ISSN: 2007-4948

https://repository.uaeh.edu.mx/revistas/index.php/xikua/issue/archive

XIKUA Boletín Científico de la Escuela Superior de Tlahuelilpan
13º Congreso Internacional de Computación
Inteligencia artificial: Presente y futuro
Red Iberoamericana de Computación
Vol. 12, Número Especial (2024) 111-116

Mejora en la clasificación de datos con desbalance de clases mediante una redistribución de clases por k-means

Improvement in the classification of data with class imbalance through a redistribution of classes by k-means

Antonio Alarcón Paredes [a], *Roberto Jhoshua Alegre-Ventura* [b], *Gustavo Adolfo Alonso Silverio* [c]

Abstract:

In the field of machine learning, there are several challenges that affect the performance of classification algorithms. Some of these include the curse of dimensionality or class imbalance. The dimensionality curse is a phenomenon that occurs when the number of features (p) in a dataset increases significantly compared to the number of samples (n) available. On the other hand, class imbalance occurs when one or more classes in a dataset have significantly less representation than other classes. This decreases the performance of a classifier since it generates classification biases towards the majority class. Microarray data is widely used to analyze and understand gene expression on a global level. These provide information about the expression of thousands of genes simultaneously and can be used to classify different conditions or diseases. Such data exhibits both dimensionality curse and class imbalance complexities.

In this work, a method to divide the majority class into two or more classes by means of the k-means clustering algorithm in microarray datasets is presented. Classification is performed using a variety of state of the art classification algorithms. The proposed method exceeds the classification performance of the original methods as it is reported, taking into account the balanced accuracy and a 5-fold cross-validation. After performing the Mann-Whitney statistical test, it was determined that the proposed mehotd obtains significantly better results than when the original algorithms are used.

Keywords:

Class imbalance, Statistical significance tests, Machine learning.

Resumen:

En el campo del aprendizaje automático, existen varios desafíos que afectan el desempeño de los algoritmos de clasificación. Algunos de estos incluyen la maldición de la dimensionalidad o el desbalanceo de clases. La maldición de la dimensionalidad es un fenómeno que ocurre cuando el número de características (p) de un conjunto de datos aumenta significativamente en comparación con el número de instancias (n) disponibles. Por otro lado, el desbalanceo de clases ocurre cuando una o varias clases en un conjunto de datos tienen una representación significativamente menor que otras clases. Esto disminuye el rendimiento del clasificador, ya que genera sesgos de clasificación hacia la clase mayoritaria. Los datos de microarreglos son ampliamente utilizados para analizar y comprender la expresión genética en un nivel global. Estos proporcionan información sobre la expresión de miles de genes

[a] Antonio Alarcón Paredes, Instituto Politécnico Nacional, https://orcid.org/0000-0002-9785-1252, Email: aalarcon@cic.ipn.mx

[b] Universidad Nacional Autónoma de México, https://orcid.org/0009-0004-3038-8402, Email: 319257836@ciencias.unam.mx

[c] Universidad Autónoma de Guerrero, https://orcid.org/0000-0002-2699-140X, Email: gsilverio@uagro.mx

Fecha de recepción: 16/04/2024, Fecha de aceptación: 22/05/2024, Fecha de publicación: 01/07/2024

DOI: https://doi.org/10.29057/xikua.v12iEspecial.12768

simultáneamente y pueden utilizarse para clasificar diferentes condiciones o enfermedades. Ese tipo de datos presentan tanto maldición de la dimensionalidad como desbalanceo de clases, por lo que su clasificación es compleja.

En este trabajo se presenta un método para dividir la clase mayoritaria dos o más clases por medio del algoritmo de agrupamiento k-means en conjuntos de datos de microarreglos. Se lleva a cabo la clasificación usando una variedad de algoritmos de clasificación en el estado del arte. Se reporta que el método propuesto supera el desempeño de clasificación de los métodos clásicos, tomando en consideración el balanced accuracy y un 5-fold cross-validation. Tras realizar la prueba estadística de Mann-Whitney se determinó que la propuesta obtiene resultados significativamente mejores que cuando se usan los algoritmos clásicos.

***Palabras Clave**:*

Desbalanceo de clases, Significancia estadística, Machine learning.

Introducción

En general, los conjuntos de datos altamente dimensionales están constituidos por muy pocas instancias (n) y por miles de características (p). Debido a la naturaleza de estos conjuntos de datos, se han identificado algunos inconvenientes en el uso de algoritmos estadísticos y de aprendizaje automático clásicos con fines de clasificación. En primera instancia, se suele presentar un bajo rendimiento de clasificación cuando existen muy pocas instancias en relación con los atributos, denominado n<<p, en el que las técnicas de aprendizaje automático no tienen suficientes observaciones para realizar una buena clasificación o predicción; esto es ampliamente conocido como la maldición. de la dimensionalidad 1.

El desbalanceo de clases es uno de los desafíos importantes en el aprendizaje automático, y está presente en diversas aplicaciones del mundo real. Estas pueden ser detección de fraudes, análisis de sentimientos, o diagnósticos médicos; en todos estos escenarios las clases de interés suelen ser mucho menos frecuentes que las demás. Existen técnicas de submuestreo y sobremuestreo que versan sobre la eliminación o replicación de instancias en las clases abundantes o minoritarias, respectivamente 2. Así también, la generación de instancias sintéticas ha sido propuesta para paliar esta situación 3.

Las tecnologías de expresión genética, como los microarrays, entran en la categoría de datos altamente dimensionales y comúnmente presentan desbalance de clases. La información contenida en los microarreglos ayuda a monitorear y medir datos relevantes para comprender diferentes aspectos biológica y facilita el análisis en contextos específicos como el diagnóstico de cáncer o la clasificación de diferentes tipos de tumores 4–6.

Aunque existen estrategias que pueden ayudar a mitigar el problema del desbalance de clases, no se ha podido resolver por completo. Por un lado, las técnicas de submuestreo pueden llevar a la pérdida de instancias importantes, mientras que el sobremuestreo puede inducir ruido o instancias que no sean representativas de la distribución real de la clase y, en casos extremos, llevar al overfitting o sobreajuste. Por otro lado, las técnicas de generación sintética de instancias, como el SMOTE presentan un desafío a la hora de capturar la distribución verdadera de la clase minoritaria 2,7,8.

Metodología propuesta

A pesar las limitaciones mencionadas, las estrategias para atacar este problema son importantes y siguen siendo aplicadas con éxito en ciertos ámbitos. Aun así, es fundamental tener en cuenta que esta problemática persiste en algunos escenarios. Por ello, aquí se presenta una estrategia que no aumenta ni disminuye el número de instancias, y tampoco genera otras de forma sintética, sino que trata de ayudar en la clasificación a través de una subdivisión de clases en la clase mayoritaria.

La presente propuesta se centra en el uso de un método de agrupamiento de datos: el popular *k*-means 9, que de forma general permite generar *k* grupos dentro de un conjunto de datos de forma que las instancias de un mismo grupo posean mayor similitud entre sí, y que no sean tan similares con las instancias de los otros grupos obtenidos. Para ello, típicamente se utiliza la distancia euclidiana, o de mahalanobis como métrica de dis-similitud, aunque puede elegirse alguna otra dentro de un gran compendio de ellas 10.

Tomando en consideración que se trabaja con un conjunto de datos con desbalance de clases, el método versa principalmente en aplicar *k*-means sobre las instancias de la clase mayoritaria con la finalidad de lograr subconjuntos de datos con una cardinalidad similar. Hasta este momento solamente se separan elementos de la clase mayoritaria en nuevos subconjuntos, sin embargo, es necesario pensar en la etapa de clasificación, por lo que estos nuevos subconjuntos de datos necesitarán ser asignados a una clase temporal, modificando el escenario de clasificación binaria en uno de clasificación multiclase. Esta idea va dirigida sobre uno de los principales problemas del desbalanceo de clase, que es cuando un método de clasificación se sobreajusta a la clase con mayor cantidad

de elementos. Así pues, ahora se cuenta con un conjunto de datos multiclase, cuyas clases son balanceadas.

Una vez logrado esto, basta con aplicar el algoritmo de clasificación que se desee sobre el nuevo conjunto de datos multiclase generado en el paso anterior. Sin embargo, al método *k*-means utilizado se le debe elegir cuántos sub-grupos deseamos obtener, lo cual presenta un problema para hacerlo de forma automática.

Para ello, se propone tomar ventaja de una medida importante en el desbalanceo de clases: la tasa de desbalance (*IR: imbalance ratio*). El *IR* se calcula dividiendo la cardinalidad del conjunto de instancias de la clase mayoritaria entre el número de elementos de la clase minoritaria (*IR*: $\frac{card\{clase\ mayoritaria\}}{card\{clase\ minoritaria\}}$), donde "card" se refiere a la cardinalidad del conjunto de instancias de la clase mayoritaria o minoritaria, según sea el caso.

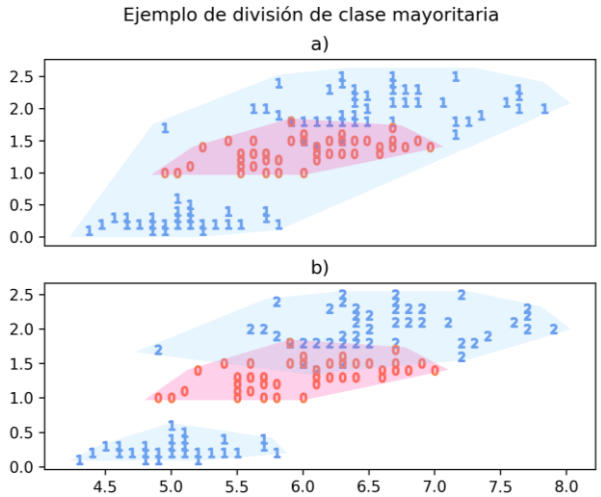

Figura 1. División de clase mayoritaria (De propia creación).

Para este trabajo se utilizaron cuatro conjuntos de datos de microarreglos, a saber: AMLALL, Colon Tumor, Medulloblastoma, y Prostate. La Tabla 1, muestra las características de dichos conjuntos de datos.

Tabla 1
Descripción de los conjuntos de datos de microarreglos

Dataset	Atributos	Instancias	Descripción
AML / ALL	7,129	72	Identificar entre leucemia aguda y leucemia mieloide aguda
Colon Tumor	2,000	62	Detección de cáncer de colon
Medullo-blastoma	7,129	60	Identificación de tumor en el sistema nervioso central
Prostate	12,600	102	Identificación de cáncer de próstata

Con la finalidad de llevar a cabo la predicción, se tomaron en cuenta algunos de los algoritmos de clasificación que son los más usados en el estado del arte. Dichos algoritmos son los siguientes: k-Nearest Neighbors (1-nn y 3-nn), Naive Bayes (bayes), Support Vector Machine con kernel lineal (svml), Decision Tree (dtree), AdaBoost (adaboost), Random Forest (randf), Extra Trees (xtree), Extreme Gradient Boosting Trees (xgboost) y Multilayer Perceptron (mlp).

Para evaluar adecuadamente el resultado de clasificación se tomó en consideración el balanced accuracy debido a que los conjuntos de datos exhiben un desbalanceo de clases. Para estos casos se hace una distinción entre las clases del conjunto de datos en clase positiva o clase 1 (P) y la clase negativa o clase 0 (N). Dicha distinción permite crear la matriz de confusión conformada por los verdaderos positivos: el número de datos P clasificados como clase P (TP); los falsos negativos: número de datos P clasificados como clase N (FN); los verdaderos negativos: número de datos N clasificados como clase N (TN); y los falsos positivos: el número de datos N clasificados como clase P (FP). Con la obtención de estos valores es posible implementar medidas de desempeño, que sean útiles en conjuntos de datos desbalanceados, como pueden ser: la sensibilidad (sensitivity: $\frac{TP}{TP+FN}$), especificidad (specificity: $\frac{TN}{TN+FP}$) y la exactitud balanceada (balanced accuracy), calculada como el promedio de las dos métricas anteriores.

Aunado a lo anterior, se considera un método de validación cruzada con 5 divisiones (5-fold cross validation).

El diagrama de la propuesta está ilustrado en la Figura 2.

Resultados

Esta sección está encargada de mostrar un estudio comparativo entre los resultados obtenidos al clasificar los datos con algoritmos de clasificación tradicionales versus los mismos algoritmos, pero utilizando el método propuesto para evitar el desbalance de clases.

La Tabla 2 muestra los resultados de clasificar los conjuntos de datos utilizando la métrica de evaluación balanced accuracy para todos los algoritmos de clasificación. Por otro lado, la Tabla 3 está enfocada a este mismo propósito, pero en este caso se aplica el método propuesto antes de la clasificación. Por supuesto, después de clasificar los conjuntos de datos con la nueva distribución de clases, se lleva a cabo un proceso para verificar que se esté clasificando correctamente las 2 clases originales del inicio. Para ello se hace una verificación como se muestra en la Figura 3.

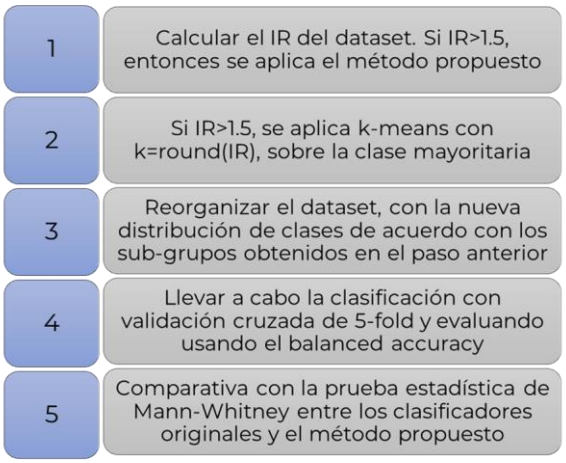

Figura 2. Metodología propuesta (De propia creación).

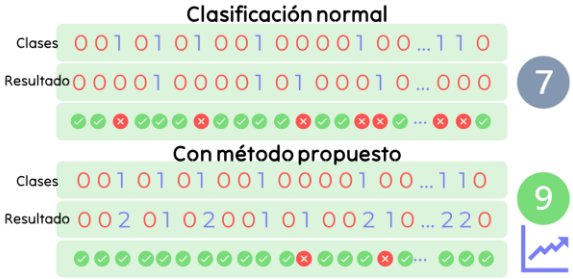

Figura 3. Ejemplo de mejora en clasificación con el método propuesto (De propia creación).

Tabla 2
Resultado de clasificación con algoritmos clásicos

Clasificador	AML ALL	Colon Tumor	Medullo-blastoma	Prostate
1-NN	0.817	0.603	0.525	0.633
3-NN	0.921	0.769	0.55	0.567
Bayes	0.8	0.433	0.525	0.567
SVM-L	0.976	0.658	0.55	0.633
DTree	0.762	0.587	0.425	0.633
AdaBoost	0.817	0.587	0.333	0.533
RandF	0.944	0.611	0.4	0.833
xTrees	0.833	0.619	0.475	0.733
XGBoost	0.952	0.595	0.375	0.867
MLP	0.952	0.658	0.55	0.567

Tabla 3
Resultado de clasificación utilizando el método propuesto

Clasificador	AML ALL	Colon Tumor	Medullo-blastoma	Prostate
1-NN	0.944	0.8	0.733	0.8
3-NN	0.944	0.833	0.633	0.8
Bayes	1	0.733	0.675	0.667
SVM-L	1	0.733	0.7	0.967
DTree	0.817	0.603	0.675	0.833
AdaBoost	1	0.633	0.675	0.933
RandF	0.976	0.8	0.633	0.9
xTrees	0.944	0.8	0.7	0.933
XGBoost	1	0.634	0.633	0.933
MLP	1	0.833	0.675	0.967

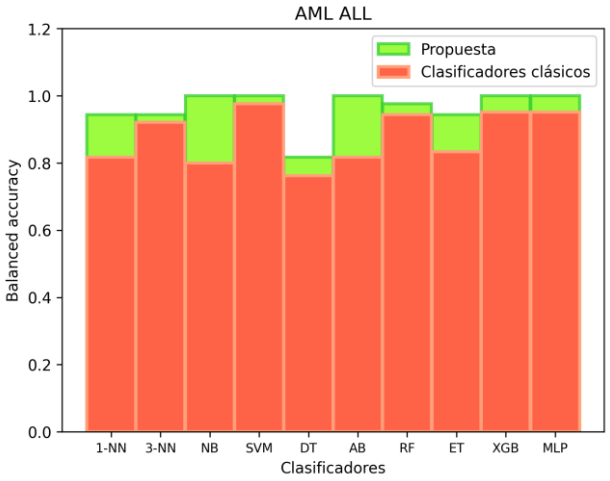

Figura 4. Resultado de clasificación con clasificadores clásicos y con el método propuesto, sobre AML ALL. (De propia creación).

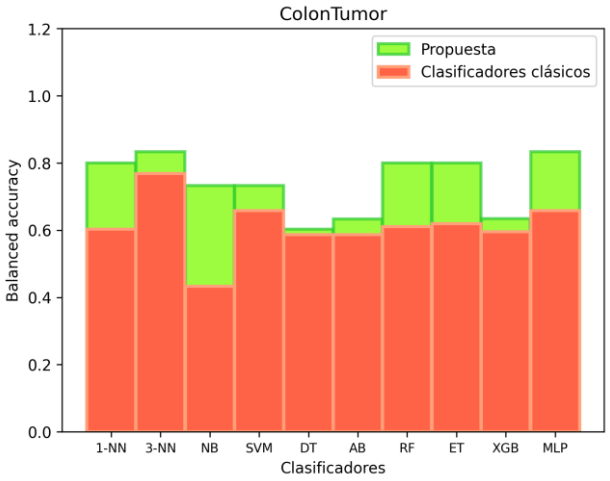

Figura 5. Resultado de clasificación con clasificadores clásicos y con el método propuesto, en ColonTumor. (De propia creación).

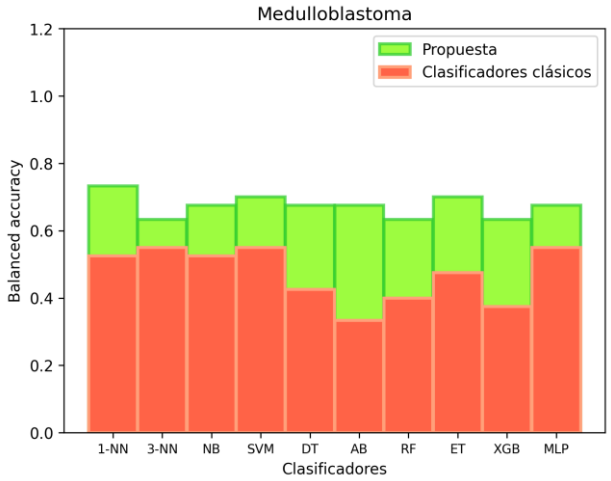

Figura 6. Resultado de clasificación con clasificadores clásicos y con el método propuesto, en Medulloblastoma. (Propia creación).

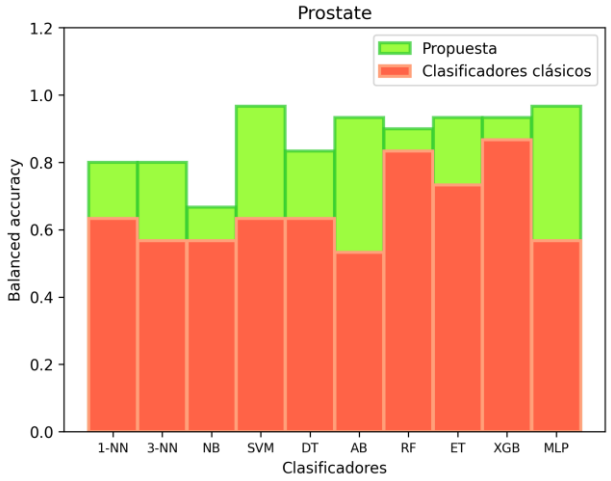

Figura 7. Resultado de clasificación con clasificadores clásicos y con el método propuesto, en Prostate. (De propia creación).

Las Tablas 2 y 3 presentan los resultados de los algoritmos de clasificación clásicos, y usando el método propuesto, respectivamente, sobre los cuatro conjuntos de datos de microarreglos, donde puede verse la mejora en la clasificación. Con la finalidad de ilustrar de mejor manera los resultados de clasificación por cada conjunto de datos, se presentan las Figuras 4 a Figura 7, donde se reporta la comparativa sin y con la propuesta para los datos AML ALL, Colon Tumor, Medulloblastoma y Prostate, respectivamente.

La prueba de Mann-Whitney es una prueba estadística no paramétrica en la que la hipótesis nula (H0) establece que la distribución de una muestra A es la misma que la distribución de una muestra B. A menudo es utilizada para determinar diferencias entre funciones de distribución para muestras independientes. En este caso particular, la muestra A consiste en el vector formado por todos los valores resultantes de la clasificación usando los algoritmos clásicos, mientras que la muestra B consiste correspondientemente en el vector formado por los valores resultantes de la clasificación utilizando el método propuesto. Así pues, la prueba estadística determinará, para cada conjunto de datos, si los resultados de la propuesta hecha aquí difieren significativamente de los resultados de algoritmos clásicos. Para la evaluación de la prueba de Mann-Whitney se toma en consideración una confianza del 95%, esto es, los resultados son significativos cuando el valor $p<0.05$.

Una vez analizados los valores de clasificación por balanced accuracy, individualmente por conjunto de datos, se presentan los resultados en la Tabla 4, donde es posible apreciar que en todos los casos la mejora obtenida con la propuesta es significativamente mejor que los resultados obtenidos con los clasificadores clásicos.

Tabla 4
Resultado de comparativa estadística usando Mann-Whitney

	AML ALL	Colon Tumor	Medullo-blastoma	Prostate
Valor *p*	0.016	0.006	0.000	0.002

Conclusiones

En el presente trabajo se ha demostrado la efectividad del método propuesto, en particular, que el hecho de subdividir la clase mayoritaria en más clases para balancear las clases del conjunto de datos funciona, al menos en este escenario que combina desbalance de clases con datos de alta dimensión. En este sentido, es notorio el hecho de que, en todos los clasificadores, el resultado obtenido con el método propuesto mejora el valor del balanced accuracy para todos los conjuntos de datos tomados en consideración.

Lo anterior fue además evidenciado al llevar a cabo un análisis de significancia estadística utilizando la prueba de Mann-Whitney, en la que los resultados obtenidos por el método propuesto son significativamente mejores estadísticamente que los obtenidos por los clasificadores clásicos.

Esta situación permite elucubrar de manera firme que, una vez subdividida la clase mayoritaria, la distribución de clases obtenida permite generar un clasificador más robusto que con los datos originales.

A partir de esta fuerte asunción se propone como trabajo a futuro que, cuando se presenten conjuntos de datos cuya clasificación sea muy compleja o que los datos de las clases estén muy mezclados, aun si no existe un desbalance de clases, se aplique una subdivisión en ambas clases para modificar la su distribución y así intentar mejorar el desempeño en la clasificación.

Referencias

[1] Narendra, P.M.; Fukunaga, K. A Branch and Bound Algorithm for Feature Subset Selection. IEEE Trans. Comput. 1977, 26, 917–922, doi:10.1109/tc.1977.1674939.

[2] Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. Learning from imbalanced data sets; Springer, 2018; Vol. 10;.

[3] Chawla, N. V; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 2002, 16, 321–357.

[4] Golub, T.R. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science (80-.). 1999, 286, 531–537, doi:10.1126/science.286.5439.531.

[5] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. Mach. Learn. 2002, 46, 389–422.

[6] Chan, W.H.; Mohamad, M.S.; Deris, S.; Zaki, N.; Kasim, S.; Omatu, S.; Corchado, J.M.; Al Ashwal, H. Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme. Comput. Biol. Med. 2016, 77, 102–115, doi:10.1016/j.compbiomed.2016.08.004.

[7] Wang, L.; Han, M.; Li, X.; Zhang, N.; Cheng, H. Review of classification methods on unbalanced data sets. IEEE Access 2021, 9, 64606–64628.

[8] Thabtah, F.; Hammoud, S.; Kamalov, F.; Gonsalves, A. Data imbalance in classification: Experimental evaluation. Inf. Sci. (Ny). 2020, 513, 429–441.

[9] Jain, A.K. Data clustering: 50 years beyond K-means. Pattern Recognit. Lett. 2010, 31, 651–666.

[10] Pfitzner, D.; Leibbrandt, R.; Powers, D. Characterization and evaluation of similarity measures for pairs of clusterings. Knowl. Inf. Syst. 2009, 19, 361–394.