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.

Data Mining Methods for Improving Business Process Modelling

Arti Mirche; Pranali Kukade & Prof. Veena Katankar
Computer Science & Engineering, R.T.M.N.U. Nagpur
artimirche6@gmail.com; pranalikukade31@gmail.com

Abstract

This paper introduces a novel methodology to extract core concepts from raw dataset. This methodology is based on data mining and analysis. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. At the data mining phase the keywords are extracted by tokenizing, removing stop-lists and generating similarity by classification and clustering. For this methodology we are going to use k-NN and k-means algorithms. We applied our methodology on large data set. Similarity based algorithm was interesting and gave us valuable knowledge about content and used for deep analysis.

Introduction

Data mining is the extraction of the hidden productive information from large databases. It is the powerful new technology with great potential to analyze important information in the data warehouse. Data mining is the search for the relationship and global patterns that exist in large databases, but are hidden among vast amount of data. Preprocessing is the process in which all the data in the datasets will be clean by tokenizing and removing stop-lists.

Classification is one of the most important areas of machine learning. Similarity-based methods, including many variants of the k-nearest neighbor algorithms, belong to the most popular and simplest methods used for this purpose. Clustering is another important process used for mining the data which creates the clusters of classified data. For large databases, especially in problems requiring real-time decisions, such “lazy approaches” relaying more on calculations performed at the time of actual classification rather than at the time of training are too slow. Training of all similarity-based methods, including kernel-based SVM approaches, also suffers from the same quadratic scaling problem. Fast methods for finding approximate neighbors can reduce this time.

In this only transformations based on similarities to the nearest k-samples scaled by Gaussian kernel features are explored, but any other similarity measures may be used in the same way. In essence this connects similarity-based methodology with deep learning techniques, creating higher-order k-nearest neighbors method with kernel features.

Existing System

- In existing system, the data mining process involves different phases : Selection, Pre-processing, Clustering and Classification.
- These phases are performed by using different algorithms for generating the efficient result.
- The algorithms are used in data mining for extraction of data are k-NN algorithm, Trans D algorithm etc.

Disadvantages of Existing System

- Performance of each mining process is done individual.
- The time required due to each and every phase is time consuming.
- This slows down the process and response time is more.

Scope & Objectives:

- A partition of a real data set generated by a clustering algorithm.
- To identify the genuine clusters from the partition.
- A classification is using for classifying clusters data in predefined groups.
- Converting a random based data into similar data for the quick extraction of data which is time consuming.
- To discuss the potential use of data mining and knowledge discovery in databases.

1. Data Selection:

Data which is relevant for mining process is retrieved from various sources of data like data warehouse or database.

2. Preprocessing:

As data are drawn from multiple sources, they may contain inconsistent, incorrect or missing information. For example, it is very much possible that the same information in different sources can be presented in different formats. Therefore this phase is concerned with the data cleaning process during which unnecessary information is removed. This ensures that the data are configured into useful and meaningful information.

The major steps involved in data preprocessing, namely, data cleaning, data integration, data reduction, and data transformation.

Data have quality if they satisfy the requirements of the intended use. There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability.

3. Classification:

Classification refers to the portioning of given data into predefined disjoined groups or classes. In such task, a model which is also known as classifier is build to predict the class of the new item, given that the items belongs to one of the classes and given past instances, which is also known as training instances of items along with the classes which they belong.

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels.

4. Clustering:

Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering.

Cluster analysis has been widely used in many applications such as business intelligence, image pattern recognition, Web search, biology, and security. In business intelligence, clustering can be used to organize a large number of customers into groups, where customers within a group share strong similar characteristics. This facilitates the development of business strategies for enhanced customer relationship management. Moreover, consider a consultant company with a large number of projects. To improve project management, clustering can be applied to partition projects into categories based on similarity so that project auditing and diagnosis (to improve project delivery and outcomes) can be conducted effectively.

Conclusion:

In our project, “Data Mining Methods for Improving Business Process Modeling” we mine the large dataset using mining methods for producing some analysis for the business model. The result is being represented using Google map.