CHAPTER 2

LITERATURE BACKGROUND

Data mining is the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data. A wide variation exists in terms of the problem domains, applications, formulations, and data representations that are encountered in real applications. Therefore, “data mining” is a broad umbrella term that is used to describe these different aspects of data processing. Data mining strategies fall into two broad categories namely Supervised Learning and Unsupervised Learning. Supervised Learning methods are deployed when there exists a field or variable with known values and about which predictions will be made by using the values of other fields or variables inputs. Unsupervised Learning methods tend to be deployed on data for which there do not exist a field or variable with known values, while fields or variables do exist for other fields or variables.

Data mining (Jiawei Han & Kambar 2006) is the extraction of interesting nontrivial, implicit, previously unknown and potentially useful information or patterns from large databases.

2.1 DATA MINING TECHNIQUES

It is one of the phases in knowledge discovery process. The main goal of this phase is pattern recognition. But this pattern recognition varies based on user interests and based on the application. There exist several mining techniques for pattern recognition which are described below.
2.1.1 Intrusion Detection

Several effective data mining techniques have been developed for detecting intrusions (Warrender et al. 1999) which perform close to or better than systems engineered by domain experts. However, successful data mining techniques are themselves not enough to create deployable IDS. Despite the promise of better detection performance and generalization ability of data mining-based IDS, there are some inherent difficulties in the implementation and deployment of these systems. These difficulties can be grouped into three general categories: accuracy (i.e., detection performance), efficiency, and usability. Typically, data mining-based IDS (especially anomaly detection systems) have higher false positive rates than traditional hand-crafted signature based (misuse detection systems) methods, making them unusable in real environments. Also, these systems tend to be inefficient (i.e., computationally expensive) during both training and evaluation.

This prevents them from being able to process audit data and detect intrusions in real time. Finally, these systems require large amounts of training data and are significantly more complex than traditional systems. In order to be able to deploy real time data mining based IDS, these issues must be addressed.

Large amount of data exists in the system which could be gathered by network personnel to detect security policy violations. With this scenario, the analysis is a tedious one and network administrators do not have the resources to analyze the data for security policy violations especially in the presence of a high number of false positives that cause them to waste their limited resources. One of the challenges of intrusion detection systems is to analyze data so that a legitimate or intrusive activity could be detected (Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham 2005). The solution is
to employ data mining techniques (Fayyad U Piatesky-Shapiro et al. 1996) in an offline environment. This kind of approach would add additional depth to the network administrator’s defenses, and allows them to more accurately determine what the threats against their network are through the use of multiple methods on data. Data mining techniques are used in classification and identification (Toussaint 1978) of new patterns from large volume of training data that are collected from KDD (Knowledge Discovery in Data Mining) CUP 1999 benchmark dataset in order to perform hybrid intrusion detection in host as well as in network. Moreover, intrusion detection has been carried out using classification and clustering algorithms integrated with feature selection (Dash & Liu 1997). The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning and database.

2.1.2 Clustering in Data mining

Clustering is a general unsupervised classification procedure that divides a set of objects in different classes. Objects from the same class should be similar to each other. There is no initial indication about the classes or about their number but only the properties of the objects in the set of data. This segment presents some of the presented works related to Clustering in Data Mining. (Bailey Kenneth 1994) highlighted the numerical taxonomy and clustering analysis of the datasets in clustering. (Jain et al. 1999) highlighted the reviews of data clustering. (Han & Kamber 2006) explained Concepts and Techniques in Data Mining. (Jain & Dubes 1988) introduced the data distribution perspective of the K-Means Clustering. (Tapas Kanungo et al. 2002) experimented that K-means Clustering Algorithm is efficient in analysis, implementation, pattern analysis and machine intelligence. (Eibe Frank et al. 2009) tested that data mining and knowledge discovery using Weka. (Liu & Motoda 1998) reviewed how to select the features in data
mining. (David KY Chiu & Yan Wang 2006) highlighted multipattern consensus regions in multiple aligned protein set of related events and their segmentation based on their interrelation. (Dr. A.Bharathi and Shilpa 2014) highlighted the equivalence measures which helps to discover unanswered crimes in crime dataset pattern. Partitioning clustering algorithm is one of the best method for finding equivalence measures.

2.1.3 K-Means and Fuzzy C means

(Soumi Ghosh 2013) highlighted the analysis of K-Means and Fuzzy C Means algorithms. K-Means algorithm used to define the clusters from the beginning itself. This algorithm has problems like getting affected by local optima, sensitivity to outliers, memory space, unknown number of iteration steps that are required to cluster. Therefore, the time complexity of the K-Means algorithm and FCM algorithm is $O(ncdi)$ and $O(nc^2i)$. From this results we may conclude that K-Means algorithm has improved performance when compared to FCM algorithm. FCM produces close results to K-Means clustering but it still requires much more computational time when compared to K-Means, due to the fuzzy measures calculations involved in the algorithm. The FCM clustering comprise the very old constituent of software computing, which is suitable for handling the issues related to identify ability of patterns, incomplete and noisy data, mixed different media of information, human communication. It can make available the nearby results very faster. So, the overall conclusion is that K-Means algorithm seems to be superior than Fuzzy C Means algorithm.

2.2 CLUSTERING IN WEKA

(Priyanka Sharma 2015) highlighted comparison analysis of different algorithms with respect to unsupervised learning in Data Mining. Comparative analysis of various clustering algorithms has been made. The
results have been certified using six datasets from various repository and observed that datasets are successfully clustered with a relatively good precision. Few of the clustering techniques have better accuracy, others take less time, and many others have a compromise between accuracy and time taken. Therefore, appropriate clustering methods can be selectively used based on their usage. The tool used to analyze, partition and cluster the data sets are done using WEKA (Waikato Environment for Knowledge Analysis). The tool can be downloaded and executed from www.cs.waikato.ac.nz/ml/weka.

2.3 DATA MINING IN TRANSACTIONS

This new algorithm could get any stock’s money flow and establish a series of quantitative trading strategies based on the new money flow model. These strategies could guide investors to grasp an upward trend or avoid a down trend. The quantitative investment trading strategies can be automated trading. It could reduce investor’s subjective operability and increase return rate. This algorithm was suitable for index trading, futures trading, stock trading, and ETF trading (Jinfei Yang & Qingzhen Xu 2016). There is a new strategy which could the analyze the relationship that exists between the attributes of the stored data in database of Palestinian Government. Based on the analysis the patterns are identified that help us to make the system more user-friendly by offering suggestions to the users during data entry process. Naive Bayes, Rule Induction, K-NN, and Decision Tree methods are applied to the stored data in order to produce a prediction model that predicts entries to the user during the entry process. The experiment result shows the Naive Bayes is the best model among the other techniques by achieving the highest accuracy of 68.41% (Basei A Alhaj & Asraf YA Meghari 2017).
2.4 BIG DATA MINING

Big data often requires tremendous computational resources. This has become a major obstacle to broad applications of big data analytics. Cloud computing allows data scientists to access computational resources on demand for building their big data analytics solutions in the cloud. However, the monetary cost of mining big data in the cloud can still be unexpectedly high. How to achieve a sufficiently satisfactory result at the lowest possible computation cost? In this paper, they explore and demonstrate the cost effectiveness of big data mining with a case study using well known k-means. With the case study, we find that achieving 99% accuracy only needs 0.32%-46.17% computation cost of 100% accuracy (Qiang He et al. 2017).

2.5 DECISION TREE WITH DATA PREPROCESSING

A decision tree is an important classification technique in data mining classification. Decision trees have proved the most valuable tools used for the classification and generalization of data. J48 is a decision tree algorithm which is used to create classification model. J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. In this paper, the method used improves the accuracy for decision tree mining with data preprocessing. Then the supervised filter discretization is applied on J48 algorithm to construct a decision tree. The obtained results applying and not applying discretization are compared. Interestingly it was found that the results are better when applying the discretization (Priyanga Chandrasekar, 2017). In this paper, focusing on the priority-based bandwidth allocation a hybrid data mining method is developed to manage the limited bandwidth in a university network more effectively. This method is composed of two main steps and uses the clustering and classification techniques. The main purpose is to detect, analyze and predict student’s behavioral patterns in a university.
network and identify the main factors that affect their tendency in using internet. The proposed method is applied on a real data of a network university. The results indicate that degree type and age are the most important factors that influence student’s tendency to use internet. The results would be also useful to predict a new student’s tendency to use internet given his/her characteristics. By analyzing the results, the IT managers can make better decisions to optimize the allocation of bandwidth resources (Elham Akhond et al. 2016).

2.6 CORRELATION MINING AMONG MULTIPLE TASKS

In this paper, a novel semi supervised feature selection algorithm is applied which mines the correlations between multiple tasks and apply the same on different multimedia applications. Instead of independently computing the importance of features for each task, this algorithm leverages shared knowledge from multiple related tasks, thus improves the performance of feature selection. The algorithm is built upon an assumption that different tasks share some common structures. The algorithm selects features in a batch mode, by which the correlations between various features are taken into consideration. Besides, considering the fact that labeling a large amount of training data in real world is both time-consuming and tedious, we adopt manifold learning, which exploits both labeled and unlabeled training data for a feature space analysis. Since the objective function is non smooth and difficult to solve, they propose an interactive algorithm with fast convergence (Xiaojun Chang & Yi Yang 2017).

2.7 WEB USAGE MINING PROCESS

Commonly the Web usage mining refers to the process of discovering user access patterns from the log of websites. The web log contains unstructured, noisy and irrelevant data. This data suitable for pattern
mining and pattern analysis when this data is passed through data preprocessing phase. Data preprocessing improves the quality of the data and also reduces the size of web log file. Data preprocessing involves several steps including data collection, data cleaning, session identification, user identification and path completion. This paper presents several data preprocessing techniques in order to prepare raw data suitable for mining and analysis tasks (Sanjay Kumar Dwivedi & Bhupesh Rawat 2015).

2.8 OPTIMIZED DATA MINING FOR SUMMARIZATION

Whenever a research scholar starts working on some innovative ideas he/she searches for the domain specific technical research articles published as research papers in most of the international journals, conferences or workshops. Reading these all related papers completely one by one to get the latest research developments in the interested domain is time-consuming, unnecessary, irrelevant, cumbersome and impossible. These problems are solved by developing an innovative solution, which optimizes and summarizes these research papers as an aid for research scholars using various Data Mining strategies. This helps research scholar for getting short, condensed, accurate and most relevant summarized information i.e. overview of domain specific topic-based contents from research papers. Data mining strategies such as extraction and clustering are used for identifying research relevant novel terms. Extracting relevant sentences from multiple papers uses ‘Maximal Marginal Relevance’ (MMR) criteria containing research relevant terms. In addition to optimized and summarized contents, this paper also informs earlier and latest research developments, progress, and challenges and future scope in particular field of study through various research categories identified such as research methods/techniques/approaches used, comparisons with existing research and providing starting material for further innovation (Sunita R Patil & Sunita M Mahajan 2012).
2.9 Data mining in prediction

Credit risk is related to the risk of the borrower that the lender will not be able to return their debt including interest. Numerous researches have been conducted in the area of credit risk, both using classical models such as Altman Z-score and using machine learning methodology. However, the research using the data from Croatian financial institutions is scarce, especially research focused on the selection of the demographic and/or behavior variables. In addition, it is important to develop robust models that estimate credit risk as accurately as possible. The goal of this research is to develop a data mining model for prediction of credit risk, using the data from Croatian financial institutions on defaulted clients (demographic and behavior data). Decision tree models are constructed for the prediction of credit risk. Different algorithms for the variable selection are evaluated based on the classification accuracy of the decision trees developed based on the selected variables (Pejic Bach et al. 2017).

2.9.1 Data mining in flight crash investigation

Data mining refers to extraction of information from huge chunks of the dataset. It’s also called information mining. It is exercised in numerous fields like medicine, environment, education, crime, etc. In this research work crash investigation and analysis of the flights are done. Flight crashes may be caused due to pilot error, mechanical failure, bad weather, sabotages or human error. This research paper investigates international flight crashes since 1908 to 2009 through K-Means clustering data mining technique and cosine similarity. Clustering helps to put objects into the same group. Cosine similarity measure helps in finding similarity among different texts. The research work is done for identifying aboard/ground fatality rate with operators and locations well as to find similarity among the plane crashes (Shagun Sharma & Sai Sabitha 2016).
2.9.2 Data mining in industries

On the way to full production control, wire bonding equipment requires data-driven condition-based maintenance of the mechanical setup. In this paper, the aspects of regular mechanical equipment setups that severely affect bonding quality and equipment health in mass production is identified. It also shows that mechanical equipment setups lower process stability in aluminum wire bonding. Typical faults in mechanical setup directly influence device quality, which is measured through the use of pull and shear tests. In this paper we demonstrate that data mining can detect mechanical setup faults by extracted features of monitored machine data. Condition based maintenance enables the verification of mechanical setup conditions after manual adjustments and individual bonding events. This novel condition-based maintenance has the potential to partially substitute pull and shear tests, and hence avoid related equipment downtime as well as additional cost (Klingert et al. 2017).

2.9.3 Volume diagnosis data mining

With decreasing feature sizes and increasing complexity of fabrication processes for manufacturing VLSI semiconductor devices, more systematic defects occur at the advanced technology nodes. Product yield ramp up is mostly determined by how fast systematic defects are identified and fixed. Given the long times and expense of Physical Failure Analysis (PFA), use PFA on a large number of failing devices to find systematic defects is becoming infeasible. For this reason, volume diagnosis data mining for root cause the identification based statistical methods to reduce turnaround time and cost to speed up the process of systematic defect identification. The identified root cause information not only can be used to improve yield analysis but also can reduce PFA cost by focusing on failing devices with systematic defects (Wu-Tung Cheng et al. 2017).
2.9.4 Data mining in health care

Data mining is an important area of research and is pragmatically used in different domains like finance, clinical research, education, healthcare etc. Further, the scope of data mining has thoroughly been reviewed and surveyed by many researchers pertaining to the domain of healthcare which is an active interdisciplinary area of research. In fact, the task of knowledge extraction from the medical data is a challenging endeavor and it is a complex task. The main motive of this review paper is to give a review of data mining in the purview of healthcare. Moreover, intertwining and interrelation of previous researches have been presented in a novel manner. Furthermore, merits and demerits of frequently used data mining techniques in the domain of health care and medical data have been compared. The use of different data mining tasks in health care is also discussed. An analytical approach regarding the uniqueness of medical data in health care is also presented (Subhash Chandra Pandey 2016).

2.9.5 Data mining in disease analysis

In this paper, the various technologies of Data Mining (DM) models for forecast of heart disease are discussed. Data mining plays an important role in building an intelligent model for medical systems to detect Heart Disease (HD) using data sets of the patients, which involves risk factor associated with heart disease. Medical practitioners can help the patients by predicting the heart disease before occurring. The large data available from medical diagnosis is analyzed by using data mining tools and useful information known as knowledge is extracted. Mining is a method of exploring massive sets of data to take out patterns which are hidden and previously unknown relationships and knowledge detection to help the better understanding of medical data to prevent heart disease. There are many DM
techniques available namely Classification techniques involving Naïve Bayes (NB), Decision Tree (DT), Neural Network (NN), Genetic Algorithm (GA), Artificial Intelligence (AI) and Clustering algorithms like KNN, and Support Vector Machine (SVM). Several studies have been carried out for developing prediction model using individual technique and also by combining two or more techniques. This paper provides a quick and easy review and understanding of available prediction models using data mining from 2004 to 2016. The comparison shows the accuracy level of each model given by different researchers.

2.10 SMART GRID DATA MINING AND VISUALIZATIONS

The power industry innovation has increasingly become a top concern for current reforms. Power systems feature scattered data storage, incapable data analysis ability, poor computing capability, and ineffective interaction interface. To resolve these issues, we need multiple data mining techniques to extract information for analytical capacity improvement. Secondly, we need visualization techniques to analyze and optimize interaction. Lastly, we need distributed technologies for unified data management to increase computing capability and system scalability. Considering China's smart grid information, this paper proposes solutions to problems, such as the existing underdeveloped power management systems, a lack of automation methods, low data visualization, and poor data management. The electric power industry has functional requirements for this research. Based on existing data mining, visualization and understanding of distributed technologies, we discussed the functions of each part of the implementation in a smart grid management system: the data mining module, visualization module and data management module (Yingyao Zhou, Ping Li Yuning Xiao, Anum Masood, Qichen Yu & Bin Shag 2016).
2.11 **K-NEAREST NEIGHBOR**

This paper presents a salary prediction system using a profile of graduated students as a model. A data mining technique is applied to generate a model to predict a salary for individual students who have similar attributes to the training data. In this work, we also made an experiment to compare five data mining techniques including Decision trees, Naive Bayes, K-Nearest neighbor, Support vector machines, and Neural networks to find the suitable technique to the salary prediction. In the experiment, 13,541 records of graduated student data were used with 10-fold cross validation method. Results showed that K-Nearest neighbor provided the best efficiency to be used as a model for salary prediction. For usage evaluation, a questionnaire survey was conducted with 50 user samplings and a result showed that the system was effective in boosting students’ motivation for studying and also gave them a positive future viewpoint. The result also informed that they found they satisfied with the implemented system since the system was easy to use, and the prediction results were simple to understand without requiring any background knowledge (Pomthep Khong Chai & Pokpong Songmuang 2016).

2.12 **DATA MINING IN SIGNAL REPRESENTATIONS**

This paper presents the author’s personal path through the signal representations of the past three decades, from the early days and excitement that surrounded the advent of wavelets and associated multi resolution representations, to the present day foray into graph signal processing and data mining. It is a tribute to Dr. John Cozzens of the NSF and his vision and support for the development of the field (Jelena Kovacevic 2017).
2.13 WEB BASED MINING

This paper gives details about web-based department automation system which will be implemented at educational institution level for maintaining faculty details and records. The proposed application aims at providing efficient and hassle-free working environment for faculty of the organization as it reduces the amount of paperwork involved. This system is based on the modern approach of data mining and web technology which overcomes the drawbacks of existing manual method used by the educational institution reducing mistakes and errors in record keeping. The system provides faster computation of queries to fetch data from a huge database. The application is designed using Eclipse (Kepler version), Apache Tomcat server 7.0 along with wamp server to host the database locally. J2EE, JSP along with Html have been used for designing the user interface whereas MySQL database is used for storing data (Usmani et al. 2017).

2.14 FP-TREE BASED ASSOCIATION RULE MINING

Based on the development and application of on-board subsystem test bench for current CTCS-3 system, this paper focuses on the approach of automatically generation of test sequence, takes the existing test sequences of ETCS-2 (European Train Control system level 2) as the train set existing relatively mature test sequence as the training set, to execute association rule mining. The whole data mining process involves data preparation (including data cleaning and data selection) firstly, providing basement for association rule, then establishes FP_tree (frequent pattern) and seeks test cases with frequent pattern through implementing FP_growth algorithm for the target database. Comparing the analysis results and experience, it shows that the association rule based on FP_tree could play an important role on the efficiency and verification of automatically generation of test sequence (Ying Gao et al. 2017).
2.14.1 PPRAM

Data mining can extract important knowledge from large database - sometimes this database is split among various parties. Here, the main aim of privacy preserving data mining is to find the global mining results by preserving the individual sites private data/information. Many Privacy Preserving Association Rule Mining (PPARM) algorithms are proposed for different partitioning methods by satisfying privacy constraints. The various methods such as randomization, perturbation, heuristic and cryptography techniques are proposed by different authors to find privacy preserving association rule mining in horizontally and vertically partitioned databases. In this paper, the analysis of different methods for PPARM is performed and their results are compared. For satisfying the privacy constraints in vertically partitioned databases, algorithm based on cryptography techniques, Homomorphic encryption, Secure Scalar product and Shamir’s secret sharing technique are used. For horizontal Partitioned databases, algorithm that combines advantage of both RSA public key cryptosystem and Homomorphic encryption scheme and algorithm that uses Paillier cryptosystem to compute global supports are used. This paper reviews the wide methods used for mining association rules over distributed dataset while preserving privacy (Fosca Giannotti et al. 2013).

2.14.2 MOOC based data mining

In today's world virtual online educational platforms emerge literally on daily bases and many offer MOOC-based courses. With the appearance of MOOC, educational platforms have gained an additional boost, a new aspect in their evolutionary process, which has opened a new field of research thanking to the extraction of logging information within the frames of data mining. It has become clear that educators will be able to tailor their
courses by merging the two previously mentioned fields and by carrying out MOOC-based data mining, targeting pedagogical aspects. This field of research seems promising and important, thus a faculty at the University of Szeged has created its own MOOC educational platform which has been set to facilitate data mining by implementing a wide range of logging algorithms. The data would be processed through a complex Artificial Intelligence program, which, in the short term, could reveal new and exciting pedagogical findings, while in the long run, the supervisors could put together a platform that would help and notify educators about relevant information. It would become possible to create adaptive educational materials, as well (Korosi & Havasi 2017).