International Scientific Journal of Contemporary Research in Engineering, Science and Management (ISJCRESM)
Vol. 1, Issue 1, March-May 2016. ISSN : 2456–1134

Exploring Clustering Methods in Data Mining

P Seetha Subha Priya
Assistant Professor, Kongu Engineering College, Perundurai-638052

Abstract

Partitioning a large set of objects into homogeneous clusters is a fundamental operation in data mining. The k-means algorithm is best suited for implementing this operation because of its efficiency in clustering large data sets. However, working only on numeric values limits its use in data mining because data sets in data mining often contain categorical values. In this research, present an algorithm, called k-modes, to extend the k-means paradigm to categorical domains and compare it with the original k-means clusters. Introduce new dissimilarity measures to deal with categorical objects, replace means of clusters with modes, and use a frequency based method to update modes in the clustering process to minimize the clustering cost function.

Keywords: k-means, data mining, homogeneous clusters

1.Introduction

Generally, data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

The objective of this paper is to evaluate the performance of clustering techniques between k-means clustering and extended k-modes clustering, developed and implemented to cluster data sets required in number of data mining applications (refer to LIC). The evaluation is done by partitioning very large heterogeneous sets of objects into a number of smaller and more manageable homogeneous subsets. This can be more easily modeled and analyzed, and detecting under represented concepts, e.g., evaluation of fraud in a large number of insurance claims.

1. Clustering Methods

k-means: The k-means algorithm [14, 1] is built upon four basic operations

(1) selection of the initial k means for k clusters,
(2) calculation of the dissimilarity between an object and the mean of a cluster,
(3) allocation of an object to the cluster whose mean is nearest to the object,
(4) Re-calculation of the mean of a cluster from the objects allocated to it so that the intra cluster dissimilarity is minimized. Except for the first operation, the other three operations are repeatedly performed in the algorithm until the algorithm converges.

The essence of the algorithm is to minimize the cost function

$$E = \sum_{l}^{k} \sum_{i}^{n} y_{i,l} d(X_i, Q_l) \qquad (1)$$

Where n is the number of objects in a data set **X**, $X_i \in \mathbf{X}$, Q_l is the mean of cluster l, and $y_{i\,l}$, is an element of a partition matrix $\mathbf{Y_{n\,x\,l}}$ n as in (Hand 1981). d is a dissimilarity measure usually defined by the squared Euclidean distance. There exist a few variants of the k-means algorithm which differ in selection of the initial k means, dissimilarity calculations and strategies to calculate cluster means [1, 4]. The sophisticated variants of the k-means algorithm include the well-known ISODATA algorithm [2] and the fuzzy k-means algorithms [20]. Most k-means type algorithms have been proved convergent [14, 3].

The k-means algorithm has the following important properties.

1. It is efficient in processing large data sets. The computational complexity of the algorithm is **O**(tkmn), where m is the number of attributes, n is the number of objects, k is the number of clusters, and t is the number of iterations over the whole data set. Usually, k, m, t << n. In clustering large data sets the k-means algorithm is much faster than the hierarchical clustering algorithms whose general computational complexity is O (n^2) [16].

2. It often terminates at a local optimum [14, 22]. To find out the global optimum, techniques such as deterministic annealing [19, 12] and genetic algorithms [6] can be incorporated with the k-means algorithm.

3. It works only on numeric values because it minimizes a cost function by calculating the means of clusters.

4. The clusters have convex shapes [1]. Therefore, it is difficult to use the k-means algorithm to discover clusters with non-convex shapes.

One difficulty in using the k-means algorithm is to specify the number of clusters. Some variants like ISODATA include a procedure to search for the best k at the cost of some performance.

The k-means algorithm is best suited for data mining because of its efficiency in processing large data sets. However, working only on numeric values limits its use in data mining because data sets in data mining often have categorical values. Development of the k-modes algorithm to be discussed in the next

section was motivated by the desire to remove this limitation and extend its use to categorical domains. The performance evaluation of k-means cluster and extended k-modes cluster is also done through experimentation on partitioning large databases.

The k-modes algorithm is a simplified version of the k-prototypes algorithm. In this algorithm have made three major modifications to the k-means algorithm, i.e.,

- using different dissimilarity measures,
- replacing k means with k modes, and
- using a frequency based method to update modes.

.

k-Modes Algorithm: The k-modes algorithm consists of the following steps:

1. Select k initial modes, one for each cluster.

2. Allocate an object to the cluster whose mode is the nearest to it according to d. Update the mode of the cluster after each allocation according to the Theorem.

3. After all objects have been allocated to clusters, retest the dissimilarity of objects against the current modes. If an object is found such that its nearest

mode belongs to another cluster rather than its current one, reallocate the object to that cluster and update the modes of both clusters.

4. Repeat 3 until no object has changed clusters after a full cycle test of the whole data set.

Like the k-means algorithm the k-modes algorithm also produces locally optimal solutions that are dependent on the initial modes and the order of objects in the data set. The initial mode selection methods can improve the clustering results. In our current implementation of the k-modes algorithm include two initial mode selection methods. The first method selects the first k distinct records from the data set as the initial k modes. The second method is implemented to make the initial modes diverse, which can result in better clustering results.

Let X, Y is two categorical objects described by m categorical attributes. The dissimilarity measure between X and Y can be defined by the total mismatches of the corresponding attribute categories of the two objects. The smaller the number of mismatches is, the more similar the two objects. Formally,

$$d(X, Y) = \sum_{j=1}^{m} \delta(x_j, y_j) \qquad (2)$$

where

$\delta(x_j, y_j) = d(X,Y)$ gives equal importance to each category of an attribute.

CONCLUSION & FUTURE DEVELOPMENT

The main benefit of the k-means algorithm in data mining applications is its efficiency in clustering large data sets. However, its use is limited to numeric values. The k-modes algorithm presented in this thesis has removed this limitation whilst preserving its efficiency. The k-modes algorithm has made the following extensions to the k-means algorithm in improving the performance of cluster formation

1. Replacing means of clusters with modes,

2. Using new dissimilarity measures to deal with categorical objects, and

3. Using a frequency based method to update modes of clusters.

These extensions allow us to use the k-means paradigm directly to cluster categorical data without need of data conversion. Another advantage of the k-modes algorithm is that the modes give characteristic descriptions of clusters. These descriptions are very important to the user in interpreting clustering results. Because data mining deals with very large data sets, scalability is a basic requirement to the data mining algorithms.

The future work plan is to develop and implement a parallel k-modes algorithm to cluster data sets with millions of objects. Such an algorithm is required in a number of data mining applications, such as partitioning very large heterogeneous sets of objects into a number of smaller and more manageable homogeneous subsets that can be more easily modeled and analyzed, and detecting under-represented concepts, e.g., fraud in a very large number of insurance claims.

References

[1] Anderberg, M. R. (1973) Cluster Analysis for Applications, Academic Press.

[2] Ball, G. H. and Hall, D. J. (1967) A Clustering Technique for Summarizing Multivariate Data, Behavioral Science, 12, pp. 153-155.

[3] Bezdek, J. C. (1980) A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2(8), pp. 1-8.

[4] Bobrowski, L. and Bezdek, J. C. (1991) c-Means Clustering with the l1 and l□□Norms, IEEE Transactions on Systems, Man and Cybernetics, 21(3), pp. 545-554.

[5] Fisher, D. H. (1987) Knowledge Acquisition Via Incremental Conceptual Clustering, Machine Learning, 2(2), pp.139-172.

[6] Goldberg, D. E. (1989) Genetic Algorithms in Search, Optimisation, and Machine Learning, Addison-Wesley. Gowda, K. C. and Diday, E. (1991) Symbolic Clustering Using a New Dissimilarity Measure, Pattern Recognition, 24(6), pp. 567-578.

[7] Gower, J. C. (1971) A General Coefficient of Similarity and Some of its Properties, BioMetrics, 27, pp. 857-874.

[8] Greenacre, M. J. (1984) Theory and Applications of Correspondence Analysis, Academic Press.

[9] Hand, D. J. (1981) Discrimination and Classification, John Wiley & Sons.

[10] Huang, Z. (1997) Clustering Large Data Sets with Mixed Numeric and Categorical Values, In Proceedings of The First Pacific-Asia Conference on Knowledge Discovery and Data Mining, Singapore, World Scientific.

[11] Jain, A. K. and Dubes, R. C. (1988) Algorithms for Clustering Data, Prentice Hall.

[12] Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983) Optimisation by Simulated Annealing, Science, 220(4598), pp.671-680.

[13] Kodratoff, Y. and Tecuci, G. (1988) Learning Based on Conceptual Distance, IEEE Transactions on Pattern Analysis and Machine Intelligence, 10(6), pp. 897-909.

[14] MacQueen, J. B. (1967) Some Methods for Classification and Analysis of Multivariate Observations, In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281-297.

[15] Michalski, R. S. and Stepp, R. E. (1983) Automated Construction of Classifications: Conceptual Clustering Versus Numerical Taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence, 5(4), pp. 396- 410.

[16] Murtagh, F. (1992) Comments on “Parallel Algorithms for Hierarchical Clustering and Cluster Validity”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(10), pp. 1056-1057.

[17] Murthy, C. A. and Chowdhury, N. (1996) In Search of Optimal Clusters Using Genetic Algorithms, Pattern Recognition Letters, 17, pp. 825-832.

[18] Ralambondrainy, H. (1995) A Conceptual Version of the k-Means Algorithm, Pattern Recognition Letters, 16, pp. 1147-1157.

[19] Rose, K., Gurewitz, E. and Fox, G. (1990) A Deterministic Annealing Approach to Clustering, Pattern Recognition Letters, 11, pp. 589-594.

[20] Ruspini, E. R. (1969) A New Approach to Clustering, Information Control, 19, pp. 22-32.

[21] Ruspini, E. R. (1973) New Experimental Results in Fuzzy Clustering, Information Sciences, 6, pp. 273-284.

[22] Selim, S. Z. and Ismail, M. A. (1984) K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality, IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(1), pp. 81-87.

[23] Shafer, J., Agrawal, R. and Metha, M. (1996) SPRINT: A Scalable Parallel Classifier for Data Mining, In Proceedings of the 22nd VLDB Conference, Bombay, India, pp. 544-555.

Author Biography

Ms P Seetha Subha Priya has completed Master of Computer Application. Her research expertise covers Image mining, data mining, cloud computing and big data. She is currently working as Assistant Professor in Kongu Engineering College.