DATA MINING AND ITS ISSUES

Satveer Kaur

Assistant Professor, Department of Computer Science, Dashmesh Khalsa College,
Zirakpur (Mohali), (India)

ABSTRACT

Data Mining is a practical field of research development which aims to extract knowledge by analyzing data sets. This field has achieved conclusive success in almost every area such as Wireless Sensor Network, Social Network, healthcare, large spatial databases etc. In all these areas a lot of issues arise from which a few are resolved and others are unresolved. In this paper, we discussed Data Mining and issues like scalability, memory, perception, privacy, missing value imputation, feature selection, outlier detection, cluster analysis of high dimensional data, imbalanced classes in classification, mining from complex/distributed data etc. regarding data mining which are faced while extracting information in the above said areas.

Keywords: Clustering, Data Mining, Knowledge Discovery, Perception, Privacy.

I INTRODUCTION

Data mining has been the one of the buzz phrases in the business intelligence research and industry communities over the last few years [1]. With the advancement of technology, the degree of information being produced and stored is developing exponentially in data marts and data warehouses. As a result, traditional and informal mixtures of statistical techniques and data management tools are not efficient for analyzing vast collection of data [2]. Therefore a quick data analysis method has been discovered known as Knowledge Discovery so that knowledge could be extracted from various databases. Data mining is one of the steps of Knowledge discovery [3]. Data mining is about automatically or semi-automatically extracting knowledge from the data. The knowledge to be searched from the data can be defined in different ways such as searching structured, frequent, approximate etc. patterns in data [4] [5], association rules [6][7][8], grouping/clustering/bi-clustering data according to one or many criteria[9]. Data mining is the central element of the knowledge discovery process. Mining information and knowledge from large data sources such as weather databases, financial data portals or emerging disease information systems has been recognized as an opportunity of major revenues from applications such as warehousing, process control, and customer services [10].

What Data mining does?

1. Data mining takes information as data and produce learning.
2. Data mining discovers new example inserted in huge information sets.
3. Data mining investigates information sets to discover unspecified connections and to outline the information which is exceptionally valuable having reasonable correlations and comparable designs [11].

The target of mining is the data which can be computer based acquisition data, observed data, large scale data and variety of data. Data mining can be portrayed as knowledge discovery from data with three expressions: Data preprocessing, Data displaying and Data post-handling [11].

1. Data pre-processing:- The first step of data mining is to develop raw information. About 60-90% of aggregate time is consumed in understanding and planning information. Under this step, numerous exercises are to be performed e.g. portrayal of information, joining of two or more information sets to structure a solitary one, decreasing the information by removing unimportant information etc.
2. Data displaying: - Following the planning of crude or raw information, relations between different information are uncovered to explore new examples. For this, different data mining calculations are utilized such as classification, regression, clustering and association etc.
3. Data post-handling:- In this step, assessment of new separated examples is carried out called as learning. For learning exposure different perceptive and graphic calculations are utilized.

II ISSUES IN DATA MINING

We will discuss data mining issues which are faced by various areas or domains where the process is being followed to extract information from large databases.

2.1 Missing Value Imputation

Many existing industrial and research data sets can contain missing values. They can be introduced due to different reasons such as equipment errors, manual data entry and incorrect measurements. Missing data is often found in most of the information resources used. The results obtained from such incomplete data sets may be very hard and of low quality. Missing values can be managed by carrying out three different ways: first is to discard the instances that are having missing values in their attributes; second is to estimate the parameters like variance and covariance based on complete data using maximum probability procedures and the third is the imputation by filling in the missing values with estimated ones. When the attributes of data set are dependent, missing values can be calculated by establishing relationships among attributes [3].

2.2 Feature Selection

Datasets that are used for mining of data may consist of various dimensions but it is possible that they all may not be relevant. Feature selection removes irrelevant, redundant or noisy data. It also removes the number of columns that results into better performance of classification, clustering and association algorithms. The process comprises of four steps named as feature subset selection, subset evaluation, stopping criteria and validation [12]. Feature can be chosen from the given data set either in the forward manner or in the backward manner [3].

2.3 Outlier Detection

There can be some instances in the data sets used for data analysis that do not conform to the general properties of the other instances that are present in the data sets. These types of entities are known as outliers or anomalies. Traditional methods to detect these anomalies have been grouped as unsupervised, supervised and semi-supervised. Existing methodologies for anomaly detection are statistical methods, neural networks, machine learning models and hybrid forms etc. but all these approaches need the data that is either categorical or cardinal or ordinal data. Outlier detection techniques need to be investigated for temporal, spatial and video data too [3].

2.4 Cluster Analysis Of High Dimensional Data

Basically clustering methods are hierarchical, partitioning based, grid based, density based, and model based [13]. But these methods are not optimal for computational biological and clinical data which are of very high dimension. For any point the difference between the distance to its nearest point or to the farthest point turns out to be insignificant as the dimensions of data grow. This process may leave the clustering algorithm results to any upset state due to noise and make the effect of clustering useless [14]. Many clusters during clustering may be present in distinct subspaces of smaller dimensions but the sets of dimensions may be overlapping or non-overlapping [12].

2.5 Imbalances Classes In Classification

A class-imbalanced classifier is a rule which is used to forecast the class members of new samples obtained from an available data set. These class members have the disagreed class sizes. When the class sizes are very different, the classification algorithms may favor the majority class which will produce poor accuracy in minority class proposition [3]. Two methods exists which deal with class imbalance problem. In first method, cluster based over-sampling is done. In the second method, clustering each class identifies sub-clusters. Afterwards, re-sampling of each sub-cluster is done to increase class-size. This will remove the class imbalance. But still class overlapping and data shift problems may be present in the data. These may lead to inaccurate results of classification [15].

2.6 Privacy Of Data

Today, major requirements of the significant areas like healthcare systems, organizational security and space research organizations is collection and analysis of large amount of information which includes spatial as well as non-spatial data [16]. Although, data mining is useful to extract information, many data holders don’t feel safe to supply their data for mining due to the fear of violating the privacy. Many data repositories may be accessed via public interfaces that allow cumulative querying. It gives an opportunity to an opponent to find perceptive details of the data using queries. Data can be distributed across different sites or computers. A single site may need the information from different data sets which are partitioned either horizontally or vertically over these sites. Whenever these sites are not desired to share their entire data, they may allow restricted data sharing by using some of protocols [3].

2.7 Mining From Complex/Distributed Data

The growth of data mining has resulted in emergence of complex data such as WWW pages, DNA representations etc. that may involve heterogeneous data sources. Here, traditional methods cannot be applied due to assumption of data storage in single file [3]. Three approaches namely multi-relational data mining, multi-database mining and combined mining are developed for complex/distributed data. First two approaches extract features from various tables into a single table. The third approach finds the clusters of patterns from multiple data sets [17].

2.8 Scalability Issues

Scalability issues involve volume of data, variety of data and response time for data. The large data do not fit into main memory for processing, except the requirements for efficient access to data as data transfer could be infeasible. While the variety of the data, the amount of features to be taken into account increases considerably e.g. contextual information. For computing analytics and historical processing of data, the processing time can possibly be large, each time more end-users need faster responses. It may have an immediate impact on the decision making. Another issue is while using real data that requires a lot of storage to handle the data [18].

2.9 Memory Issues

Another issue that arises when mining large sets of data such as the log files of a big campus is the amount of data in input/output operations and the collected memory. The memory issue requires analyzing three options very carefully while mining large sets of data [18].

- If data is loaded completely into memory then it is called in-memory. This option allows fast access to data.
- If the data access is not as fast as in-memory and depends on the database server, the database can be distributed. By doing so, large amounts of data can be stored.
- If data access is fast, the data is accessed through Hadoop. By this way, large amount of data can also be stored.

2.10 Issues By Online Mining of Data Streams

Online data mining is attracting interest. Many internet-based applications generate data streams that could be of interest of mining for pattern discovery. This mining is different from the mining large data sets that are already stored in databases, data warehouses or file systems. Online data mining has to deal with the incoming flow of data. Another issue is the data variety i.e. structured or unstructured, low or high volume that can appear with the data stream. Window-based sampling, chain sampling techniques etc. are used to deal with the incoming flow of data [18].

2.11 Issues On Mining of Speech, Audio And Dialog

With the advent of inexpensive storage space and fast processing since the past decade, data mining research has started to grow in areas of speech and audio processing as well as spoken language dialog. It has been increased by a lot of audio data that is becoming available from multimedia resources such as webcasts, conversations, music, meetings, voice messages, lectures, television and radio etc. Efficient techniques for mining speech, audio, and dialog data can effect various business and government applications. The technology for monitoring conversational speech to discover patterns, capture useful trends is essential for intelligence and law enforcement organizations. It is useful for analyzing, monitoring and tracking customer preferences and interactions to better establishments of customized sales and technical support strategies. It is an essential tool for media content management [19].

2.12 Perception Issues

Perception is essential for data mining. When dealing with real-world problems e.g. expert systems analyzing financial data etc., there is a difference between the real world and what the user or expert system perceives to be the real world. Data mining retrieves the perceived data and the problem is to bring together this perceived data with the real world. Perceived data can be in many distinct ways, depending upon the main point of interest of a given user or a group of users. Granulation of information is needed that supports perception by creating percepts that promote this interpretability and channels pursuits of data mining towards more efficient and feasible processing cost at some level of perception [20].

III CONCLUSION

Data mining has been an active and practical field of research for a long time. Before applying the techniques of data mining such as clustering, classification etc. one has to preprocess the data to obtain good quality results and well known about its issues. This paper has discussed data mining concept and its issues. The future of data mining is very vivid.

REFERENCES

[1] A.G. Buchner, M. Baumgarten, M.D. Mulvenna, R. Bohm, S.S.Anand, Data Mining and XML: Current and Future Issues, Proc. First International Conference on Web Information Systems Engineering, 2000, 131-135.

[2] M. Sushmita, P.K. Sankar, M.Pabitra, Data Mining in Soft Computing Framework: A Survey, *IEEE Transaction on Neural Networks*, 13(1), 2002, 3-14.

[3] Archana Purwar, Sandeep Kumar Singh, Issues in Data Mining: A comprehensive Survey, IEEE International Conference on Computational Intelligence and Computing Research, 2014, 1-6.

[4] R.J.Bayrado, Efficiently mining long patterns from databases, Proc. ACM-SIGMOD International Conference Management of Data, Seattle, 1998, 8593.

[5] J.Han, J.Pei, Y.Yin, R.Mao, Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach, *Data Mining and Knowledge Discovery*, 8(1), 2004, 53-87

[6] R. Agrawal, T.Imielinski, A.Swami, Mining association rules between sets of items in large databases, Proc. ACM-SIGMOD International Conference on Management of Data, Washington, 1993, 207216

[7] R.Agrawal and R.Srikant, Fast algorithms for mining association rules, Proc. International Conference on Very Large Data Bases, Santiago, 1994, 487499.

[8] avasere, A.,Omiecinski, E., and Navathe, S., An efficient algorithm for mining association rules in large databases, Proc. International Conference on Very Large Data Bases, Zurich, 1995, 432443.

[9] J.Han, M.Kamber, *Data Mining –Concepts and Techniques* (Morgan Kaufmann Publishers, 2000).

[10] Matthias Klusch, Stefano Lodi, Gianluca Moro, IEEE/WIC International on Intelligent Agent Technology, 2003, 211-217.

[11] Bhavya, Mahak, Pooja Mittal, Data Mining in Medicine: Current Issues and Future Trends, International Conference on Advances in Computer Engineering and Applications, 2015, 979-983.

[12] H.Liu and L.Yu, Toward Integrating Feature Selection Algorithms for Classification and Clustering, *IEEE Transactions on Knowledge and Data Engineering*, 17(4), 2005, 491-502.

[13] J.Han, M.Kamber, *Data Mining –Concepts and Techniques* (2nd ed. Morgan Kaufmann Publishers, 2006).

[14] H.Kriegel, E.Muller and A.Zimek, Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based clustering and Correlation Clustering, *ACM Transactions on Knowledge Discovery from Data*, 3(1), 2009.

[15] T.Jo and N.Japkowicz, Class Imbalances versus Small Disjuncts, *SIGKDD Explorations*, 6(1), 2004, 40-49.

[16] Dr. P.Kamakshi, a survey on Privacy Issues and Privacy Preservation in Spatial Data Mining, International Conference on Circuit, Power and Computing Technologies, 2014, 1759-1762

[17] L.Cao, H.Zhang, Y.Zhao, D.Luo and Chengqi Zhang, Combined Mining: Discovering Informative Knowledge in Complex Data, *IEEE Transactions on Systems, Man, and Cybernetics: Part B*, 41(3), 2011, 699-712.

[18] Vladi Kolici, Fatos Xhafa, Leonard Barolli, Algenti Lala, Scalability, Memory Issues and Challenges in Mining Large Data Sets, International Conference on Intelligent Networking and Collaborative Systems, 2014, 268-273.

[19] Mazin Gilbert, Roger K. Moore, Geoffrey Zweig, Introduction to the Special Issue on Data Mining of Speech, Audio, and Dialog, *IEEE Transactions on Speech and Audio Processing*, 13(5), 2005, 633.

[20] Michael H. Smith and Witold Pedrycz, Perception Issues in Data Mining, International Conference on Systems, Man and Cybernetics. E-Systems and e-Man for Cybernetics in Cyberspace, 2001, 2553.