Data Extension Tool

[1]Amit Kumar Singh, [2]Kartik nigam, [3]Pratiksha Mishra, [4]Gaurav Bajpai

[1,2,3]*Student,* [4]*Faculty*

[1,2,3,4]*PSIT college of engineering, Kanpur*

Abstract-Data mining is an important field of computer science and information technology which involves extraction of some meaningful and relevant information from an existing database or other data sources. Data mining is used in various fields such as business intelligence, artificial intelligence, machine learning etc. The main function of the tool is to provide data mining operations over the data using various algorithms. There are some basic operations in Data Mining, like classification, clustering, handling noisy data etc. that are used very frequently. The project entitled Data extension Tool is a project based on java and it provides convenience to the user to operate over data and carry out basic data mining tasks like smoothing, clustering, association rule mining, normalization and summarization. In this Utility tool ,user selects the operation to be done over the data and the mode of input of data and selects the options which are to be processed and output of the operation are provided. The output can be obtained into a file or console based on users choice.

I. INTRODUCTION

The project Data Extension Tool, provides some basic data mining operations like summarization using five point summary, clustering using k-means algorithm, association rule mining using apriori algorithm, smoothening of noisy data using binning technique and normaliza- tion. The introduction part of the report gives basic description regarding above operations and related algorithms. Following pages attempt to give a complete detail of object-oriented analysis, design, and programming applied to ourproject.

Apriori Algorithm

The Apriori Algorithm is an efficient algorithm for mining data sets that occurs frequently for Boolean association rules. The algorithm can be suitable for most commercial products, and is also called Associate Rule Algorithm.. This algorithm is more effective on large item set property. It is predefined that subset of a large item set should be large. The large item sets are also called to be downward closed, it is because if an item set satisfies the minimum support requirements, so must be true for all of its subsets. On the other hand, the contra positive of above, if there is a small set item, there is no need to generate and supersets of them as candidates since they also must be small. The basic idea of the Apriori algorithm is generating candidate item sets of a desired size and then scanning the data base to count them to see if they are large. Afterwards, frequent individual items is identified in database and are extended to larger and larger item sets until those item sets converges sufficiently in the database. Association rules, which effectively highlight general trends in database, is determined by frequent item sets, which in turn is determined by Apriori. It has vast applications in domains such as market basket analysis. Apriori is designed in a way to operate on databases facilitating transactions. Apriori uses bread th-first search and a Hashtree structure for counting the candidate item sets accurately.

Support

The support supp(X) of an items et X is defined as the proportion of transactions in the data set which contain the item set.

supp(X)=no. of transactions which contain the item set X/total no. of transactions.

Confidence

The confidence of rule is determined by :

$$Conf(X \rightarrow Y) = supp(X \cup Y) / Supp(X)$$

Binning Algorithm

To transform the numerical variables into categorical counterparts is a process, which is known as Binning or Discretization. Numerical variables are generally discretized in the modelling methods based onfrequency tables(e.g.,decisiontrees).Itismainlyusedforreducing the minor observation errors and its effects. The original data values falls in a bin, a small given interval, and can be replaced by a value representative of that given interval, commonly the central value. Bin can be visualized as a form of quantization. Moreover, binning also reduces non-linearity and noise and hence effective in improving the accuracy of the predictive models. Binning also provides easy identification of outliers, and counters missing and invalid values of numerical variables. Unsupervised and supervised are two classifications of Binning.

Supervised Binning. This method is used for transforming numerical variables into categorical counterparts and used to refer the target (class) information while selecting discretisation cut points, by using training data sets. Entropy-based binning is an example of this method.

Unsupervised Binning. This method is used to transform numerical variables into categorical counterparts without using the target (class) information, and thus, not using training data set. Equal Width and Equal Frequency are examples of this method. Equal- width (distance) partitioning: It divides the range into M intervals of equal size: uniform grid. If A and B are the highest and lowest values of the attribute, the width of these intervals will be:

W = (A-B)/M. The interval boundaries are: A + w, A + 2w, ... , A + (M-1)w. But outliers may or may not be dominating presentation. Skewed data are not handled very well. Equal-depth (frequency) partitioning: 10 It divides the range into M intervals, each contains approximately equal amount of samples. It provides good data scaling and good handing of skeweddata.

Equal Width Binning

- It divides the equally-sized data into p intervals. The interval widthis:

  w = (maximum - minimum)/p

- Theintervalboundariesare:

  minimum+w, min+2w, ... , min+(p-1)w

Equal Frequency Binning

It divides the data into p groups with each group containing same amount of values approximately. For both methods, the best way to determine p is by trying different intervals or groups, while observing his to grams.

II. NORMALIZATION

In data mining and statistical data analysis, before models are build or algorithms are used, the data should be prepared In this context, to ease the algorithm's job, prepared data are transformed before analysis. Often, to verify the hypothesis on which algorithm are based, the data are altered, while preserving the information intact at the same time. Normalization is the most basic transformation technique. Basically, normalizing generallymeanstransforminginordertoscalethedatato fit into desired range. In deviated cases, normalization refers to more peculiar adjustments where the main intentionisbringingtheentireprobabilitydistributionsof adjusted values into alignment Three methods are there in Normalization which areMinimum-maximumnormalization,Z-score normalization and Decimal scaling based normalization.

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Min-max Normalization .It is the method of taking data measure din engineering units and transforming it to a value within a given range, that is, [Min, Max].

Generally,0and1areminimumandmaximumvaluesinwhic hthegivendataisnormalized.Forcomparing input values, it isaneasy method.

Z-scoreNormalization.Az-score(alsoknownasz-value,standardscore,ornormalscore)I same a sure of the divergence of an individual experimental result from the most probable result, the mean .Z is expressed in terms of the number of standard deviations from the meanvalue.Z-scores assuming

the sampling distribution of the test statistic(meaninmostcases)isnormaland transform the sampling distribution into a standard normal distribution.

$$z = \frac{X - mean\ (x)}{stdev\ (x)}$$

Five Point Summary

The five-number summary is a descriptive statistic which provides basic knowledge about a particular set of observations. It has mainly five basic classification of sample percentiles, which are: The sample minimum (smallest observation), The upper quartile or third quartile, The lower quartile or first quartile, The median (middle value) and The sample maximum (largest observation). Concise summary of the distribution of the observations is provided by five-number summary method. One can avoid the need to decide on the most appropriate summary statistic by repeating five numbers .It also provides information about the spread, location and observations range. Five-number summary is adequate for ordinal measurements, interval and ratio measurements, as it reports the order statistics. By comparing their five-number summaries, we can quickly compare several sets of observations, which can be represented graphically using a box plot. The least value in the set is known as the minimum value of set and, the greatest value in the set is known as maximum value of set. the range of a data set is the distance between the maximum and minimum value. To compute the range of data set, we subtract the minimum from themaximum

range = maximum – minimum

The distance between the two quartiles is interquartile range of a data set.

interquatilerange(IQR) = q3 −q2

Clustering

Clustering is a method of partitioning a set of data (or objects) into clusters, which is a set of meaningful sub-

classes. A cluster of data objects is treated as a onesingle group. In cluster analysis, firstly the set of data is partitioned into groups on the basis of data similarityand secondly, assigning the label to the groups. Clustering is better than classification in a way that, it is adaptable to changes and help to single out useful features that distinguished different group. There are many types of Clustering algorithms. Some of them are: connectivity based cluster- ing (hierarchical clustering), centroid- based clustering, distribution-based clustering, density- based clustering and recent developments. In ourproject, partition/centroid based, that is, k-means algorithm, is used.

Centroid based Clusturing. In this type of clustering, a central vector represents the cluster, which may or may not be a member of the data set. When the number of clusters is fixed to k, k-means clustering provides a formal definition as an optimization problem: Find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

The optimization problemitselfisknowntobeNP-hard,andthusthebasic method is to search only for approximate solutions. A particularly well known 13 approximate method is Lloyd's algorithm often actually referred to as "k-means algorithm". In K-means, items are moved among sets of clusters until the desired set is reached, which means itis an iterative clustering algorithm. It can be assumed as a type of squared error algorithm, although the convergence criteria need not be defined on the squared error. A high degree of similarity among the elements in clusters are attained, whereas a high degree of dissimilarity among elements in different clusters is obtainedsimultaneously.

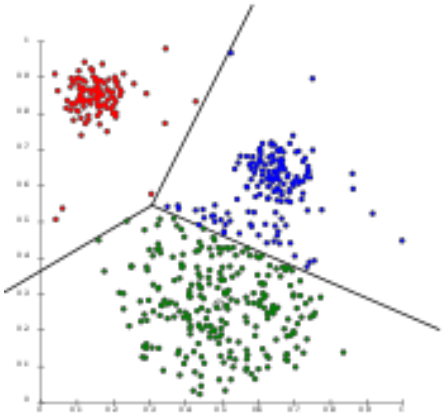

Fig:-K-means separates data into Voronoi-cells, which assumes equal-sized clusters (not adequate here)

Fig:-K-means cannot represent density-based clusters

III. CONCLUSION

As we have studied that Data mining algorithm works well in each and every field of research work like soft computing, image processing and cloud computing etc. so well, So we planned it to propose all the data mining algorithm at a single platform using single user interface namely data extension tool with some additional features like Output in different formats (like pdf and xls) and also use our techniques into the fast growing research are namely time series financial forecasting as well wherever the data set concept will arise.

REFERENCES

Book References:

Book References:

[1] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concepts and Techniques, 3rd edition
[2] Maragaret H. Dunham, Data Mining Introductory and Advanced Topics.

Book Chapter:

[3] Jiawei Han, Micheline Kamber, Jian Pei . "Data Preprocessing" in Data Mining Concepts and techniques, 2nd, volume.
[4] World Wide Web
[5] http://www.saedsayad.com/unsupervised_binning.htm
[6] http://www.damienfrancois.be/blog/pivot/entry.php?id=8
[7] http://www.dataminingblog.com/standardization-vs-normalization/
[8] http://en.wikipedia.org/wiki/Five-number_summary
[9] http://www.saedsayad.com/binning.htm
[10] http://theglobaljournals.com/ijsr/file.php?val=MTEyNg

[11] http://en.wikipedia.org/wiki/Cluster_analysis
[12] http://www.cs.gordon.edu/courses/cs211/AddressBookExample/index.html
[13] http://www.visualparadigm.com/product/vpuml/features/behavioralmodeling.jsp(tool for drawing sequence diagram)