International Journal of Research
ISSN: 2348-6848 Vol-3, Special Issue-3
International Conference on Research and Recent Trends in Engineering and Technology. (ICRRTET)
Held on 27th January 2016 organized by **Sai Polytechnic College,** Kinhi Jawade, Yavatmal, Maharastra, India.

Clustering High Dimensional Data Using Fast Algorithm

Pooja M. Salwe; Diksha P. Wasnik & Prof. Megha Goel
Computer Science & Engineering, R.T.M.N.U. Nagpur
poojasalwe@gmail.com; dikshawasnik4595@gmail.com

Abstract

Real world data may contain hundreds of attributes in dataset many of which maybe irrelevant to the mining. Whenever we want to extract data from dataset that may beincomplete, inconsistent or contain noise because dataset collect and store data from various external sources. To overcome this problem clustering is mainly use to simplify the data, detecting the data patterns and identifying features of pattern. Feature selection in data mining is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy and improving result.

Feature selection is a process which selects the subset of attributes from original dataset by removing irrelevant and redundant attribute. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Clustering is a technique in data mining which groups the similar objects into one cluster and dissimilar object into other cluster. Feature selection reduces the computational time greatly due to reduced feature subset and also improves clustering quality. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is used and experimentally evaluated in this project. The FAST algorithm works in two steps. In the first step, features are divided into clusters. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) using the Kruskal's Algorithm clustering method. The efficiency and effectiveness of the FAST algorithm is evaluated through the study.

Introduction

With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. They can be divided into four broad categories: the Embedded, Wrapper, Filter, and Hybrid approaches. The embedded methods incorporate feature selection as a part of the training process and are usually specific to given learning algorithms, and therefore may be more efficient than the other three categories. Traditional machine learning algorithms like decision trees or artificial neural networks are examples of embedded approaches. The wrapper methods use the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets, the accuracy of the learning algorithms is usually high. However, the generality of the selected features is limited and the computational complexity is large. The filter methods are independent of learning algorithms, with good generality. Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. The hybrid methods are a combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the subsequent wrapper. They mainly focus on combining filter and wrapper methods to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. The wrapper methods are computationally expensive and tend to over fit on small training sets. The filter methods, in addition to their generality, are usually a good choice when the number of features is very large. Thus, we will focus on the filter method in this paper. With respect to the filter feature selection methods, the application of cluster analysis has been demonstrated to be more effective than traditional feature selection algorithms. Pereira et al. Baker et al. and Dhillon et al. employed the distributional clustering of words to reduce the dimensionality of text data. In cluster analysis, graph-theoretic methods have been well studied and used in many applications. Their results have, sometimes, the best agreement with human performance. The general graph-theoretic clustering is simple: Compute a neighborhood graph of instances, then delete any edge in the graph that is much longer or shorter than its neighbors. The result is a forest and each tree in the forest represents a cluster. In our study, we apply index based clustering methods to features. In particular, we adopt the minimum spanning tree (MST) based clustering algorithms.

Background and Related Work

The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Clustering

Clustering is the process of making a group of abstract objects into classes of similar objects. A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain

insight into structures inherent to populations.

- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Proposed System

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. This is because irrelevant features do not contribute to the predictive accuracy and redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s). Of the many feature subset selection algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features**.** The FAST algorithm falls into the second group. Traditionally, feature subset selection research has focused on searching for relevant features.

In our proposed system implements, semisupervised learning has captured a great deal of attentions. Semisupervised learning is a machine learning paradigm in which the model is constructed using both labeled and unlabeled data for training typically a small amount of labeled data and a large amount of unlabeled data. In this proposed system it retrieve the data from training data or labeled data and extract the feature of the data and compare with labeled data and unlabeled data to. In clustering process, semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. Semisupervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy.

Modules

A. USER MODULE

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

B. CIRCULATED CLUSTERING

Cluster is nothing but it is a combination of various features including text subsets, it has been used to cluster words into groups based either on their contribution in particular grammatical relations with other words. Here the distributed clustering focuses on the cluster with various text subsets. In this module the system can manage the cluster with various classifications of data.

C. TEXT DETACHMENT

The text detachment is the filtration process, which filters the combination of various subsets present in the cluster into matching clusters belonging to the text head with the help of k Means algorithm. A novel algorithm which can efficiently and effectively deal with both inappropriate and superfluous features, and acquire a good feature subset.

D. ASSOCIATION RULE MINING

Association rule mining is the best method for discovering interesting relations between variables in large databases or data warehouse. It is intended to identify strong rules discovered in databases using different measures of interestingness. With the help of this rule mining the system can manipulate and associates the text cluster into the respective heads based on the internal features of data.

E. TEXT ORGANIZATION

The Text organization contains many problems, which has been widely studied in the data mining, machine learning, database, and information retrieval communities with applications in a number of diverse domains, such as target marketing, medical diagnosis, news group filtering, and document organization. The text organization technique assumes categorical values for the labels, though it is possible to use continuous values as labels. The latter is referred to as the regression modeling problem. The problem of text organization is closely related to that of classification of records with set-valued features. However, this model assumes that only information about the presence or absence of words is used in a document. In reality, the frequency of words also plays a helpful role in the classification process and typical domain size of text data (the entire size) is much greater than atypical set valued classification problem.

System Architecture:

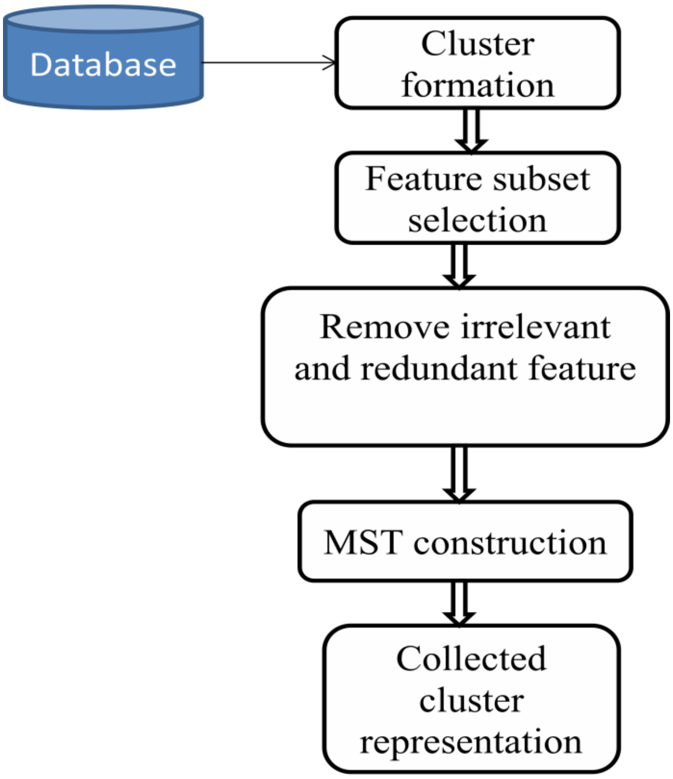
Database
Cluster formation
Feature subset selection
Remove irrelevant and redundant feature
MST construction
Collected cluster representation

Fig:System Architecture

Advantages of Proposed System

- It is easy to understand.
- This method is relatively scalable.
- It is fast searching method.
- It detects irrelevant and redundant data.

Future scope

For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space.

Conclusion

In this project, we used a novel clustering-based FAST algorithm for searching high dimensional data. The desired output obtained in very less time. It is efficient in processing large dataset.

References

[1] A Survey on Clustered Feature Selection Algorithms for High Dimensional Data, International Journal of Computer Science and Information Technologies, Vol. 5, March 2014.

[2] Subset Selection Algorithm for High Dimensional Data,International Journal of

Emerging Trends & Technology in Computer Science (IJETTCS), June 2014.

[3] A Methodology for Direct and Indirect Discrimination Prevention in Data Mining, IEEE transactions on knowledge and data engineering Vol.25 no.7 July 2013.

[4] A Survey on Feature Selection Algorithm for High Dimensional Data Using Fuzzy Logic, the International Journal of Engineering And Science (IJES), Sept 2013.

[5] High Dimensional Unsupervised Clustering Based On Wrapper And Filter Approach, International Journal of Engineering Science and Technology (IJEST), Vol 4 May 2012.