Enliven Archive
Intensifying Thoughts

Research Article www.enlivenarchive.org Enliven: Journal of Genetic, Molecular and Cellular Biology

ISSN: 2379-5700

Analysis of Gene Expression Data to Validate and Estimate the Number of Entropy-Based Cluster

Vincent Marcov

Department of Pediatrics and Genetics, Ain Shams University, Nour Mosque, 38 Abbasia Next, El-Mohamady, Al Waili, Egypt

*Corresponding author: Vincent Marcov, Department of Pediatrics and Genetics, Ain Shams University, Nour Mosque، 38 Abbasia Next، El-Mohamady, Al Waili, Egypt, E-mail: vincentmarcov@protonmail.com

Received Date: 9th January 2018
Accepted Date: 20th March 2018
Published Date: 21st March 2018

Citation: Marcov V (2018) Analysis of Gene Expression Data to Validate and Estimate the Number of Entropy-Based Cluster. Enliven: J Genet Mol Cell Biol 5(2): 002.

With regards to gene expression data, several internal and external validity measures have been recommended to approximate the number of their clusters. Nonetheless, in most cases, the experts fail to take into account the analysis related to how a clustering algorithm can produce stability of the groupings [1]. The article recommends a new selection process and standard of cluster validation to establish the appropriate number of clusters. The method evaluates the predictive stability or power of partitioning [2]. Furthermore, the validity measure is developed by estimating the consensus matrix's "clearness" which can be considered as the outcome or product of resampling a consensus clustering or resampling clustering scheme [3]. Additionally, the analysis of the distance between the validity plots for permutated and initial data sets helps in selecting the number of clusters [4]. The researchers employed the selection process to approximate the clustering results on numerous datasets as shown below [5].

Therefore, the recommended procedure creates a comprehensive and correct estimation of the number of clusters which are consistent with the gold standards and biological knowledge of cluster quality. At present, it is highly expected to concurrently assess the expression level of genes using the microarray chip technology. It has greatly contributed towards advancement of new techniques of computational intelligence as well as new concerns in bioinformatics research [6-7]. Structuring massive amount of data so that it would be possible to develop or extract knowledge from them has been problematic [8]. Clustering has proved to be most popular exploring method that assists in establishing the important trends of co-expressed genes as well as facilitates significant grouping of the gene expression [9]. Furthermore, by detecting a set of gene clusters, scientists can now understand the function of a cell by exploring the mechanisms of gene interaction and regulation as well as defining the roles of certain genes which were formerly unrecognized [10]. Alternatively, analysis of diverse tumor samples to determine their gene expression profiles has led to improvements in medical treatment and detection of new unidentified tumor subtypes [11]. Furthermore, microarray data is being explored using several all-purpose and specialty clustering algorithms. The algorithms used for clustering gene expression data have different pros and cons [12]. The diverse clustering algorithms can lead to inconsistent clustering owing to the certain type of grouping of the data items and inner bases [13]. The major issue is the evaluation of cluster assignments for each sample, the chosen number of clusters, and the confidence of the clustering outcome. The researchers recommended a validity measure to facilitate estimation data clustering stability which ensures the confidence of the clustering result is increased [14]. The validity measure adopts the computations of the consensus matrix entropy as well as relies on past scholarly projects on resample-based consensus clustering is used to develop the validity measure. The article recommends the selection process to establish a suitable number of clusters largely on the basis of the entropy calculations. Furthermore, it is evident that clustering algorithms and new methods to the validation of the clustering outcomes as a substitute for the independent and manual verification are constantly developed [15]. There are two main categories of the techniques being employed for the validation of the clustering outcomes. They include the external and internal validation measures [16].

According to Natalia Novoselova and Igor Tom, the external measures use the gold standard for clustering or recognized set of class labels are used to assess the clustering results [17]. They evaluate the consensus between the gold standard and a partitioning on the basis of the contingency table of data items' pairwise assignments for instance, the FM score, Jaccard coefficient, entropy measure, and the adjusted Rand Index [18].

Alternatively, the internal measures are usually approximate the clustering results' correspondence to the internal data structure and they are mostly based on data alone. Majority of the internal measures attempt to achieve

maximization of the cluster separability, for instance, the Silhouette Width and Dunn-like Indices as well as to reduce the cluster compactness [19, 20]. The authors recommend to introduce a new measure of clustering validity through the estimation of consensus matrix entropy. Additionally, a comparison between the gold standard and the validity measure founded on empirical cumulative distribution (CDF) is also made [21, 22]. Furthermore, the authors try to establish the number of clusters by recommending a selection process which links to the best stable clustering outcome [23-25]. The researchers recommend a validity measure which approximates the stability of the clustering outcome by computing the entropy's consensus matrix.

References

1. Novoselova N, Tom I (2012) Entropy-Based Cluster Validation and Estimation of the Number of Clusters in Gene Expression Data. J Bioinform Comput Biol 10: 1250011.
2. Ramoni MF, Sebastiani P, Kohane IS (2002) Cluster Analysis of Gene Expression Dynamics. Proc Natl Acad Sci U S A 99: 9121-9126.
3. Junjie Wu, Xiong H, Chen J (2009) Adapting the Right Measures for K-Means Clustering. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2: 877-886.
4. Eisen MB, Spellman PT, Brown PO, Botstein D (1999) Cluster Analysis and Display of Genome-Wide Expression Patterns. Proc Natl Acad Sci U S A 95: 14863-14868.
5. Swathi M (2018) Enhancement of K-Mean Clustering for Genomics of Drugs. Enliven: J Genet Mol Cell Biol 5: 001.
6. Hestilow TJ, Huang Y (2009) Clustering of Gene Expression Data Based on Shape Similarity. EURASIP J Bioinform Syst Biol 195712.
7. Dagliyan O, Uney-Yuksektepe F, Kavakli IH, Turkay M (2011) Optimization Based Tumor Classification from Microarray Gene Expression Data. PLoS One 6: e14579.
8. Mark Junjie Li, Michael K. Ng, Yiu-ming Cheung, Joshua Zhexue Huang (2008) Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters. IEEE Transactions on Knowledge and Data Engineering 20: 1519-1534.
9. Swathi M (2017) Clustering Enhancement Using Similarity Indexing to Reduce Entropy. Enliven: Bioinform 4: 001.
10. Maulik U, Mukhopadhyay A, Bandyopadhyay S (2009) Combining Pareto-Optimal Clusters Using Supervised Learning for Identifying Co-Expressed Genes. BMC Bioinformatics 10: 27.
11. Anil KJ (2010) Data Clustering: 50 Years beyond K-Means. Pattern Recognition Letters 31: 651-666.
12. Sandrine D, Fridlyand J (2002) A Prediction-Based Resampling Method for Estimating the Number of Clusters in a Dataset. Genome Biol 3: 3-6.
13. Damaris P, Pla F, Sanchez JS (2010) Cluster Validation Using Information Stability Measures. Pattern Recognition Letters 31: 454-461.
14. Alexander S, Ghosh J (2002) Cluster Ensembles - a Knowledge Reuse Framework for Combining Multiple Partitions. Journal of Machine Learning Research 3: 583-617.
15. Lee Yongjin, Seungjin Choi (2004) Minimum Entropy, K-Means, Spectral Clustering. 2004 IEEE International Joint Conference on Neural Networks 1: 117-122.
16. Swathi M (2017) Drug Prediction of Cancer Genes Using SVM. Enliven: Pharmacovigilance and Drug Safety 4: 001.
17. Haifeng Li, Keshu Zhang, Tao Jiang (2004) Minimum entropy clustering and applications to gene expression analysis Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004. 1: 142-151.
18. Bertoni A, Valentini G (2006) Randomized Maps for Assessing the Reliability of Patients Clusters in DNA Microarray Data Analyses. Artif Intell Med 37: 85-109.
19. Moftah HM, Azar AT, Al-Shammari ET, Ghali NI, Hassanien AE (2014) Adaptive K-Means Clustering Algorithm for MR Breast Image Segmentation. Neural Computing and Applications 24: 1917-1928.
20. Novoselova N, Tom I (2012) Entropy-Based Cluster Validation and Estimation of the Number of Clusters in Gene Expression Data. J Bioinform Comput Biol 10: 1250011.
21. Lange T, Roth V, Braun ML, Buhmann JM (2004) Stability-Based Validation of Clustering Solutions. Neural Comput 16: 1299-1323.
22. Ulrike Von Luxburg (2010) Clustering Stability: An Overview. Foundations and Trends in Machine Learning 2: 235-274.
23. Rokach, Lior, Oded Maimon (2005) Clustering Methods. In Data Mining and Knowledge Discovery Handbook 321-352.
24. Thomas MC, Thomas JA (2001) Entropy, Relative Entropy and Mutual Information. Elements of Information Theory 3: 1-55.
25. Maria Halkidi, Batistakis Yannis, Vazirgiannis Michalis (2002) Cluster Validity Methods: Part I. ACM Sigmod Record 31: 40-45.
26. Julia H, Knowles J (2006) Feature Subset Selection in Unsupervised Learning via Multiobjective Optimization. International Journal of Computational Intelligence Research 2: 217-238.
27. Kaijun W, Baijie W, Liuqing Peng (2009) CVAP: Validation for Cluster Analyses. Data Science Journal 8: 88-93.
28. Keke C, Ling L (2004) VISTA: Validating and Refining Clusters Via Visualization. Information Visualization 3: 257-270.
29. Daxin J, Tang C, Zhang A (2004) Cluster Analysis for Gene Expression Data: A Survey. IEEE Transactions on Knowledge and Data Engineering 16: 1370-1386.