Dependence of the data mining techniques on the decision problem nature

Hoda A. Abdelhafez*

Faculty of Computers & Informatics,
Information Systems & Decision Support Department,
Suez Canal University,
Old Campus, ElShikh Zayed, Ismaila, Egypt
Email: Hodaabdelhafez@gmail.com
*Corresponding author

Kamal Abdel-Raouf ElDahshan

Faculty of Sciences,
Computer Science and Information Systems Department,
AlAzhar University,
Nasr City, Cairo, Egypt
Email: dahshan@gmail.com

Abstract: Data mining is a valuable tool for many industries, including finance, manufacturing, retail, medical health, insurance and telecommunications. The aim of this paper is to investigate and demonstrate that each class of decision problems needs a suitable or set of data mining techniques. Each category of decision problems applies some data mining techniques that may differ from the other categories. These techniques are determined based on the type of the problem, the scale of the dataset, as well as the dynamic changes of these data. The results illustrated that the choice of the data mining technique is decision problem-dependent meaning that for each class of decision problems there is a suitable class of data mining techniques.

Keywords: data mining; decision problems classification; data mining techniques; data mining problems; data science; fraud detection; customer behaviour; risk management; fault patterns; customer churn.

Reference to this paper should be made as follows: Abdelhafez, H.A. and ElDahshan, K.A-R. (2015) 'Dependence of the data mining techniques on the decision problem nature', *Int. J. Data Science*, Vol. 1, No. 2, pp.103–117.

Biographical notes: Hoda Ahmed Abdelhafez is an Assistant Professor of Information Systems & Decision Support in Faculty of Computers & Informatics, Suez Canal University. She holds a PhD in Information Technology from Alexandria University. She has taught graduate and undergraduate courses in knowledge management, advanced e-commerce, management information systems, systems analysis and design, databases and information storage and retrieval. She has co-authored over 13 publications in decision support systems, data warehouse, data mining, e-learning, e-government and big data. She is a co-author of the book 'Information Resources & Centers' published in 2001. She has two book chapters published by IGI Global in 2014.

Kamal Abdel-Raouf ElDahshan is a Professor of Computer Science and Information Systems at Al-Azhar University in Cairo, Egypt. He obtained his Doctoral degree from the Université de Technologie de Compiègne in France, where he also taught for several years. He also worked at the prestigious Institute National de Télécommunications in Paris. His extensive international research, teaching and consulting experiences have spanned four continents and include academic institutions as well as government and private organisations. He was a Visiting Professor at Virginia Tech; he was a Consultant to the Egyptian Cabinet Information and Decision Support Centre; and he was a Senior Adviser to the Ministry of Education and Deputy Director of the National Technology Development Centre. He is a Professional Fellow on Open Educational Resources as recognised by the United States Department of State.

This paper is a revised and expanded version of a paper entitled 'The decision problem nature affects the choice of the data mining technique' presented at *Proceedings of 37th International Conference on Computers and Industrial Engineering*, Alexandria, Egypt, 20–23 October, 2007.

1 Introduction

Data mining helps managers to make knowledgeable decisions and gain competitive advantage through providing them with useful information extracting from huge amounts of databases. It is a process, not a particular technique or algorithm and its goal is to predict and generalise a pattern to other data (Edelstein, 2003). The development of data mining is related to the availability of large databases that are collected not primarily for analysis and prediction (Friedman, 1997; Kardaun and Alanko, 1998; Seifert, 2005). Models and patterns represent two kinds of structures for data mining activities. A model is a global summary of relationships between variables that helps to understand phenomenon and allows predictions. A pattern is a characteristic structure exhibited by a few number of points. The problem in pattern discovery is to be sure that patterns are real and useful, and the probability of finding a given pattern increases with the size of the database (Hand, 2000).

The major data mining (DM) techniques are statistics, association rules for transactional database, artificial intelligent (AI) techniques, decision trees, genetic algorithm (GA) and visualisation. Statistics is used to evaluate the results of data mining, and it can also deal with missing data using estimation techniques. Mining transactional or relational databases derive a set of strong association rules and a repeated scanning through a massive transactional database might be required. AI techniques such as neural networks are widely used in addition to case-based reasoning and intelligent agents. Decision tree approach can generate rules for the classification of a dataset. Genetic algorithms are appropriate for problems that require optimisation with respect to some computable criteria. Visual data mining techniques help in exploring data analysis, and provide a good potential for mining large database (Lee and Siau, 2001). While data mining products can be very powerful tools, they are not self-sufficient applications. Thus, data mining requires skilled technical and analytical

specialists who can structure the analysis and interpret the output that is created (Seifert, 2005).

Problem statement in this research is to investigate and demonstrate that each class of decision problems needs a suitable or set of data mining techniques. We focus on using data mining is the analysis of business activities to reveal hidden patterns and trends in order to solve decision problems. For example, businesses use data mining for performing customer behaviour analysis to do better forecasting model and profile customers with more accuracy, and finding the cause of manufacturing problems to prevent losses in revenue. Our proposed methodology depends upon collecting information about real decision problems and data mining techniques used to solve these problems. Then, we use the information to:

- classify decision problems into categories, i.e., make a decision problem taxonomy
- focus on the selected data mining techniques to solve each class of decision problem
- determine the reasons behind the selection of each particular technique.

The results showed that each category of decision problems applies some data mining techniques, which may differ from the other categories. These techniques are determined based on the type of the problem, the scale of the dataset and the dynamic changes of these data.

2 Previous studies

One of the studies presented the choice of appropriate data mining tools to solve a problem using several criteria. First, consider a problem to address and the results required to obtain. Secondly, select tools that can process the amount of the data that must be analysed (Brooks, 1997). Trial and error method was also used to determine some of the criteria that are important in determining the data mining technique. This method is not suitable for real world data because data are constantly changing, but a robust model of the best data mining tools could provide a great deal of analysis (Berson et al., 2005). According to this survey, data mining is an application-dependent issue and the classification scheme was presented a clear picture on different data mining requirements and techniques. The survey has taken into consideration an overview of the data mining methods from database perspective and discovering several kinds of knowledge from data mining. It did not discuss how data mining technique is an application dependent (Chen et al., 1996).

3 Data mining problems classification

Data mining can solve a lot of problems; in this section we classify decision problems or data mining problems into seven categories. These categories are analysing customer behaviour, customer churn and determining profitability, Fraud detection, risk management, planning and scheduling, identifying fault patterns and process-improvement and saving cost.

3.1 Analysing customer behaviour

Analysing customer behaviour leads to better understanding of the customer and better forecasting model. In banking and other financial institutions, hidden patterns of customer behaviour are becoming clear focus with use of new data mining tools (Fabris, 1998). Analysis consumer behaviour of credit card database using data mining tools supports banks in establishing a best customer relationship and increasing customer loyalty and revenue (Hsieh and Chu, 2009).

In telecommunications industry, it is often useful to profile customers based on their patterns of phone usage, which can be extracted from the call detail data (Weiss, 2005). Understanding and exploiting natural buying patterns required to find trends across large numbers of transactions. This helps the retailers in decision process to improve efficiency in inventory management and financial forecasting (Moxon, 1996; Raorane and Kulkarni, 2011).

3.2 Customer churn and determining profitability

Customer churn is a significant problem because of the association loss of revenue and the high cost of attracting new customers. In the mobile telephony and telecommunication industry in general, customer churn represents a continuous problem because of competitive market. Therefore, churn prediction models are important for improving the recognition rates of possible churners. After the possible churners are identified, intervention strategies should be made for retaining loyal customers (Kirui et al., 2013; Sharma and Panigrahi, 2011; Oseman et al., 2010). On the other hand, keeping customers and their loyalty has become important factor in the insurance companies; determining the reasons of churning is a challenge. The cost of acquiring new customers is higher than keeping customers' satisfaction and loyalty thus, data mining tools are used to manage customer churn (Soeini and Rodpysh, 2012).

Furthermore, the retailers measure customer activity and expected lifetime by determining profitability and churn rate. They could use these findings to determine customer acquisition goals needed to meet future revenue or profitability objectives in retail/marketing (Groth, 1999).

3.3 Fraud detection capability

Data mining techniques are used successfully for fraud detection in different fields. Detect fraudulent claims in insurance firms (Tauhert, 1997) and (Gayle, 2002), detect fraud of credit lines especially credit card payments (Chan et al., 1999; Brause et al., 1999; Groth, 1999) and identify patterns of fraud in telecommunication (Weiss, 2005). Telecommunication fraud represents unauthorised use or manipulation of cell phone or service. It is also deliberate abuse of voice and data networks (Akhter and Ahamad, 2012). Providing a fair market and protecting market participants from fraudulent practices is a challenge for regulators in securities Market (Golmohammadi and Zaiane, 2012).

Also in medical and scientific, the fraud can occur at various levels. It can occur in clinical trials, and also occur in a more commercial context. In Computer intrusion fraud, hackers can find passwords, read and change files, alter source code, read emails and so on (Bolton and Hand, 2002).

3.4 Risk management

Credit card business has been released by many banks with high return, high spread and easy-to-apply in order to satisfy consumers' needs. To encourage the customers, some banks simplify the credit rating, which in turn has increased credit risk. Therefore, credit risk auditing is critical in the successful management of credit card business (Chen and Huang, 2011). Data mining assesses the risk of a bank's entire portfolio of loans that allow credit risk managers to allocate optimal loan loss reserves-funds in order to cover bad loans which is important to profitability in banking and other financial institutions (Anonymous, 1998; Fabris, 1998; Yager, 1996; Hormozi and Giles, 2004).

In medical healthcare, deeper analysis of member populations and better assessment of at-risk populations can reduce risk and better serve their members' needs (Hagland, 2004; Cumming, 2003). Moreover data mining in insurance companies, identify risk factors that predict profits claims and loses, and improve predictive accuracy (Gayle, 2002).

3.5 Planning and scheduling

Examples of applying data mining for planning and efficient schedules are:

Honda Motors Company in the USA is using data mining to predict at what age or mileage various components of cars are likely to fail. The resulting information allows engineers to plan maintenance schedules (Collier and Held, 2000). The automobile industry is one of the important sectors in the German economy. Therefore, the efficient planning based on reliable forecasts is an important contribution for business management in the German automobile market (Brühl et al., 2009; Hulsmann et al., 2012). Also data mining is used for effective production schedules to reduce the bottlenecks and to smooth assembly line setups and breakdowns (Grupe and Simon, 2004). Furthermore, British Telecomm (BT) employs around 20000 engineers in UK who provide services for business and resident customers. To manage its resources efficiently and effectively, a sophisticated dynamic scheduling system using data mining techniques is built to proposed sequence of work for field engineers (Ho and Azvine, 2001).

3.6 Fault patterns

Data mining has been applied for identifying fault patterns in different areas as follows:

Rapid detection faults are necessary to minimise the undesirable effects of detection and reconfiguration delays (Gaeid and Mohamed, 2010). Faults diagnosis especially bearing problems in induction motors are very important for many industries. Detecting bearing fault avoids fatal breakdowns of the machines (Zarei, 2012). Faults in the gearboxes and mechanical faults in industrial robots as well as other machines often show their presence as audible deviations compared with a normal sound profile. As a part of the end-test of industrial robots, a subjective condition monitoring based on hearing is used in order to detect audible deviations (Bengtsson et al., 2004). For managing communication networks, alarms must be analysed automatically in order to identify network faults before they occur and degrade network performance (Weiss, 2005). Feeder patrols in Taiwan Power Company have identified the fault locations, the abnormal observations of the feeder, and the conditions in the surrounding environments (Peng at el., 2004). In manufacturing process of industrial conveyor belts, faulty quality

categories is identified for the products being used and causal relationship is extracted between manufacturing parameters and product quality measures (Hou et al., 2003).

3.7 Process improvement and saving cost

Florida Hospital has applied data mining to find patterns in tests and medications most likely to be ordered by physicians. The results are compared with applicable treatment standards in order to improve patient care (Baldwin, 2000).

In semiconductor manufacturing, the yield of a silicon wafer is affected by micro-contaminants which can damage some of the chips on the wafer during the fabrication process, thus reducing yield. As device geometry continues to shrink, these micro-contaminants have an increasingly negative impact on yield. By diminishing the contamination problem, semiconductor manufacturers will significantly improve the wafer yield (Braha and Shmilovici, 2002).

The payment errors during processing claims increase the health insurance costs. These costs are due to the increased administrative costs for extra administrative effort, for instance, the administrative costs for healthcare in the USA represent 186 billion dollars. Therefore, reducing these errors and generating explanations will help the auditors to correct these claims and will allow health insurance companies saving millions of dollars every year (Kumar et al., 2010).

American Skiing Company is using data mining to identify persons who would respond to cross-promotional efforts involving midweek stays at both Steamboat Springs and Killington in order to reduce the advertising cost (Grupe and Simon, 2004). Furthermore, Western Digital, a leading hard disk manufacturer in California applies data mining to reduce the rate of field failures in hard disk drivers and improve quality by identifying root causes (Collier and Held, 2000).

4 Applied data mining techniques in each decision problem classification

In each data mining problem or decision problem category, data mining techniques are used to solve this problem as shown below.

For analysing customer behaviour, bank databases of Taiwanese major credit card insurer use neural network and decision tree technique. Neural network grouped customers with shared customer behaviour and customer value. Decision tree identified relevant knowledge and built customers profiles to determine the customers with higher customer values. These customers might be the target customer groups of precedence. From this, marketers can make better decisions on each target group of customers for specific marketing strategies (Hsieh and Chu, 2009). The customers' profiles in the telecommunication companies are extracted from the call detail data based on their patterns of phone usage. These profiles can then be used for better understand the customer behaviour or marketing purposes. To do so, neural network is used to predict the probability of a customer being a business or resident, based on the distribution of calls by time of day. Then, the probability estimate generated by the neural network is used as an input to the decision tree learner. The decision tree generated rules to classify a customer as being a resident or business (Weiss, 2005). In retailing business, market-basket analysis is employed to discover customer purchasing patterns. This technique examines customer buying habits by findings association relationship among different

items that customers place in their shopping baskets. The aim is to find trends across large numbers of transaction records that can be used to understand and exploit natural buying patterns. This information can be used to introduce targeted promotional activities to increase overall sales, adjust inventories, or move specific products (Moxon, 1996; Raorane and Kulkarni, 2011).

In customer churn and determining profitability, neural network is applied in cellular network services to predict churn and prevent the customer's turnover. The churn dataset deals with customers of cellular service provider and the data pertinent to the voice calls they make. The generated neural net calculates two new fields (the predicted Churn and a confidence value for the prediction) for every record in the input database. The output layer contains two neurons corresponding to the two values of the output fields churn being true or false (Sharma and Panigrahi, 2011). In mobile telephony industry, two probabilistic data mining algorithms Naïve Bayes and Bayesian Network are applied to build churn prediction model. A set of new features was presented to improve the recognition rates of possible churners. These features were derived from customer profile data and call details, and they were evaluated using Naïve Bayes and Bayesian Network algorithms. As a result of improving prediction rates, mobile telephony industry can reduce the various costs associated with customer churn (Kirui et al., 2013). The decision tree is another technique that is used as a churn prediction model in Malaysian telecommunication industry. From the decision tree analysis, the first classification attribute that contribute to churn is the area of the subscribers. This area is related to the lengths of services and total of minutes for customer churn. If area is rural and length of service more than 20 years, then the subscribers are not likely to churn, and if area is sub-urban and the total of minutes that customer engages in line less than 10 min, the subscribers are likely to churn (Oseman et al., 2010). Furthermore, *K*-Means clustering method and CART decision tree are implemented in insurance industry. *K*-Means identifies the characteristics of customers and determining churn rate. CART decision tree extracts patterns about the customers' churn and determines the reason for churning in each group of customers. These tools can help insurance companies to make suitable strategies in order to prevent churn of customers (Soeini and Rodpysh, 2012). In retail, Dovetail Solutions' methodology is the Value, Activity, and Loyalty™ Method, or VAL™, which uses transactional data to extract information about customer activity, churn rate, and expected future purchases. Customer value is not only determined by past revenues, but also by the customer's expected future purchasing behaviour. This can be measured by customer activity and expected lifetime. Activity gauges the probability that a customer will purchase again, while lifetime measures how long a customer is expected to remain active. This would allow a retailer to measure the loyalty of its customer based on determining the profitability and churn rate (Groth, 1999).

In fraud detection, in fraud detection, neural network for credit card detected system is designed to facilitate real time transaction entry and react to a suspicious transaction. The architecture of the neural network was based on unsupervised method. It was applied four clusters of low-, high-risky and high-risk clusters. This detected fraudulent use of a card faster and efficiently (Bolton and Hand, 2002; Ogwueleka, 2011). A combined rule-based systems and neuro-adaptive approach are also implemented to detect credit card transactions. Each transaction has symbolic and analogue data. Mining the symbolic data is based on rules which represent the misuse transactions. In mining the analogue data, learning process of neural network is applied for dynamic classification. The combined information of rule-based association system and the analogue neural expert are served as

a criterion for fraud diagnosis. This increases the probability for the diagnosis fraud to be correct and therefore increases the confidence and decreases the number of false alarms (Brause et al., 1999). In telecommunications fraud, rule-based detection systems apply rules such as calls that appear to overlap in time, very high value and very long calls. At a higher level, statistical summaries of call distributions (often called profiles) are compared with thresholds determined by experts to known fraud or non-fraud cases. The main fraud detection software of the Fraud Solutions unit of Nortel Networks uses a combination of neural network and profiling (Bolton and Hand, 2002). Moreover, neural networks, mixture models and Bayesian networks are implemented in telecom fraud detection based on call records stored for billing. First, a feed-forward neural network based on supervised learning is used to classify subscribers using summary statistics. Secondly, Gaussian mixture model is used to model the probability density of subscribers' past behaviour, so the probability of current behaviour can be calculated to detect any abnormalities from the past behaviour. Lastly, Bayesian networks are used to describe the statistics of a particular user and the statistics of different fraud scenarios (Taniguchi et al., 1998). Visualisation for mining very large datasets is also used in telecom fraud detection. In visualisation method, the human pattern recognition skills interact with graphical computer display of quantities of calls between different subscribers in various geographical locations. A possible future scenario would be to code into software the patterns, which humans detect (Bolton and Hand, 2002). Generally, neural network represents a better method for detecting telecommunications fraud, due to its speed and efficiency as well as its inherent ability to adapt (Akhter and Ahamad, 2012). In the area of insurance fraud, neural network is commonly used. Insurer in fraud detection can use neural networks uncover trends in physicians' or claimants' behaviour such as excesses in the amounts billed for standard treatments (Tauhert, 1997). For insurance fraud detection, specific analytical techniques adept at finding such subtleties are also used including market basket analysis, cluster analysis and predictive models to look for unusual associations' anomalies or outlying (Gayle, 2002). The Canadian securities market auditors and regulators found great benefits in data visualisation. It helps them in identifying frauds through see the patterns within the data or other information not readily discernable. A visual analytics framework for fraud detection includes 3D Treemap and social network visualisation. The 3D Treemap is used for real-time monitoring of stock market performance and identifying a particular stock that produced an unusual trading pattern. The social network visualisation aims to analysis the trading network in order to identify suspected pattern (Golmohammadi and Zaiane, 2012; Huang et al., 2009). Medical scientific fraud implements genetic algorithms and k-nearest neighbour methods to classify the practice profiles into classes from normal to abnormal. The genetic algorithm determines the optimal weighting of the features in the profile of either the doctor or patient. The KNN algorithm uses weights and examines both the majority rule and the Bayes rule to drive a classification from *k*-nearest neighbours. The results indicate that this classification methodology achieved good generalisation in classifying profiles and detecting fraud (He et al., 2000). Sequence analysis is applied in computer intrusion because the hacker's activities are the sequence of commands that is used when compromising the system (Bolton and Hand, 2002). Lee and Stolfo construct detection model including association rules algorithm and frequency episodes algorithm. These algorithms can compute audit record patterns that are important for describing program or user behaviour. The experiments on sendmail system call data and network tcpdump data are used to compute

classifiers that can recognise anomalies and known intrusions. The discovered patterns can guide the audit data gathering process and facilitate feature selection (Lee and Stolfo, 1998). Another application is performing profiling by training a neural network on the process data as well as referenced other neural approaches. Experimental evaluation on real-world data shows that neural network intrusion detector (NNID) can learn to identify users by what commands they use and how often identification can be used to detect intrusions in a network computer system (Ryan et al., 1997).

In risk management, a model of artificial neural network and decision tree is built for determining the risk factors in lending decisions in credit card business. The credit auditing data are analysed to identify feature variables such as demographic data, debt data and payment rating. After that, neural network is utilised to predict customer's regular pattern of consumption, payment and bad debt. The decision tree developed a set of credit granting principle to improve credit checking effect and credit risk control. This model could enhance the stability and profitability of the bank's credit card business (Chen and Huang, 2011). Axciom group built a customised model of linear and logistical regression for credit risk to achieve client business needs. The company identifies clusters or group by demographics, credit use and other factors. The size of the group is the function of what is profitable for the client. The eight or ten clusters are a good balance for helping card issuers target rewards and customise products. HNC Corporation uses hybrid models in credit card industry to manage clients' transaction data instead of summary data for managing risk. Transaction data can help companies know what they want, but the summary data does not yield subtleties (Anonymous, 1998). Fuzzy logic and genetic algorithm are used to solve problems concerning financial risk management. Fuzzy logic is useful in soliciting information on user perceptions of risk factors. Genetic algorithm helps clarify how and when user preferences affect the perceived desirability of a particular outcome. It also helps in tuning the parameters of fuzzy multiple criteria decision models (Yager, 1996). Moreover, the predictive modelling applies rule-based clinical logic to the data from medical, laboratory, pharmacy and health-risk assessment to find the statistical relationships between current use patterns and future outcomes (i.e., clinical and economic). Using analytic technique of predictive modelling can help in predict future healthcare cost and medical outcomes for a given population. This information can have an impact on identification and management of high-risk members (Cumming, 2003). Neural network as a very powerful modelling tool is used to predict more accurately risk factors, which can help insurance companies to set rates more accurately. For instance, an Irish auto insurer applied a data mining neural network tool to look at sub-groupings within high risk customer groups (Gayle, 2002; Tauhert, 1997).

For planning and scheduling, Honda motor in the USA implements statistics using Weibull analysis to predict at what age or mileage various components of cars are likely to fail. The resulting information allows engineers to plan maintenance schedules and design cars that will last longer. This careful analysis and the feedback of its findings into production have enabled Honda to achieve some of the highest value for cars in the USA (Collier and Held, 2000). Support vector machine (SVM) and decision trees represent suitable data mining tools for sales forecast in German automobile market. The database of the forecast model includes the main time series, which represent the registrations of new automobiles and the secondary time series (exogenous parameters). In the quarterly data, SVM is very reliable method for the present forecasting workflow due to its non-linearity, while in case of the monthly data, decision trees is the reliable and explicable one using absolute exogenous parameters (Brühl et al., 2009; Hulsmann et al., 2012).

Genetic algorithm is also applied in the most effective production schedule. Bierwirth and Mattfeld present a genetic algorithm to solve the job shop scheduling problem. GA is tested in a dynamic environment under different workload situations. Thus, a highly efficient decoding procedure is proposed to improve the quality of schedules, and it is tested for scheduling and rescheduling in a non-deterministic environment. The GA approach produces better results at a reasonable runtime (Bierwirth and Mattfeld, 1999). Integrating visualisation tools with predictive model including regression tree, linear function, and set of fuzzy rules enable British Telecom to schedule tasks and activities for field engineers in accordance with predetermined rules. This integration helps in the process of mining travel data, and the use of the visualiser adds another dimension to visual data mining. Therefore, providing multiple maps in a single view allow user to recognise complex dependencies among many attributes (Ho and Azvine, 2001).

Several data mining techniques deal with identifying fault patterns. Neural network is used to detect bearing faults of induction motors. The data was collected in three conditions including healthy, outer race defect and inner race defect. The input is the domain features that are obtained from direct processing of the signal segments. The time domain features are lead to accurately diagnosis of various motors bearing faults (Zarei, 2012). Gearboxes and mechanical faults in industrial robots often show their presence as audible deviations compared with a normal sound profile. A hybrid case-based reasoning method is implemented using a nearest neighbour approach for diagnosing audible faults on industrial robots. The reason for using this method is the feasibility of applying old knowledge of problem solving to solve new problem in this type of industrial applications (Bengtsson et al., 2004). Telecommunication networks use Timeweaver, a genetic-algorithm-based data mining system. Timeweaver is capable of operating directly on the raw network-level time-series data and other time-series data, thereby making it unnecessary to re-represent the network level data. Timeweaver searches through the space of possible patterns, which includes sequence and temporal relationships, to find predictive patterns. The system is especially designed to perform well when the target event is rare, which is critical since most network failures are rare (Weiss, 2005). Taiwan Power Company implements rough set theory to model the causal relationships between the faulty equipment and the evidences of observations during feeder outages and the surrounding environments. Rough set theory is a useful data mining tool for diagnosing the faulty equipment and thus inferring the fault location on the distribution feeder. The feeder patrols can quickly locate the fault location and find the faulty equipment through the derived inference rules. The attributes derived by a rough set algorithm form different rules with the corresponding predictive power and accuracy rates (Peng et al., 2004). Neural networks and rough set theory are applied in manufacturing system that makes conveyor belts. Neural network accurately classified quality faults, such as wrinkles and uneven thickness and the rough set determined the casual relationship between manufacturing parameters and output quality measures. The integration of neural networks and a rough set approach not only provides information about what is going to happen, but also reveals why it is happening and how to recover from the abnormal condition with specific guidelines on process parameter setting (Hou et al., 2003).

For process-improvement and saving cost, Florida Hospital uses statistical techniques to solve the problem of improving the hospital's collection rate on delinquent self-pay accounts. To improve patient care, Florida Hospital also uses statistics to find patterns in

tests and modifications most likely to be ordered by physicians. The results are compared with applicable treatment standards (Baldwin, 2000). Classification-based data mining methods including decision tree induction, neural networks, and composite classifiers are employed in semiconductor manufacturing. These classification methods enhance the understanding of the laser cleaning mechanisms and identify the attributes that are significant in the cleaning process yield. The composite classifiers yield higher accuracy than the accuracy of each individual classifier on its own (Braha and Shmilovici, 2002). To reduce health insurance costs, linear SVM is used as a robust for large data mining tasks with large feature sets. Linear SVM predicts claims that will need to be reworked and generating explanations to help the auditors correct these claims. This would have a big effect on the healthcare costs and help make the healthcare process smoother (Kumar et al., 2010). American Skiing Company applies genetic algorithm to identify the most likely families to visit their ski resort in Steamboat Springs Colorado. The solution identifies persons who would respond to cross-promotional efforts involving midweek stays at both Steamboat Spring and Killington, to reduce the cost of advertising to a small subset of potential customers (Grupe and Simon, 2004). On the other hand, Western Digital, a hard disk manufacturer has applied neural network, decision tree, and regression algorithm into three partitioned datasets. The Enterprise Miner manipulated the field failure dataset through running regression models, decision trees, and neural networks. As a result of this analysis, Western Digital focused efforts on its component suppliers to improve component quality. It also began analysing its financial data to assess the cost and value of each of its supplier relationships (Collier and Held, 2000).

5 Data mining technique choice is a decision problem dependent

Seven categories of data mining problems or decision problems in many industries were discussed in the previous section and the data mining techniques that have been implemented in each category. Table 1 represents the data mining techniques that map to the decision problem. Each category of decision problem implements different DM techniques. Among the techniques mentioned in this research, neural network is implemented in all categories of decision problems except planning and scheduling category. Neural network is suitable in the large and complex datasets. Its strength lies in discovering problems; it has the ability to learn over a period of time. Decision tree is applied in most of the decision problems categories except fault patterns and fraud detection. Decision tree cannot provide learning abilities compared with neural networks but its goal is to look for certain values through mining datasets that are sequentially related. Genetic algorithm is also used in most of the decision problems categories except two categories analysis customer behaviour and the customer churn. This means that genetic algorithm could identify the best solution for business problems based on the specified criteria, but it cannot provide learning such as neural network. GA is capable of securing optimum or nearby optimum solutions to the business problems. Moreover, the table illustrates that the predictive modelling and statistics are used in three categories of data mining problems. Predictive modelling helps in detecting fraud, managing risk and planning and scheduling for predicting the future events. Its goal is to identify a model or set of models that can be used to predict some response of interest. The three decision problems that apply statistics are customer churn, planning and scheduling and process improvement and saving cost. In these problems, statistics can discover patterns and build

predictive models. Finally, in Table 1, we can see other data mining techniques for solving specific decision problems. These techniques are association rules, visualisation, nearest neighbour and SVM. Association rules, visualisation and nearest neighbour are used for fraud detection while visualisation and SVM are used in planning and scheduling.

Table 1 The implemented data mining techniques in each category of decision problem

Decision problem / *DM technique*	*Fraud detection*	*Analysis customer behaviour*	*Customer churn*	*Risk management*	*Fault patterns*	*Planning and scheduling*	*Process improvement and saving cost*
Neural network	*	*	*	*	*		*
Decision tree		*	*	*		*	*
Rule-based systems	*						
Genetic algorithm	*			*	*	*	*
Visualisation	*					*	
Statistics			*			*	*
Clustering			*				
Predictive models	*			*		*	
Association rules mining	*	*					
Regression				*			*
Case-based reasoning					*		
Nearest neighbour	*				*		
Rough set					*		
Hybrid models				*			
Support vector machine (SVM)						*	*
Set of fuzzy rules						*	
VAL			*				
Sequence analysis	*						
Gaussian mixture model	*						

On the other hand, each category of decision problems implements some of the data mining techniques, which are different from the other categories of decision problems classification. For example, neural networks, rule-based, genetic algorithm, visualisation, mixture models, nearest neighbour, predictive modelling, sequence analysis and association rules are applied in fraud detection while in process improvement and saving cost, the applied techniques are neural networks, genetic algorithm, statistics, decision tree, regression and SVM.

We conclude that the data mining technique choice is a decision problem dependent. This means that for each category of decision problems there is a suitable data mining

techniques. These techniques are determined based on the type of the problem, the scale of the dataset and the dynamic changes of these data. Moreover, many companies apply more than one data mining technique in order to solve the decision problems.

6 Conclusion

Data mining is vital in many industries for decision making and company's competitive position. This research has developed decision problem classification of data mining problems to demonstrate the adequate data mining techniques and determined the reasons behind the selection of each particular technique. The research has shown that two different decision problems might need two different data mining techniques, which means that decision problem and dataset affects the choice of data mining technique. Moreover, most of organisations apply more than one data mining techniques to yield higher accuracy than the accuracy of implementing individual technique.

References

Akhter, M. and Ahamad, M. (2012) 'Detecting telecommunication fraud using neural networks through data mining', *International Journal of Scientific & Engineering Research*, Vol. 3, No. 3, pp.1–5.

Anonymous (1998) 'Shifting to a profit mode', *Credit Card Management*, pp.26–31.

Baldwin, F. (2000) *Data Mining Comes to Health Care: A Way to Excavate New Revenue and Savings*, Health Care Finance Online, http://www.highbeam.com/doc/1G1-67930396.html

Bengtsson, M., Olsson, E., Funk, P. and Jackson, M. (2004) 'Technical design of condition based maintenance system technology: a case study using sound analysis and case based reasoning', *Proceeding of 8th Congress of Maintenance and Reliability Conference*, Knoxville, TN, USA, pp.1–12.

Berson, A., Smith, S. and Threarling, K. (2005) *An Overview of Data Mining Techniques*, White Paper, http://www.thearling.com/text/dmtechniques/dmtechniques.htm

Bierwirth, C. and Mattfeld, D. (1999) 'Production scheduling and rescheduling with genetic algorithms', *Evolutionary Computation*, Vol. 7, No. 1, pp.1–17.

Bolton, R. and Hand, D. (2002) 'Statistical fraud detection: a review', *Statistical Science*, Vol. 17, No. 3, pp.235–255.

Braha, D. and Shmilovici, A. (2002) 'Data mining for improving a cleaning process in semiconductor industry', *IEEE Transactions on Semiconductor Manufacturing*, Vol. 15, No. 1, pp.91–101.

Brause, R., Langsdorf, T. and Hepp, M. (1999) 'Neural data mining for credit card fraud detection', *Proceeding of the 11th IEEE International Conference on Tools with Artificial Intelligence*, Chicago, IL, pp.103–106.

Brooks, P. (1997) 'Data mining today', *DBMS*, Vol. 10, No. 2, p.59.

Brühl, B., Hulsmann, M., Borscheid, D., Friedrich, C. and Reith, D. (2009) 'A sales forecast model for the german automobile market based on time series analysis and data mining methods', in Perner, P. (Ed.): *ICDM 2009*, LNCS, Vol. 5633, Springer, Berlin, pp.146–160.

Chan, P., Fan, W., Prodronidis, A. and Stolfo, S. (1999) 'Distributed data mining in credit card fraud detection', *IEEE Intelligent Systems*, pp.67–74.

Chen, M., Han, J. and Yu, P. (1996) 'Data mining: an overview from a database perspective', *IEEE Transaction on Knowledge and Data Engineering*, Vol. 8, No. 6, December, pp.866–883.

Chen, S. and Huang, M. (2011) 'Constructing credit auditing and control & management model with data mining technique', *An International Journal: Expert Systems with Applications*, Vol. 38, No. 5, pp.5359–5365.

Collier, K. and Held, G. (2000) 'Data mining quality in manufacturing data: best practices approach to the manufacturing industry', *SAS E-Intelligence*, KPMG Consulting and SAS Institute Inc.

Cumming, R. (2003) 'Predictive modeling in health care: going beyond predication to prevention and care', *Disease Management and Quality Improvement Report*, Vol. 3, No. 5, pp.1–9.

Edelstein, H. (2003) 'Data mining in depth: description is not prediction', *DM Review Magazine*, March, Available at: http://www.information-management.com/issues/20030301/6388-1.html

Fabris, P. (1998) 'Advanced navigation: marketing secrets from the financial sector show how data mining charts a profitable course to customer management', *CIO Magazine*, Vol. 11, No. 15, pp.50–55.

Friedman, J. (1997) *Data Mining and Statistics: What's The Connection?*, http://www-stat.stanford.edu/~jhf/ftp/dm-stat.pdf

Gaeid, K. and Mohamed, H. (2010) 'Diagnosis and fault tolerant control of the induction motors techniques a review', *Australian Journal of Basic & Applied Sciences*, Vol. 4, No. 2, pp.227–246.

Gayle, S. (2002) *Data Mining in the Insurance Industry*, SAS White Paper.

Golmohammadi, K. and Zaiane, R. (2012) 'Data mining applications for fraud detection in securities market', *European Intelligence and Security Informatics Conference (EISIC)*, pp.107–114.

Groth, R. (1999) *Data Mining: Building Competitive Advantage*, 2nd ed., Prentice Hall, Upper Saddle River, NJ.

Grupe, F. and Simon, J. (2004) 'Genetic algorithm: a business perspective', *Information Management and Computer Security*, Vol. 12, No. 3, pp.289–298.

Hagland, M. (2004) *Data Mining: Strong Computer Tools Allow Deeper Analysis of Medical Research Patient Care and Insurance Data*, Health Care Informatics Online, http://ehealthecon.hsinetwork.com/HI_mining_2004-04.pdf

Hand, D. (2000) 'Methodological issues in data mining', *Compstat 2000: Proceedings in Computational Statistics*, pp.77–85.

He, H., Hawkins, S., Graco, W. and Yao, X. (2000) 'Application of genetic algorithm and *k*-nearest neighbour method in real world medical fraud detection', *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 4, No. 2, pp.130–137.

Ho, C. and Azvine, B. (2001) 'Mining travel data with a visualizer', *The International Workshop on Visual Data Mining at ECML/PKDD (conjunction with the 12th European Conference on Machine Learning (ECML'01) and 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01))*, Freiburg, Germany.

Hormozi, A. and Giles, S. (2004) 'Data mining: a competitive weapon for banking and retail industries', *Information Systems Management*, Vol. 21, No. 2, pp.62–71.

Hou, T., Liu, W. and Lin, L. (2003) 'Intelligent remote monitoring and diagnosis of manufacturing process using an integrated approach of neural networks and rough sets', *Journal of Intelligent Manufacturing*, Vol. 14, pp.239–253.

Hsieh, N. and Chu, K. (2009) 'Enhancing consumer behavior analysis by data mining techniques', *International Journal of Information and Management Sciences*, Vol. 20, pp.39–53.

Huang, M., Liang, J. and Nguyen, Q. (2009) 'A visualization approach for frauds detection in financial market', *13th International Conference of Information Visualization*, pp.197–202.

Hulsmann, M., Borscheid, D., Friedrich, C. and Reith, D. (2012) 'General sales forecast models for automobile markets and their analysis', *Transactions on Machine Learning and Data Mining*, Vol. 5, No. 2, pp.65–86.

Kardaun, J. and Alanko, T. (1998) 'Exploratory data analysis and data mining in the setting of national statistical institutes', *Proceedings of NTTS98 International Seminar on New Techniques and Technologies for Statistics*, Sorrento, Italy.

Kirui, C., Hong, L., Cheruiyot, W. and Kirui, H. (2013) 'Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining', *International Journal of Computer Science Issues (IJCSI)*, Vol. 10, No. 2, pp.165–172.

Kumar, M., Ghani, R. and Mei, Z. (2010) 'Data mining to predict and prevent errors in health insurance claims processing', *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp.65–74.

Lee, S. and Siau, K. (2001) 'A review of data mining techniques', *Industrial Management and Data Systems*, Vol. 101, No. 1, pp.41–46.

Lee, W. and Stolfo, S. (1998) 'Data mining approaches for intrusion detection', *Proceedings of the 7th USENIX Security Symposium*, San Antonio, TX, USENIX Association, Berkeley, CA, pp.79–93.

Moxon, B. (1996) *Defining Data Mining*, DBMS Data Warehouse Supplement.

Ogwueleka, F. (2011) 'Data mining application in credit card fraud detection system', *Journal of Engineering Science and Technology*, Vol. 6, No. 3, pp.311–322.

Oseman, K., Shukor, S., Haris, N. and Bakar, F. (2010) 'Data mining in churn analysis model for telecommunication industry', *Journal of Statistical Modeling and Analytics*, Vol. 1, pp.19–27.

Peng, J., Chien, C. and Tseng, T. (2004) 'Rough set theory for data mining for fault diagnosis on distribution feeding', *IEE Proceedings Gener. Transm. Distrib*., Vol. 151, No. 6.

Raorane, A. and Kulkarni, R. (2011) 'Data mining techniques: a source for consumer behavior analysis', *International Journal of Database Management Systems*, Vol. 3, No. 3, pp.45–56.

Ryan, J., Lin, M. and Miikkulainen, R. (1997) 'Intrusion detection with neural networks', *AAAI Workshop on AI Approaches to Fraud Detection and Risk Management*, AAAI Press, Menlo Park, CA, pp.72–79.

Seifert, J. (2005) *Data Mining: An Overview*, CRS Report for Congress.

Sharma, A. and Panigrahi, P. (2011) 'A neural network based approach for predicting customer churn in cellular network services', *International Journal of Computer Applications*, Vol. 27, No. 11, pp.26–31.

Soeini, R. and Rodpysh, K. (2012) 'Applying data mining to insurance customer churn management', *International Proceedings of Computer Science and Information Technology (IPCSIT)*, Vol. 30, pp.82–92.

Taniguchi, M., Haft, M., Hollmén, J. and Tresp, V. (1998) 'Fraud detection in communication networks using neural and probabilistic methods', *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.1241–1244.

Tauhert, C. (1997) 'Neural networks: not just a black box', *Insurance and Technology*, Vol. 22, No. 4, pp.30–32.

Weiss, G. (2005) 'Data mining in telecommunications', *Data Mining and Knowledge Discovery Handbook*, Springer, USA, pp.1189–1201.

Yager, R. (1996) 'Fuzzy logic and genetic algorithms for financial risk management', *Proceedings of the IEEE/IAFE Conference*, 24–26 March, 1996, pp.90–95.

Zarei, J. (2012) 'Induction motors bearing fault detection using pattern recognition techniques', *Expert Systems with Applications*, Vol. 39, No. 1, pp.68–73.