A GENERIC STUDY ON DM TECHNIQUES AND APPLICATIONS FOR KNOWLEDGE DISCOVERY

Ms. A. Malarvizhi
Research Scholar and Assistant Professor, PG and Research Department of Computer Science,
H.H. The Rajah's College (A), Pudukkottai, Tamilnadu- 622001
Email:malarselvamtnau@gmail.com

Dr. S. Ravichandran
Assistant Professor and Head, PG and Research Department of Computer Science,
H.H. The Rajah's College (A), Pudukkottai, Tamilnadu- 622001
Email:rajahsravis@gmail.com

ABSTRACT - Data mining (DM) [1] is identifying significanttrends and correlations in large amounts of stored data. A main task in DM is exploration of data for analysis. The need for automated extraction of beneficial knowledge is widely recognized. DM techniques identify similarity in data for necessary inferences. DM can be used in medicinal, marketing, business, social media and any other field having voluminous data. DM or Knowledge discovery in data is a growing field driven by research interests and needs. This paper details on the DM techniques and applications prevalently found for knowledge discovery.

Key words: Data Mining, KDD, DM Techniques

1. INTRODUCTION

Data is information in symbolic form which can be processed on a computer. Organizations accumulate data in different formats. Emergence of electronic data management has led to powerful database systems which collect and manage data. Organizations computerize operations as they grow leading to accumulation of data. They use DM techniques to summarize stored data into meaningful information for decision making for organizational needs or forecasts [2]. DM isessentially used in various fields, thus being an impetus for advancements in the fields of Statistics, Machine Learning and Databases, Pattern identification, Artificial Intelligence and many more [3]. Data holds hidden trends and patterns and once discovered can be used to optimize and improve business. The amount of accumulated data makes it impossible to be analyzed manually, thus turning towards automated tools. DM is a multi-step process that can contain tasks like Association Rule learning, Anomaly Detection, Classification, Regression and Summarization. DM systems can be classified based on data source types, databases, knowledge discovered and DM techniques. Data mining also involves several processes before the application of DM techniques like removing inconsistent data, combining multiple data sources, retrieving relevant data and transforming data into an appropriate form.KDD and DM have been used mainly in research andexperimental environments. Currently generation of data is exponentially increasing due to innovations in the field of computers and introduction of new technologies [4]. Data is ubiquitous and generated from multiple sources like social networking sites, e-commerce websites, data obtained from smartphones and sensors etc.Further, the market size for data mining market will continue to grow. Therefore, Data mining and DMtechniques have become an increasingly important research area[5]. A major goal of this paper is to provide an overview knowledge discovery and data mining and details on different DM techniques and applications.

2. DATA MINING AND KNOWLEDGE DISCOVERY

Data analysis using DM techniques has been adapted in many domains in addition to statistical analysis. Medical field uses DM techniques for knowledge discovery of identifying successful prescription patterns for diseases, prediction using computer aided diagnosis or expert learning. In Health Care Insurance DM is used for analysis of medical procedures, identification of potential business and detection of risky customers. In banking and finance, DM techniques can be used detect fraudulent credit card usage, find hidden correlations in financial indicators, analysis of stock trading based on historical data and buying behavior patterns in marketing. Integrating data mining with forecasting can provide dependable and highquality forecasts. Business leaders can take better decisions with theadvanced forecasting processes offered by DM and Knowled Discovery in Databases (KDD). Prediction of diseases using DM techniques is a motivating task for increasing diagnostic accuracy. Organized DM tools helps decrease cost and time in terms capability and of human resources. Knowledge discovery from medical data is a complicated task, mainly due to noises and irrelevantdata. Using more than one DM techniquefor predicting diseases results in better accuracy. Data mining is a small part of

KDD [6].KDD is the process of turning low-level data into high-level knowledge. It is a nontrivial process of identifying valid, potentially usefuland understandable patterns in data [7]. KDD involves developing an understanding of the application domain and the goals of the data mining processes, selecting a target data set and Integrating the data set, reprocessing, and transformation of data for model building, choosing a suitable DM technique, visualization and interpretation of the results and using this discovered knowledge.

3. DATA MINING ISSUES

Data mining becomesa complicated task due to many reasons. Data may not always be available at a single source and has to be integrated from heterogeneous data sources. User needs on knowledge may vary creating an issue of non-standardization. DM techniques need to cover a broader range of knowledge discovery while mining data. Interactive data mining is needed for users focusing on search patterns and refining mining requests dynamically based on the returned results. Not all DM techniques are interactive. Discovery process needs to be guided with background knowledge to discover patterns. DM techniques need concise terms which have to be abstracted from generic knowledge creating an issue. Though Data Mining Query language allows the user to define mining tasks, it needs to be integrated with data warehouse queries and optimized for efficiency and flexibility. The presentation and visualization of data mining results is another challenge as the discovered patters need to be expressed in high level languages. Incomplete data is another major issue of data mining and data cleaning methods are required to handle incomplete information. Absence of data cleaning methods result is lesser levels of accuracy. Generally, patters discovered are uninteresting as they represent common knowledge and may lack novelty. Performances of DM techniques is also an issue and can be related to efficiency and scalability of data mining algorithms. Other factors like database size, data distributionand complexity of data enhance the challenges in data mining. A data store may contain complex and multimedia data types making it impossible to mine at a single go. Further, data source may be structured, semi structured or unstructured, thus adding challenges to mine from these data sources.

3.1 Data Mining Tasks

DM methods extracting patterns from data are the core of KDD processes. Methods may have different goals and depend on the intended outcome of KDD making it imperative to apply different methods for the needed result. For example, potential customers can be identified by grouping while customer behavior can be predicted using regression models. Data mining goals can be categorized as detailed below.

- Data Processing: It is the selection, sampling, filtering, cleaning or transforming data based on the goals and requirements of KDD.
- Prediction: predicts a specified value in the data attribute. Prediction is also used to validate hypothesis.
- Regression: It is the analysis inter attribute dependency on values.
- Classification: It is determining the data items belonging in a set of pre-defined classes.
- Clustering: It is grouping or partitioning data items, in a set of classes.
- Associations: It is identifying relationships between attributes for patterns.
- Model Visualization: It is the Visualization of discovered knowledge. The techniques may range from scatter plots/histogram to movies.
- Exploratory Data Analysis: It is an interactive exploration of data without dependence on assumptions and identify interesting patterns.

4. DATA MINING METHODOLOGY

Data mining methods can be classified into the following groups.

Statistical Methods: Statistical approaches focuson testing of preconceived hypotheses. They rely on an explicit underlying probability model. These methods require human intervention is for the generation of candidate hypotheses and models.

Case-Based Reasoning (CBR):It tries to solve a given problem by using past experiences and solutions from cases, where a case is a specific previously encountered problem is solved. New problems are added to the case base for future reference.

Neural Networks**:** They are a class of systems modelling human brain with simulated neurons that are connected to each other. These interconnections change in response on the algorithm and obtainoutput, thus enabling learning.

Decision Trees: A decision tree has non-terminal nodes representing a test or decision from the input data. A branch is chosen for the outcome of a test. It starts at the root node and follows assertions until a terminal node (or leaf) is reached and decisions are made. The results can be interpreted as a hierarchical organization of rules.

Rule Induction: They state a statistical correlation between attribute occurrences in data items in a data set. An association rules are formed with confidence and significance values..

***Bayesian Belief Networks*:** They are an acyclic graphof probability distributions matching co-occurrence counts in the data. The nodes represent attribute variables and probabilistic dependencies between the attribute variables are the edges. Conditional probability distributionrelationships between the node and its parentsare associated with each node.

Fuzzy Sets: It is a key methodology processing uncertainties in inconsistency, non-specificity or vagueness. Fuzzy sets is a powerful approach that is helpful in developing uncertain models on uncertain data for better performances traditional systems. They are used on data that are vague as they offer robust and noise tolerant models, while predicting from imprecise inputs.

Rough Sets: They are fuzzy sets with a three-valued membership functions of yes, no and perhaps which deal dela with uncertain data mathematically. They are defined by lower bound (member of the set)and upper bound (non-member of the set) for a set. The upper bound is a union between the lower bound and boundary region(aprobable member of the set). Therefore, rough sets may be viewed as. Like fuzzy sets, rough sets are a ([42]). Both fuzzy sets and rough sets smoothened outputs are used by other methods like classification or clustering for results.

5. DM ALGORITHMS FOR DATA ANALYSIS

DM algorithms create models for calculations. They segment, associate or examine data for processing. DM techniques based on Algorithms fall into one the tasks detailed in the above section. This section details on a few prevalent and regularly used DM algorithms with a comparison of performances.

K-Means Clustering: Data points are clustered based on its distance from the cluster centroid. The accuracy is dependent on the value of K and determining the appropriate number of clusters in K-means clustering is a complex affair.

***Nearest Cluster Algorithm**:* This is a miniature K-nearest neighbor clustering algorithm. The input is a set of cluster centers generated from the training data set using standard clustering algorithms.

***ID3 Algorithm*:** It is specific machine-learning algorithm that constructs a decision tree based on training data sets. It is a useful Inductive Logic Programming technique.

***J48 Algorithm**:* This algorithm addresses loop holes in ID3 and an enhanced version of C4.5 algorithm.

Partial Decision Tree(PART):This algorithm is derived from C4.5 and RIPPER algorithms and generates rules by repeatedly producing partial decision trees. This algorithm is. It does not perform global optimization. Once a decision tree is generated, it is then transformed it into a rule set and simplified.

Support Vector Machines (SVM): This classifier is considered for binary classification. It transforms the training data into a vast feature space dimension. It first maps the input vector into a higher dimensional feature space and then obtains the optimum. Simplification depends on the geometrical characteristics of training data.

Fuzzy Logic (FL):It processes network inputs and detects anomalies. It is based on the concept that the set values for reasoning series between 0 and 1. The truth of a statement can range between 0 and 1 and not constrained to two accuracy values.

Naïve Bayes: This is a simple approach and based on the inferences of probabilistic graphic models specifying probable dependencies using a graph structure A probabilistic graphical form is a graph in which nodes represent accidental variables, and the arcs symbolize restricted confidence assumptions. It provides a compact representation of combined probability distributions.

Random Forest Classifier (RFC): These processes combinerandom selection of featuresand trapping to construct a group of decision trees with restricted dissimilarity. A Decision Table and Random Forest classifiers classify accurately.

Apriori: This is an iterative and confidence-based Association Rule Mining algorithm. It produce frequent item sets and then scans those frequent item sets to distinguish most frequent items. The process is iterative. Frequent item sets are generated from one step by constructing another item sets and by joining with previous frequent item sets. Table 1 lists comparative performances of DM Algorithms.

Table 1 Comparative Performance of DM Algorithms

S.No	Algorithm	PercentageofSuccessful Prediction(PSP)%	Training Time(TT) Sec.
1	K-Means	88.95	51.2
2	Nearest Cluster	91.32	9.73
3	ID3	92.52	19.61
4	J48	92.06	15.85
5	Partial Decision Tree	44.77	165
6	Support Vector Machines	80.48	221.38
7	Fuzzy Logic	93.90	872.8
8	Naïve Bayes	77.42	4.67
9	Random Forest Classifier	91.91	490
10	Apriori	86.70	19

Table 1 shows the accuracy levels of methods on data sets. Also, the time complexity in seconds of various classifiers to build a model for training data [8] [9] is listed.The results indicate higher accuracy inID3, Nearest Cluster and J48, while K-means and Fuzzy Logic have an acceptable level of accuracy.

6. CONCLUSIONS

Data mining and knowledge discovery applications have important significance in decision making and are an essential component in many organizations and fields.DM techniques are used to extract knowledge or interesting patters from large amounts of data. Researcherscan focus on issues of data mining and be the impetus for creation of analytical tools in data mining. Software based on data mining algorithms should hide complexities from end-users and create use cases with tight constraintsindesign. In crime detection, DM techniques have scope for improvement in visualand investigative techniques and can be useful detection of crime on social networks interrelationships. Prediction of non-academic factors are non-cognitive and DM techniques can be used to measure and monitor these factors. This paper has detailed on the issues, methods, techniques of data mining and knowledge discovery while comparing the performances of selected classifiers. It is concluded that knowledge discovery using DM techniques has a lots of scope for improvement as it a growing and evolving field in the field of data science.

REFERENCES

[1] Gorunescu, F, Data Mining: Concepts, Models, and Techniques, Springer, 2011

[2] Han, J., and Kamber, M., Data mining: Concepts and techniques, Morgan-Kaufman Series of Data Management Systems San Diego:Academic Press, 2001.

[3] NeelamadhabPadhy, Dr. Pragnyaban Mishra and RasmitaPanigrahi, “The Survey of Data Mining Applications and Feature Scope, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)”, vol.2, no.3, June

[4] Ha, S., Bae, S., & Park, S. (2000). Web mining for distance education. In IEEE international conference on management of innovation and technology (pp. 715–719).

[5] Fayyad, U., Djorgovski, S. G., & Weir, N. (1996). Automating the analysis and cataloging of sky surveys. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 471–494). Cambridge, MA: MIT Press

[6] Brachman, R., and Anand, T. The process of knowledge discovery in databases: A human-centered Approach. In Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge, 1996.

[7] Piatetsky-Shapiro , Fayyad, U., G., and Smyth, P. From Data Mining to Knowledge Discovery: An Overview. In Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining, MIT Press, 1-36, Cambridge, 1996

[8] Nikita Jain, Vishal Srivastava,” DATA MINING TECHNIQUES: A SURVEY PAPER”, IJRET: International Journal of Research in Enginee ring andTechnology, Volume: 02 Issue: 11, eISSN: 2319-1163, Nov-2013

[9] Ms. Ruth D, Mrs. LovelinPonnFelciah M,” A Survey on Intrusion Detection System with Data Mining Techniques”, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 3, ISSN 2348 – 7968, May 2014.