ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

Survey on Data Mining Related Methods / Techniques and Text Mining

Monika Kohli[1], Rohit Tiwari[2]

Assistant Professor, Computer Engineering & Information Technology Department,

K.J. Institute of Engineering & Technology, Savli, Vadodara, Gujarat, India[1]

Assistant Professor, Computer Engineering & Information Technology Department,

K.J. Institute of Engineering & Technology, Savli, Vadodara, Gujarat, India[2]

Abstract: This paper emphasis on Data Mining techniques/methods and text mining concepts. Data Mining can be considered as a procedure to excerpt consequential information or patterns out of a very gigantic volume of data. Knowledge is also excerpted out of such extensive batch of data as well. Data mining develops imperative relationship among variables that are saved in large data set or data warehouse. This paper provides useful information regarding the techniques/methods, and applications of data mining in various fields making the businesses to give better results; including some basic concepts of text based mining.

Keywords: Data mining Techniques, Association Rule, Classification, Data mining, Knowledge discovery, Data mining application, Text mining.

I. INTRODUCTION

In recent world massive amount of data are available which is related to diverse field such as agriculture, educational, medical, enterprise and various other areas. Such data may deliver information as well as knowledge for further decision making. Data mining is a strong concept for analysing data of diverse fields and it's a process of finding interesting patterns from massive amount of data which is stored in different operational systems such as databases, relational databases, data warehouses, web based data warehouses and other external sources [2]. Interesting patterns are easy to understand and various patterns are used for further processing on extracted knowledge from data mining methodology viz. if you want to find out alumni students data in a college/university or sales data of a shopping store. Data can be précised and analysed, so it can easily meet the specified requirements. Data mining is also used as a core concept used to extract hidden patterns from large amount of databases and data warehouses [3].
The aim of data mining includes fast retrieval of information, knowledge discovery from data and to identify hidden patterns from databases. Such patterns are helpful in reducing complexities and time saving. Data mining refers to extracting useful and meaningful information and knowledge form huge amount of data which is called as Knowledge Discovery from Data (KDD). It is a knowledge extraction or data/pattern analysis. As stated earlier in this paper, data mining is a persistent procedure to excerpt consequential information or patterns out of a very extensive volume of data, in order to find out crucial and appropriate information. The aim of data mining is to search hidden and interesting patterns that are not known. Once these interesting patterns are found, they are used for decision making to grow business effectively and efficiently.

KDD is an iterative procedure, which contains following steps:

- Data Cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

Data Cleaning: Removal of noise, incompleteness, uncertainty and inconsistency of data from the gathering in performed in data cleaning phase.

Data Integration: In this phase, multiple operational system's data such as Data Warehouses, DBMS, RDBMS, MDBMS etc., are combined that are stored in heterogeneous manner.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

Data Selection: In data selection phase, data relevant to the task are retrieved from the databases or other operational sources.

Data Transformation: Data transformation is an essential function of KDD process. In this, the conversion of the data is performed when multiple forms of the data is available, which needs to be transformed in the required form for further processing. In this data transformation phase, the data is consolidated into forms which are appropriate for mining, and summarization or aggregation operation is also performed.

Data Mining: It is the essential phase of KDD process in which various intelligent methods are used to extract interesting or hidden patterns that are potentially useful for mining.

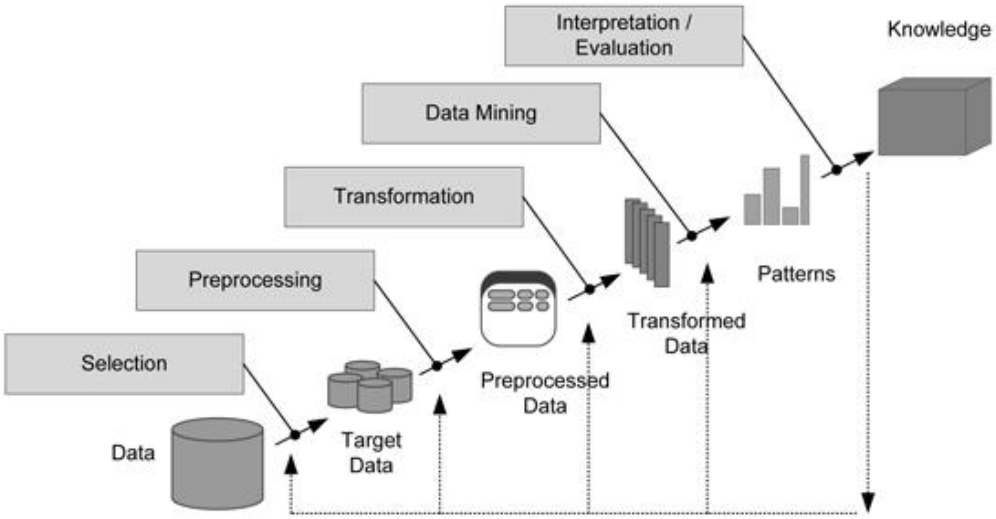

Fig. 1 Knowledge Discovery from Data (KDD)

Pattern Evaluation: In this phase truly interesting patterns are used to represent knowledge on the basis of measures.

Knowledge Presentation: It is the final phase of KDD process where various visualization and knowledge representation methods/techniques are used to present final result or outcome to the user in the form of graphs, charts, canvas, projections etc.

II. DATA MINING METHODS/TECHNIQUES

A. *Classification*

In classification, test data are used to evaluate the accuracy of the classified data. The rules are applicable over the new classified data only when the accuracy of such data is confirmed. The pre-classified interpretations imperative for convenient discrimination are being used by various classifier training algorithms to ascertain the set of parameters [5]. There are number of classification models used, such as Classification by Decision Tree Induction, Bayesian Classification, Support Vector Machines (SVM), and Classification Based on Associations.

For better understanding, the classification is describe as:

Age (Z, "Youth") AND Income (Z, "High") ⟶ Class (Z, "A")

Age (Z, "Youth") AND Income (Z, "Low") ⟶ Class (Z, "B")

Age (Z, " Middle _age") ⟶ Class (Z, "C")

B. *Statistics*

Statistics is a technique to identify data set that contain objects that do not comply normal behavior or model of data. These objects are known as outliers. The analysis of outlier data is called as outlier analysis and it is used to identify records using mean and mode methods.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

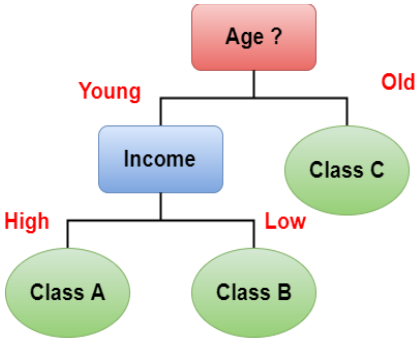

Fig. 2 Classification by Decision Tree

C. *Clustering*

Unlike classification, which analyses class labelled data sets, clustering is a technique in which data objects are analysed without using class labels. Class labels can be constructed for a set of data with the application of clustering. Depending upon the fundamental of maximizing the intraclass and minimizing the interclass, all such data objects are clustered collectively. There are various types of methods are used for clustering such as Hierarchical Agglomerative (divisive) methods, Partitioning Methods, Density based methods, Grid-based methods, Model-based methods, Density based methods.

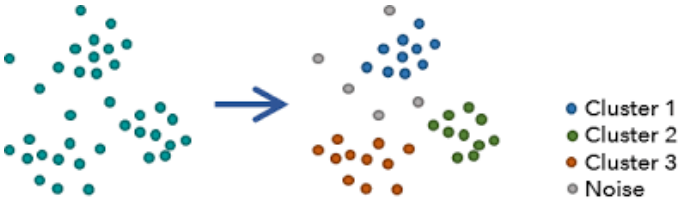

Fig. 3 Clustering

D. *Prediction*

Prediction is a method which defines the further state of the process rather than the current state of the process. Prediction technique is useful for pertaining success ratio high and drop out ratio less. Regression method can be adopted from prediction and regression analysis can be used to build the relationship among one or more autonomous parameters and dependent variables [6]. In data mining, autonomous parameters are the attributes that are known already and respond to those variables that we want to predict before the process completion. Awkwardly so many real world related problems are not easily predictable, such as volume of sales, stock exchange rates, failure rates of product etc. These are difficult to predict problems, because such problems may depend on many complex factors and facts. To predict future standards, there will be a requirement of some intricate techniques such as decision trees, neural networks and logistic regression. Prediction includes various methods such as Linear Regression, Nonlinear Regression, Multivariate Linear Regression and Multivariate Nonlinear Regression.

E. *Association Rule*

Association and correlation is used to find out frequent item set from large data set, which means how frequently the items are purchased together. This type of strategy is helpful in business related decisions like designing of catalogue, cross market analysis and customers behaviour of shopping analysis.

Two types of Association rules are there, one is single dimensional and other is multi-dimensional.

Single dimensional Association rule describe as:

Buys (A, “Computer”) ⟶ Buys (A, “Software”)
Multi-dimensional Association rule describe as:

Age (A, “20-25”) ^ Salary (A, “40000-45000”) ⟶ Buys (A, “Laptop”)

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 7, Issue 8, August 2018

III. Data Mining Applications in Various Areas

Data mining is applied in many areas, including some of such areas combining data mining with statistics, patterns recognition, hidden patterns finding and other important tools. Data mining is used to find out hidden patterns and connections that are normally difficult in finding. Data mining is a standard technology followed by many companies to increase business, as it allows them to study more about their buyers and helps them to making smart shopping decisions and marketing strategies [7]. Data mining is applied in diverse areas such as Marketing and Sales, Agriculture, Healthcare and Insurance, Transportation Surveillance, National Security Agency, Customer analytics, Educational Data Mining, Financial data analysis, Biomedical and DNA data analysis, Retail industry, Telecommunication industry, Quantitative structure.

IV. Text Mining

Immense quantity of unstructured text data can be analysed as well as explored using the process of Text Mining. There are various software available to identify patterns, concepts, keywords and attributes in the data. Text mining also known as text analytics. Text mining is an amalgamation of three independent processes known as text summarization, text categorization and text clustering. Text mining can be considered as a procedure where the data is being synthesized through analysing relations, patterns and discrete practices over unstructured or semi-structured text based data. Text summarization is the process of extracting partial content from whole contents of text automatically [10]. The user can predefine a text and a category can be assigned to such predefined text through text categorization. The text can be segregated into number of clusters with the process of text clustering. There are various approaches of text mining such as keyboard based association analysis, document classification analysis, and document clustering analysis.

V. Conclusion

The Data mining is an important technology regarding finding hidden patterns, discovering knowledge, extraction of information from huge amount of data. It is applied in diverse business areas. Data mining algorithms and techniques like prediction, classification, clustering, association rule are helpful in finding hidden patterns, which facilitates decision making for business growth in future. Data mining technology is focused on diverse areas such as business, retail, education, medical, government, industries and individuals. Today data mining technique is used in every industry, where large amount of data is formed, to extract the useful information which is helpful further in tactical decision making for business. Text mining is the process of summarization of unstructured text into structure text by using various approaches.

ACKNOWLEDGMENT

This work was supported by **Dr. Ashok Kumar Jetawat**, the authors thank to, for his kind guidance in the research and as reviewer to this research paper. The author would also thank the KJIT management for their astonishing support in the research.

References

[1] R Agrawal ,T 1 mielinski, A Swami. Database Mining: A Performance Perspective[J]• IEEE Transactions on Knowledge and Data Engineering, 1993,12:914-925.
[2] Y. Peng, G. Kou, Y. Shi, Z. Chen (2008). "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery" International Journal of Information Technology and Decision Making, Volume 7, Issue 4 7: 639 – 682. doi:10.1142/S0219622008003204.
[3] Han, J. & M. Kamber, Data mining: concepts and techniques, San Francisco: Morgan Kaufman (2001).
[4] D Ramesh , B Vishnu Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013.
[5] Gupta GK (2012) Introduction to data mining with case studies PHI, New Delhi
[6] Kumar R, Kapil AK, Bhatia (2012) A Modified tree classification in data mining. Global Journals Inc. 12, 12: 58-63.
[7] Jain AK (2010) Data Clustering: 50 Years Beyond K- Means. Pattern Recognition Letters, 31(8): 651-666.
[8] Radaideh Q, Nagi E (2012) Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance. IJACSA 3:144-151.
[9] Combes C, Azema J (2013) Clustering using principal component analysis applied to Autonomy – disability of elderly people. Decision Support Systems 55:578–586.
[10] M. Li Lin, Longbing Cao, Jiaqi Wang, Chengqi Zhang, "The Applications of Genetic Algorithms in Stock Market Data Mining Optimisation", Proceedings of Fifth International Conference on Data Mining, Text Mining and their Business Applications,pp- 593-604,sept 2005.
[11] Abbas OA (2008) Comparisons between Data Clustering Algorithms. International Journal of Information Technology 5: 320-325.
[12] Kriegel HK, Borgwardt KM, Kröger P, Pryakhin A, Schubert M, Zimek A (2007) Future trends in data mining. Data Mining and Knowledge Discovery 15:87–97.