ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

THE SCINTILLATION OF ALGORITHMIC APPROACH IN FITNESS BASED CLUSTERING AND OUTLIER ANALYSIS

[1]Payal Joshi *[2] Arvind Selwal*

Department Of Computer Science Engineering

Ambala College of Engineering & Applied Research, Devasthali, Ambala-133101

Ambala City, India

ABSTRACT

Clustering is the process of grouping a set of physical objects into classes of similar objects or homogenous pattern. Clustering groups the data into sets in such a way that the intra-cluster similarity is maximized while the inter-cluster similarity is minimized. This paper presents a new algorithm for clustering and outlier detection for categorical data. An Effective Algorithmic Approach has been developed with multiple parameters and assorted machine learning techniques to fetch the rule and intelligence based results. The projected algorithm makes use of the multilayered approach for the clustering and classification of the data sets in multiple real life applications. The algorithm analyzes the input data set using a dynamic threshold

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

parameter for clustering and rule based classification. Using this technique and multiphase algorithmic approach, efficient and more meaningful clusters are formed

KEYWORDS- Categorical Data**,** Clustering, Data Mining, Fitness Value, Outlier Detection, Similarity Measure

INTRODUCTION

Clustering is to partition a set of objects into cluster so that objects in the same cluster are more similar between them objects from other cluster according to the established criterion [6]. The clustering technique has been extensively studied in many fields such as data mining, pattern recognition, customer segmentation, similarity search and trend analysis [3]. Research community have proposed different clustering algorithm and many are suitable for clustering numerical data. In real world scenario, data in database are categorical in nature, which are raw or unsummarized data, where the attributes cannot be pre-arranged as numerical values. Clustering categorical; data is a major challenge in data mining [6]. In addition, the notion of similarity can differ depending on the particular domain, data set, or task at hand [5].

An outlier in a dataset is defined informally as an observation that is considerably different from the remainders as if it is generated by a different mechanism. Mining for outliers is an important data mining research with numerous applications. Including credit card fraud detection, discovery of criminal activities in electronic commerce, weather prediction, marketing and customer segmentation [2].

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

In the existing approach, Data Intensive Similarity Measure for Categorical Data(DISC) is presented. Enhanced Squeezer works in similar fashion as that of Squeezer algorithm except the similarity measure. DISC captures the semantics of the data without any help from domain experts for defining similarity [5]. The main components of Enhanced Squeezer algorithm which makes use of DISC similarity measure are as follows:

1) Categorical Information Table

2) Similarity Measure

3) Cluster Formation

4) Cluster Validation

Categorical Information Table :- Categorical Information Table is a data structure which serves for the purpose of quick reference for information related to co-occurrence statistics.

Similarity Measure :- In order to group data points into cluster, similarity between two data points must be computed. DISC is used for measuring the similarity among data points.

Cluster Formation :- . 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.

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)

International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

Cluster Validation :- The result of Enhanced Squeezer is evaluated to prove the degree of confidence of the results [6].

LITERTURE REVIEW

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.

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

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

clusters formed are not purely based on frequency distribution as many similarity measures do.

PROPOSED APPROACH

The proposed approach is shown below with the help of the flow diagram. The training dataset is selected either randomly or sequentially from the data warehouse. Then calculate fitness value of each tuple. Further clusters and outliers are detected with the help of filtering module. Here filtering module is the similarity measure function. With the help of similarity measure comparison of cluster and tuple will take place. After the generation of clusters and outlier detection final statistics and report is generated.

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

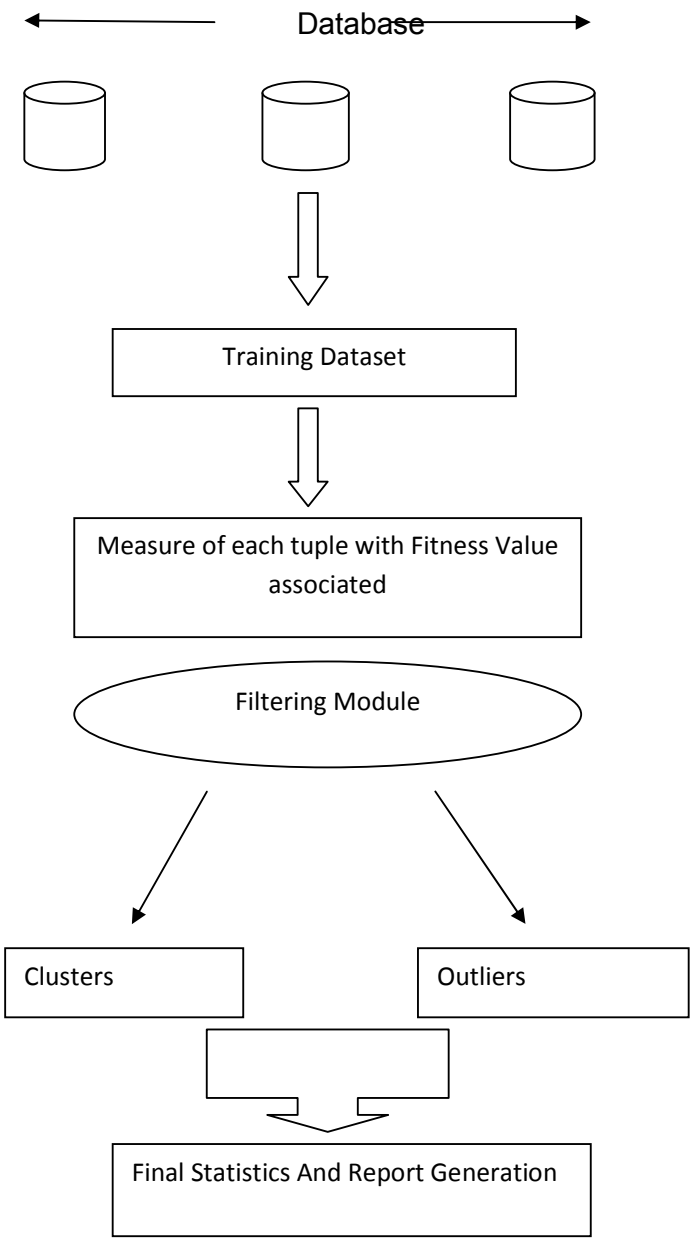

Figure 1:Flow Diagram

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

In the algorithm given below, for clustering our first step will be the generation of data set from a huge data ware house eiher sequentially or randomly. Here, T denotes the set of generated tuples. Then, we will assign fitness value to each tuple in the set T. Here, F_i denotes the fitness value assigned to each tuple. Now, generation of the set of random clusters denoted by C will take place. Further threshold value will be assigned to the generated clusters. Thereafter, tuple T_i will be compared with cluster Ci based on fitness value and associated parameters. If Cluster C_i is null and the fitness value of tuple T_i is null then the algorithm terminates else if tuple T_i is assigned to cluster C_i then assign initial threshold value based on the application otherwise goto step 1.

In the next part of the algorithm, outlier will be detected from the clusters if any. In the first step of outlier detection, data item is read from cluster C_i. If the fitness value F_i does not matches with the threshold value of the cluster C_i to maximum extent then this data item is considered as outlier. If the outliers detected are similar then a new cluster is created of these outliers. Statistics are generated from the above results.

Further, in the third part of the algorithm we will compare the existing approach and the proposed approach and calculate complexity of both the algorithms.

PROPOSED ALGORITHM / PSEUDOCODE

1. CLUSTERING

1. Generate Dataset (Sequentially or Randomly) / Tuple series (from huge data warehouse)

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

$|T|= \{ T_i \mid i \epsilon (1,N) \}$

2. Assign Fitness Value (F_i) to each tuple based on the Acceptance / Rejection of the Data Item for joining the Cluster

$T= \{ T_i[F_i] \mid i \epsilon (1,N) \}$

3. Generate the set of random clusters (if already exists)

$C= \{ C_i \mid i \epsilon (1,N) \}$ and assign Threshold

4. Compare T with C based on Fitness value and Associated Parameters

5. If (C_i == NULL) AND $T_{fitness}$==NULL GoTo Step 7

Else

If (C_i= First Cluster)

Assign initial threshold based on the application

Else

GoTo Step 1

6. End

2. OUTLIER DETECTION

1. Read First / Next Data Item from C_i

If F_i not matching to predefined threshold to maximum extent

Similarity ($F_i \mid F(C_i)$)

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)

International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

Then

Separate the tuple and mark it as Outlier

2. If Similar Outliers Occurs

Then

Create Clusters of these outlier entries

3. Get Statistics

4. End

3. FINAL INVESTIGATION

Generate Final Results from both algorithms and calculate complexity.

CONCLUSION AND FUTURE SCOPE

The proposed algorithm is used for the generation of clusters and detecting outliers. The clusters generated will be high quality clusters of different shapes and sizes. Further association mining rules may be applied. The further course of work shall be to simulate the algorithm on sample data set for clustering and classification. The web based simulated environment shall be used to extract the huge data sets and multiple cluster formation will be accomplished.

REFERENCES

ISSN (Online) : 2319 - 4618
International Journal of Reviews, Surveys and Research (IJRSR)
International Refereed and Indexed Journal for Research Scholars and Practitioners

International Manuscript ID : ISSN23194618-V2I2M1-052013

[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