International Journal of Innovative Research in Engineering
Volume 3, Issue 5 (September-October 2022), PP: 127-131.
www.theijire.com
ISSN No: 2582-8746

Applications and Trends in Data Mining

Vaishnavi S[1], Atchaya B[2], Sri Suvetha R M[3], Swetha K[4]
[1,2,3,4] *Computer Science, Sri G.V.G Visalakshi College for Women, Udumalpet, Tamilnadu, India.*

How to cite this paper:
Vaishnavi S[1], Atchaya B[2], Sri Suvetha R M[3], Swetha K[4], "Applications and Trends in Data Mining", IJIRE-V3I05-127-131.

Abstract: *Studied principles and methods for mining relational data, data ware houses, and complex types of data. Datacollected in the banking and financial industries are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability, and popularity of business conducted on the Web, or e-commerce. The integration of telecommunication, computer network, internet, and numerous other means of communication and computing is also underway. Biological data mining has become an essential part of new research field called bioinformatics. Scientific applications are shifting from the "hypothesize-and-test" paradigm toward a "collected and store data, mine for new hypotheses, confirm with data or experimentation" process. Misuse detection searches for patterns of program or user behavior that match know intrusion scenarios, which are stored as signatures.*

Keyword*: Data Mining Applications, Data Mining for Financial Data Analysis, Data Mining for the Retail Industry, Data Mining for the Telecommunication Industry, Data Mining for Biological Data Analysis, Data Mining in Other Scientific Applications, Data Mining for Intrusion Detection.*

I. INTRODUCTION

As a young research field, data mining has made broad and significant progress since its early beginnings in the 1980s.Today, data mining is used in a vast array of areas, and numerous commercial data mining systems are available. Many challenges still remain. Study applications and trends in data mining. Begin by viewing data mining application in business and in science. Tips on what to consider when purchasing a data mining software system. Additional themes in data mining are described, such as theoretical foundations of data mining, statistical techniques for data mining, visual and audio mining, and collaborative recommender systems that incorporate data mining techniques. The social impacts of data mining are discussed, including ubiquitous and invisible data mining and privacy issues. Finally, Examine current and expected data mining trends that arise in response to challenges in the field.

II. DATA MINING APPLICATION

We have studied principles and methods for mining relational data, data warehouses and complex types of data (including stream data, time-series and sequence data, complex structured data, spatiotemporal data, multimedia data, heterogeneous multi database data, text data and web data). Because data mining is a relatively young discipline with wide and diverse applications, there is still a nontrivial gap between general principles of data mining and application- specific, effective data mining tools. Examine a few application domains and discuss how customized data mining tools should be developed for such application.

Data Mining for Financial Data Analysis:

Most banks and financial institutions offer a wide variety of banking services and investment services. Some also offer insurance services and stock investment services. Financial data collected in the banking and financial industries are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis.

Design and construction of data warehouses for multidimensional data analysis and data mining: Like many other application, data warehouses need to be constructed for banking and financial data. Multidimensional data analysis methods should be used to analyze the general properties of such data. For example, one may like to view the debt and revenue changes by month, by region, by sector, and by other factors, along with maximum, minimum, total, average, trend and other statistical information. Data warehouses, data cubes, multi feature and discovery-driven data cubes, characterization and class comparisons, and outlier analysis all play important roles in financial data analysis and mining. Loan payment prediction and customer credit analysis are critical to the business of bank. Many factors can strongly or weakly influence loan payment performance and customer credit rating. Data mining methods, such as attributes selectionand attribute relevance ranking, may help identify important factors and eliminate irrelevant ones. For example, factors relating to the risk of loan payments include loan –to value ratio, term of the loan, debt ratio (total amount of monthly debtversus the total monthly income), payment-to payment ratio, customer income level, education level, residence region,and credit history may find that, say payment-to-income ratio is a dominant factor, while education level and debt ratio are not. The bank may then decide to adjust its loan-granting policy so as to grant loans to those customers whose applications were previously denied but whose profiles show relatively low risks according to the critical factor analysis.

Classification and clustering of customers for targeted marketing: Classification and clustering methods can be used

for customer group identification and targeted marketing. For example, we can use classification to identify the most crucial factors that may influence a customer's decision regarding banking. Customers with similar behavior regarding loan payments may be identified by multidimensional clustering techniques. These can help identified customer groups, associate a new customer with an appropriate customer group, and facilitate targeted marketing.

Detection of money laundering and other financial crimes:

To detect money laundering and other financial crimes, it is important to integrate information from multiple databases as long as they are potentially detecting unusual patterns, such as large amounts of cash flow at certain periods, by certain groups of customers. Useful tools include data visualization tools, linkage analysis tools. Data Mining Applications classifications tools detect unusual amounts of fund transfers or other activities.

Data Mining for the Retailed Industry:

The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption, and service. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability, and popularity of business conducted on the web, or e- commerce. Today, many stores also have websites where customers can make purchases on-line. Some business, such as amazon.com exist solely on-line, without any brick-and-mortar (i.e., physical) store locations. Retail data provide a rich source for data mining. Retail data mining can help identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business. A few examples of data mining in the retail industry are outlined as follows.

Multidimensional analysis of sales, customers, products, time, and region:

The retail industry requires timely information regarding customer needs, product, sales, trends and fashions, as well as the quality, cost, profit, and service of commodities. It is therefore important to provide powerful multidimensional analysis and visualization tools, including the construction of sophisticated data cubes according to the needs of data analysis. The multi feature data cube, introduced in a useful data structure in retail data analysis because it facilitates analysis on aggregates with complex conditions. Analysis of the effectiveness of sales campaigns: The retail industry conducts sales campaigns using advertisements, coupons, and various kinds of discounts and bonuses to promote products and attract customers. Careful analysis of the effectiveness of sales campaigns can help improve companyprofits. Multidimensional analysis can be used for this purpose by comparing the amount of sales and the number of transactions containing the sales items during the sales period versus those containing the same items before or after the sales campaign. Moreover, association analysis may disclose which items are likely to be purchased together with the items on sale, especially in comparison with the sales before or after the campaign.

III.DESIGN AND CONSTRUCTION OF DATA WAREHOUSES BASED ON THE BENEFITS OF DATA MINING

Because retail data cover a wide spectrum (including sales, customers, employees, goods transportation, consumption, and services), there can be many ways to design a warehouse for this industry. The levels of detail to include may relate to the study. Multiple data analysis tools can also vary substantially. The outcome of preliminary data mining exercises can be used to help guide the design and development of data warehouse structures. This involves deciding which dimensions and levels to include and what processing to perform in order to facilitate effective data mining.

Data Mining for the Telecommunication Industry:

The telecommunication industry has quickly evolved from offering local and long distance telephone services to provide many other comprehensive communication services , including fax, pager, cellular phone, internet messenger, images, e-mail, computer and web data transmission, and other data traffic. The integration of telecommunication, computer network, internet and numerous other means of communication and computing is also underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand for data mining in order to help understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of services. The following are a few scenarios for which data mining may improve telecommunications.

Multidimensional analysis of telecommunication data: Telecommunication data are intrinsically multidimensional, with dimensions such as calling-time, duration, location of caller, location of call, and type of call. The multidimensional analysis of such data can be used to identify and compare the data traffic, system workload, resource usage, user group behavior, and profit. For example, analysts in the industry may wish to regularly view charts and graphs regarding calling source, destination volume and time-of-day usage patterns. Therefore, it is often useful to consolidate telecommunication data into large data warehouses and routinely perform multidimensional analysis using OLAP and visualization tools. Fraudulent pattern analysis and identification of unusual patterns: Fraudulent activity costs the telecommunication industry millions of dollars per year. It is important to identify potentially fraudulent users and their typical usage patterns; detect attempts to gain fraudulent entry to customer accounts; and discover unusual patterns that may need special attention, such as busy-hour frustrated call attempts, switch and route congestion patterns, and periodic calls from automatic dial-out equipment (like fax machines) that have been improperly programmed. Many of these patterns can be discovered by multidimensional analysis,

cluster analysis, and outlier analysis. Multidimensional association and sequential pattern analysis: The discovery of association and sequential patterns in multidimensional analysis can be used to promote telecommunication services. For example, suppose you would like to find usage patterns for a set communication services by customer group, by month, and by time of day.

Data Mining for Biological Data Analysis:

The past decade has been an explosive growth in genomics, proteomics, functional genomics, and biomedical research. Examples range from the identification and comparative analysis of the genomes of human and other species (by discovering sequencing patterns, gene functions, and evolution paths) to the investigation of genetic networks and protein pathways, and the development of new pharmaceuticals and advances in cancer therapies. Biological data mining has become an essential part of a new research field called bioinformatics. Since the field of biological data mining is broad, rich, and dynamic, it is impossible to cover such an important and flourishing theme in one subsection. Here we outline only a few interesting topics in this field, with an emphasis on genomic and proteomic data analysis. A comprehensive introduction to Biological data mining could fill several books. A good set of bioinformatics and biological data analysis books have already been published, and more are expected to come. Proteins are essential molecules for any organism. They perform life functions and make up the majority of cellular structures. The approximately 25,000 human genes give rise to about 1 million proteins through a series of translational modifications and gene splicing mechanisms. Amino acids are the building blocks of proteins. There are 20 amino acids, denoted by 20 different letters of the alphabet. Each of the amino acids is coded for by one or more triplets of nucleotides making up DNA. The end of the chain is coded for by another set of triplets. Thus, a linear string of sequence of DNA is translated into a sequence of amino acids, forming a protein. A proteome is the complete set of protein molecules present in a cell, tissue, or organism. Proteomics is the study of sequences. Proteomes are dynamic, changing from minute to minute in response to tens of thousands of intra- and extracellular environmental signals.

Chemical properties of the amino acids cause the protein chains to fold up into specific three dimensional structures. This three-dimensional folding of the chain determines the biological function of a protein**.** Genes make up only about 2% of the human genome. The remainder consists of noncoding regions. Recent studies have found a lot that of a non coding DNA sequences may also have played crucial roles in protein generation and species evolution. The identification of DNA or amino acid sequence patterns that play roles in various biological functions, genetic diseases, and evolution in challenging. This requires a great deal of research in computational algorithms, statistics, mathematical programming, data mining, machine learning, information retrieval, and other disciplines to develop effective genomic and proteomic data analysis tools: Data mining may contribute to biological data analysis in the following aspects: Genomic and proteomic data sets are often generated at different labs and by different methods. They are distributed, heterogeneous, and of a wide variety. This mantic integration of such data is essential to the cross-site analysis of biological data. Moreover, it is important to find correct linkages between research literature and their associated biological entities. Such integration and linkage analysis would facilitate the systematic and coordinated analysis of genome and biological data. This has promoted the development of integrated data warehouses and distributed federated databases to store and manages the primary and derived biological data. Data cleaning, data integration, reference reconciliation, classification, and clustering methods will development: Currently, many studies have focused on the comparison of one gene to another. However, most diseases are not triggered by a single gene but by a combination of genes acting together. Association analysis method can be used to help determine the kinds of genes that are likely to co-occur in target samples.Such analysis would facilitate the discovery of groups of genes and the study of interactions and relationships between them. While a group of genes may contribute to a disease process, different genes may become active at different stages of the disease. If the sequence of genetic activities across the different stages of disease development can be identified, it may be possible to develop pharmaceutical interventions that target the different stages separately, therefore achieving more effective treatment of the disease. Such path analyses is expected to play an important role in genetic studies. Visualization tools in genetic data analysis: Alignments among genomic or proteomic sequences and the interactions among complex biological structures are most effectively presented in graphic forms, transformed into various kinds of easy-to-understand visual displays. Such visually appealing structures and patterns facilitate pattern understanding, knowledge discovery, and interactive and data exploration. Visualization and visual data mining therefore play an important role in biological data analysis.

Data Mining in Other Scientific Application:

Scientific data analysis tasks tended to handle relatively small and homogeneous data sets. Such data were typically analyzed using a "formulate hypothesis, build model, and evaluate results" paradigm. In these cases, statistical techniques were appropriate and typically employed for their analysis.

Storage technologies have recently improved, so that today, scientific data can be amassed at much higher speedsand lower costs. This has resulted in the accumulation of huge volumes of high dimensional data, containing rich spatial and temporal information. Consequently, scientific applications are shifting from the "hypothesize-and-test" paradigm towards a "collect and store data, mine for new hypothesis, conform with data or experimentation" process. This shift brings about new challenges for data mining. Vast amounts of data have been collected from scientific domains (including geo science, astronomy, meteorology) using sophisticated telescopes, multispectral high resolution remote satellite sensors and global positioning systems. Large datasets are being generated due to fast numerical simulations in various fields, such as climate and eco system modeling, chemical engineering, fluid dynamics and structure mechanics. Other areas requiring the analysis of large amounts of complex data include telecommunications and biomedical engineering. The challenges brought about by emerging scientific applications of data mining, such as the following: Data warehouses and preprocessing. Data warehouses

are critical for information exchange and data mining. In the area of geo spatial data, however, no two geo spatial data exists today. Creating such a warehouse require finding means for resolving geo graphic and temporal data incompatibilities such as reconciling semantics, referencing systems, geo metric, accuracy and precision. For scientific applications in general, methods are needed for integrating data from heterogeneous sources (such as data covering different time periods) and for identifying events.

For climate and ecosystem data, for instance (which are spatial and temporal),the problem is that there are too many events in the spatial domain and too few in the temporal domain.(For example, EL Nino events occur only every four to seven years, and previous data might not have been collected as systematically as today.) Methods are needed for the efficient computation of sophisticated spatial aggregates and the handling of spatial-related data streams. Mining complex data types: Scientific data sets are heterogeneous in nature, typically involving semi-structured and unstructured data, such as multimedia data and geo reference stream data. Robust methods are needed for handling spatiotemporal data,related concept hierarchies, and complex geographic relationships (e.g., non-Euclidian distances). Graph-based mining: It is often difficult or impossible to model several physical phenomena and processes due to limitations of existing modelingapproaches. Alternatively, labeled graphs may be used to capture many of the spatial, topological, geometric, and other relational characteristics present in scientific data sets. In graph modeling, each object to be mined is represented by a vertex in a graph, and edges between vertices represent relationships between objects. For example, graphs can be used to model chemical structures and data generated by numerical simulation, such as fluid -flow simulations. The success of graph-modeling, however, depends on improvements and efficiency of many classical data mining tasks, such as classification, frequent pattern mining, and clustering. Visualization tools and domain- specific knowledge: High level graphical user interfaces and visualization tools are required for scientific data mining systems. These should be integrated with existing domain-specific information systems and database systems to guide researchers and general users in searching for patterns, interpreting and visualizing discovered patterns, and using discovered knowledge in their decision making.

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. An intrusion can be defined as any set of actions that threaten the integrity, confidentiality, or availability of a network resource (such as user accounts, file systems, system kernels, and so on). Most commercial intrusion detection systems are limiting and do not provide a complete solution. Such systems typically employ a misuse detection strategy. Misuse detection searches for patterns of program or user behavior that match known intrusion scenarios, which are stored as signatures. These hand-coded signatures are laboriously providedby human experts based on their extensive knowledge of intrusion techniques. If a pattern match is found, this signals an event for which an alarm is raised. Human security analysts evaluate the alarms to decide what action to take, whether itbe shutting down part of the system, alerting the relevant internet service provider of suspicious traffic, or simply noting unusual traffic for future reference .An intrusion detection system for a large complex network can typically generate thousands or millions of from several network location in order to detect these distributed attacks.

IV.CONCLUSION

Data mining is a relatively young field with many issues that still need to be researched in depth, many off- the –shelf datamining system products and domain specific data mining application software are available. As a discipline, data mining has a relatively short history and is constantly evolving-new data mining systems appear on the market every year; new function, feature and visualization tools are added to existing systems on a constant basic; and efforts toward the standardization of data mining language are still underway. Therefore, it is not our intention in this book to provide a detailed description of commercial data mining systems. Instead, we describe the features to consider when selecting a data mining product and offer a quick introduction to a few typical data mining systems. Reference articles, websites and recent surveys of data mining systems are listed in the bibliographic notes.

References

[1]. Dolado, J. J., D. Rodriguez, and J. Riquelme. "A Two Stage Zone Regression Method for Global Characterization of a project Database."(2007):13.web.5 Apr.2013.
[2]. Berzal, Fernando, Juan-Carlos Cubero and Nicol as Mar n. "Building multi-way decision trees with numerical attributes."31. Web.5 Apr.2013.
[3]. Frank, Eibe. "Pruning Decision Trees and Lists." (2000): 218. Web. 5 Apr. 2013.
[4]. Korting, Thales 5. "C4.5 algorithm and Multivariate Decision Trees." 5. Web 2 Feb. 2013.
[5]. Quinlan, J. R. "Improved Use of Continuous Attributes in C4.5," 14. Web. 11 Jan. 2013.
[6]. JUNEJA, DEEPTI, et al. "A novel approach to construct decision tree using quick C4.5 algorithm." Oriental Journal of Computer Science & Technology Vol. 3(2), 305-310 (2010) (2010): 6. Web. 18 Feb, 2013.
[7]. Ittner, Andreas, et al. "Non-Linear Decision Tree - NDT." In: Proceeding of 13th international conference on machine learning (ICML"96) 6. Web. 16 Mar. 2013.
[8]. Moertini, Veronica S. "TOWARDS THE USE OF C4.5 ALGORITHM FOR CLASSIFYING BANKING DATASET." Vol. 8 No. 2, October 2003 (2003): 12. Web. 24 Jan. 2013.
[9]. Utgoff, Paul E. "Linear Machine Decision Tree." (1991): 15. Web. 6 Feb. 2013.
[10]. Rokach, Lior, and Oded Maimon. "DECISION TREES." 28. Web. 1 Feb. 2013.
[11]. "Data Mining" from Wikipedia the free Encyclopedia. Web.

[12]. Term "INTRODUCTION OF DATA MINING , "Data Mining: What is Data Mining", source from http://www.anderson.ucla.edu/faculty/jason.frand /teacher/technologies/palace/datamining.htm.
[13]. Rokach, Lior. "Data Mining with Decision Trees: Theory and Applications." 69 (2008): Web. 3 Feb. 2013.
[14]. Ga-sperin, Matej. "Case Study on the use of Data Mining Techniques in Food Science using Honey Samples." (February 2007): 18. Web. 8 May 2013.
[15]. Ozer, Patrick. "Data Mining Algorithms for Classification." (January 2008): 27. Web. 5 May 2013.
[16]. Gholap, Jay. "PERFORMANCE TUNING OF 148 ALGORITHM FOR PREDICTION OF SOIL FERTILITY." 5. Web. 2 May 2013.