Data Mining Methods for Improving Business Process Modelling

Shital Nighot; Pallavi Dhage & Kajal Dhabale
Computer Science & Engineering, R.T.M.N.U. Nagpur
shitalnighot11@gmail.com; pallavidhage44@gmail.com; kajaldhabale176@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.

Literature survey

R. Ahmed & Karypis proposed Data mining with using clustering and classification algorithms in 2011 i

David Brownstone proposed the role of apriori algorithm for finding the association rules in data mining in 2013 in which they gives the brief specification of Generate association rules for frequent/large item-sets. In this algorithm they find some difficulties in finding all item-sets with adequate supports.

F. casati, E. Shan, u. Dayal, and M. Shan proposed A survey on clustering data mining Techniques in 20

G. Alonso, F. Casati, h. kuno, and V. Machiraju proposed Analyzing Collective Behavior from Blogs Using Swarm Intelligence, Knowledge and Information Systems in 2014 in which they works on the methodology named as Dynamic correlation technique . Issues during this methodology are the ability of commonly used software tools to capture, manage and process the data within a tolerable time.

Proposed System

In selection, Data which is relevant for mining process is retrieved from the database and preprocessing is concerned with the data cleaning process during which unnecessary information is removed. This ensures that the data are configured into useful and meaningful.

A cluster is a collection of data objects that are similar to one another within a same cluster and dissimilar to the objects in other cluster. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. The K-means algorithm is used for clustering.

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.

In proposed system, the two algorithms k-NN and k-Means which are used for clustering and classification are worked together parallel for fast result and the task will be distributed among these algorithms which helps to reduce the workload.

Thus the system generates the output in less amount of time as compare to existing system.

In graph plotting, the result of cluster analysis is displayed on the world map for the relevant input.

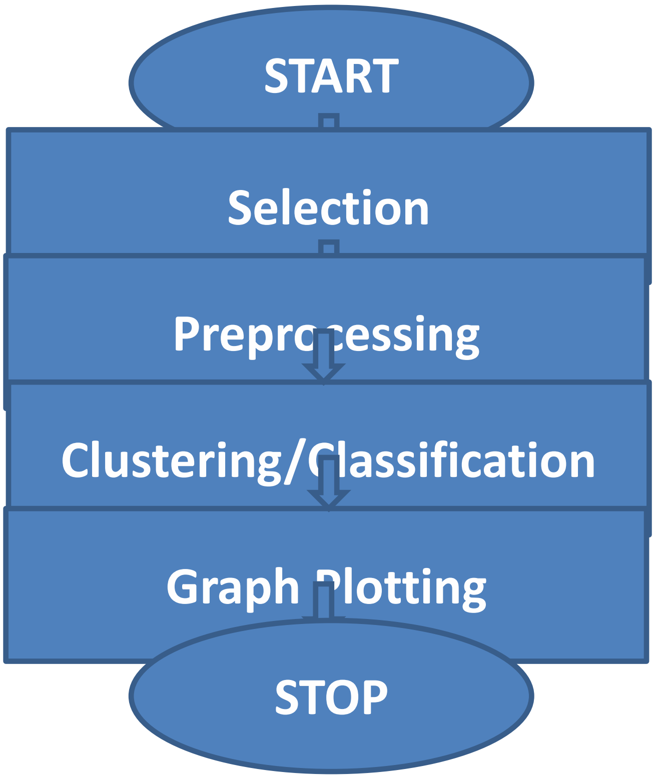

Figure 1: Proposed System

Modules

1. Data Selection:

Data which is relevant for mining process is retrieved from the 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.

Algorithm: k-Nearest-Neighbor

Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all the training tuples are stored in an n dimensional pattern space. When given an unknown tuple, a k nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "nearest neighbors" of the unknown tuple.

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.

Algorithm: k-Means

The k-means algorithm defines the centroid of a cluster as the mean value of the points within the cluster. First, it randomly selects k of the objects in D, each of which initially represents a cluster mean or centre. The k-means algorithm then iteratively improves the within-cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration. All the objects are then reassigned using the updated means as the new cluster centres. The iterations continue until the assignment is stable, that is, the clusters formed in the current round are the same as those formed in the previous round.

System Architecture

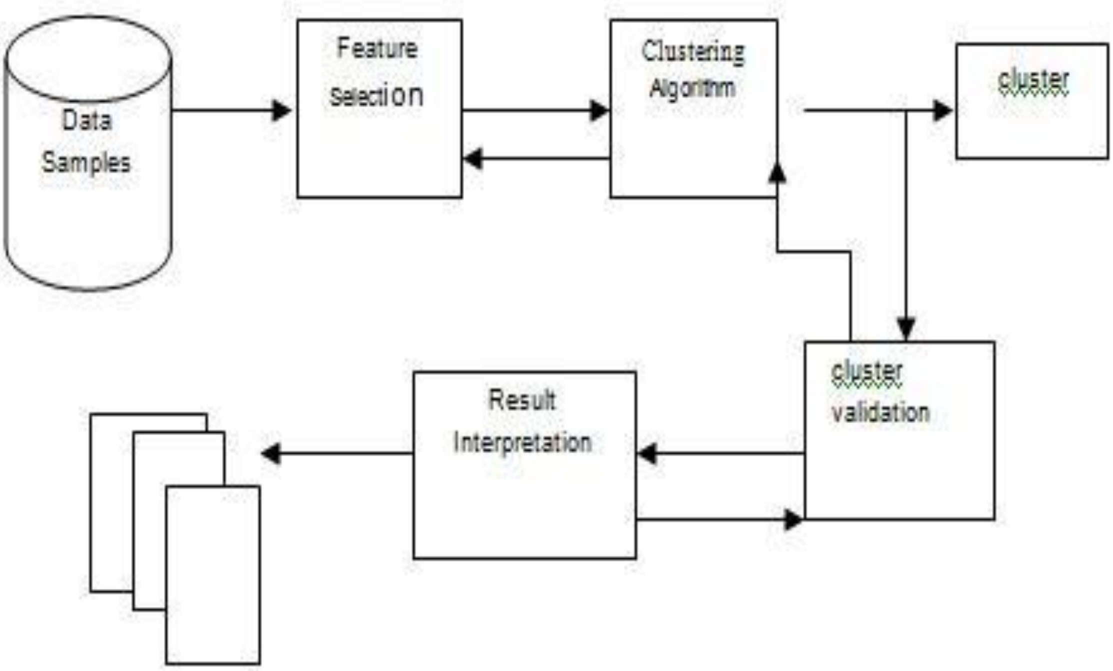

Fig. : System Architecture

Advantages

- Converts the random based data into similar data which helps for fast extraction of data.
- Combining two algorithms together reduces the response time i.e. time consuming.
- Task will be divided among both algorithms so the workload is also divided among them.
- By using two algorithms in parallel will optimized the data mining process and help in fast analysis.

Disadvantages

- Using two algorithms in a combination will be increases the complexity level of programming.
- We handle only simple dataset not big data.

Applications

- Financial Banking data analysis
- Healthcare and Insurance
- Biological data analysis
- Transportation and Medical
- Transportation and Medical

Future scope

- Big data working, using Hadoop concept.
- Time complexity by running the clustering and classification phase parallel.

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.

References:

[1] Yuh-Jyh Hu, Min-Che Yu, Hsiang-An Wang, and Zih-Yun Ting (2015) A Similarity-Based Learning Algorithm Using Distance Transformation. Proc. Of IEEE Transaction On Knowledge And Data Engineering, VOL.27, NO. 6

[2] Data Mining Concepts and Techniques, Third Edition, Jiawei Han University of Illinois at Urbana-Champaign, Micheline Kamber, Jian Pei, Simon Fraser University

[3] R. Ahmed & Karypis proposed Data mining with using clustering and classification algorithms, k-means and k-medioids in 2011

[4] Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C" Application of k- means Clustering algorithm for prediction of Students Academic Performance" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, 2010.

[5] Varun Kumar and Nisha Rathee, ITM University, "Knowledge discovery from database Using an integration of clustering and classification", International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.

[6] David Brownstone proposed The role of apriori algorithm for finding the association rules in data mining , Generate association rules for frequent/large item-sets in 2013.

[7] McKinsey Global Institute (2011) Big Data: The next frontier for innovation, competition and productivity.

[8] F.casati,E.Shan,u.Dayal, and M. Shan proposed A survey on clustering data mining Techniques in 2014 for Datasets to find relationships ANN, Fuzzy sets-theory Approximate Reasoning.

[9] Chen, H., Chaing, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, 36, 4, pp. 1165-1188.

[10] G.Alonso, F.Casati, h.kuno, and V. Machiraju proposed Analyzing Collective Behavior from Blogs Using Swarm Intelligence, Knowledge and Information ,Dynamic correlation technique in 2014.

[7] The Theory of Attributes from Data Mining Prospect.