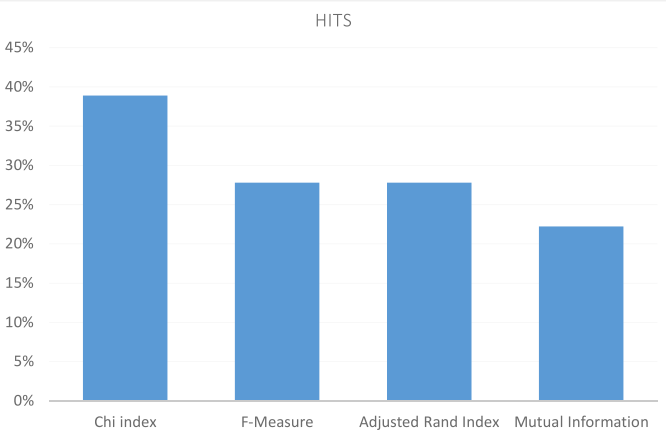
El test de Quade es una prueba estadística no paramétrica que evalúa las diferencias entre más de dos muestras estableciendo un ranking entre ellas. En nuestro caso, las muestras que vamos a evaluar son los CVI que vamos a comparar. Cuanto menor sea el *p-value*, mayor es la confianza de que un CVI funciona correctamente y, por lo tanto, se obtiene una mejor clasificación en el test de Quade.

Después de comprobar que los rankings medios son significativamente diferentes con un valor $\alpha = 0,05$, y siempre que el test de Quade rechace la hipótesis nula, se realizará el test post-hoc de Holm para evaluar el rendimiento relativo de las CVI estudiadas frente a un índice de control, en nuestro caso, tomaremos el que obtenga mejor puntuación en el ranking de Quade.

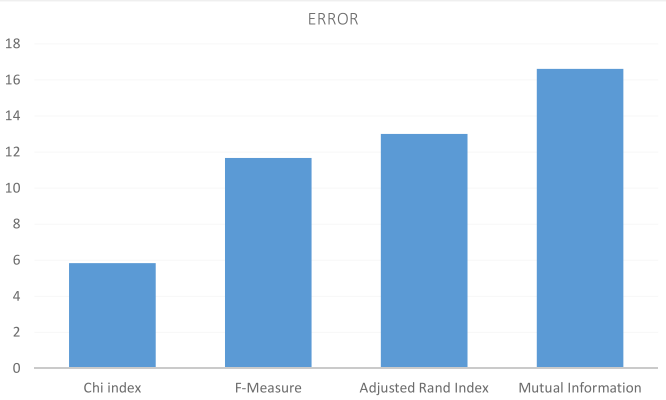
D. Resultados de los experimentos

Esta sección presenta los resultados realizados. La figura 4a muestra la media de aciertos para cada CVI en orden ascendente. *Chi index* ha alcanzado el valor más alto de aciertos (39%) con una diferencia significativa frente a sus competidores. Los índices de la literatura obtuvieron tasas similares de aciertos, que oscilaron entre el 28% en el caso de *F-Measure* y el 22% en el caso de *Mutual Information*.

Por otro lado, la Figura 4b presenta el error medio al cuadrado de cada CVI. Cabe señalar que *chi index* obtuvo el porcentaje de error más bajo (6%) quedando en segunda



(a) Media de aciertos para cada CVI.



(b) Error medio al cuadrado por CVI.

Fig. 4: Comparativa de resultados de los CVI.

posición *F-Measure* con casi el doble de puntos de error. Esto significa que *chi index* acierta el número óptimo de clusters la mayoría de las veces y en caso de error, el valor que indica no se aleja la solución.

1) Análisis estadístico: La Tabla V muestra el ranking del test de Quade para cada CVI. Como se muestra en el ranking, *chi index* ha quedado en primera posición con una puntuación de 1,807. El siguiente índice en el ranking fue *F-Measure* con una puntuación de 2,152, a 0,3 puntos del primer puesto. En última posición ha quedado *Mutual Information* con una puntuación en el ranking de 3,464.

TABLE V: Ranking del test de Quade.

CVI	Ranking
Chi index	1,807
F-Measure	2,152
Adjusted Rand Index	2,576
Mutual Information	3,464

El estadístico de Quade fue de 10,915, distribuida según una distribución F con 3 y 51 grados de libertad. El valor p de Quade fue 0,0 que fue inferior a 0.05. Por lo tanto, rechazó la hipótesis nula de que todos ellos se comportaron de manera similar con un nivel de significación de $\alpha = 0,05$.

Se ha realizado una prueba post-hoc por pares para verificar que nuestra propuesta es significativamente diferente al resto.

TABLE VI: Análisis post-hoc usando el procedimiento de Holm y tomando como índice de control a *chi index*.

CVI	p	z	α_{Holm}
Mutual Information	0,0058	2,760	0,0167
Adjusted Rand Index	0,2003	1,280	0,025
F-Measure	0,5530	0,574	0,050

La tabla VI muestra los *p-values*, el valor z y α_{Holm}, utilizando *chi index* como índice de control al ser el que mejor ranking obtuvo. Como puede verse, la hipótesis nula fue rechazada para todos los CVI. Por lo tanto, podemos concluir que *chi index* generó los mejores resultados (ya que obtuvo el mejor ranking) y fue estadísticamente diferente al resto de CVI.

2) Discusión: Los resultados del análisis de los datasets del repositorio de la UCI muestran que nuestro índice externo mejora la tasa de aciertos en 11% (Figura 4a). Además, en el caso de no poder alcanzar el número correcto de clusters, nuestro índice obtuvo una tasa de 6 puntos menos que los índices de la literatura (Figura 4b). Basado en la prueba de Quade (Tabla V), nuestro índice propuesto mejora los resultados en 2 puntos.

Es interesante resaltar que *chi index* indica la solución de clustering óptima de manera directa y concisa. A diferencia de los CVI de la literatura que necesitan ser interpretados siguiendo el método del codo y localizando máximos o mínimos locales, *chi index* indica la solución de clustering óptimo de manera directa y concisa como vimos en la Sección III.

V. Conclusiones

En este trabajo, se ha propuesto un nuevo índice de validación de clustering externo implementado en Spark para ser aplicado en datasets sin importar su tamaño. Hemos mostrado las diferencias entre nuestra propuesta y los índices de la literatura. El índice propuesto se basa en el test estadístico de chi cuadrado.

El estudio experimental indica que nuestro índice externo es muy competitivo. Hemos probado su efectividad en datasets públicos con un tamaño que podrían ser considerados big data en los que varían el número de clusters, sus características y el número de instancias. Los principales logros obtenidos son los siguientes:

- Un CVI externo basado en el test estadístico de chi cuadrado.
- Nuestro índice nos permitió estimar el número óptimo de clusters basado en la clase del dataset.
- Los resultados de *chi index* son directos y no requieren ser interpretados.
- El índice propuesto está listo para trabajar con conjuntos de datos que pueden ser considerados big data.
- El software de esta contribución se puede encontrar como un paquete de Spark en http://spark-packages.org/package/josemarialuna/externalValidity.

- El código fuente del índice de chi y los otros índices de la literatura se pueden encontrar en https://github.com/josemarialuna/ExternalValidity

Actualmente estamos aplicando este *chi index* en el análisis de datos de empleo y los resultados son prometedores. *Chi index* también se está aplicando en datos eléctricos en colaboración con una compañía eléctrica española. Como trabajo futuro, sería interesante ampliar la aplicación del índice en bases de datos que tengan multiclase.

Agradecimientos

Este trabajo ha sido apoyado por el Ministerio de Economía y Competitividad bajo el proyecto TIN2014-55894-C2-R. J.M. Luna-Romera es becario FPI del Ministerio de Economía y Competitividad.

References

[1] M. Ghane'i-Ostad, H. Vahdat-Nejad, and M. Abdolrazzagh-Nezhad, "Detecting overlapping communities in lbsns by fuzzy subtractive clustering," *Social Network Analysis and Mining*, vol. 8, no. 1, 2018, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85044313063&doi=10.1007%2fs13278-018-0502-5&partnerID=40&md5=96282433476a98b7bfc029f892772fc7

[2] F. Ginot, I. Theurkauff, F. Detcheverry, C. Ybert, and C. Cottin-Bizonne, "Aggregation-fragmentation and individual dynamics of active clusters," *Nature Communications*, vol. 9, no. 1, 2018, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042217142&doi=10.1038%2fs41467-017-02625-7&partnerID=40&md5=0d3b0fa19ff161e129139ac5de367f18

[3] R. Perez-Chacon, R. L. Talavera-Llames, F. Martinez-Alvarez, and A. Troncoso, "Finding electric energy consumption patterns in big time series data," in *Distributed Computing and Artificial Intelligence, 13th International Conference*, S. Omatu, A. Semalat, G. Bocewicz, P. Sitek, I. E. Nielsen, J. A. García García, and J. Bajo, Eds. Cham: Springer International Publishing, 2016, pp. 231–238.

[4] X. Wu, J. Zhu, B. Wu, J. Sun, and C. Dai, "Discrimination of tea varieties using ftir spectroscopy and allied gustafson-kessel clustering," *Computers and Electronics in Agriculture*, vol. 147, pp. 64–69, 2018, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042208353&doi=10.1016%2fj.compag.2018.02.014&partnerID=40&md5=bbb11cb2c0d295517af2c497192a4d43

[5] J. Rojas-Thomas, M. Santos, and M. Mora, "New internal index for clustering validation based on graphs," *Expert Systems with Applications*, vol. 86, pp. 334 – 349, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417304104

[6] J. Hämäläinen, S. Jauhiainen, and T. Kärkkäinen, "Comparison of internal clustering validation indices for prototype-based clustering," *Algorithms*, vol. 10, no. 3, 2017, cited By 0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85029738018&doi=10.3390%2fa10030105&partnerID=40&md5=15f131f750705ed3aad57f2d3dbba8b1

[7] D. Campo, G. Stegmayer, and D. Milone, "A new index for clustering validation with overlapped clusters," *Expert Systems with Applications*, vol. 64, pp. 549 – 556, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417416304158

[8] Z. Zhang, H. Fang, and H. Wang, "A new mi-based visualization aided validation index for mining big longitudinal web trial data," *IEEE Access*, vol. 4, pp. 2272–2280, 2016, cited By 7. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84979828536&doi=10.1109%2fACCESS.2016.2569074&partnerID=40&md5=cb527399d9c0c9990be218434656d657

[9] A. Spark, "Clustering - Spark 2.2.0 Documentation," https://spark.apache.org/docs/2.2.0/ml-clustering.html, 2018, [Online; accessed 6-april-2018].

[10] M. Rezaei and P. Fränti, "Set matching measures for external cluster validity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2173–2186, Aug 2016.

[11] Y. Zhao and G. Karypis, "Criterion functions for document clustering: Experiments and analysis," Tech. Rep., 2002.

[12] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '99. New York, NY, USA: ACM, 1999, pp. 16–22. [Online]. Available: http://doi.acm.org/10.1145/312129.312186

[13] L. A. Goodman and W. H. Kruskal, *Measures of Association for Cross Classifications*. New York, NY: Springer New York, 1979, pp. 2–34. [Online]. Available: https://doi.org/10.1007/978-1-4612-9995-0_1

[14] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971. [Online]. Available: http://www.jstor.org/stable/2284239

[15] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 1073–1080. [Online]. Available: http://doi.acm.org/10.1145/1553374.1553511

[16] R. Sokal and P. Sneath, *Principles of Numerical Taxonomy*, ser. Books in biology. W. H. Freeman, 1963. [Online]. Available: https://books.google.es/books?id=3Y4aAAAAMAAJ

[17] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, 1983.

[18] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec 1985. [Online]. Available: https://doi.org/10.1007/BF01908075

[19] A. Ben-Hur and I. Guyon, *Detecting Stable Clusters Using Principal Component Analysis*. Totowa, NJ: Humana Press, 2003, pp. 159–182. [Online]. Available: https://doi.org/10.1385/1-59259-364-X:159

[20] M. Meilă, "Comparing clusterings by the variation of information," in *Learning Theory and Kernel Machines*, B. Schölkopf and M. K. Warmuth, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 173–187.

[21] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von mises-fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, Dec. 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1046920.1088718

[22] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[23] A. Alok, S. Saha, and A. Ekbal, "Development of an external cluster validity index using probabilistic approach and min-max distance," vol. 6, pp. 494–504, 06 2012.

[24] J. Rodríguez, M. Medina-Pérez, A. Gutierrez-Rodríguez, R. Monroy, and H. Terashima-Marín, "Cluster validation using an ensemble of supervised classifiers," *Knowledge-Based Systems*, vol. 145, pp. 1–14, 2018.

[25] M. Rezaei and P. Fränti, "Set matching measures for external cluster validity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2173–2186, Aug 2016.

[26] C. Liu, W. Wang, M. Konan, S. Wang, L. Huang, Y. Tang, and X. Zhang, "A new validity index of feature subset for evaluating the dimensionality reduction algorithms," *Knowledge-Based Systems*, vol. 121, pp. 83 – 98, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705117300291

[27] J. Antoch, "A Guide to Chi-Squared Testing : Greenwood, P.E. and Nikulin, M.S. New York: John Wiley & Sons, Inc.,.pp. 280 +XII, ISBN 0-471-55779-X. AMS 1991 Classification: 62-02, 62F03, 62H15," *Computational Statistics & Data Analysis*, vol. 23, no. 4, pp. 565–566, February 1997. [Online]. Available: https://ideas.repec.org/a/eee/csdana/v23y1997i4p565-566.html

[28] J. A. Parejo, J. Garcia, A. Ruiz-Cortes, and J. C. Riquelme, "Statservice: Herramienta de análisis estadístico como soporte para la investigación con metaheurísticas," in *Actas del VIII Congreso Expañol sobre Metaheurísticas, Algoritmos Evolutivos y Bio-inspirados*, 2012.