ISSN:0975-9646

Fuzzy based clustering of High Dimensional Datasets

M.Pavani[1] M.Pravallika[2], M. SaiChandan[3], M.Lokesh[4], A.snehitha[5]
Lendi Institute of Science and Technology, Visakhapatnam, India

Abstract. Now-A-Days Fuzzy Association Rule Mining (ARM) on high dimensional data is a high topic in many fields of data mining. There are various data repositories such as relational database, transactional data base etc., which stores the data in different dimensions. Applying the association rules on high dimensional data is challenging issue. The main strength of fuzzy ARM is completeness but the strength is overcome by the major drawback i.e., to handle the large number of datasets. Fuzzy ARM is traditional process it is not much efficient. So, FAR-HD process the Fuzzy frequent item sets in a DFS manner.

In our paper we are particularly concentrated on problems of Fuzzy ARM on high dimensional data and classification of the high dimensional data.

Keywords: Fuzzy Association Rule in High Dimensional data sets (FAR-HD), SURF (Speed up Robust Feature) Algorithm, Fuzzy C-mean, Fuzzy Association Classification (FAC).

1.INTRODUCTION

In the Data mining, since from many years mining Association rules is an extreme focused research area. Many association rule algorithms like Apriori exist for mining association rules and frequent item sets based on given support and confidence. Association rule mining with fuzzy logic is used to further data mining tasks for classification and clustering. Traditional Fuzzy ARM algorithms are failed to mine rules from high-dimensional data efficiently, since those algorithms are meant to deal with relatively much less number of attributes. Fuzzy ARM with high-dimensional data is a challenging issue with respect to memory constraints. To overcome the problem we used a fast and efficient Fuzzy ARM algorithm FAR-HD especially for large high-dimensional datasets.

The SURF algorithm is similar to the SIFT, but it uses a different scheme and should provide better results: it works much efficient than SIFT. Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are not similar. In fuzzy clustering every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. Fuzzy classification is the process of grouping elements into a fuzzy set whose membership function is defined by the truth value of a fuzzy propositional function.The advantages of associative classifiers are that frequent item sets capture all dominant relationships between items in a dataset, and that they deal only with statistically significant associations.

2.RELATED WORKS

In previous they worked on some of the algorithms such as

2.1 Fuzzy Apriori: To implement the fuzzy association rule mining procedure, we used a modified version of the Apriori algorithm. The algorithm is much more economical by treating negative items as new database attributes. It is also very much preferable to the approach for mining negative association rules which involves the costly generation of infrequent as well as frequent item sets. Regarding the quality of the mined association rules, we observed that most of them are negative.

2.2 Fuzzy Cluster-Based Association Rules (FCBAR): The FCBAR method is to create cluster tables by scanning the database once and cluster the data until the kth cluster table, where k is the length of a record. Not only that the fuzzy large item sets are generated by contrasts with the partial cluster tables. It needs less amount of time to perform data scans and requires less contrast. Experiments with the real-life database show that FCBAR outperforms fuzzy Apriori like algorithm, a well–known and widely used association rules algorithm. FCBAR method for discovering the fuzzy large item sets, it requires a single scan of the transaction database, in contrast with the partial cluster tables. Not only this it tends to considerable amount of data reducing the time needed to perform data scan and requiring less contrast, but also mined results are ensured as correct.

2.3 Mining Fuzzy Association Rules in Large High-Dimensional Datasets[1]: Fuzzy Association Rule Mining (ARM) has been extensively used in relational or transactional datasets having less to medium number of attributes. The mined fuzzy association rules are not only used for manual analysis by database administrator, but difficult to drive further mining tasks like classification and clustering which automate decision-making. FAR-HD which is a Fuzzy ARM algorithm designed specifically for large high-dimensional datasets. FAR-HD processes fuzzy frequent item sets in a DFS manner . It also uses a byte-vector representation of tidlists, it stored in the main memory in a compressed form. Additionally, FAR-HD uses Fuzzy Clustering to convert each numerical vector of the original input dataset to a fuzzy cluster based representation, which is ultimately used for the actual Fuzzy ARM process. FAR-HD has been compared experimentally with Fuzzy Apriori ,in which FAR-HD is best.

2.4 SIFT (Scale - Invariant Feature Transform): Scale-invariant feature transform (SIFT) is an algorithm in computer vision to detect and describe local features in images. SIFT key points of objects are first extracted from a set of reference images and stored in a database. If two images are compared individually then each feature from the new image to this database and finding candidate matching features based on Euclidean distance of their feature vectors. From the set of matches, subsets of key points which are agree on the object , location, scale, and orientation, the new image are identified to filter out good matches.

3.PROPOSED WORK

In this section, consider the two images such that image2 is similar that of image1.Now we have to find the interesting points from that image by applying SURF (Speed Up Robust Feature) algorithm. After that we perform the clustering by using Fuzzy C-mean. FAR-HD is the process which is used for high dimensional datasets to generate frequent item sets. By using FAC (Fuzzy Association Classification) we did the classification of the above clusters

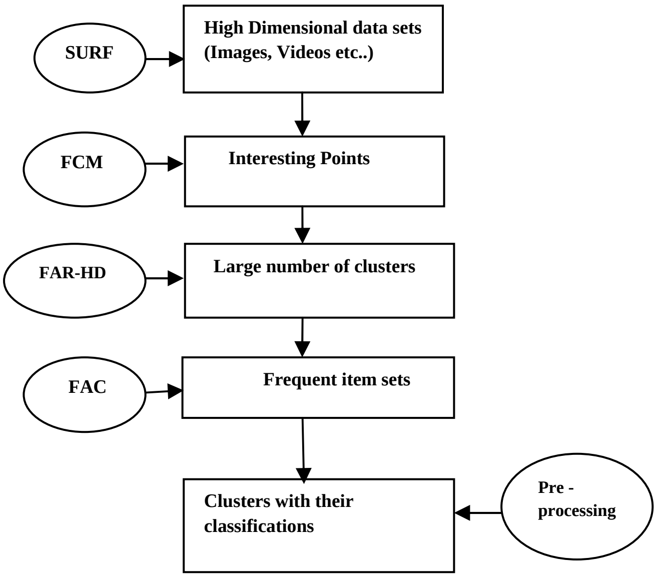


Fig.1. System Architecture

3.1Representation of SURF[3]: In this section first we consider high dimensional data sets such as images, videos etc. by applying SURF algorithm we get the interesting points. Traditionally we use SIFT to get the Interesting points. But the SURF is extension of the SIFT which is several times faster than the SIFT.

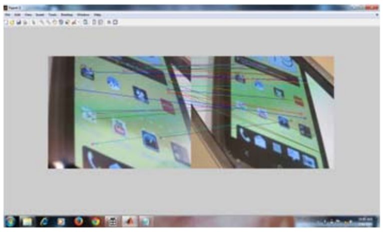
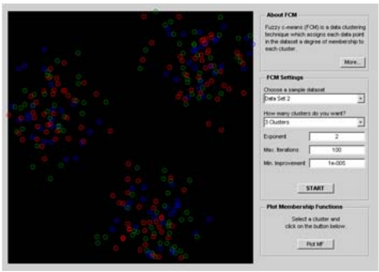


Fig.2. Interesting points using SURF algorithm

3.2Fuzzy C-Mean: The Fuzzy C-Mean (FCM) algorithm is commonly used for clustering the performance of the FCM algorithm depends on the selection of initial cluster. If the initial cluster is good then the final cluster can be found very quickly and the processing time can be drastically reduced. It is a data clustering technique in which a dataset is grouped into n clusters with every data point in the dataset belonging to every cluster to a certain degree. Consider an example, a certain data point that lies close to the center of a cluster will have a high degree of belonging to that cluster and another data point that lies far away from the center of a cluster will have a low degree of belonging to that cluster.



About FCM

Fuzzy c-means (FCM) is a data clustering technique which assigns each data point in the dataset a degree of membership to each cluster.

FCM Settings

Choose a sample dataset: Data Set 2

How many clusters do you want? 3 Clusters

Exponent: 2

Max. Iterations: 100

Min. Improvement: 1e-005

Plot Membership Functions

Select a cluster and click on the button below.

Fig.3. Fuzzy c-mean clusters

3.3 Fuzzy Association Rule in High Dimensional Datasets (FAR-HD): FAR-HD is a process which is used for high dimensional data sets to generate frequent item sets. FAR-HD[2] uses Fuzzy Clustering to convert each the original input dataset to a fuzzy based clustering representation, which is ultimately used for the actual Fuzzy ARM process. By using the FAR-HD we can mine the data from large datasets to generate the frequent item sets[5].

3.4 Fuzzy Association Classification (FAC)[4]: Fuzzy classification is the process of grouping elements into a fuzzy set whose membership function is defined by the truth value of a fuzzy propositional function. It is used for the Classification purpose. After clustering we did the classification for that clusters. The frequent item sets which are generated are classified by using FAC.

3.5 Pre-processing techniques: Data pre-processing is often neglected but important step in data mining. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the accessing of data is more difficult. We apply some of the pre-processing techniques to remove the noisy, incomplete and inconsistent data.

CONCLUSION

In this project association rule mining was studied for very large high-dimensional data in the image domain. The framework by name “Fuzzy Associative Classifier using images” was implemented by using matlab. The SURF algorithm is capable of finding interesting points from two similar image dataset and we are using some of the clustering and classification techniques to the datasets to increase the speed and efficiency in large high-dimensional data sets.

REFERENCES

1. A. Mangalampalli and V. Pudi, “FAR-HD: A Fast And Efficient Algorithm For Mining Fuzzy Association Rules In Large High-Dimensional Datasets” in Fuzzy Systems (FUZZ), 2013 IEEE.
2. A. Mangalampalli and V. Pudi, “FAR-miner: a fast and efficient algorithm for fuzzy Association rule mining,” IJBIDM, vol. 7, no. 4, pp. 288–317, 2012.
3. H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, “Speeded-up robust features (SURF),” Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346–359, 2008.
4. F. A. Thabtah, “A review of associative classification mining,” Knowledge Eng. Review,vol. 22, no. 1, pp. 37–65, 2007.
5. R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in largedatabases”, in VLDB, 1994, pp. 487–499.
6. H. H. Malik and J. R. Kender, “High quality, efficient hierarchical document clustering using closed interesting itemsets,” in ICDM, 2006, pp. 991–996.
7. M. Delgado, N. Marn, D. S´anchez, and M. A. V. Miranda, “Fuzzy association rules: General model and applications,” IEEE Transactions on Fuzzy Systems, vol. 11, pp. 214–225, 2003.
8. B. C. M. Fung, K. Wang, and M. Ester, “Hierarchical document clustering using frequent Item sets,” in SDM, 2003.

10. Ashish Mangalampalli, Vineet Chaoji, Subhajit Sanyal “I-FAC: Efficient Fuzzy Associative Classifier for Object Classes in Images” in Fuzzy Systems (FUZZ), 2010 IEEE.