ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, Issue 1, January 2013

A Comprehensive Study of Data Mining and Application

Dheeraj Agrawal

Abstract— Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis. Data mining applications can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering and forecasting. An application compared to other data analysis applications such as structured queries or statistical analysis software. Illustration of the data mining application that offer opportunities for research. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. In this paper present a comprehensive study for data mining, models, issue, and focuses its application.

***Index Term*—Data mining Application, Issues, Models, Machine Learning, OLAP, Warehousing.**

I. Introduction

Data Mining or knowledge discovery is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data [2].

A. Knowledge Discovery in Databases

Non-trivial process of identifying valid, novel, potentially Useful & understandable patterns & relationships in data (knowledge = patterns & relationships).

a) Pattern: expression describing facts about data set

b) Relation: expression describing dependencies between data and/or patterns.

c) Process: KDD is multistep process, involving data preparation, data cleaning, and data mining.

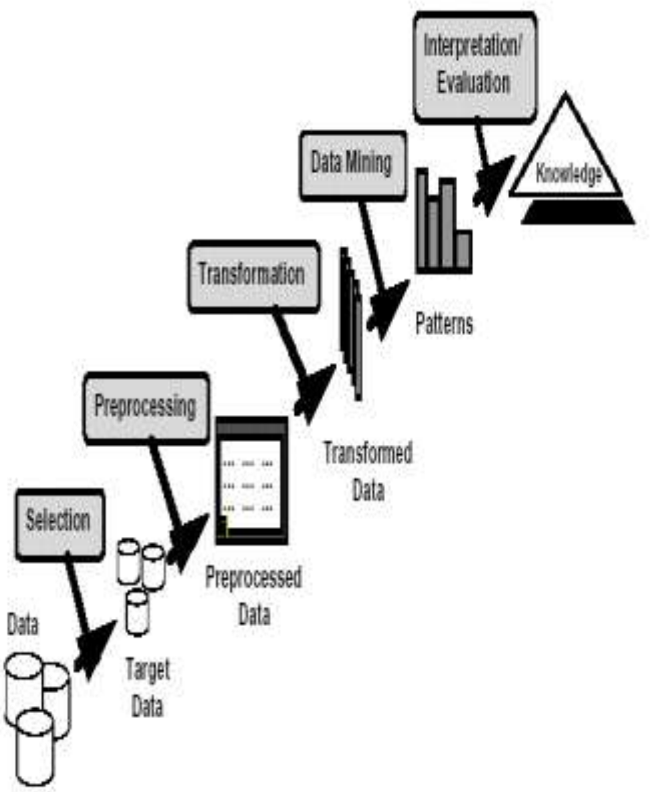
d) Valid: discovered patterns, relationships should be valid on new data with some certainty (or correctness, below error level)

e) Novel: not yet known (to KDD system)

Potentially useful: should lead to potentially useful actions

Understandable: provide knowledge that is understandable to Humans or that leads to a better understanding of the data set.

II. Steps in Knowledge Discovery



1. Understanding the problem domain.

In this step one works closely with domain experts to define the problem and determine the project goals, identifies key people, and learns about current solutions to the problem.

2. Understanding the data.

This step includes collection of sample data, and deciding which data will be needed including its format and size

3.Preparation of the data.

Data cleaning and preprocessing: remove noise handle missing data & unlabeled data. This is the key step upon which the success of the entire knowledge discovery process depends; it usually consumes about half of the entire project effort. In this step, we decide which data will be used as input for data mining tools of step four.

4. Data mining.

This is another key step in the knowledge discovery process. Although it is the data mining tools that discover

Dheeraj Agrawal, Department of Computer Science, *SSCET Bhilai(CG),India ,*

new information, their application usually takes less time than data preparation.

5. Evaluation of the discovered knowledge.

This step includes understanding the results, checking whether the new information is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.

6. Using the discovered knowledge.

This step is entirely in the hands of the owner of the database. It consists of planning where and how the discovered knowledge will be used. The application area in the current domain should be extended to other domains.

III. Data Mining Issues

As data mining initiatives continue to evolve, there are several issues may decide to consider related to implementation and oversight.

a) Data Quality

Data quality is a multifaceted issue that represents one of the biggest challenges for data mining. Data quality refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed.

b) Interoperability

Related to data quality, is the issue of interoperability of different databases and data mining software. Interoperability refers to the ability of a computer system and/or data to work with other systems or data using common standards or processes.

c) Mission creep

Mission creep is one of the leading risks of data mining cited by civil libertarians, and represents how control over one’s information can be a tenuous proposition. Mission creep refers to the use of data for purposes other than that for which the data was originally collected. This can occur regardless of whether the data was provided voluntarily by the individual or was collected through other means.

d) Privacy

As additional information sharing and data mining initiatives have been announced, increased attention has focused on the implications for privacy. Concerns about privacy focus both on actual projects proposed, as well as concerns about the potential for data mining applications to be expanded beyond their original purposes.

IV. THE DATA MINING MODELS

The data mining models are of two types Predictive and Descriptive.

The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc.

The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. Ex. Clustering, Summarization, Association rule, Sequence discovery etc.

a) Classification

This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules.

b) Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification.

c) Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict.

d) Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one.

e) Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs.

V. DATA MINING AND DATA WAREHOUSING

The construction of a data warehouse, which involves data cleaning and data integration, can be viewed as an important pre-processing step for data mining. However, a data warehouse is not a requirement for data mining. Building a large data warehouse that consolidates data from multiple sources, resolves data integrity problems, and loads the data into a database. Data mining uses the data warehouse as the source of information for knowledge data discovery (KDD) systems through an amalgam of artificial intelligence and statistics-related techniques to find associations, sequences, classifications, clusters, and forecasts.

VI. DATA MINING AND OLAP

The capability of OLAP to provide multiple and dynamic views of summarized data in a data warehouse sets a solid foundation for successful data mining. The term OLAP, standing for Online Analytical Processing, is often used to describe the various types of query- driven analysis that are undertaken when analyzing the data in a database or a data warehouse OLAP provides for the selective extraction and viewing of data from different points of view; these views are generally referred to as dimensions. A powerful paradigm that integrates OLAP with data mining technology is OLAM (Online Analytical Mining) which is sometimes referred to as OLAP mining. OLAM systems are particularly important because most data mining tools need to work on integrated, consistent, and cleaned data, which again, requires costly data cleaning, data transformation and data integration as pre-processing steps.

VII. DATA MINING AND MACHINE LEARNING

Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience. Machine learning aims to provide increasing levels of automation in the knowledge engineering process, replacing much time-consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data.

VIII. DATA MINING AND STATISTICS

The disciplines of statistics and data mining both aim to discover structure in data. So much do their aims overlap, that some people regard data mining as a subset of statistics. But that is not a realistic assessment as data mining also makes use of ideas, tools, and methods from other areas particularly database technology and machine learning, and is not heavily concerned with some areas in which statisticians are interested. Most of the learning algorithms use statistical tests when constructing rules or trees and also for correcting models that are over fitted. Statistical tests are also used to validate machine learning models and to evaluate machine learning algorithms.

IX. APPLICATION OF DATA MINING

A. Spatial Data Mining

Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases.

A spatial database stores a large amount of space-related data, such as maps, preprocessed remote sensing or medical imaging data, and VLSI chip layout data.

Spatial Data Cube Construction and Spatial OLAP

As with relational data, we can integrate spatial data to construct a data warehouse that facilitates spatial data mining. A spatial data warehouse is a subject-oriented, integrated, time variant and nonvolatile collection of both spatial and non spatial data in support of spatial data mining and spatial-data-related decision-making processes.

There are three types of dimensions in a spatial data cube:

a) A non spatial dimension
b) spatial-to-non spatial dimension
c) spatial-to-spatial dimension

B. Spatial Clustering Methods

Spatial data clustering identifies clusters, or densely populated regions, according to some distance measurement in a large, multidimensional data set.

Spatial Classification and Spatial Trend Analysis Spatial classification analyzes spatial objects to derive classification schemes in relevance to certain spatial properties, such as the neighborhood of a district, highway, or river.

C. Mining Raster Databases

Spatial database systems usually handle vector data that consist of points, lines, polygons (regions), and their compositions, such as networks or partitions. Typical examples of such data include maps, design graphs, and 3-D representations of the arrangement of the chains of protein molecules.

D. Multimedia Data Mining

A multimedia database system stores and manages a large collection of multimedia data, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages Similarity Search in Multimedia Data When searching for similarities in multimedia data, we can search on either the data description or the data content approaches:

a) Color histogram–based signature
b) Multifeature composed signature
c) Wavelet-based signature

E. Multidimensional Analysis of Multimedia Data

To facilitate the multidimensional analysis of large multimedia databases, multimedia data cubes can be designed and constructed in a manner similar to that for traditional data cubes from relational data.

A multimedia data cube can contain additional dimensions and measures for multimedia information, such as color, texture, and shape.

F. Classification and Prediction Analysis of Multimedia Data

Classification and predictive modeling can be used for mining multimedia data, especially in scientific research, such as astronomy, seismology, and geoscientific research.

G. Mining Associations in Multimedia Data

a) Associations between image content and nonimage content features.
b) Associations among image contents that are not related to spatial relationships.
c) Associations among image contents related to spatial relationships.

H. Audio and Video Data Mining

An incommensurable amount of audiovisual information is becoming available in digital form, in digital archives, on the World Wide Web, in broadcast data streams, and in personal and professional databases, and hence a need to mine them.

I. Text Mining

Text Data Analysis and Information Retrieval Information retrieval (IR) is a field that has been developing in parallel with database systems for many years. Basic Measures for Text Retrieval: Precision and Recall.

J. Mining the World Wide Web

The World Wide Web serves as a huge, widely distributed, global information service center for news, advertisements, consumer information, financial management, education, government, e-commerce, and many other information services. The Web also contains a rich and dynamic collection of hyperlink information and Web page access and usage information, providing rich sources for data mining.

K. Scientific Applications

Data collection and storage technologies have recently improved, so that today, scientific data can be amassed at much higher speeds and lower costs. This has resulted in the accumulation of huge volumes of high-dimensional data, stream data, and heterogeneous data, containing rich spatial and temporal information. Consequently, scientific applications are shifting from the “hypothesize-and-test” paradigm toward a “collect and store data, mine for new hypotheses, confirm with data or experimentation” process. This shift brings about new challenges for data mining

L. *Data Mining for Intrusion Detection*

The security of our computer systems and data is at continual risk. The extensive growth of the Internet and increasing availability of tools and tricks for intruding and attacking networks have prompted intrusion detection to become a critical component of network administration.

M. *Visual and Audio Data Mining*

Visual data mining discovers implicit and useful knowledge from large data sets using data and/or knowledge visualization techniques.

In general, data visualization and data mining can be integrated in the following ways:
a) Data visualization
b) Data mining result visualization
c) Data mining process visualization
d) Interactive visual data mining

N. *Data Mining and Collaborative Filtering*

A collaborative filtering approach is commonly used, in which products are recommended based on the opinions of other customers. Collaborative recommender systems may employ data mining or statistical techniques to search for similarities among customer preferences.

O. *Security of Data Mining*

Data security–enhancing techniques have been developed to help protect data. Databases can employ a multilevel security model to classify and restrict data according to various security levels, with users permitted access to only their authorized level. Privacy sensitive data mining deals with obtaining valid data mining results without learning the underlying data values.

X. Limitations of Data Mining

Data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personal related, rather than technology-related.

Data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns.

Another limitation of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship.

XI. CONCLUSION

Data mining applications can use a variety of parameters to examine the data as an application, compared to other data analysis applications, such as structured queries or statistical analysis Software, data mining represents a difference of kind rather than degree. Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets Data mining is becoming increasingly common in both the private and public sectors. Data mining applications in various fields use the variety of data types. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Efficient and effective data mining in large database poses numerous requirements and great challenges to researchers and developers. The dramatically increasing demand for better decision support is answered by an extending availability of knowledge discovery, and data mining is one step at the core of the knowledge discovery process.

REFERENCES

[1]Joyce Jackson "DATA MINING: A CONCEPTUAL OVERVIEW", Communications of the Association for Information Systems (Volume 8, 2002) 267-296.

[2]Larose, D. T., “Discovering Knowledge in Data: An Introduction to Data Mining”, ISBN 0-471-66657-2,John Wiley & Sons, Inc, 2005.

[3]Dunham, M. H., Sridhar S., “Data Mining: Introductory and Advanced Topics”, Pearson Education,

[4]New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.Mr. S. P. Deshpande "DATA MINING SYSTEM AND APPLICATIONS: A REVIEW"International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010.

[5] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[6] Chen, M.-S., Jan, J., Yu, P.S. (1996) “Data Mining: An Overview from a Database Perspective”, IEEE Transactions on Knowledge and Data Engineering”.

[7] Agrawal, R., Imielinski, T., Swami, A.(1993), “Database Mining: A Performance Perspective, IEEE Transactions Knowledge and Data Engineering, (5), 914-925.

[8] Jeffrey W. Seifert"Data Mining: An Overview"CRS Report for Congress,2004.

[9] Dunham, M. H., Sridhar S., “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006.

[9] Maciag, T., Hepting, D. H., Slezak, D., Hilderman, R. J., “Mining Associations for Interface Design”.Lecture Notes in Computer Science, Springer Berlin / Heidelberg, Volume 4481, pp. 109-117, June 26, 2007.

[10] Pazzani, M.J., Knowledge discovery from data?, IEEE Intelligent Systems, pp.10-13, March/April 2000.

[11] Kraft, M. R., Desouza, K. C., Androwich, I., “Data Mining in Healthcare Information Systems: Case Study of a Veterans’ Administration Spinal Cord Injury Population”. IEEE, Proceedings of the 36th Hawaii International Conference on System Sciences, 0-7695-1874-5/03, 2002.

[12] Spangler, W. E.; May, J. H., Vargas, L. G. (1999), “Choosing Data-Mining Methods For Multiple Classification: Representational And Performance Measurement Implications For Decision Support “,Journal of Management Information Systems, Summer,37-62.

[13] Moore, A., Lee, M. S. (1998), “Cached Sufficient Statistics for Efficient Machine Learning with Large Datasets”, Journal of Artificial Intelligence Research, (8), 67.91.

[14]Han, J., Kamber, M. (2001), Data Mining: Concepts and Techniques, Morgan-Kaufmann Academic Press, San Francisco.

[15]Ye, Jianming (1998), “On Measuring and Correcting the Effects of Data Mining and Model Selection”, Journal of the American Statistical Association, (93:441), 120-131