REPRESENTATION OF DATA AND KNOWLEDGE WITH INTEGRATED APPROACH

Ekata Gupta[1] Research Guide, GNIM

Harleen Kaur[2] Student, GNIM

ABSTRACT: *Data and knowledge representation describes the process of how data and knowledge can be represented in the data mining process. We focus on how data and knowledge can be extracted through various challenges and how to overcome from these problems so that proper data and knowledge discovery process can be delivered in the data mining process. There are many data mining techniques, but we consider Clustering technique, which has many advantages but some shortcomings also exist. So there is a new technique, in which clustering technique is based on a supervised learning Decision treetechnique called CLTree i.e. Clustering based on decision tree. The new CL Tree technique is able to overcome many of the shortcomings. The key idea is to partition the data space into cluster regions and empty regions by using decision tree technique. This technique is able to find clusters of similar type in large high dimensional space efficiently. It is suitable for clustering in the full dimensional space as well as in subspaces. It also provides descriptions of the resulting clusters.*

Keywords: Classification, Integrating Constraints, Knowledge Management.

1. Introduction to Data Mining

Knowledge and data representation is the knowledge discovery process of datawhich means the extraction of essential information from large volume of data. We often see data as a number of bits, symbol, objects etc. Data mining is a field concerned with the development and the use of the techniques for identification of useful information and knowledge patterns in large databases or data warehouses. There are many application areas where data mining techniques are used extensively, e.g. marketing research, financial and medical databases. The field of data and knowledge representation mining is based on artificial intelligence and statistics disciplines. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information.

It is the efficient discovery of valuable, non-obvious information from a large collection of data. It is a knowledge discovery process helps us to understand the substance of the data in special unsuspected way with the extraction of interesting (non-trivial, implicit, previously unknown and potentiallyuseful) patterns or knowledge from huge amount of data.

1.1 KDD Process

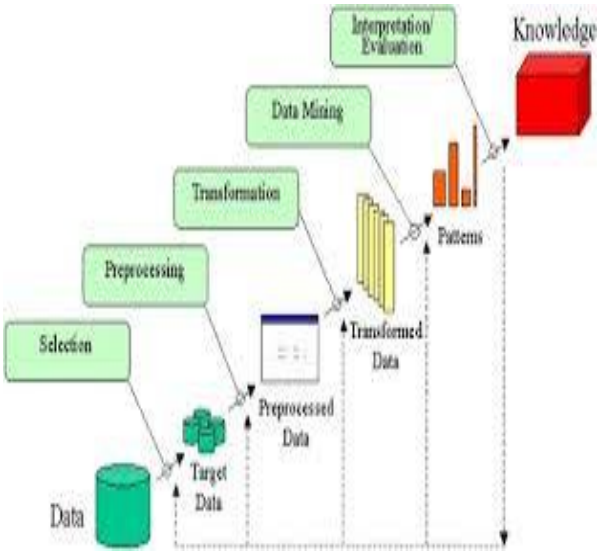

Figure 1. Knowledge Discovery Process [2]

The KDD process includes data cleaning, data integration, data selection, preprocessing, transformation, data mining, pattern evaluation, and knowledge presentation.

KDD describes the whole process of extraction of knowledge from data. Data mining should be used exclusively for the discovery stage of the KDD process. Some Data mining functionalities are characterization, discrimination, association, classification, clustering, outlier and trend analysis.

The alternative names of data mining can be Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc. One can mine tremendous amount of "patterns" and knowledge.

2. Techniques of Data Mining

Data Mining techniques are used to carry out data mining functions. There are some techniques of data mining which determine the information and knowledge patterns from large databases. Some techniques are as follows:

2.1 Decision Tree

It is a tree where internal nodes are simple decision rules on one or more attributes and leaf nodes are predicted class labels. It is easy to interpret and can be re-represented as if-then-else rules. It does not require any prior knowledge of data distribution, works well on noisy data.

2.2 Neural Networks

Neural network mimic human brain and require trained dataset and patterns for classification and prediction. Its algorithms are effective when data is shapeless and lacks pattern. It is useful for learning complex data like handwriting, speech and image recognition.

2.3 Memory-Based Reasoning (MBR)

It is based on the concept of similarity. MBR results are based on analogous situations in the past, in this we have to choose appropriate historical data for use in training.

2.4 Clustering

Clustering is a kind of unsupervised learning. Clustering is a method of grouping data that share similar trend and patterns. K means clustering is an effective algorithm to extract a given number of clusters of patterns from a training set. Once done, the cluster locations can be used to classify patterns into distinct classes.

2.5 Link Analysis

This algorithm is extremely useful for finding patterns from relationships. The link analysis technique mines relationships and discovers knowledge. For example, if you look at the supermarket sale transactions for one day, why are skim milk and brown bread found in the same transaction. Link analysis algorithms discover combinations. Depending upon the types of knowledge discovery, link analysis techniques have three types of applications: association's discovery, sequential pattern discovery, and similar time sequence discovery.

2.6 Genetic Algorithm

GAs is inspired by biological evolution. It acts like bacteria growing in a ptri dish. Many operators mimic the process of the biological evolution including Natural selection, Crossover, Mutation.

Data Mining Technique	Underlying Structure	Basic Process
Clustering	Distance calculations in n-vector space	Grouping of values in the same neighborhood
Decision Tree	Binary Tree	Splits at decision points based on entropy
Memory-based Reasoning (MBR)	Predictive structure based on distance and combination functions	Association of unknown instances with known instances
Neural Networks	Forward propagation network	Weighted inputs of predictors at each node
Link Analysis	Based on propogation of variables	Discover links among variables by their values
Genetic Algorithm	Not applicable	Survival of the fittest on mutation of derived values

Table 1.1Comparison of Data Mining Techniques [3]

3. CLUSTERING

Clustering is a technique useful for exploring data. Clustering is done to give the end user a high level view of what is going on in the database. It is particularly useful where there are many cases and no obvious natural groupings. Here, clustering data mining algorithms can be used to find whatever natural groupings may exist. Clustering analysis identifies clusters embedded in the data. A good clustering method produces high-quality clusters to ensure that the inter-cluster similarity is low and the intra-cluster similarity is high.

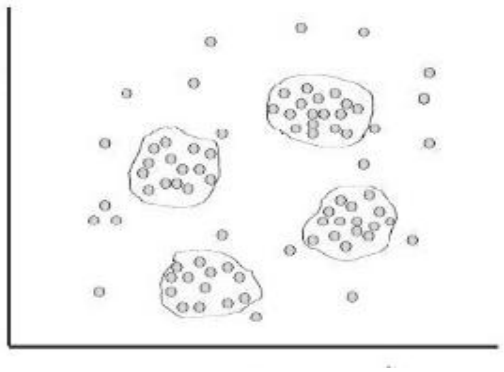

Figure 2. Clusters with two variables [3]

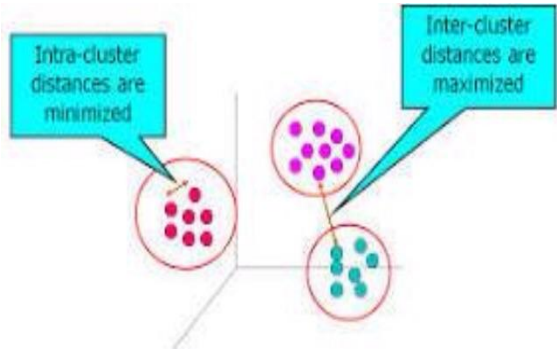

Figure 3. Data Visualization- Intra-cluster distances are minimum and inter-cluster distances are maximum [4]

2.1 Advantages of Clustering Technique

- Clustering is an unsupervised learning that means no predefined classes.
- Its typical application is to act as a stand-alone tool to get insight into data distribution.
- It is also a preprocessing tool for Summarization, Compression, Finding K-nearest neighbors, for Outlier detection.
- Relatively efficient and easy to implement.
- The clusters are non-hierarchical and they do not overlap.

Clustering have many advantages. But their exists some algorithms of clustering which have some shortcomings.

2.2 Disadvantages of Clustering Technique:

- It is sensitive to initialization. It is difficult to compare with different number of clusters.
- Needs to specify the number of clusters in advance by prior assumptions.
- It is unable to handle noisy data or outliers.
- The interpretation of how interesting a clustering is will be application-dependent and subjective to some degree.
- Clustering techniques suffer from the fact that once a merge or a split is committed, it cannot be undone or refined.
- Sometimes clustering is performed not so much to keep records together as to make it easier to see when one record sticks out from the rest.
- Clustering is statistically separating records into smaller unique groups based on common similarities.

2.3 How to overcome from shortcomings of Clustering technique

To overcome shortcomings of clustering technique, we have to apply decision tree technique on it. So in the new technique, clustering technique is based on a supervised learning decision tree technique called CLTree. The new technique is able to overcome many of these shortcomings. The basic idea to use a decision tree is to overcome from the main shortcoming i.e., to partition the data space into cluster regions and empty regions which produce

some outliers. The technique is able to find clusters in large high dimensional spaces efficiently.

3. DECISION TREES

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Decision trees can be used in order to predict the outcome for new samples. The decision tree technology can be used for exploration of the dataset and business problem. Another way that the decision tree technology has been used is for preprocessing data for other prediction algorithms.

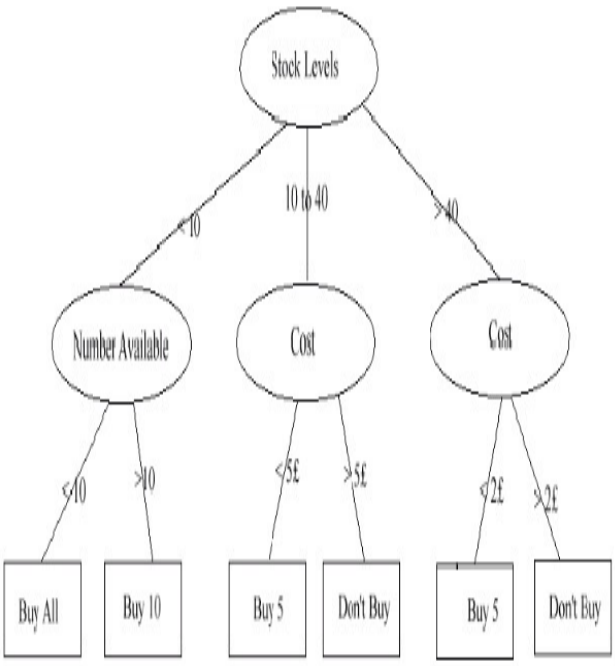

Figure 4. Decision Tree for stock levels [5]

3.1 Advantages of Decision Trees

- Reasonable training time
- Fast application
- Easy to interpret
- Easy to implement
- Can handle large number of features
- It partitions the data space into different class regions

4. ALGORITHM OFCLTree technique

For classifying data of various different classes we use a most popular technique of data mining called Decision Tree. To partition the data space into different class regions its algorithm uses a purity function. As datasets for clustering have no pre-assigned class labels so for this reason only this technique is not directly applicable to clustering.

Steps for the Algorithm are as follows:

1. We assume that each data record in the dataset to have class Y.

2. There is also another type of points called non-existing points, so we assume that these points are uniformly distributed with data space. We give them the class N.

3. With these N points added to the original data space, our problem of partitioning the data space into empty regions & data regions becomes a classification problem.

4. So now decision tree algorithm can be applied to solve this classification problem.

5. This technique only works if there are clusters in the data because the data points cannot be uniformly distributed in the entire space.

6. The most well known task of decision tree technique is to add some N points to isolate the clusters because within each cluster region, there are more Y points than N points.

7. How many N points should be added, it depends on the number of changes as the tree grows. Physically adding N points increases the size of the dataset and also the running time.

8. The issue is that it is unlikely that we can have points uniformly distributed in a very high dimensional space because we would need an exponential number of points.

We propose a technique to solve the classification problem, which guarantees that N points are uniformly distributed. This is done by not adding any N point to the space but computing them

when needed. Hence, CLTree can produce the partition by using decision tree technique with no N points added to the original data.

5. The proposed CLTree technique consists of two steps

5.1 Cluster tree construction: For constructing a cluster tree apply decision tree algorithm with a purity function to capture the natural distribution of the data without making any prior assumptions to it [6].

5.2 Cluster tree pruning: To find useful clusters in the data and to simplify the tree, a pruning step is performed after the tree is built [6].

6. CONCLUSION

In this paper, we proposed a flexible knowledge representation of data by applying integrated approach to it. Also a clustering technique, called CLTree i.e., Clustering which is based on decision Tree. CLTree performs clustering by partitioning the data space into data and empty regions at various levels. To make this tree algorithm work for best partition of clustering we designed a new purity function and also we proposed a new technique to apply non-existingpoints to the data space. Also in order to find useful clusters we use a constructing and pruning method of cluster tree. For finding outliers it produces empty or sparse regions. All these results are directly useful for the clustering technique.

7. REFERENCES

[1] Journal of Knowledge Management Practice, Vol.8, No.2, June 2007: Understanding Data, Information, Knowledge and their Inter-relationships. Anthony Liew, Walden University.

[2] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview".

[3] Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals, Paulraj Ponniah,2001. Introduction to Data Mining, KDD Process, Techniques of Data Mining.

[4] "Data Visualization" by Andrei Pandre. Cluster analysis.

[5] Decision Trees, Symmetric Rules and Transitive Rules: Chris Huyck. Decision Trees.

[6] Clustering Via Decision Tree Construction. Bing Liu, Yiyuan Xia and Philip S.Yu.Clustering, Clustering shortcomings, How to overcome from decision tree technique, Algorithm of cl tree, Proposed techniques of clustering.