

Review Article

Data Mining Application for Finding Patterns: Survey of Large Data Research Tools

Aive Islam, Tamzidl Amin

Department of Computer Science & Engineering, United International University, Dhaka, Bangladesh

Email address:

Tamzidlamin1976@gmail.com (T. Amin)

To cite this article:

Aive Islam, Tamzidl Amin. Data Mining Application for Finding Patterns: Survey of Large Data Research Tools. *American Journal of Neural Networks and Applications*. Vol. 3, No. 2, 2017, pp. 14-21. doi: 10.11648/j.ajjna.20170302.11

Received: November 1, 2017; **Accepted:** November 21, 2017; **Published:** December 5, 2017

Abstract: Data Mining is now a common method for mining data from databases and finding out patterns from the data. Today many organizations are using data mining techniques. In this paper concepts and techniques such as Neural Network, Decision Tree, Clustering, Association Rule, Clustering and many more techniques of Data Mining is reviewed. This paper focuses how different techniques of Data Mining are used in different applications for finding out patterns from the data taken from the data base.

Keywords: Neural Network, Decision Trees, Rule Induction Technique, Association Rules, Clustering, K-means

1. Introduction

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods such as neural networks or decision trees. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data. Finding useful patterns in data is known by different names (e.g., knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing)

The term “data mining” is primarily used by statisticians, database researchers, and the business communities. The term KDD (Knowledge Discovery in Databases) refers to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process [2, 3]. The steps in the KDD process, such as data preparation, data selection, data cleaning, and proper interpretation of the results of the data mining process, ensure that useful knowledge is derived from the data. Data mining is an extension of traditional data analysis and statistical

approaches as it incorporates analytical techniques drawn from various disciplines like AI, machine learning, OLAP, data visualization, etc.

2. Materials and Methods

The study was based on secondary data. The secondary data used were collected from different studies of the Library of United International University, Dhaka during from December 2016 to till March 2017.

These secondary data were United International University reports, different thesis, Books and other national and international journals.

3. Results and Discussion

E-business technology: The persons whose profile suggest that they are likely to provide a high lifetime value to the firm will be provided opportunities that will differ from those that are offered to consumers with less attractive profiles. Analytic software allows marketers to combine through data collected from multiple customers touch points to find patterns that can be used to segment their customer base. It was developed a more comprehensive understanding of data mining by examining the application of this technology in the

marketplace [4]. As more firms shift more of their business activities to the web, increasingly more information about consumers and potential customers is being captured in web server logs. It was examined issues related to social policy that arises as the result of convergent developments in e-business technology and corporate marketing strategies [4]. About consumers and potential customers is being captured in web server logs. Sophisticated analytic and data mining software tools enable firms to use the data contained in these logs, to develop & implement a complex relationship management strategy. Artificial neural networks are designed to model human brain functioning through the use of mathematics. Like neural network data mining through the use of decision tree algorithms discerns patterns in the data without being directed. According to a decision trees work like a game of 20 questions”, by automatically segmenting data into groups based on the model generated when the algorithms were run on a sample of the data. Web generated data includes information collected from forms, transactions as well as from click stream records.

Decision tree models: Decision tree models are commonly used to segment customers into “statistically significant” groups that are used as a point of reference to make predictions. Both neural networks & decision trees require that one knows where to look in the data for patterns, as a sample of data is used as a training device. The use of market basket analysis & clustering techniques does not require any knowledge about relationships in the data, knowledge is discovered when these techniques are applied to the data. Market basket analysis tools sift through data to let retailers know what products are being purchased together. Clusters prove to be most useful when they are integrated into a marketing strategy. The software companies that market personalization products that use data mining techniques for knowledge discovery, speaks to their potential clients in a language that make the benefits of these systems unmistakable. However customer relationship management appears to be the philosophy that will drive marketing strategies in the 21st century. Customer relationship management focuses not on share market, but on share of customer. Marketing strategists have been able to demonstrate that a firm’s profitability can increase substantially by focusing marketing resources on increasing a firm’s share of its customers business rather than increasing its number of customers [5]. One of the basic tenets behind customer relationship management is the Pareto Principle, the notion that 80% of any firms profit is derived from 20% of its customers. Engaging in a dialogue with that 20% in order to ascertain what their needs are & offering goods & services to meet those needs are said to be what customer relationship management is all about. Data mining technologies have allowed firms to discover and predict whom their most profitable customers will be by analyzing customer information aggregated from previously disparate database. The web has created a forum for firms to engage in a one to one dialogue with particular segments of their customer base in order to ascertain what the needs of those segments are.

Neural Network: A study was made about understanding the

brain requires a broad range of approaches and methods from the domains of biology, psychology, chemistry, physics & mathematics. The fundamental challenge is to decipher the “neural choreography” associated with complex behaviors and functions including thoughts, memory, actions and emotions. National Institute of Health recently launched the Human Connectome Project and awarded grants to two consortia. The consortium led by Washington University of Minnesota aims to characterize whole brain circuitry & its variability across individuals in 1200 healthy adults. Neuroimaging & behavioral data from the HCP will be made freely available to the neuroscience community via a database & a platform for visualization & user friendly data mining. The informatics effort involves major challenge owing to the large amounts of data, the diversity of data types & the many possible types of data mining. Some investigators will drill deeply by analyzing high resolution connectivity maps between all gray matter locations. It is inefficient for individual researcher to sequentially visit and explore thousands of databases. To promote the discovery and use of online databases, the NIF created a portal through which users can search not only the NIF registry but also the content of multiple databases simultaneously. The current NIF federation includes more than 65 databases accessing 30 million records in major domain of relevance to neuroscience. Beside every large genomic collection, there are nearly 1 million antibody records, 23000 brain connectivity records and more than 50,000 brain activation coordinates. Many of these areas are converted by multiple databases, which the NIF knits together into a coherent view. The NIF users should be able not only to locate answers that are known but to mine available data in ways that spur new hypothesis regarding what is not known. Perhaps the single biggest roadblock to this higher order data mining is the lack of standardized frameworks for organizing neuroscience data. Individual investigators often use terminology or spatial coordinate systems customized for their own particular analysis approach. This customization is a substantial barrier to data integration, requiring considerable human effort to access each resource, understand the context and content of the data and determine the conditions under which they can be compared to other data sets of interest. Neurolex terms are accessible through a wiki that allows users to view, augment and modify these concepts. The goal is to provide clear definitions of each concept that can be used not only by human but by automated agents, such as NIF, to navigate the complexities of human neuroscience knowledge. A key feature is the assignment of unique resource identifier to make it easier for search algorithms to distinguish among concepts that share the same label. So in future good planning and future investment are needed to broaden & harden the overall framework for housing, analyzing, and integrating future neuroscience knowledge. The international neuroinformatics coordinating facility plays an important role in coordinating and promoting this framework at a global level.

Data mining: Data mining techniques provide a popular & powerful tool set to generate various data driven classification

systems. It had already studied about combined use of self organizing maps & nonsmooth, nonconvex optimization techniques in order to produce a working case of a data driven risk classification system [6]. The optimization approach strengthens the validity of self organizing map results. This study is applied to cancer patients. Cancer patients are partitioned into homogenous groups to support future clinical treatment decisions. Most of the different approaches to the problem of clustering analysis are mainly based on statistical, neural network, machine learning techniques. The global optimization approach to clustering and demonstrate how the supervised data classification problem can be solved via clustering. The objective function in this problem is both nonsmooth and nonconvex and has a large number of local minimizers. Due to a large number of variables and the complexity of the objective function, general purpose global optimization techniques, as a rule fail to solve such problem. It is very important therefore, to develop optimization algorithm that allow the decision maker to find "deep" local minimizers of the objective function. Such deep minimizers provide a good enough description of the data set under consideration as far as clustering is concerned. Some automated rule generation methods such as classification and regression trees are available to find rules describing different subsets of the data. When the data sample size is limited, such approaches tend to find very accurate rules that apply to only a small number of patients. It was demonstrated that data mining techniques can play an important role in rule refinement even if the sample size is limited. For that first stage methodology is used for exploring and identifying inconsistencies in the existing rules, rather than generating a completely new set of rules. K-mean algorithm lies in the improved visualization capabilities resulting from the two dimensional map of the cluster. It was developed self organizing maps as a way of automatically detecting strong features in large data sets. Self organizing map finds a mapping from the high dimensional input space to low dimensional feature space, so the clusters that form become visible in this reduced dimension ability. The software used to generate the self organizing maps is ViscoverySOMine (www.eudaptics.com), which provides a colorful cluster visualization tool, & the ability to inspect the distribution of different variables across the map.

Cluster analysis: The subject of cluster analysis is the unsupervised classification of data & discovery of relationship within the data set without any guidance. The basic principle of identifying this hidden relationship is that if input patterns are similar, they should be grouped together. Two inputs are regarded as similar if the distance between these two inputs is small. This study demonstrates that data mining techniques can play an important role in rule refinement, even if the sample size is limited. It was demonstrated that both self organizing maps & optimization based clustering algorithms can be used to explore existing classification rules, developed by experts and identify inconsistencies with a patient database [6]. As the proposed optimization algorithm calculate clusters step by step and the form of the objective function allow the user to significantly reduce the number of instances in a data

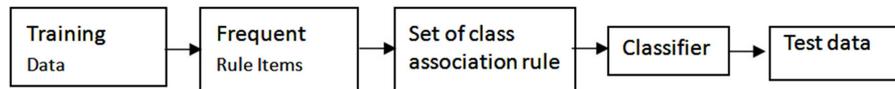
set. A rule based classification system is important for the clinicians to feel comfortable with the decision. Decision tree can be used to generate data driven rules but for small sample size these rules tend to describe outliers that do not necessarily generalize to larger data sets.

Following suggestions are given by Huda Akil, Maryann E. Martone, David C Van Essen:

Neuroscientists should as much as feasible, share their data in a form that is machine accessible, such as through a web based database or some other structured form that benefits from increasingly powerful search tools. Database spanning a growing portion of the neuroscience realm need to be created, populated and sustained. This effort needs adequate support from federal & other funding mechanism. Because databases become more useful as they are more densely populated, adding to existing database may be preferable to creating customized new one. Some type of published data should be reported in standardized table formats that facilitate data mining. Cultural changes are needed to promote wide spread participation in this endeavor. In a study said that the size & the complexity of the data from scientific simulations, observations and experiments becoming a major impediment to their analysis. To enable scientists to address this problem of data overload and benefit from their improved data collecting abilities, the Sapphire project team has been involved in the research, development & application of scientific data mining techniques for nearly a decade. According to team the raw data available for analysis was in the form of images, structured or unstructured mesh data with physical variables at each mesh point or time series data collected by different sensors. They designed and built a software toolkit with separate modules for different tasks such as denoising, background subtraction to identify moving objects in video, dimension reduction to identify key characteristics of objects, pattern recognition for clustering and classification. Recognizing that for each task different methods were likely to be appropriate based on the data; we used object oriented techniques to provide a uniform interface to the algorithms. Data from first survey are available in two forms of image maps & catalog. Goal of this scientific application is not to build a predictive model but to discover a set of features that may provide insights into the phenomena of interest. Study made by Fadi Thabtah about associative classification mining said associative classification integrates two known data mining tasks, association rule discovery and classification to build a model for the purpose of prediction. Classification and association rule discovery are similar tasks in data mining, with the exception that the main aim of classification is the prediction of class labels, while association rule discovery describes correlations between items in a transactional database. Rule induction approach such as IREP and RIPPER derive local sets of rules in a greedy manner [7]. The derived rules are local because when a rule is discovered, all training data objects associated with it are discarded & the process continues until the rule found has an acceptable error rate. This means rules are discovered from participations of the training data set & not from the whole

training data set once. The search process for the rules is greedy as most rule induction algorithms normally look for the rule that maximizes a statistical measure. Example:- The IREP rule induction algorithm constructs the rules based on first order inductive learner, gain measure [8]. This means that the

Associative classification steps:



In the Apriori association rule discovery algorithm the discovery of frequent itemsets is accomplished in levels, where in each level Apriori uses itemsets found to be frequent in the previous level to produce new candidate itemsets. Apriori utilizes the downward closure property with aim of speeding up the search process by reducing the number of candidate itemsets at any level. The downward closure property ensures that all subsets of a frequent itemset must be frequent as well. If an itemset is infrequent at any level, it will be removed because any addition of items to it will not make it frequent. Apriori uses this property to prune candidate itemsets that have infrequent subsets before counting their support at any level. This should reduce the time to produce and compute the support for all items combinations in the transactional database. The problem is defined in terms of automatic constructions of index pages based on logfile data, with the goal of constructing index pages that provide users access to information that they are likely to view [9]. The algorithm proposed has four steps: 1. Processing the logfiles into user visits. 2. Computing co-occurrence frequencies between pages and creating a similarity matrix. 3. Creating a graph from this matrix & then finding cliques in this graph. 4. Creating index pages corresponding to each clique in the graph. The algorithm is evaluated based on how often users simultaneously visit the pages in the index page. Another interesting approach was proposed to the problem of automatically synthesizing web pages. They address the issue of comparing the “expected location” of various pages with the actual location where the page is, and they propose an algorithm that automatically finds such pairs of web pages. The main idea is examine user session in the logfiles and identify backtracking points where a user backtracks to a previous page & continues the session. An objective of website design can be to maximize navigational simplicity by minimizing the average number of clicks required to get to any page. Further there may be several constraints on the design. Yet another constraint may require the sports & finance pages to be explicitly linked to each other based on input from the marketing department. However it is difficult to provide user specific constraint a priori. Data Mining can help to address this problem. First the website is structured as per the solution to the optimization problem. As the user navigates this site, data on user’s access pattern is gathered. Data Mining can thus identify additional constraint using patterns identified by this data. Data Mining can find most user access the finance & sports pages together. This information can then be used as a constraint. The optimization problem is then solved

attribute value in the training data set with the best FOIL gain is chosen first as a member of the current rule left hand side (antecedent) and the process is repeated until a stopping condition is met.

incorporating the additional constraint. This process can be iteratively done until some user defined stopping criteria. Most of the approaches developed for solving this type of recommendation problem use statistical & data mining approaches and usually tries to determine “good” products to recommend to the customer. Various existing recommendation methods were classified into content based, collaborative & hybrid approaches and have been reviewed in survey [10, 11]. This example demonstrates the opportunity for new research to identify similar problems that can be formulated in the context of website design. In this mode, problem in website design will be mainly formulated as optimization problems. However Data Mining will play a key role in the specification of the optimization step and will also play a part in the specification of the constraints. Another interesting customer interaction issue is the provision of personalized services to customers, including recommendations.

Example: - Amazon.com may want to determine which best 10 books to put on the customer’s welcome page. The challenge for this problem is that the rating function is usually partially specified. Once the optimization problem is defined, data mining can contribute to its solution by learning additional constraints with data mining methods, thereby significantly reducing the search space. For example, in the case of online wine store, we can learn that a customer usually prefers to buy red inexpensive wines, however on certain occasions, such as his wife’s birthday, he usually buys midrange white Chablis. By storing this information in the customer’s profile and invoking it on these special occasions, much tighter constraints on recommendations can be specified. In this section, we reviewed some of the existing approaches to solving three main analytical e_CRM problems. First is specifying effective performance metrics and finding optimal solutions based on them. Second is developing effective customer analysis & third is a customer interaction method. Another study is made about how data mining approach develop bankruptcy prediction model suitable for normal & crisis economic condition. It observes the dynamics of model change from normal to crisis condition and provides interpretation of bankruptcy classification.

Bankruptcy prediction: The bankruptcy prediction model revealed that the major variables in predicting bankruptcy were “cash flow to total assets” and “productivity of capital” under normal conditions and “cash flow to liabilities”, “Fixed assets to stockholders equity and long term liabilities” under crisis conditions. The accuracy rate of final prediction models

in normal conditions & in crisis conditions were found to be 83.3 % and 81.0 % respectively. When the normal model was applied in crisis situations, prediction accuracy dropped significantly in the case of bankruptcy classification (from 66.7 percent to 36.7). As we are in the 21st century, corporate bankruptcy in the world, especially in East Asia, has reached an unpredictable level. Corporate bankruptcy brings with it economic losses to management, stockholders, employees, customers and others together with great social and economical cost to the nation. Thus accurate prediction of bankruptcy has become an important issue in finance. Since the seminal study of Altman on bankruptcy prediction, numerous follow up studies have tried to further develop appropriate models, by applying data mining techniques including multivariate discriminate analysis, logistical regression analysis, profit analysis, genetic algorithms, neural networks, decision trees and other statistical & computational methods. It is worth noting however that all of these bankruptcy models assume “normal” economic conditions.

Z- score model: The detection of corporate failures is a subject that has been particularly amenable to financial ratio analysis. The first study was done during the great depression era, then showed that failing firms exhibit significantly different ratio than do successful firms. Beaver then applied univariate analysis of financial ratios to predict corporate bankruptcy, while others strongly recommended multivariate analysis. The breakthrough bankruptcy prediction model, the Z- score model developed by Altman came in the late 1960s. The five variables Z - score model using multiple discriminant analysis showed by very strong predictive power. Most of the studies including that of Altman used relatively small firms in their samples; generalization of research results was hard to accept. Altman, Haldeman and Narayanan therefore developed the ZETA model to be applied to larger firms, not limited to specific industries. The study made by HE Zengyou, XU Xiafei and DENG Shengchun said that clustering is an important KDD technique with numerous applications, such as marketing & customer segmentation. Clustering typically groups data into sets in a way that intra - cluster similarity is maximized, while the inter cluster similarity is minimized. Many efficient clustering algorithms such as ROCK, C2P, DBSCAN, BIRTH, CURE, CHAMELEON, wavecluster & CLIQUE have been proposed by the database research community. Most previous clustering algorithm focus on numerical data whose inherent geometric properties can be exploited naturally to define distance function between data points. However many of the data in databases are categorical, where attribute values cannot be naturally ordered as numerical values. An example of categorical attribute is shape whose values including circle, rectangle, ellipse etc. Due to the special properties of categorical attributes, the clustering of categorical data seems more complicated than that of numerical data. They present Squeezer, a new clustering algorithm for categorical data. The basic idea of Squeezer is simple. Squeezer repeatedly read tuples from a dataset one by one. When the first tuple arrive, it forms a cluster alone. The consequent tuples are either put into

existing clusters or rejected by all existing clusters to form a new cluster by given a similarity function defined between a tuple and a cluster. The Squeezer algorithm only makes one scan over the dataset, thus is highly efficient for disk resident datasets where the I/O cost becomes the bottleneck of efficiency. The main objective of Squeezer is the combination of efficiency and scalability. Experimental result shows that Squeezer achieves both high quality clustering results & scalability. Main contribution of the study are the algorithm is suitable for clustering data streams, where given a sequence of points, the objective is to maintain consistently good clustering of the sequence so far, using a small amount memory & time, outliers can be handled efficiently and directly, the algorithm does not require the number of desired clusters as an input parameter. This is very important for the user who usually does not know this number in advance. The only parameter to be pre - specified is the value of similarity between the tuple and the cluster, which incorporates the users expectation that how close the tuples in a cluster should be. Analytical eCRM as a four step iterative process consisting of collecting & integrating online customer data, analyzing this data, building interactions with customers based on this analysis such that certain performance metrics such as LTV are optimized. A typical performance metric used in many CRM applications is the LTV of a customer. One of the key questions in CRM is how to develop proactive customer interaction strategies that maximize LTV. This key problem is viewed by some as the “holy grail” of CRM. Traditionally the problem of estimating LTV is divided into two components. One is estimating how long a customer will stay & second is estimating the flow of revenue from the customer during this period. Estimating the flow of revenue during customer’s lifetime was done with parametric models in the marketing literature. With respect to estimating customer tenure, provide analytical models that determined whether a customer at any point in time is “active”, by identifying a set of qualitative criteria that capture when a customer is more likely to be active [12]. The concept of customer retention is a natural extension of this approach, because it deals with trying to prevent a customer from becoming inactive and hence it is a proactive method of increasing LTV. A more general framework for inactive marketing for LTV, suggesting the key notion that customers are “addressable” and can be engaged in interaction [13]. This notion of customer interaction is further explored in which the problem of optimal communication to maximize LTV is addressed [14]. They show that both too little and too much communication can result in a firm’s failure to capture adequate value from its customers.

This analysis provides useful insights into the LTV optimization problem. However this work considers only optimal communication frequency with customers & the model does not take into account the existence of data on customer behavior. In contrast the Data Mining community studied the LTV problem in the presence of large volumes of customer data [15]. In particular Mani et al. (1999) predict customer tenure using classical survival analysis methods by building a neural network & training it on past customer data.

However Mani et al. (1999) do not address the problem of computing optimal parameters for LTV models. Rosset et al. (2002) compute LTV based on large volumes of customer data by focusing on using Data Mining to estimate customer churn & future revenues. They use statistical and data mining methods to estimate future revenues. They point out that it is difficult to solve an LTV optimization problem in the presence of large volumes of data and Rosset et al. (2002) stopped short of doing this. Similar Dreze and Bonfrer (2002) studied only an optimization problem & did not deal with large volumes of data. Therefore developing new algorithms computing optimal LTVs and utilizing optimization & data mining methods in the presence of large volumes of data constitutes an important & challenging problem for operations research or management science researchers. One way to deal with the complexity of the general LTV optimization problem is to reduce it to simpler types of problems. One such reduction may consider various heuristics producing higher LTV values, rather than attempting to find an optimal solution. For example, we can study customer attention problems & determine policies that will result in lower attrition rates and therefore higher customer LTVs. Another way to deal with this complexity is to seek to optimize simpler performance measures that can serve as proxies for LTV. We can use customer satisfaction rates with various offerings as one such measure. Customer analysis includes two main steps in the eCRM context. First is preprocessing data that tracks various online activities of the customers – this involves starting with individual user clicks on a site and constructing logical users “session” and summary variables. Second is building customer profiles from this and other data. At a general level, a profile is a set of patterns that describe a user. Data Mining is used to learn these patterns from data. Customer profiles are then built from these results. Current literature considers heuristic methods for analyzing click stream data generated by websites. One of the most important problem is the session identifies problems, which determines how to group consecutive clicks into session. This is an important business problem because most of the users tracking system developed by such companies such as Epiphany and Blue Martini provided only ad hoc solutions to the session identification problem. Some popular session identification methods include session level characterization that aggregates user clicks into session, a fixed length sliding window method that breaks a session into several sliding windows, and different types of clipping methods that break a session into windows of different sizes using various splitting methods [16]. Various session identification methods can produce radically different conclusions derived from the same data. The main reason for preprocessing click stream data is to build a model, such as one that would predict the likelihood that a current user’s session would result in a purchase. Accurate model are crucial for such problems, requiring optimal preprocessing of the click stream data. To partition the click stream data into session, one needs to specify optimization criteria. This can be done by first identifying “similar” groups of consecutive pages in the click stream, which can then be partitioned into

sessions to maximize intrasession similarities and intersession differences. One way to specify these similarities & difference is to identify the variance of some browsing measure, such as the time spent viewing a page. Building rich & accurate profiles of customer based on their transactional histories is also crucial in many CRM applications, including recommendation applications, one to one marketing & personalized web content delivery. These user profiles can contain factual information about the users such as demographic & psychological data, a set of rules capturing behavior of the user. Because many of the discovered rules can be superiors, irrelevant or trivial, one of the main problem is how to select an optimal set of rules for each customer from the set of rules previously discovered with Data Mining methods. In such applications personalized content needs to be delivered in real time, while the customer is waiting online. Therefore content delivery discussion should be driven by only a few rules to guarantee real time delivery and relevant content for the customer. Consequently optimization can help customer profiling and other applications to select a small number of the most important patterns from the set of previously discovered patterns. The aim of the study was to demonstrate the epidemiological use of multiple correspondence analyses as applied to tuberculosis (TB) data from North East Landon primary care trusts between the years 2002 & 2007 was used. TB notification data were entered for Multiple Corresponding Analysis allowing display of graphical data output. There has been rise in tuberculosis notifications in the UK since 1987. However excluding TB in Landon, rates of TB in the UK are relatively low & stable. In the context of North East Landon, high rates of TB are observed in some primary care trust areas whilst in others rates are relatively low. This demonstrates the complexity of TB epidemiology in the UK and Landon & is suggestive of a range of factors that give rise to high rates of TB in specific geographical areas. The enhanced tuberculosis surveillance (ETS) system was introduced in 1999 to aid notification. These collected data show the different demographic and clinical profiles of patients observed in NE Landon & may account for variations in TB rates. This requires appropriate statistical support & effective communication to decision makers. However analysis of large volume of data with a large proportion of categorical or nominal data that can display multiple associations may prove to be difficult to interpret if bivariate comparison is made. Factor analysis & Principle Component Analysis are inappropriate methods of analysis for these data which include a mix of continuous & categorical data. Multiple correspondence analyses are an analytical method that allows analysis of multiple categorical variables. We demonstrate the use of MCA as a tool for performing epidemiological mapping of TB patient variables. This may prove to be useful in identifying commissioning priorities in NE Landon. Multiple Correspondences Analysis is a multivariate extension of Correspondence Analysis that allows explanation of relationships between two or more variables. By including two or more variables in this type of analysis the complexity is increased, relationships between

variables are described in terms of the variance of data. Partitioning approach organizes the data into tree structure that can be processed independently. The performance of these methods, depend on the size of original database. Previous work on the solution to analytical electronic customer relationship management problem has used either data mining or optimization methods, but has not combined the two approaches. By leveraging the strength of both approaches, the eCRM problems of customer analysis, customer interaction & the optimization of performance metrics can be better analyzed. In particular many eCRM problems have been traditionally addressed using DM methods. There are opportunities for optimization to improve these methods describes these opportunities

Prior research has used optimization methods for solving data mining problems and used DM methods for solving optimization problems. In this survey we systematically explore how optimization and DM can help one another for certain customer relationship management (CRM) applications in e-commerce, termed analytical eCRM. Analytical eCRM includes customer analysis, customer interactions and optimization of various performance metrics such as customer lifetime value in web based e-commerce. To illustrate how optimization and DM interact in eCRM settings, consider the following two important eCRM problems. The first deal with finding the optimal lifetime value of a customer by determining proactive customer interaction strategies resulting in maximal lifetime profits from that customer. These eCRM applications include Max. of customer LTV, Customer analysis, including preprocessing click stream data & building profiles, Customer interaction methods including website design & personalization. By considering previous studies authors find out the scope to go for research in market basket analysis using three different algorithms namely Association Rule Mining, Rule Induction Technique and Apriori Algorithm.

4. Conclusion

In some cases accuracy is called the confidence and coverage is called the support. Accuracy refers to the probability that if the antecedent is true that the precedent will be true. High accuracy means that this is a rule that is highly dependable. Coverage refers to the number of records in the database that the rule applies to. High coverage means that the rule can be used very often and also that it is less likely to be a spurious artifact of the sampling technique or idiosyncrasies of the database. Assume minimum accuracy and minimum coverage. Authors will make a comparative study of three techniques and adopt the best conclusion. In Association Rule Mining, we will generate association rules and calculate support and confidence. Assume minimum support and minimum confidence. The rules satisfying both the criteria of minimum support & minimum confidence is true otherwise false. Rule induction technique retrieves all interesting patterns from the database. In rule induction systems the rule itself is of the simple form of "if this and this and this then this". The rules satisfying both the criteria of minimum

accuracy & minimum coverage is true otherwise false. The theory of Apriori algorithm is that "All nonempty subsets of a frequent itemset must also be frequent." This property prune the candidate which is not in any of the category & thus to reduce number of candidates. Authors will collect the data from shopping mall and will apply the data mining algorithms to find out the association between the products. It can be concluded that data mining has a vast range of applications and techniques to use in the fields of business, industries and medical facilities. In the reviews we can see how far Data mining has come and how it is being used for predicting the uncertain future. Different algorithms take different resources and need different input. But together they accomplish the task as per cycle of Data mining goes they train the data to make effective and accurate decision and classifications.

References

- [1] BalajiPadmanabhan and Alexander Tuzhilin, Institute for Operation Research and Management Science: 2011, On the use of optimisation for data mining:Theoretical Interaction and eCRM opportunities.
- [2] Shakil Ahmed, FransCoenen, Paul Leng, Knowledge Information System: 2006, Tree based partitioning of data for association rule mining.
- [3] RachidAnane, Computer and the humanities: 2001, Data mining and serial documents.
- [4] Anthony Danna, Oscar H. Gandy, Journal of business ethics: 2002, All that glitters is not gold: Digging Beneath the surface of data mining.
- [5] Peppers, D., & Rogers, M. (1993). The new marketing paradigm: one-to-one. *American Advertising*, 9(4), 93-102.
- [6] Leonid Churilov, AdylBagirov, Daniel Schwarta, Kate Smith, Michael Dally, Journal of management information system: 2005, Data mining with combined use of optimization techniques and self-organizing maps for improving risk grouping rules: application to prostate cancer patients
- [7] Fürnkranz, J., & Widmer, G. (1994). Incremental reduced error pruning. In *Proceedings of the 11th International Conference on Machine Learning (ML-94)* (pp. 70-77). Morgan Kaufmann.
- [8] Quinlan, J. R., & Cameron-Jones, R. M. (1993, April). FOIL: A midterm report. In *European conference on machine learning* (pp. 1-20). Springer, Berlin, Heidelberg.
- [9] Perkowitz, M., & Etzioni, O. (1998, July). Adaptive web sites: Automatically synthesizing web pages. In *AAAI/IAAI* (pp. 727-732).
- [10] Balabanović, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3), 66-72.
- [11] Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial intelligence review*, 13(5-6), 393-408.
- [12] Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting Your Customers: Who-Are They and What Will They Do Next?. *Management science*, 33(1), 1-24.

- [13] Blattberg, R. C., & Deighton, J. (1991). Interactive marketing: Exploiting the age of addressability. *Sloan management review*, 33(1), 5.
- [14] Dreze, X., & Bonfrer, A. (2002). Lifetime Value Maximization through Optimal Communication Timing.
- [15] Mani, I., & Maybury, M. T. (Eds.). (1999). *Advances in automatic text summarization* (Vol. 293). Cambridge, MA: MIT press.
- [16] Theusinger, C., & Huber, K. P. (2000, August). Analyzing the footsteps of your customers. In *Proc. of the Sixth ACM SIGKDD Internat. Conf. on Web KDD 2000* (pp. 44-52).