Data Mining, Machine Learning Approaches and Medical Diagnose Systems: A Survey

1. N.Satyanandam, *Associate Professor-CSE, Bhoj Reddy Engineering College for Women(BRECW), Hyderabad, India.*
2 Dr. Ch. Satyanarayana, *Associate Professor - CSE, JNTUK, Kakinada, India.*
3. Md.Riyazuddin, *Assist.Professor-IT, MuffakhamJah College of Engineering and Technology (MJCET) , Banjara Hills, Hyderabad, India.*
4. Amjan.Shaik, *Professor-CSE, ECET, Hyderabad, India.*

Abstract:
Data mining technology provides a user- oriented approach to novel and hidden patterns in the data**.**Data mining is a process which finds useful patterns from large amount of data. This technology has been successfully applied in Engineering and Technology, Science, Health Care Systems, Medical Diagnose Systems, Marketing and Finance to assist new discoveries and fortify markets. Some of the organizations have adapted this technology to progress their businesses and found outstanding results. In this paper we discussed a broad overview of some of the data mining techniques, their use in various emerging algorithms and applications. It provides an impression of the development of smart data analysis inmedicine from a machine learning irrespective.

Keywords:Machine learning, Data Mining, Clustering, Classification, Healthcare System.

I.INTRODUCTION

The Healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation**.** Data mining is often used during the knowledge discovery process and is one of the most important subfields in knowledge management. Data mining aims to analyze a set of given data or information in order to identify novel and potentially useful patterns(Fayyad et al., 1996). These techniques, such as Bayesian models, decision trees, artificial neural networks, associate rule mining, and genetic algorithms, are often used to discover patterns or knowledge that are previously unknown to the system and the users (Dunham, 2002; Chen and Chau, 2004). Data mining has been used in many applications such as health care, marketing, customer relationship management, engineering, medicine, crime analysis, expert prediction, Web mining, and mobile computing, among others[1,5,7].
The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

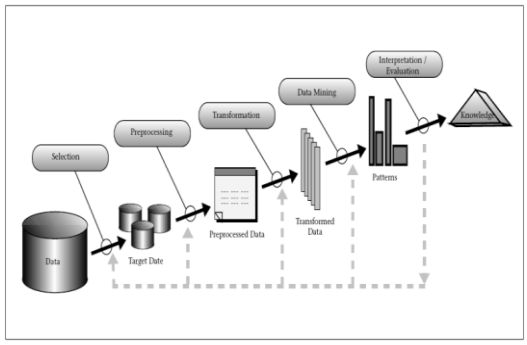

Figure 1: Knowledge Discovery Process

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses. Three steps involved are
1. Exploration
2. Pattern identification
3. Deployment
In data exploration, data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined. Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction. Patterns are deployed for desired outcome[2,6,8].
The remaining sections of the paper are organized as follows: In Section II, a brief overview of some of the Data Mining Algorithms and Techniques are presented. An introduction about the Data Mining Applications is given in Section III. Machine Learning Approaches in Section IV. Evaluation Methodologies in Section V**.** Data Mining for Health care and Knowledge Management in Health

Care is presented in Section VI,VIIand the conclusions are summed up in Section VIII.

II.DATA MINING ALGORITHMS AND TECHNIQUES

An assortment of algorithms and techniques are used for Knowledge Discovery from databases. They are Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor Method etc.[17].

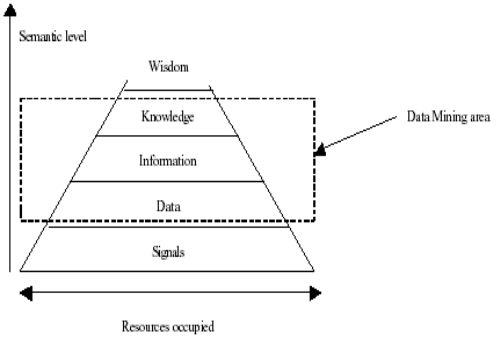

Figure: 1 Information pyramid

A. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules[12,13,17]. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis.

The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination .The algorithm then encodes these parameters into a model called a classifier[3,11,15]. The types of classification models are:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

B. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. The types of clustering methods are:

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

C. Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models[10,14]. The types of regression methods are:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

D. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. The types of association rules are:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

E. Neural Networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. Back Propagation is a type of neural networks [8,9,13].

Artificial neural networksattempt to achieve human-like performance by modeling the human nervous system. A neural network is a graph of many active nodes (neurons) that are connected with each other by weighted links (synapses). While knowledge is represented by symbolic descriptions such as decision trees and production rules in symbolic learning, knowledge is learned and remembered by a network of interconnected neurons, weighted synapses, and threshold logic units (Rumelhart et al., 1986a; Lippmann, 1987). Based on training examples, learning algorithms can be used to adjust the connection weights in the network such that it can predict or classify unknown examples correctly. Activation algorithms over the nodes can then be used to retrieve concepts and knowledge from the network (Belew, 1989; Kwok, 1989; Chen and Ng, 1995). Many different types of neural networks have been developed, among which the feedforward/backpropagation modelis the most widely used. Back propagation networks are fully connected, layered, feed-forward networks in which activations flow from the input layer through the hidden layer and then to the output layer (Rumelhart et al., 1986b). The network 10 MEDICAL INFORMATICSusually starts with a set of random weights and adjusts its weights according to each learning example. Each learning example is passed through the network to activate the nodes. The network's actual output is then compared with the target output and the error estimates are then propagated back to the hidden and input layers. The network updates its weights incrementally according to these error estimates until the network stabilizes. Other popular neural network models include Kohonen'sself-organizing mapand the Hopfield network. Self-organizing maps have been widely used in unsupervised learning, clustering, and pattern recognition (Kohonen, 1995); Hopfield networks have been used mostly in search and optimization applications (Hopfield, 1982). Due to their performances (in terms of predictive power and classification accuracy), neural networks have been widely used in experiments and adopted for critical biomedical classification and clustering problems[18].

III. DATA MINING APPLICATIONS

Data mining is a relatively new technology that has not fully matured. Despite this, there are a number of industries that are already using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions. Here is overview of business problems and solutions found using data mining technology.

A. FBTO Dutch Insurance Company

The Challenges for the company are :

- To reduce direct mail costs.
- Increase efficiency of marketing campaigns.
- Increase cross-selling to existing customers, using inbound channels such as the company's sell center and the internet a one year test of the solution's effectiveness[17].

The Outcomes for the company are:

- Provided the marketing team with the ability to predict the effectiveness of its campaigns.
- Increased the efficiency of marketing campaign creation, optimization, and execution.
- Decreased mailing costs by 35 percent.
- Increased conversion rates by 40 percent [17].

B. ECtel Ltd., Israel

The Challenges for the company are :

- Fraudulent activity in telecommunication services.

The Outcomes for the company are:

- Significantly reduced telecommunications fraud for more than 150 telecommunication companies worldwide.
- Saved money by enabling real-time fraud detection [17].

C. Provident Financial's Home credit Division, United Kingdom

The Challenges for the company are :

- No system to detect and prevent fraud.

The Outcomes for the company are:

- Reduced frequency and magnitude of agent and customer fraud.
- Saved money through early fraud detection.
- Saved investigator's time and increased prosecution rate.

D. Standard Life Mutual Financial Services Companies

The Challenges for the company are :

- Identify the key attributes of clients attracted to their mortgage offer.
- Cross sell Standard Life Bank products to the clients of other Standard Life companies.
- Develop a remortgage model which could be deployed on the group Web site to examine the profitability of the mortgage business being accepted by Standard Life Bank.

The Outcomes for the company are:

- Built a propensity model for the Standard Life Bank mortgage offer identifying key customer types that can be applied across the whole group prospect pool.
- Discovered the key drivers for purchasing a remortgage product.
- Achieved, with the model, a nine times greater response than that achieved by the control group.
- Secured £33million (approx. $47 million) worth of mortgage application revenue [17].

E. Shenandoah Life insurance company United States.

The Challenges for the company are :

- Policy approval process was paper based and cumbersome.
- Routing of these paper copies to various departments, there was delays in approval.

The Outcomes for the company are:

- Empowered management with current information on pending policies.
- Reduced the time required to issue certain policies by 20 percent.
- Improved underwriting and employee performance review processes.

F. Soft map Company Ltd., Tokyo

The Challenges for the company are :

- Customers had difficulty making hardware and software purchasing decisions, which was hindering online sales.

The Outcomes for the company are:

- Page views increased 67 percent per month after the recommendation engine went live.
- Profits tripled in 2001, as sales increased 18 percent versus the same period in the previous year[17].

IV. MACHINE LEARNING APPROACHES:

Machine learning algorithms can be classified as supervised learning or unsupervised learning. In supervised learning, training examples consist of input/output pair patterns. Learning algorithms aim to predict output values of new examples based on their input values. In unsupervised learning, training examples contain only the input patterns and no explicit target output is associated with each input. The unsupervised learning algorithms need to use the input values to discover meaningful associations or patterns. Many successful machine learning systems have been developed over the past three decades in the computer science and statistics communities. Chen and Chau (2004) categorized five major paradigms of machine learning research, namely probabilistic and statistical models, symbolic learning and rule induction, neural networks, evolution-based models, and analytic learning and fuzzy logic. We will briefly review research in each of these areas and discuss their applicability in biomedicine[13,15,16].

A. Probabilistic and Statistical Models

Probabilistic and statistical analysis techniques and models have the longest history and strongest theoretical foundation for data analysis. Although it is not rooted in artificial intelligence research, statistical analysis achieves data analysis and knowledge discovery objectives similar to machine learning. Popular statistical techniques, such as regression analysis, discriminant analysis, time series analysis, principal component analysis, and multi-dimensional scaling, are widely used in biomedical data analysis and are often considered benchmarks for comparison with other newer machine learning techniques. One of the more advanced and popular probabilistic models in biomedicine are the Bayesian model. Originating in pattern recognition research (Duda and Hart, 1973), this method was often used to classify different objects into predefined classes based on a set of features. A Bayesian model stores the probability of each class, the probability of each feature, and the probability of each feature given each class, based on the training data. When a new instance is encountered, it can be classified according to these probabilities (Langley et al., 1992). A variation of the Bayesian model, called the Naïve Bayesian model, assumes that all features are mutually independent within each class. Because of its simplicity, the Naïve Bayesian model has been adopted in different domains (Fisher, 1987; Kononenko, 1993). Due to its mathematical rigor and modeling elegance, Bayesian learning has been widely used in biomedical data mining research, in

particular, genomic and micro array analysis. machine learning technique gaining increasing recognition and popularity in recent years is the support vector machines(SVMs). SVM is based on statistical learning theory that tries to find a hyper plane to best separate two or multiple classes (Vapnik, 1998). This statistical learning model has been applied in different applications and the results have been encouraging. For example, it has been shown that SVM achieved the best performance among several learning methods in document classification (Joachims, 1998; Yang and Liu, 1999). SVM is also suitable for various biomedical classification problems, such as disease state classification based on genetic variables or medical diagnosis based on patient indicators[18].

B. Symbolic Learning and Rule Induction

Symbolic learningcan be classified according to its underlying learning strategy such as rote learning, learning by being told, learning by analogy, learning from examples, and learning from discovery (Cohen and Feigenbaum, 1982; Carbonell et al., 1983). Among these, learning from Knowledge Management, Data Mining and Text Mining 9 examples appears to be the most promising symbolic learning approach for knowledge discovery and data mining. It is implemented by applying an algorithm that attempts to induce a general concept description that best describes the different classes of the training examples. Numerous algorithms have been developed, each using one or more different techniques to identify patterns that are useful in generating a concept description. Quinlan's ID3 decision-tree building algorithm (Quinlan, 1983) and its variations such as C4.5 (Quinlan, 1993) have become one of the most widely used symbolic learning techniques. Given a set of objects, ID3 produces a decision tree that attempts to classify all the given objects correctly. At each step, the algorithm finds the attribute that best divides the objects into the different classes by minimizing entropy (information uncertainty). After all objects have been classified or all attributes have been used, the results can be represented by a decision tree or a set of production rules. Although not as powerful as SVM or neural networks (in terms of classification accuracy), symbolic learning techniques are computationally efficient and their results are easy to interpret[18]. For many biomedical applications, the ability to interpret the data mining results in a way understandable to patients, physicians, and biologists is invaluable. Powerful machine learning techniques such as SVM and neural networks often suffer because they are treated as a "black-box" [14,17,18,19].

C. Evolution-based Algorithms

Evolution-based algorithmsrely on analogies to natural processes and Darwinian survival of the fittest. Fogel (1994) identifies three categories of evolution-based algorithms: genetic algorithms, evolution strategies, and evolutionary programming. Among these, genetic algorithms are the most popular and have been successfully applied to various optimization problems. Genetic algorithms were developed based on the principle of genetics (Holland, 1975; Goldberg, 1989; Michalewicz, 1992). A population of individuals in which each individual represents a potential solution is first initiated. This population undergoes a set of genetic operations known as crossover and mutation. Crossover is a high-level process that aims at exploitation while mutation is a unary process that aims at exploration. Individuals strive for survival based on a selection scheme that is biased toward selecting fitter individuals (individuals that represent better solutions). The selected individuals form the next generation and the process continues. After some number of generations the program converges and the optimum solution is represented by the best individual. In medical informatics research, genetic algorithms are among the most robust techniques for feature selection problems (e.g., identifying a subset of genes that are most relevant to a disease state) due to their stochastic, global-search capability[18].

D. Analytic Learning and Fuzzy Logic

Analytic learningrepresents knowledge as logical rules and performs reasoning on such rules to search for proofs. Proofs can be compiled into Knowledge Management, Data Mining and Text Mining 11more complex rules to solve similar problems with a smaller number of searches required. For example, Samuelson and Rayner (1991) used analytic learning to represent grammatical rules that improve the speed of a parsing system. While traditional analytic learning systems depend on hard computing rules, there is usually no clear distinction between values and classes in the real world. To address this problem, fuzzy systems and fuzzy logichave been proposed. Fuzzy systems allow the values of False or True to operate over the range of real numbers from 0 to 1 (Zedah, 1965). Fuzziness has been applied to allow for imprecision and approximate reasoning. In general, we see little adaptation of such approaches in Healthcare systems[18].

E. Hybrid Approach

As Langley and Simon (1995) pointed out, the reasons for differentiating the paradigms are "more historical than scientific." The boundaries between the different paradigms are usually unclear and many systems have been built to combine different approaches. For example, fuzzy logic has been applied to rule induction and genetic algorithms (e.g., Mendes et al., 2001), genetic algorithms have

been combined with neural network (e.g., Maniezzo, 1994; Chen and Kim, 1994), and because neural network has a close resemblance to probabilistic model and fuzzy logic they can be easily mixed (e.g., Paass, 1990). It is not surprising to find that many practical biomedical knowledge management, data mining, and text mining systems adopt such a hybrid approach[18].

V. EVALUATION METHODOLOGIES

The accuracy of a learning system needs to be evaluated before it can become useful. Limited availability of data often makes estimating accuracy a difficult task (Kohavi, 1995). Choosing a good evaluation methodology is very important for machine learning systems development. There are several popular methods used for such evaluation, including holdout sampling, cross validation, leave-one-out, and bootstrap sampling(Stone, 1974; Efron and Tibshirani, 1993). In the holdout method, data are divided into a training set and a testing set. Usually 2/3 of the data are assigned to the training set and 1/3 to the testing set. After the system is trained by the training set data, the system predicts the output value of each instance in the testing set. These values are then compared with the real output values to determine accuracy. In cross-validation, a data set is randomly divided into a number of subsets of roughly equal size. Ten-fold cross validation, in which the data set is divided into 10 subsets, is most commonly used. The system is trained and tested for 10 iterations. In each iteration, 9 subsets of data are used as training data and the remaining set is used as testing data. In rotation, each subset of data serves as the testing set in exactly one iteration. The accuracy of the system is the average accuracy over the 10 iterations. Leave-one-out is the extreme case of cross-validation, where the original data are split into *n* subsets, where nis the size of the original data. The system is trained and tested for *n* iterations, in each of which n–1 instances are used for training and the remaining instance is used for testing. In the bootstrap method, *n*independent random samples are taken from the original data set of size *n*. Because the samples are taken with replacement, the number of unique instances will be less than *n*. These samples are then used as the training set for the learning system, and the remaining data that have not been sampled are used to test the system (Efron and Tibshirani, 1993). Each of these methods has its strengths and weaknesses. Several studies have compared them in terms of their accuracies. Hold-out sampling is the easiest to implement, but a major problem is that the training set and the testing set are not independent. This method also does not make efficient use of data since as much as 1/3 of the data are not used to train the system (Kohavi, 1995). Leave-one-out provides the most unbiased estimate, but it is computationally expensive and its estimations have very high variances, especially for small data sets (Efron, 1983; Jain et al., 1987). Breiman and Spector (1992) and Kohavi (1995) conducted independent experiments to compare the performance of several different methods, and the results of both experiments showed ten-fold cross validation to be the best method for model selection. In light of the significant medical and patient consequences associated with many biomedical data mining applications, it is critical that a systematic validation method be adopted. In addition, a detailed, qualitative validation of the data mining or text mining results needs to be conducted with the help of domain experts (e.g., physicians and biologists), and therefore this is generally a time-consuming and costly process[17,18,19].

VI. DATA MININGFOR HEALTH CARE

Because of their predictive power, data mining techniques have been widely used in diagnostic and health care applications. Data mining algorithms can learn from past examples in clinical data and model the oftentimes non-linear relationships between the independent and dependent variables. The resulting model represents formalized knowledge, which can often provide a good diagnostic opinion. Classificationis the most widely used technique in medical data mining. Dreiseitl et al. (2001) compare five classification algorithms for the diagnosis of pigmented skin lesions. Their results show that logistic regression, artificial neural networks, and support vector machines performed comparably, while k-nearest neighbors and decision trees performed worse [6,8,10]. This is more or less consistent with the performances of these classification algorithms in other applications (e.g., Yang and Liu, 1999). Classification techniques are also applied to analyze various *signals* and their relationships with particular diseases or symptoms. For example, Acir and Guzelis (2004) apply support vector machines in automatic spike signal detection in ElectroEncephaloGrams (EEG), which can be used in diagnosing neurological disorders related to epilepsy. Kandaswamy et al. (2004) use artificial neural network to classify lung sound signals into six different categories (e.g., normal, wheeze, and rhonchus) to assist diagnosis. Data mining is also used to extract rules from health care data. For example, it has been used to extract diagnostic rules from breast cancer data (Kovalerchuk et al., 2001). The rules generated are similar to those created manually in expert systems and therefore can be easily validated by domain experts. Data mining has also been applied to clinical databases

to identify new medical knowledge (Prather et al., 1997; Hripcsak et al., 2002).

VII. KNOWLEDGE MANAGEMENT IN HEALTH CARE

It has been generally recognized that patient record management systemsis highly desired in clinical settings (Heathfield and Louw, 1999; Jackson,2000; Abidi, 2001). The major reasons include physicians' significantinformation needs (Dawes and Sampson, 2003) and clinical informationoverload. Hersh (1996) classified textual health information into two maincategories: patient-specific clinical information and knowledge-basedinformation, which includes research reported in academic journals, books,technical reports, and other sources. Both types of information are growingat an overwhelming pace.Although early clinical systems were mostly simple data storage systems,knowledge management capabilities have been incorporated in many ofthem since the 1980s. For example, the *HELP* system, developed at theLatter Day Saints Hospital in Utah, provides a monitoring program on top ofa traditional medical record system. Decision logic was stored in the systemto allow it to respond to new data entered (Kuperman et al., 1991). The*SAPHIRE* system performs automatic indexing of radiology reports byutilizing the UMLS Metathesaurus (Hersh et al., 2002). The clinical datarepository at Columbia-Presbyterian Medical Center (Friedman et al., 1990)is another example of a database that is used for decision support (Hripcsak,1993) as well as well as physician review. The clinical data repository at theUniversity of Virginia Health System is another example (Schubart andEinbinder, 2000). In their data warehouse system, clinical, administrative,and other patient data are available to users through a Web browser. Casebasedreasoning also has been proposed to allow physicians to access bothoperative knowledge and medical literature based on their medicalinformation needs (Montani and Bellazzi, 2002). Janetzki et al. (2004) use a
natural language processing approach to link electronic health records toonline information resources. Other advanced text mining techniques alsohave been applied to knowledge management in health care and Medical Systems [18,19].

VIII. CONCLUSIONS

Thehistorical progressofmachine learningand itsapplicationsinmedicaldiagnosisshowsthatfroms impleandstraight forward to use algorithms,systemsandmethodologyhaveemergedt hatenableadvancedandsophisticateddata analysis. Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge in different domains. Data mining techniques and algorithms such as classification, clustering helps in finding the patterns to decide upon the future businesses trends to grow. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information
Technology. In this paper we discussed a broad overview of some of the data mining techniques, their use in various emerging algorithms and applications and also the protagonist of these in Health and Medical Systems.

REFERENCES

[1].BaimP.W., AMethod forAttributeSelectionInductive Learning Systems, *IEEE Trans. on PAMI*, Vol.10,No.6,1988,pp.888-896.

[2].BevkM., KononenkoI., ZrimecT., Relation between energetic diagnoses and GDVimages, *Proc. NewScience ofConsciousness:3rdIntConf. onCognitive Science*, Ljubljana,October 2000,pp. 54-57.

[3].BratkoI.,MozetičI.,LavračN.,*KARDIO:A study indeepandqualitative knowledgeforexpert systems*, Cambridge,MA: MITPress, 1989.

[4].BratkoI.,MulecP., AnExperimentinAutomaticLearning ofDiagnostic Rules, *Informatica*,Ljubljana, Vol.4,No.4,1980,pp. 18-25.

[5].BreimanL.,FriedmanJ.H., OlshenR.A.,StoneC.J. (1984)*Classification andRegression Trees*,WadsforthInternationalGroup.

[6[.CatlettJ., Onchanging continuous attributesinto ordered discrete attributes,*Proc. European WorkingSession onLearning-91*, Porto,March4-6,1991,pp. 164-178.

[7].CestnikB., EstimatingProbabilities: ACrucial Task in Machine Learning, *Proc. European Conf. onArtificial Intelligence*, Stockholm, August, 1990,pp. 147-149.

[8].CestnikB.,KononenkoI.&BratkoI.,ASSISTANT 86:Aknowledgeelicitationtoolforsophisticatedusers, in:I.Bratko, N.Lavrac(eds.): *ProgressinMachine learning*,Wilmslow: SigmaPress, 1987.

[9].Chan K.C.C. &Wong A.K.C., AutomaticConstructionofExpert Systems fromData: AStatistical Approach, *Proc. IJCAIWorkshop onKnowledgeDiscoveryinDatabases*,Detroit,Michigan,August, 1989,pp.37-48.

[10].Clark P. &BoswellR., RuleInductionwith CN2: SomeRecentImprovements,*Proc. European WorkingSession onLearning-91*, Porto,Portugal,March, 1991,pp.151-163.

[11].CravenM.W.andShavlikJ.W., Learningsymbolicrulesusingartificialneuralnetworks, *Proc. 10thIntern. Conf.onMachine Learning*, Amherst, MA,MorganKaufmann,1993,pp.73-80.

[12].DiamondG.A.andForesterJ.S.,Analysisofprobabilityasanaidintheclinicaldiagnosisofcoronaryartery disease,*NewEnglandJ.ofMedicine*, 300:1350,1979.

[13].ElomaaT.,HolstiN.,AnExperimentalComparison ofInducing DecisionTreesandDecisionListsinNoisyDomains, *Proc. 4thEuropean WorkingSession onLearning*, Montpeiller, Dec. 4-6,1989,pp.59-69.

[14].GoodI.J., *ProbabilityandtheEeighingofEvidence*. London: Charles Griffin,1950.GoodI.J., *TheEstimationofProbabilities.*

[15]. Jiawei Han and MichelineKamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd edition.

[16].Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

[17].Bharati M. Ramageri / Indian Journal of Computer Science and Engineering,Vol. 1 No. 4 301-305.

[18].”Knowledge Management, Data Mining and Text Mining in Medical Informatics”Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and WilliamHersh, Medical Informatics and Clinical Epidemiology, Portland, Oregon 97239-3098.
[19].Abidi, S. S. R. (2001). “Knowledge Management in Healthcare:Towards ‘Knowledgedriven’Decision support Services,” *International Journal of Medical Informatics*, 63, 5-18.

ABOUT THE AUTHORS

N.Satyanandamis working as an Associate Professor in the Department of CSE, Bhoj Reddy Engineering College for Women, Hyderabad, India. He received B.Tech(CSSE) in 1996 and MBA (MM) in 1999 both from Andhra University and M.Tech(CSE) in 2004 from Jawaharlal Nehru Technological University, Hyderabad. At present he is a Research Scholar of JNTUH, Hyderabad, India. His main research interests are Data Mining and Warehousing, Digital Image Processing, Computer Networks, Software Engineering and Natural Language Processing. He is a Member of ISTE.

Dr. Ch. Satyanarayana is working as an Associate Professor in the Department of CSE, JNTUK,Kakinada,India. He received B.Tech(CSE) in 1996 and M.Tech(CST) in 1998 both from Andhra University. He has been working in Jawaharlal Nehru Technological University for the last 12 years. He has published 26 research papers in various International Conferences and Journals. His main research interests are Pattern Recognition, Image Processing, Speech Processing, Computer Graphics, Data Mining and Warehousing and Compiler Writing. He is a member of different technical bodies like ISTE, IETE and CSI.

Md. Riyazuddinis working as an Assistant Professor , Department of Information Technology at Muffakhamjah College of Engineering and Technology (MJCET), Banjarahills, Hyderabad, India. He has received B.Tech (IT) from Kakatiya University, Warangal and M.Tech (CSE) from JNTUH Hyderabad. He has been published 2 Research Papers in International Journals. His main research interests are Data Mining, Cloud Computing, Software Engineering, Natural Language Processing andComputer Networks.

Amjan.Shaikis working as a Professor in the Department of Computer Science and Engineering, Ellenki College of Engineering and Technology(ECET),Hyderabad, India. He has received M.Tech.(Computer Science and Technology) from Andhra University and PGDM from MKU . He has been published and presented good number of Research and Technical papers in International Journals , International Conferences and National Conferences. His main research interests are Software Metrics, Software Engineering, Software Testing, Software Quality, Object Oriented Design and NLP. He is a member of ISTE and CSI.