Enliven Archive
Intensifying Thoughts

Review Article www.enlivenarchive.org Enliven: Pharmacovigilance and Drug Safety

ISSN: 2378-5411

Method Based on the Different Entropy Principles

Aldin Paul

Lecturer, Department of Pharmaceutical Sciences, Chulalongkorn University, 254 Phayathai Road, Pathumwan, Bangkok 10330, Thailand

*Corresponding author: Aldin Paul, Lecturer, Department of Pharmaceutical Sciences, Chulalongkorn University, 254 Phayathai Road, Pathumwan, Bangkok 10330, Thailand, E-mail: paulaldin@tutanota.com

Received Date: 16th September 2018

Accepted Date: 22nd November 2018

Published Date: 23rd November 2018

Citation: Paul A (2018) Method Based on the Different Entropy Principles. Enliven: Pharmacovigilance and Drug Safety 5(1): 002.

Internal clustering validation indexes (CVIs) which can be assumed as the objective function of clustering algorithms is usually relied upon to assess the quality of different clustered partitions with the intention of ascertaining the local clustering outputs in an unsupervised way [1]. The researchers started by investigating numerous popular internal clustering validation indexes used for categorical data clustering [2]. In addition, they validated the ineffectiveness of assessing the partitions of diverse sets of clusters while disregarding any inter-cluster assumptions or separation measures [3]. The correctness of separation as well as how it coordinates with the intra-cluster compactness measures negatively impact performance. Consequently, the researcher suggested a new internal clustering validation index known as CUBAGE which could assess both the partition's separation and the compactness [4]. The study results proved that the CUBAGE performed excellently as compared to other internal clustering validation indexes whether in the presence or absence of the number of clusters [5].

Clustering analysis involves subdividing a set of data into clusters with the aim of grouping similar and dissimilar objects into their respective clusters. There are two types of clustering methods namely the hard and soft techniques [6]. In this case, the researchers adopted the hard former where all the objects belonged to a single cluster. The partitions obtained after clustering largely varied with the criteria of dissimilarity or similarity, clustering methods, and the parameter settings [7]. For instance, the clustering technique's mechanism such as the random initialization also lead to variations in the clustering outcome or output. The researchers sought to determine the final result from several likely partitions by performing numerous clustering processes with diverse schemes correspondingly before they chose the highest quality partition [8]. For the researchers, they focused on exploring the use of internal and external clustering validation indexes to help in defining and measuring the quality of partitions [9].

The internal CVIs can be relied on to assess the compliance of the clustered and prior partitions. Alternatively, the external CVIs rely on information obtained from outside sources to measure the quality of the clustering output. For example, in the case of the presence of prior knowledge, it can be utilized to analyze the compliance of the previous and clustered partitions. Since the unsupervised scenario lacks prior knowledge, the external CVIs cannot be applied [10]. The internal clustering validation indexes do not need previous experience, and usually, they can be applied in different areas or discipline such as data mining, biological engineering, image and text analysis, as well as information retrieval [11]. The inter-cluster isolation or separation and intra-cluster compactness are relied upon to test the quality of clustering results internally [12]. The isolation indicates how the data sets in a single cluster are dissimilar to others while the compactness signifies the level of equivalence or uniformity of data sets in the same cluster [13].

Furthermore, since numerous internal clustering validation indexes such as the Calinski-Harabasz, the Silhouette, the I, and the Dunn indexes are regarded as inept for categorical data clustering since they utilize intuitive geometric information to analyze the partitions [14]. There is need for advance or additional studies in internal clustering validation indexes for categorical data since there is a massive amount of categorical data being practically applied as well as the problematic concerns that have failed to be fully handled in the literature [15]. Consequently, the researchers restricted the scope of the study to offer perspectives and improvements of the internal clustering validation indexes for categorical data [16]. They sought to determine if the internal CVIs for categorical data without isolation measures disregard the isolation, whether they demonstrate monotonicity with regards to the total number of groups, and what can be done to improve the performance of the internal CVIs [17].

The researchers investigated five popular internal clustering validation indexes used for validating categorical data clustering. They include two objective functions of clustering with subjective factors and slope (R and Cloper respectively), the category utility function (CU), the k-modes objective function (F), and the information entropy function (E) [18]. They focused on examining the separation or isolation and compactness measures to determine the characteristics of the five internal CVIs [19]. Furthermore, they sought to determine if monotonicity in some particular situations can be shown by the compactness measures for categorical data as well as how

isolation measures can help to offset the monotonicity [20]. Additionally, they proposed the CUBAGE which analyzes the dataset's reciprocal entropy to measure compactness and AGE to assess the isolation [21]. The CUBAGE performed excellently in the experimental studies as compared to other indexes proving that the compactness and separation measures are correct; besides, they perfectly coordinated in majority of the datasets.

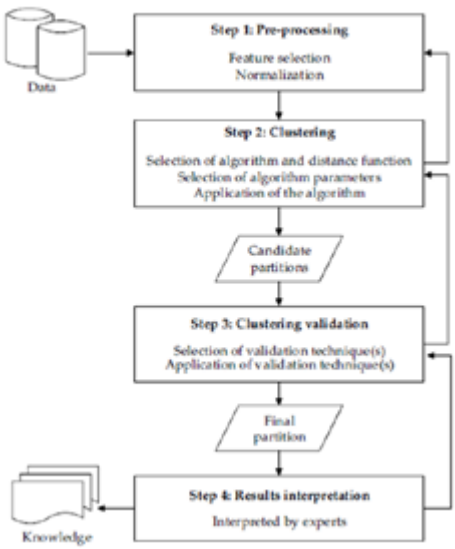

Figure 1: Clustering procedure consisting of four steps with a feedback pathway. (Source: James C. Bezdek, Objective Function Clustering (Boston: Springer, 1981) 67.)

References

1. Aldana-Bobadilla E, Kuri-Morales A (2015) A clustering method based on the maximum entropy principle. Entropy 17: 151-180.
2. Karthik S, Sinha A, Deb K, Miettinen K (2009) Local search based evolutionary multi-objective optimization algorithm for constrained and unconstrained problems. In IEEE Congress on Evolutionary Computation 2919-2926.
3. Kruskal JB (1964) Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika 29: 1-27.
4. Rokach Lior, Oded Maimon (2005) Clustering methods. In Data mining and knowledge discovery handbook 321-352.
5. Dunn JC (1974) Well-Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybernetics 4: 95-104.
6. Bezdek JC (1981) Objective function clustering. In Pattern recognition with fuzzy objective function algorithms 43-93.
7. Tian Z, Ramakrishnan R, Livny M (1996) BIRCH: an efficient data clustering method for very large databases. In ACM Sigmod Record 25: 103-114.
8. Martin E, Kriegel HP, Sander J, Xu X (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Kdd 96: 226-231.
9. Masashi S, Niu G, Yamada M, Kimura M, Hachiya H (2014) Information-maximization clustering based on squared-loss mutual information. Neural Computation 26: 84-131.
10. Yan M (2005) Methods of determining the number of clusters in a data set and a new clustering criterion. PhD diss., Virginia Tech.
11. Liu Y, Li Z, Xiong H, Gao X, Wu J (2010) Understanding of internal clustering validation measures. 2010 IEEE International Conference on Data Mining 911-916.
12. Swathi M (2017) Clustering Enhancement Using Similarity Indexing to Reduce Entropy. Enliven: Bioinform 4: 001.
13. Cheng CH, Fu AW, Zhan Y (1999) Entropy-based subspace clustering for mining numerical data. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining 84-93.
14. Xin L, Mak MW, Li CK (1999) Determining the Optimal Number of Clusters by an Extended RPCL Algorithm. JACIII 3: 467-473.
15. Larsen Bjornar, Chinatsu Aone (1999) 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 16-22.
16. Ying Z, George Karypis, Ding-Zhu Du (2005) Criterion functions for document clustering. University of Minnesota.
17. Swathi M (2017) Drug Prediction of Cancer Genes Using SVM. Enliven: Pharmacovigilance and Drug Safety 4: 001.
18. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence 2: 224-227.
19. Erendira R, Garcia R, Abundez I, Gutierrez C, Gasca E, et al. (2008) Niva: A robust cluster validity. In WSEAS International Conference 12: 23-25.
20. Thomas B, Schwefel HP (1993) An overview of evolutionary algorithms for parameter optimization. Evolutionary computation 1: 1-23.
21. Kuri-Morales AF, Aldana-Bobadilla E, López-Pena I (2013) The Best Genetic Algorithm II. In Mexican International Conference on Artificial Intelligence 16-29.
22. Hino H, Murata N (2014) A Nonparametric Clustering Algorithm with a Quantile-Based Likelihood Estimator. Neural Computation 26: 2074-2101.
23. Jenssen R, Hild KE, Erdogmus D, Principe JC, Eltoft T (2003) Clustering Using Renyi's Entropy. In Neural Networks, 2003. Proceedings of the International Joint Conference on 1: 523-528.
24. Noam S, Atwal GS, Tkacik G, Bialek W (2005) Information-based clustering. Proceedings of the National Academy of Sciences 102: 18297-18302.
25. Swathi M (2018) Enhancement of K-Mean Clustering for Genomics of Drugs. Enliven: J Genet Mol Cell Biol 5: 001.