Enliven Archive
Intensifying Thoughts

Short Note www.enlivenarchive.org Enliven: Bioinformatics

ISSN: 2376-9416

The Correlation between Categorical Clustering and Entropy-Based Criterion

John Howes

Independent Researcher, University of Waterloo, 200 University Avenue West Waterloo, ON, Canada

*Corresponding author: John Howes, Independent Researcher, University of Waterloo, 200 University Avenue West Waterloo, ON, Canada, E-mail: howesjohn@outlook.com

Received Date: 10th January 2018
Accepted Date: 12th March 2018
Published Date: 13th March 2018

Citation: Howes J (2018) The Correlation between Categorical Clustering and Entropy-Based Criterion. Enliven: Bioinfo 5(1): 004.

Abstract

Earthworm is also called as Oligochaeta. It was the first database which can provide the information about the earthworm species in India. We identified 68 genus and 515 species of earthworms all over India. The database consists of family, taxonomy, length, segments, diameter, food habit, habitat, casting, description and distribution of the Indian species. The database was developed by MySQL running in Windows operating system. The database interface was developed by PHP, and HTML.

Availability: http://www.mscwbif.org/earthworm/home.html

Keywords: Earthworm; Oligochaeta; Earthworm habitat; Earthworm distribution

There are very few definite clustering algorithms despite the increasing demand for cluster analysis of categorical data. The significance of data clustering to perform data analysis has progressively gained prominence over the last few years. Notably, the clustering algorithms are used to group all the comparable items into clusters using the similarity measure [1]. According to Aggarwal, Charu, Cecilia Procopiuc, and Philip there are clear differences between the clustering techniques for numerical and categorical data particularly with regards to providing the definition of the similarity measure [2]. With regards to the numerical clustering techniques, they delineate the similarity measure by using the distance function such as the Euclidean distance. Alternatively, it has been proven that between the categorical values, there is no integral distance meaning [3]. Customarily, data preprocessing phase is usually used to merge the numerical and categorical data clustering whereby a domain-based knowledge is used to define conceptual similarity between data or the categorical data are manipulated to construct or extract the numerical features [4]. Nonetheless, given the little experience about the data the initial stages of the analysis process it is difficult to extract any meaningful conceptual similarity or numerical features [5]. Furthermore, it has been extensively recognized that most applications require direct clustering of the raw categorical data, for instance, the network intrusion analysis, protein or DNA sequence analysis, market basket data analysis, and environmental data analysis.

Cluster validation methods need to be adopted to assess the quality of the clustering outcomes since the diverse clustering algorithms rarely produce similar results for a single dataset [6]. Officially, cluster validation is faced by two major concerns first, how to establish the numbers of clusters (the "best K") which specify the essential clustering structures of the dataset [7]. Secondly, while considering the fixed K number of clusters, it is concerned with how to assess the quality of varied partition schemes being produced by the diverse clustering algorithms for some dataset [8].

With regards to the numerical data, the clusters' density and geometry are mostly used to validate the clustering structure. The density-based methods are naturally applied into the clustering when the distance function is provided for the numerical data [9]. Therefore, the density concepts and the distance functions are critical with regards to ensuring the numerical clustering results are validated [10]. The different visualization-based and statistical cluster validation methods which are based on the density property and geometry have been recommended for numerical data. The effectiveness of such cluster validation methods is enhanced by the density distribution and geometry [11]. An excellent example frequently observed in clustering literature includes the assessment of the clustering results of the 2-Dimensional (2D) experimental dataset through visualization. It involves using cluster visualization to test and validate how the clustering results equal the density distribution and geometry of the points [12].

Despite the categorical data lacking the distance meaning, the tactics utilized in cluster validation for numerical data cannot be applied for categorical data. The general distance functions are mostly non-spontaneous and irrelevant, due to the absence of practical numerical feature construction/extraction for a particular categorical dataset [13]. The methods have failed to adequately tackle the main issue related to categorical clustering, that manipulates the categorical dataset to establish the best K number of clusters. The traditional cluster validation methods which are based on the density distribution and geometry shape cannot be used to answer the question as the categorical data lacks the inherent distance function [14].

The researchers sought to explore entropy property of the categorical data then recommend a BkPlot technique that can help to determine "best Ks" as a set of the candidate [15]. Furthermore, a hierarchical clustering algorithm is used to get experimental results which prove the method known as HierEntro can successfully obtain the significant clustering structures [16]. Available studies regarding categorical clustering have largely focused on adding knowledge on algorithms only. In this case, the researchers sought to identify the best Ks for categorical data clustering by developing an entropy-based cluster validation method. The experimental outcomes indicate that the strategy taken can successfully establish significant clustering structures [17]. The method proposes analyzing the "Entropy Characteristic Graph (ECG)" to establish the best Ks. Besides, the Entropy Characteristic Graph (ECG) can be used to characterize the clustering structure of categorical data [18]. Additionally, the significant points located on the ECG can be conveniently be found using the Best-K plot (BkPlot) [19]. Diverse algorithms usually produce the BkPlot but they may perform differently with regards to identifying the significant clustering structures.
It is evident that the HierEntro which is an entropy-based agglomerative hierarchical algorithm is used to get a comprehensive BkPlot for experimental data sets in comparison to other types of entropy-based algorithms namely the Cool cat and Monte-Carlo algorithm [20]. Furthermore, high clustering results in relation to entropy criterion can also be obtained using the HierEntro. Consequently, with regards to categorical datasets, it is evident that combining HierEntro algorithm and the BkPlot validation method to analyze their significant clustering structures [21].

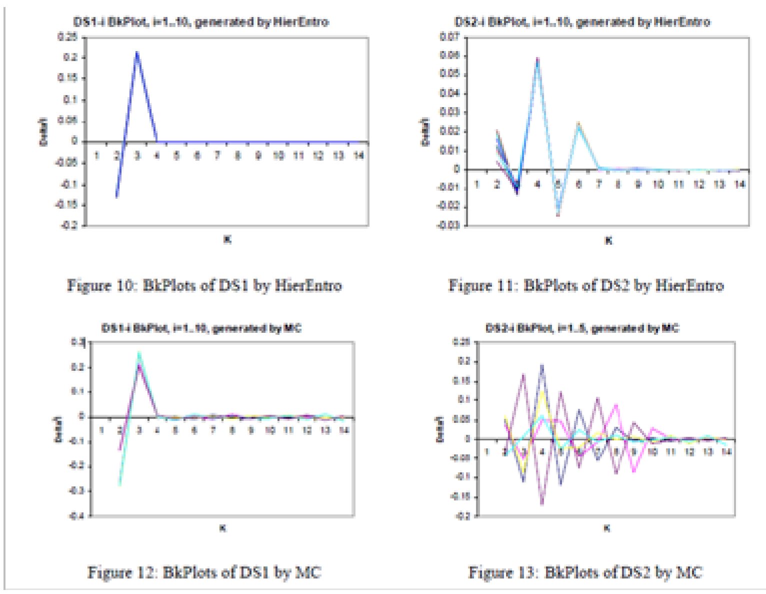

Figure 1: Different BkPlots based on various clustering methods [15].

References

1. Chen K, Liu L (2004) The 'Best K' for Entropy-Based Categorical Data Clustering. Information Visualization 3: 257-270.
2. Aggarwal CC, Procopiuc C, Yu PS (2002) Finding Localized Associations in Market Basket Data. IEEE Transactions on Knowledge and Data Engineering 14: 51-62.
3. Swathi M (2017) Drug Prediction of Cancer Genes Using SVM. Enliven: Pharmacovigilance and Drug Safety 4: 001.
4. Periklis Andritsos, Panayiotis Tsaparas , Renee J. Miller, Kenneth C. Sevcik (2004) LIMBO: Scalable Clustering of Categorical Data. International Conference on Extending Database Technology 123-146.
5. Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jorg Sander (1999) OPTICS: Ordering Points to Identify the Clustering Structure. ACM Sigmod Record 28: 49-60.
6. Steven Noel (2002) Modern Intrusion Detection, Data Mining, and Degrees of Attack Guilt. Applications of Data Mining in Computer Security 1-31.
7. David Gibson, John Kleinberg, Prabhakar Raghavan (2000) Clustering Categorical Data: An Approach Based on Dynamical Systems. The VLDB Journal 8: 222-236.
8. Swathi M (2018) Enhancement of K-Mean Clustering for Genomics of Drugs. Enliven: J Genet Mol Cell Biol 5: 001.
9. Daniel Barbara, Yi Li, Julia Couto (2002) COOLCAT: An Entropy-Based Algorithm for Categorical Clustering. Proceedings of the Eleventh International Conference on Information and Knowledge Management: 582-589.
10. Tao L, Sheng Ma, Mitsunori (2004) Entropy-Based Criterion in Categorical Clustering. In Proceedings of the Twenty-First International Conference on Machine Learning 8: 68.
11. Andreas D. Baxevanis, Francis Ouellette BF (2004) Bioinformatics: APractical Guide to the Analysis of Genes and Proteins. 504.
12. Baulieu FB (1997) Two Variant Axiom Systems for Presence/Absence Based Dissimilarity Coefficients. Journal of Classification 14: 159-170.

13. Andreas D. Baxevanis, Francis Ouellette BF (2004) Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.
14. Bock HH (2009) Probabilistic Aspects in Cluster Analysis. Conceptual and Numerical Analysis of Data. 12-44.
15. Gilles Celeux, Gerard Govaert (1991) Clustering Criteria for Discrete Data and Latent Class Models. Journal of classification 8: 157-176.
16. Keke Chen, Ling Liu (2004) VISTA: Validating and Refining Clusters via Visualization. Information Visualization 3: 257-270.
17. Swathi M (2017) Clustering Enhancement Using Similarity Indexing to Reduce Entropy. Enliven: Bioinform 4: 001.
18. Chun-Hung C, Ada WF, Yi Zhang (1991) Entropy-Based Subspace Clustering for Mining Numerical Data. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 84-93.
19. Thomas M. Cover, Joy A. Thomas (2001) Entropy, Relative Entropy and Mutual Information. Elements of Information Theory 3, no. 2: 1-55.
20. Inderjit S. Dhilon, Subramanyam Mallela, Dharmendra S. Modha (2013) Information-Theoretic Co-Clustering. ACM SIGKDD International Conference On Knowledge Discovery and Data Mining: 89-98.
21. Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis (2002) Clustering Validity Checking Methods: Part II. ACM Sigmod Record 31: 19-27.
22. Ying Z, George K (2004) Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. Machine Learning 55: 311-331.
23. Zhexue H (1997) A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. DMKD 3: 34-39.
24. Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis (2002) Cluster Validity Methods: Part I. ACM Sigmod Record 31: 40-45.
25. Matthew B (2009) An Entropic Estimator for Structure Discovery. In Advances in Neural Information Processing Systems 4: 723-729.