AN EFFECTIVE ALGORITHMIC APPROACH FOR CLUSTERING AND BOUNDARY ANALYSIS IN HETEROGENEOUS DATABASE APPLICATIONS

Payal Joshi*

Arvind Selwal**

Anuradha Sharma***

Abstract: *Clustering or Categorization is mandatory in every Knowledge Discovery in Databases (KDD) applications. Classical clustering methods or algorithms works based on similarity measure. Similarity based approaches determines the association/rejection of the object from other objects. These approaches are not efficient for sensitive information including cyber security, business problems, medical sciences and many other. Existing algorithms makes use of numerical data/parameters and then find association or similarity. Existing algorithms fails in case of non-numerical or independent data. Existing approaches add or remove the tuple from the existing clusters. Major limitation in the algorithm is single scan. In single scan of tuples or objects the reliability of incoming training dataset should not be measured. In this paper, a new approach to form clusters and to detect the outliers is devised and proposed.*

Keywords: *Clustering, Data Mining, Fitness Function, Outlier Detection, Similarity measure*

*Department Of Computer Science Engineering, Ambala College of Engineering & Applied Research, Devasthali, Ambala, India

INTRODUCTION

Data mining has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated AI methods are also used. Data mining requires identification of a problem, along with collection of data that can lead to better understanding, and computer models to provide statistical or other means of analysis [8].

Data comes in, possibly from many sources. It is integrated and placed in some common data store. Part of it is then taken and pre-processed into a standard format. This 'prepared data' is then passed to a data mining algorithm which produces an output in the form of rules or some other kind of 'patterns' [7].

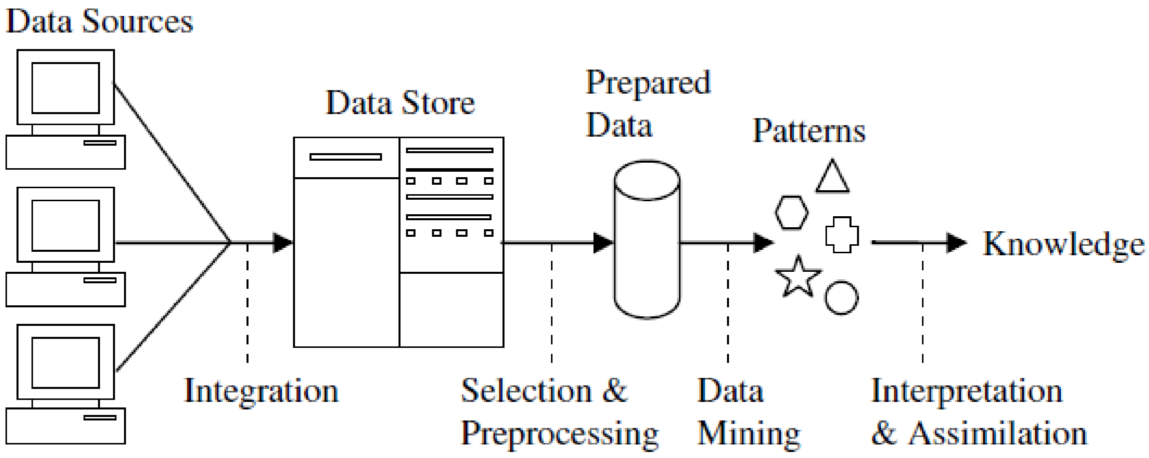


Figure 1: The Knowledge Discovery Process [7]

Clustering is an important KDD technique with numerous applications, such as marketing and customer segmentation. Clustering typically groups data into sets in such a way that the intra-cluster similarity is maximized and while inter-cluster similarity is minimized [1]. Clustering is an unsupervised learning. Clustering algorithms examines data to find groups of items that are similar. For example, an insurance company might group customers according to income, age, types of policy purchased, prior claims experience in a fault diagnosis application, electrical faults might be grouped according to the values of certain key variables [7]

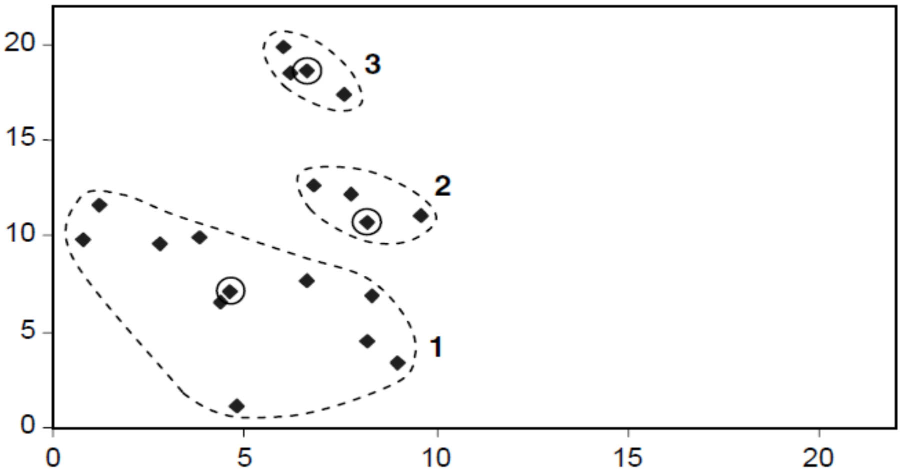


Figure 2: Clustering of Data [7]

Most previous clustering algorithms focus on numerical data whose inherent geometric properties can be exploited naturally to define distance functions between data points. However, much of the data existed in the databases is categorical, where attribute values can't be naturally ordered as numerical values. Due to the special properties of categorical attributes, the clustering of categorical data seems more complicated than that of numerical data [3]. To overcome this problem, several data-driven similarity measures have been proposed for categorical data. The behaviour of such measures directly depends on the data [4].

Outlier detection refers to the problem of finding patterns in data that do not conform to expected normal behaviour [9].

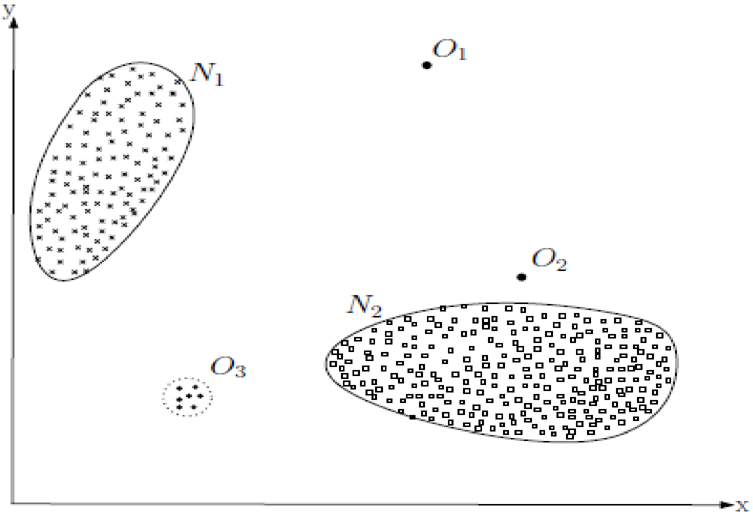


Figure 3: Outliers in two-dimensional dataset [9]

Figure 3 illustrates outliers in a two=dimensional dataset. The data has two normal regions, N1 and N2. O1 and O2 are two outlying instances while O3 is an outlying region. The outlier instances are the ones which do not lie within the normal regions [9].

LITERATURE SURVEY AND PARADIGMS

Zengyou He et al [1] proposed Squeezer algorithm, a clustering algorithm for categorical data. It takes n tuples as input and produces clusters as output. Initially, the first tuple is read and cluster structure is constructed. Read subsequent tuples one after another. For each tuple, compute its similarities with all existing clusters. Select the largest similarity value. If the largest similarity value is greater than threshold ‘s’, the tuple is inserted into the existing cluster else new cluster is formed. The Cluster Structure (CS) will be updated for each iteration. Squeezer algorithm makes use of Cluster Structure which consists of cluster information and summary information.

Zengyou He et al [3] proposed NabSqueezer algorithm, an improved Squeezer algorithm. NabSqueezer algorithm gives more weight to uncommon attribute value matches for finding similarity in similarity computation of Squeezer algorithm. In this algorithm weight of each attribute is precalculated using More Similar Attribute Value Set (MSFVS) method.

Zengyou He et al [2] proposed FindCBLOF Algorithm for detecting outliers. This algorithm computes the value of CBLOF for each record which determines the degree of record’s deviation. This algorithm is efficient for handling large datasets.

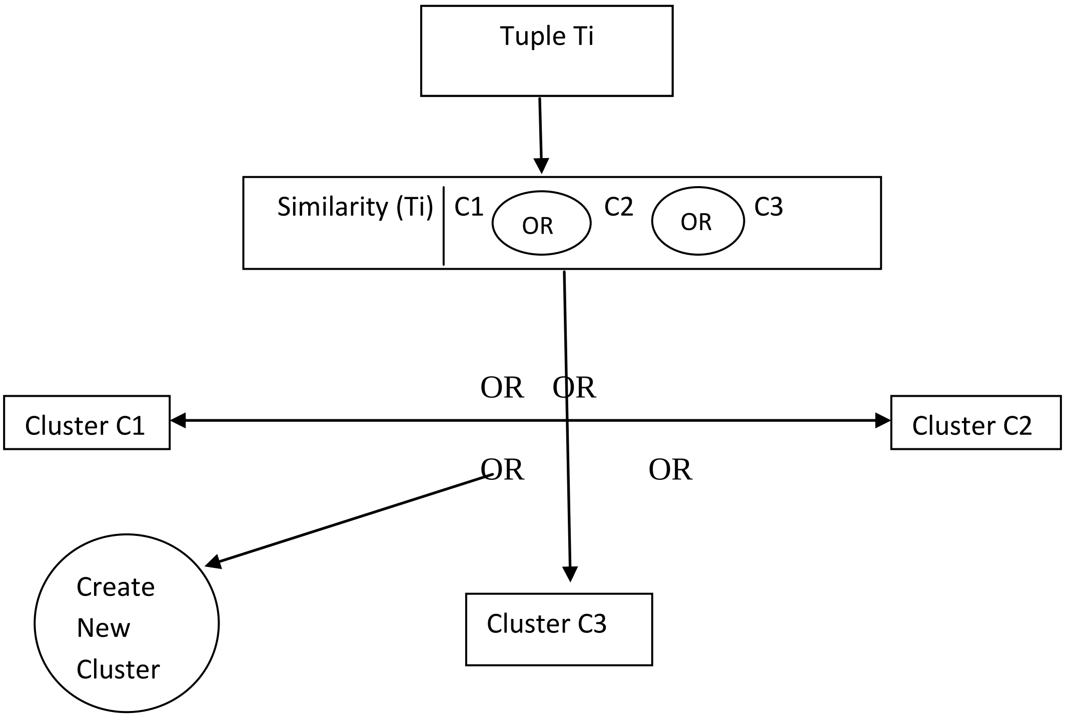
Shyam Boriah et al [4], the author presents a comparative study on number of similarity measures such as Goodall, Occurence Frequency, Overlap, Inverse Occurence Frequency, Burnbay, Gambaryan, Smirnov. In this paper we have studied the performance of a variety of similarity measures in the context of a specific data mining task: outlier detection.

Aditya Desai et al [5], use similarity which are neighbourhood-based or incorporate the similarity computation into the learning algorithm. These measures compute the neighbourhood of a data point but not suitable for calculating similarity between a pair of data instances X and Y.

R.Ranjani et al [6] proposed Enhanced Squeezer algorithm, which incorporates Data-Intensive Similarity Measure for Categorical Data (DISC) in Squeezer Algorithm. DISC measure, cluster data by understanding domain of the dataset, thus clusters formed are not purely based on frequency distribution as many similarity measures do.

PROPOSED APPROACH AND DESCRIPTION

Existing Algorithm repeatedly reads tuples one by one from the dataset. When the first tuple arrives, a new cluster is formed. The consequent tuples are either put in the existing clusters or rejected by the existing clusters to form a new cluster based on the similarity measure between a tuple and a cluster. In existing approach each tuple belongs to one cluster only.



In proposed algorithm, suppose there are n tuples. A fitness value is assigned to each tuple using the fitness function. Based upon this fitness value the tuples will be assigned to the clusters. If the fitness value of the tuple is equal to or nearly equal to the threshold value of the generated set of random clusters then only the tuple will be assigned to the cluster otherwise tuple is assigned to the outlier cluster. If there are many clusters in the outlier cluster then a similarity is calculated among theses clusters and outlier is detected. In this approach, tie can also occur i.e. if a tuple belongs to two clusters then we can arbitrarily assign this tuple to any one cluster.

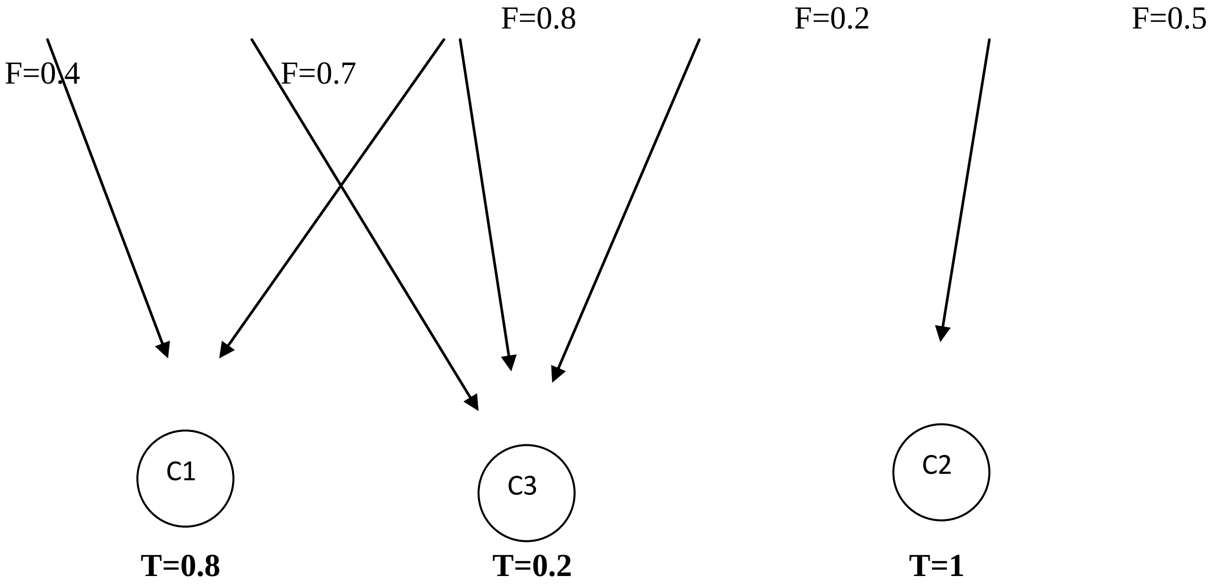


Figure 5: Proposed Approach

T= Threshold

F= Fitness Value

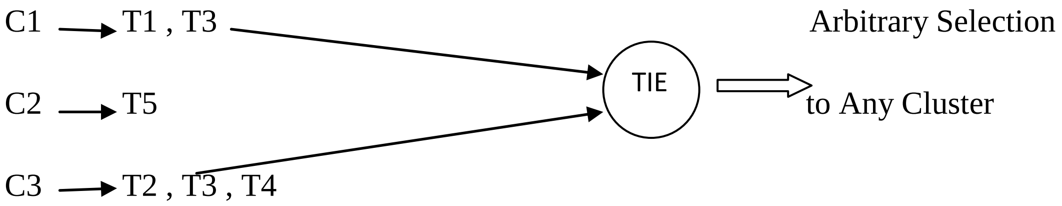


Figure 6 : Analysis of Fitness and Selection

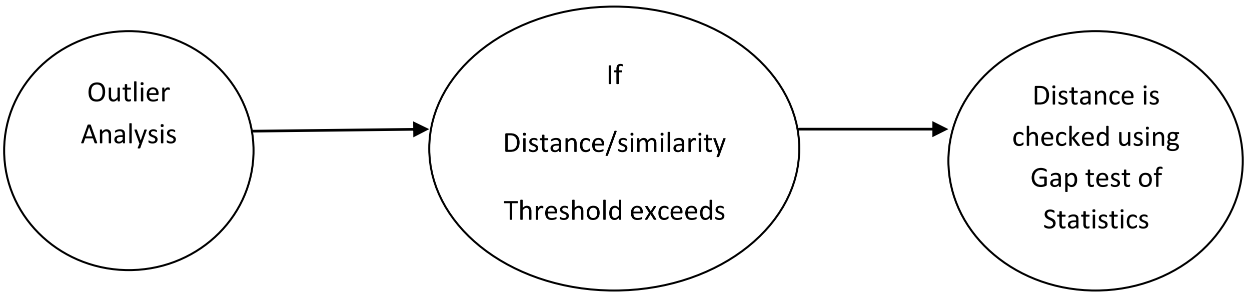


Figure 7: Outlier Detection

Let us consider an example of Employee Database:

ID	DESIGNATION	SALARY	CITY	FITNESS
1	A	15000	X	M
2	B	25000	Y	N
3	B	22000	X	O

Table 1: Employee Database

Suppose the tuples whose fitness value is ‘m’ and ‘n’ belongs to clusters C1 and C2 respectively. The tuple whose fitness value is ‘o’ doesn’t belong to any cluster therefore it is considered as outlier.

CONCLUSION

This paper presents a new approach for clustering and outlier detection. This new approach uses fitness function for the categorization of data. Existing approach uses similarity measure for the same. This approach is better than existing approach because it is multi-directional whereas existing methods uses single directional approach. The proposed approach is more efficient due to multi-directional, the reliability of incoming data set can be measured.

REFERENCES

[1] He Zengyou, Xu Xiaofei, Deng Shenchun, 2002. Squeezer: An Efficient Algorithm for Clustering Categorical Data

[2] He Zengyou, Xu Xiaofei, Deng Shenchun, 2003. Discovering Cluster Based Local Outliers

[3] He Zengyou, Xu Xiaofei, Deng Shenchun, 2006. Improving Categorical Data Clustering Algorithm by Weighting Uncommon Attribute Value Matches

[4] Shyam Boriah, Varun Chandola, Vipin Kumar, 2008. Similarity Measures for Categorical Data: A Comparative Evaluation

[5] Aditya Desai, Himanshu Singh, Vikram Pudi, 2011. DISC: Data-Intensive Similarity Measure for Categorical Data

[6] R.Ranjini, S.Anitha Elavarasi, J.Akilandeswari.2012. Categorical Data Clustering Using Cosine Based Similarity for Enhancing the Accuracy of Squeezer Algorithm

[7] Principles of Data Mining by Max Bramer

[8] Advanced Data Mining Techniques by David L. Olson and Dursun Delen

[9] Varun Chandola, Arindam Banerjee, Vipin Kumar. Outlier Detection: A Survey