International Journal of Engineering Innovation & Research
Volume 1, Issue 6, ISSN: 2277 – 5668

Outlier Finding from Clusters using Categorical Data

Anamika Malaiya
AISECT University Bhopal

Pradeep Chouskey
Deptt. of Computer Engineering
TIT, Bhopal

Gajendra Vaiker
AISECT University Bhopal

***Abstract* – A New Categorical Data algorithm which combines a systematic method for finding initial centroids and an efficient way for assigning data points to clusters. This method ensures the entire process generate cluster without sacrificing the accuracy. It always generate good cluster by reducing the mean square error.**

Our clustering algorithm serves as a good benchmark to monitor the progression of student's performance in institute. It also enhances the decision making by academic planners to monitor the student's performance semester by semester by improving on the future academic results in the subsequence academic session.

***Keywords* – Outlier Finding, Clusters, Algorithm.**

I. INTRODUCTION

Clustering [17] is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering.

A simple example of clustering would be the clustering that most people perform when they do the laundry - grouping the permanent press, dry cleaning, whites and brightly colored clothes is important because they have similar characteristics. And it turns out they have important attributes in common about the way they behave (and can be ruined) in the wash. To "cluster" your laundry most of your decisions are relatively straightforward. There are of course difficult decisions to be made about which cluster your white shirt with red stripes goes into (since it is mostly white but has some color and is permanent press). When clustering is used in business the clusters are often much more dynamic - even changing weekly to monthly and many more of the decisions concerning which cluster a record falls into can be difficult.

The business user can get a quick high level view of what is happening within the cluster. Once the business user has worked with these codes for some time they also begin to build intuitions about how these different customers clusters will react to the marketing offers particular to their business. For instance some of these clusters may relate to their business and some of them may not. But given that their competition may well be using these same clusters to structure their business and marketing offers it is important to be aware of how you customer base behaves in regard to these clusters.

II. CONCLUSION

Our clustering algorithm serves as a good benchmark to monitor the progression of student's performance in institute. It also enhances the decision making by academic planners to monitor the student's performance semester by semester by improving on the future academic results in the subsequence academic session.

REFERENCES

[1] Dechang Pi, Xiaolin Qin and Qiang Wang, "Fuzzy Clustering Algorithm Based on Tree for Association Rules", International Journal of Information Technology, vol.12, No. 3, 2006.

[2] Fahim A.M., Salem A.M., "Efficient enhanced k-means clustering algorithm", Journal of Zhejiang University Science, 1626 – 1633, 2006.

[3] Fang Yuag, Zeng Hui Meng, "A New Algorithm to get initial centroid", Third International Conference on Machine Learning and cybernetics, Shanghai, 26-29 August,1191 – 1193, 2004.

[4] Friedrich Leisch1 and Bettina Gr un2, "Extending Standard Cluster Algorithms to Allow for Group Constraints", Compstat 2006, Proceeding in Computational Statistics, Physica verlag, Heidelberg, Germany,2006

[5] J. MacQueen, "Some method for classification and analysis of multi varite observation", University of California, Los Angeles, 281 – 297.

[6] Maria Camila N. Barioni, Humberto L. Razente, Agma J. M. Traina, "An efficient approach to scale up k-medoid based algorithms in large databases", 265 – 279.

[7] Michel Steinbach, Levent Ertoz and Vipin Kumar, "Challenges in high dimensional data set", International Conference of Data management, Vol. 2,No. 3, 2005.

[8] Parsons L., Haque E., and Liu H., "Subspace clustering for high dimensional data: A review", SIGKDD, Explor, Newsletter 6, 90 -105, 2004.

[9] Rui Xu, Donlad Wunsch, "Survey of Clustering Algorithm", IEEE Transactions on Neural Networks, Vol. 16, No. 3, may 2005.

[10] Sanjay garg, Ramesh Chandra Jain, "Variation of k-mean Algorithm: A study for High Dimensional Large data sets", Information Technology Journal5 (6), 1132 – 1135, 2006.

[11] Vance Febre, "Clustering and Continues k-mean algorithm", Los Alamos Science, Georgain Electonics Scientific Journal: Computer Science and Telecommunication, vol. 4, No.3, 1994.

[12] Zhexue Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining".

[13] Prof. Brian D. Ripley, "Study of the pure interaction dataset with CART algorithm", Professor of Applied Statistics

[14] Brin, S., Motwani, R., Ullman Jeffrey D., and Tsur Shalom. Dynamic itemset counting and implication rules for market basket data. SIGMOD. 1997.