International Journal of Computing Academic Research (IJCAR)
ISSN 2305-9184 Volume 2, Number 4 (August 2013), pp. 155-166
© MEACSE Publications
http://www.meacse.org/ijcar

Technical Analysis of the Learning Algorithms in Data Mining Context

Hisham S, Katoua

Faculty of Economics & Administration
Management Information Systems Dept.
King Abdulaziz University ,Jeddah, Kingdom of Saudi Arabia

Abstract

In this paper, the various learning algorithms that used in the knowledge discovery process as data mining techniques, have been discussed, e.g. visualization techniques, query tools, OLAP tools, K-nearest neighbor, decision trees, association rules, neural networks, genetic algorithms, fuzzy sets. Then a comparison between some of them have been done in two dimension to represent the advantage and disadvantage of this learning algorithms depending on number of qualities as a first dimension (quality of input, output, and performance) and some of the learning algorithms as a second dimension.

Keywords: Knowledge Engineering, Learning Algorithms, Intelligent Decision Making, Intelligent Information Systems, Databases Science, Data Mining.

1. Introduction

Data mining deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases. It is currently regarded as the key element of a much more elaborate process called knowledge discovery in databases (KDD), which is closely linked to another important development - data warehousing. A data warehouse is a central store of data that has been extracted from operational data. The information in a data warehouse is subject-oriented, non-volatile, and of an historic nature, so data warehouses tend to contain extremely large data sets. The combination of data warehousing, decision support, and data mining indicates an innovative and totally new approach to information management. Until now, information systems have been built and operated mainly to support the operational processes of an organization. KDD and data warehousing view the information in an organization in an entirely new way - as a strategic source of opportunity [7].

One of the major challenges facing many scientists today involves the organization and analysis of the phenomenal explosion of data that has been provided by recent computer and data-collection technology. Since data are collected and stored at a very large acceleration these days, there has

become an urgent need for a new generation of robust software packages to extract useful information or knowledge from large volume of data. Researchers in many fields are looking for automated systems to help them with the management and analysis of their rapidly growing databases. These automated systems, tools and packages are the subject of a new discipline called knowledge discovery in databases [1].

This paper is organized as follow, section one serves as introduction to data mining. Section two introduce data mining techniques e.g. association rule, OLAP - online analytical processing, case-based learning (k-nearest neighbor), visualization , genetic algorithms, decision trees, statistical techniques, query tools and neural network. Section three explores data mining tasks e.g. classification, clustering, summarization and regression. Section four explain clustering task of data mining. Section five discusses aspects of the evaluation of machine-learning algorithms. Section six ends up with conclusions.

2. Data Mining Techniques

Various different techniques are used for different data-mining tasks, in the following those that are of interest:

1- Association Rules

Association rules play a major role in determining how two items are similar in a database. An example of this is the market basket analysis: is it likely that a customer buying milk from a grocery store will also buy eggs? Mining for association rules extracts different sets of items that are predictors of each other that is with probability p a user who buys A will buy B for all B in the set. Often, these associations can help generating new sales. For example, if a customer who buys eggs also buys milk, then to promote the sales of milk, the storeowner may consider putting eggs on sale because this might persuade the customer to also buy milk.

2- OLAP - Online Analytical Processing

OLAP queries involve taking a certain set of records, and performing an aggregation computation on that set. Finding the total sales by region over different time periods is an example of an OLAP query. The problem with performing OLAP queries is the magnitude of the data. Records may be stored in different database tables and joining these tables can be an expensive operation if it must be done for many subset possibilities over groups.

3- Case-Based Learning (k-nearest neighbor)

Nearest neighbor method is a technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique. In case-based reasoning (CBR) system expertise is embodied in a library of past cases, rather than being encoded in classical rules. Each case typically contains a description of the problem, plus a solution and/or the outcome. The problem description states the world at the time the case was happening. The outcome is the resulting state of the world when the solution was carried out. The knowledge and reasoning process used by an expert to solve the problem is not recorded.

4- Visualization

Visualization techniques are a very useful method for discovering patterns in data sets, and may be used at the beginning of a data mining process making it possible for the analyst to gain deeper, intuitive understanding of the data by presenting a picture for users. For example, a graphical image presenting four variables would present a large amount of information in a concise manner, which will make groups of data versioned as peaks or valleys.

These simple methods can provide us with a wealth of information, an elementary technique that can be of great value is the so-called scatter diagram; in this technique, information on two attributes is displayed in a Cartesian space. Scatter diagrams can be used to identify interesting sub sets of the data sets so that we can focus on the rest of the data mining process.

There is a whole field of research dedicated to the search for interesting projections of data sets, this is called projection pursuit. There are other reasons to conceive records as points in a multidimensional data space. The space metaphor is very useful in a data mining context. Using this metaphor we can determine the distance between two records in this data space: records that are close to each other are very alike, and records that are very far removed from each other represent individuals that have little in common.

5- Genetic Algorithms

Genetic algorithms are random search algorithm that imitates natural evolution with Darwinian survival of the fittest approach. Genetic algorithms are optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution [5].

6- Decision Trees

Decision trees have a tree organization with intermediate nodes labeled by split attributes, the branches starting in an intermediate node labeled by split predicates involving the corresponding split predicate and leaves labeled by class labels. Prediction using decision trees is made by navigating the tree on true predicates until a leaf is reached, when the corresponding label is returned. The construction of classification trees proceeds in two phases. In the growing phase an oversized tree is build recursively, at every step a split attribute and split predicates involving this attribute are selected in order to maximize the goodness of split criteria. A large number of such criteria have been proposed in the literature [6].

7- Statistical Techniques

Statistical methods are used in many data mining problems, particularly in the areas of model building and pattern detection that capture key relationships between variables in the data. *Statistical techniques* can be used in several data mining stages: data cleaning ,data selection and data analysis. The statistical techniques of data mining are familiar. They include linear and logistic regression, multivariate analysis, principal components analysis and others.

8- Query Tools

The first step in a data-mining project should always be a rough analysis of the data set using traditional query tools. Just by applying simple structured language (SQL) to a data set, you can obtain a wealth of information. However, before we can apply more advanced pattern analysis algorithm, we need to know some basic aspects and structure of the data set .With SQL we can uncover only shallow data, which is information that is easily accessible from the data set, yet although we can find hidden [7].

9- Neural Networks

Artificial neural networks (NN) are non-linear predictive models that learn through training and resemble biological neural networks in structure. These are collections of connected nodes with inputs, outputs and processing at each node. A number of hidden processing layers exist between the visible input and output layers. The neural model has to train the net on a training dataset and then use it to make predictions. Neural nets typically cannot be trained on very large databases, but, with suitable sampling methods, the net can produce reasonable accuracy on small and medium sized data sets. The problem with neural networks is that no explanation of the results is provided (black box operation). This inhibits confidence, acceptance and application of results. However, there are some proprietary neural net products, which can translate the neural model into a set of understandable rules [4].

Various learning NN-based algorithms and their associated network architectures are summarized in Table 1. It is noticed that both the supervised and unsupervised learning paradigms employ learning rules based on error-correction, and competitive learning.

3. Data Mining Tasks

Data mining is supported by hosting models or tasks that capture the characteristics of data in several different ways such as: clustering model, regression model, classification model and summarization model.

1- Clustering

Clustering segments a database into subsets or clusters. Clustering is a data-mining task in which records are grouped based on similarities in their attributes. Often it is not known which attribute or attributes will group the records into the tightest groups so clustering algorithms must try many attributes combinations. The key objective of clustering is to find natural groupings (clusters) in highly dimensional data. Closely related to clustering is the method of probability density estimation, which consists of techniques for estimating from data the joint multi-variant probability density function of all of the variables/fields in the database. Clustering is an example of unsupervised learning, and it is a part of pattern recognition.

Table 1: Well-known NN-based learning algorithms

Learning paradigm	Learning rule	Architecture	Learning algorithm	Task
Supervised	Error-correction	Single-or multilayer perceptron	Perceptron learning algorithms Backprobagation	Pattern classification unction approximation control
	Boltzmann	Recurrent	Boltzmann learning algorithm	Pattern classification
	Hebbian	Multilayer feed forward	Linear Discriminant Analysis	Data analysis Pattern classification
	Competitive	Competitive	Learning Vector Quantization	Within-class categorization Data compression
		ART network	ARTMAP	Pattern classification Within-class categorization
Unsupervised	Error-correction	Multilayer feed forward	Sammon's projection	Data analysis
	Hebbian	Feed forward or competitive	ftmcipleCpniponent Analysis	Data analysis Da& compression
		Hopfield net	Associative memory learning	Associative memory
	Competitive	Competitive	Vector Quantization	Categorization Data compression
		Kohonen SOM	Kohonen's SOM	Categorization Data analysis
		ART networks	ART1,ART2	Categorization
Hybrid	Error-correction and competitive	RBF network	RBF learning algorithm	Pattern classification Function approximation control

2- Classifications

Classification concerns with learning that classifies data into one of several predetermined classes. If the data that is being analyzed is comprised of various attributes and a class label, then we may be able to build a classification model to help predicting the class labels of future records.

3- Regressions

Regression models or regressors are functional mappings from the cross product of the domains of predictor attributes $X_1 \ldots\ldots X_m$ to the domain of the continuous predicted attribute, Y. They only differ from classifiers in the fact that the predicted attribute is real valued.

4- Summarization

In summarization model a compact description for a subset of data is found, e.g., the derivation of summary or association rules and the use of multivariate visualization techniques.

5-Clustering

Clustering is usually viewed as a process of grouping physical or abstract objects into classes of similar objects. According to this view, in order to cluster objects, one needs to define a measure of similarity between the objects and then apply it to determine classes. Classes are denned as collections of objects whose intraclass similarity is high and interclass similarity is low. Because the notion of similarity between objects is fundamental to this view, clustering methods based on it can be tailed similarity-based methods. Many such methods have been developed in numerical taxonomy, a field developed by social and natural scientists, and in cluster analysis, a subfield of pattern recognition (qv).

Another view recently developed in AI postulates that objects should be grouped together not just because they are similar awarding to A given measure but because as a group they represent a certain conceptual class. This view, called conceptual clustering, states that clustering depends on the goals of classification and the concept available to the clustering system for characterizing collections of entities. For example, if the goal is to partition a configuration of points into simple visual groupings, one may partition them into those that form a T-shape, an L-shape, and so on, even though the density distributions and distances between the points may suggest different groupings. A procedure that uses only similarities (or distances) between the points and is unaware of these simple shape types clearly can only accidentally create clustering corresponding to these concepts. To create such clustering, these descriptive concepts must be known to the system. Another example of conceptual clustering is the grouping of visible stars into named constellations.

Clustering is the basis for building hierarchical classification schemes. For example, by first partitioning the original set of entities and then repeatedly applying a clustering algorithm to the classes generated at the previous step, one can obtain a hierarchical classification of the entities (a divisive strategy). A classification schema is obtained by determining the general characteristics of the classes generated.

Building classification schemes and using them to classify objects is a widely practiced intellectual process in science as well as in ordinary life. Understanding this process, and the mechanisms of clustering underlying it is therefore an important domain of research in AI and other areas. This process can be viewed as a cousin of the "divide and conquer" strategy widely used in problem solving (qv). It is also related to the task of decomposing any large-scale engineering system into smaller subsystems in order to simplify its design and implementation.

4. Data Mining Uses and Applications

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. This section presents a brief account of the well known DM uses and applications (see table 2).

Table 2: Data Mining Applications in Private Sector

Application domain	Task
Insurance and banking industries	detect fraud and assist in risk assessment (e.g., credit scoring)
Medical community	predict the effectiveness of a procedure or medicine
Pharmaceutical firms	use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases.
Retailers	assess the effectiveness of product selection and placement decisions, coupon offers, and which products are often purchased together.
Telephone service providers	to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor

DM approaches have grown also to be used for purposes such as measuring and improving program performance. It has been reported that data mining has been able to assess crime patterns and adjust resource allotments. In addition, data mining can be used to predict demographic changes in the constituency it serves so that it can better estimate its budgetary needs. Data mining can be used also to review plane crash data to recognize common defects and recommend precautionary measures.

Recently, data mining has been increasingly cited as an important tool for homeland security efforts. Some observers suggest that data mining should be used as a means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records.

Based on the analysis of published papers during the five years, figure 1 shows the emerging data mining research areas.

• Graph mining **• Data mining in bioinformatics** **• Privacy-aware data mining** **• Large scale data mining** **• Temporal pattern mining** **• Stream data mining** **Mining moving object data, RFID data, and data from sensor networks** **• Ubiquitous knowledge discovery**	**• Mining multi-agent data** **• Mining and link analysis in networked settings: web, social and computer networks, and online communities** **• Mining the semantic web** **• Data mining in electronic commerce** **• Data mining in e-Learning** **•Web search, advertising, and marketing task**

Fig. 1 emerging applications of data mining techniques

5. Aspects of Evaluation of Machine-Learning /Data Mining Techniques and Algorithms

Based on our analysis of the recent published results, table 3 shows the machine learning/data mining techniques and the appropriate mining tasks. We see for instance, that neural networks are somewhat better at classification tasks whereas genetic algorithm perform better at problem solving tasks. Inductive logic programming has a high score in the knowledge engineering area. One fact can be learned from this that there is no single best machine learning technique: different tasks pre-suppose different kinds of techniques. A knowledge environment discovery therefore must support these different types of techniques, and such environment is called hybrid. In some cases, k-nearest neighbor as a technique does well; in other cases decision trees give a better analysis of the data set. This illustrates the validity of a multiple strategy approach to data mining.

Table 3: Data Mining Tasks and Techniques

Data Mining Task	**The Appropriate Data Mining Technique/Algorithm**
Classification	Neural Networks Support Vector Machine Decision Trees Genetic Algorithms Rule induction
Clustering	K-Means
Regression and prediction	Support Vector Machine Decision Trees

	Rule induction, NN
Association and Link Analysis (finding correlation between items in a dataset)	Association Rule Mining
Summarization	Multivariate Visualization

In Fig 2 a concise overview of the advantages and disadvantages of various learning algorithms in the data-mining context is given [7]. In one dimension of the figure there are main five algorithms areas. The other dimension represents a number of relevant qualities of data mining algorithms ordered into three sub-groups: quality of input, quality of output and performance. These sub-groups represent the main area of attention when selecting a data-mining algorithm. An important aspect of the input is the number of records, since some algorithms are better at handling large number of records than others. Another important aspect is the number of attributes: the performance of neural networks and genetic algorithms deteriorates considerably as the number of attributes grows. Last but not least, the types of attribute play an important role. Not all algorithms can deal with numeric attributes or strings, so this may be a decisive element in the choice of the optimal algorithm.

Another approach to the selection of a learning algorithm might be an analysis of the quality of the desired output. One key issue here is where or not the algorithm is capable of learning rules. Some algorithm, such as k-nearest neighbor and neural networks, give a yes or no answer but provide no explanation of their responses. In other cases, it might be important to select a machine-learning algorithm that is able to learn incrementally, when new information become available, these kinds of algorithm are able to revise their theories, which means that we do not have to start the learning process all over again, and this can be of great relevance in situations where we have extremely large data sets. A final point of interest in the evaluation of the output is the ability to estimate the statistical significance of the results that are found. In the case of genetic algorithms and neural networks, it is often very difficult to evaluate the results from a statistical point of view, although for certain organizations, this might prove to be a decisive factor when selecting the algorithms they want to work with. A final aspect of the evaluation of machine-learning algorithms is their general performance. We are interested in the efficiency of our algorithms in two different situations: (a) the learning stage and (b) the actual application stage of the algorithm.

On average, algorithms that learn quickly are somewhat slow in the application stage, and vice versa. In Figure 2 we give estimates for disk load as well as CPU load in both the learning and the application stage. Note that an algorithm is considered to be good if both the disk load and the CPU load are low. At a quick glance, we can immediately see why decision trees and association rules are so popular in data mining applications – on average; they score high on input/output as well as performance issues. Of course, an overview like the one in Figure 2 has to be handled with care as it gives only a rough impression of the weak and strong areas of the algorithm. There are ways of rectifying the disadvantages of all the individual algorithms, such as speeding up nearest neighbor in the application phase, and there are methods of applying genetic algorithms to numeric areas, and so on. In general, however, Figure 2 provides a good starting point

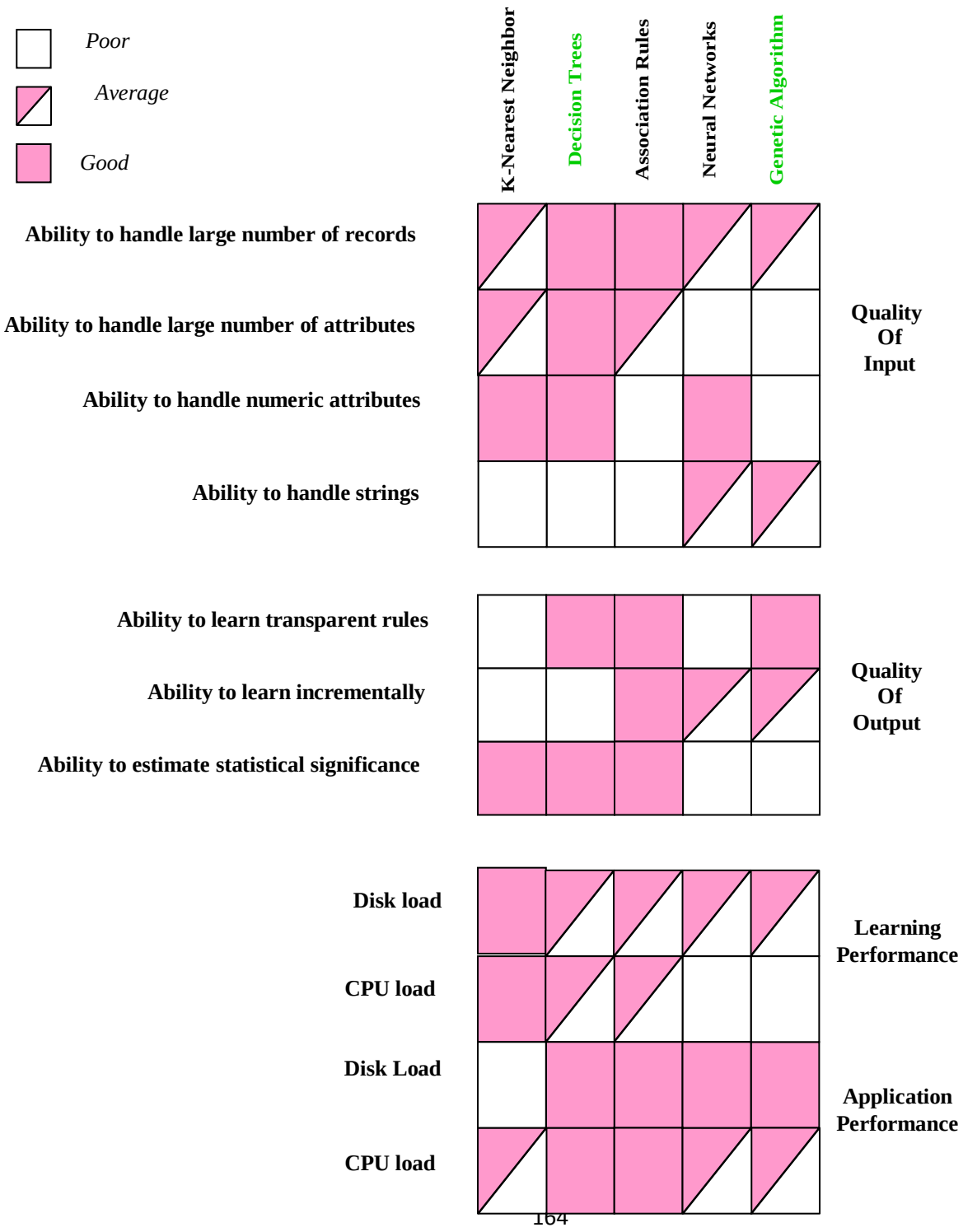

Figure 2 Data mining algorithm selection

6. Summary and Conclusion

In this paper we have presented the learning algorithms and tasks of the data mining technology. Data mining methodology states that in the optimal situation, data mining is an ongoing process. Organizations should continually work on their data, constantly identifying new information needs and trying to improve the data to make it match the goals better. In this way any organization will become a learning system.

Data mining is not so much a single technique as the idea that there is more knowledge hidden in the data than shows itself on the surface. Any technique that helps extract more out of the data is useful, so data mining techniques form a heterogonous group. The paper presents and discusses the varies different techniques which are used for different purposes, e.g. query tools, statistical techniques, visualization, online analytical processing, case-based learning, decision trees, association rules, neural networks and genetic algorithms. Although the paper discusses the main aspects of the evaluation of the data mining algorithm which are , (a) quality of the input, (b) quality of the output and (c) performance. The number of recodes is an important aspect for the input, since some algorithms are better at handling large number of records than others. Not all algorithms are good at handling numeric attributes or strings, so this may be a decisive element in the choice of optimal algorithm. Concerning the second aspect, some algorithms, e.g. k-nearest neighbor and neural network, give a yes or no answer but provide no explanation of their responses. Another point of interest in the evolution of the output is the ability to estimate the statistical significance of the results that are found. The third aspect of the evaluation is the, performance of the algorithm, depends on the learning stage and the actual application stage of the algorithm.

References

1. Moustakis, VS., Lehto M. and Salvendy, G. , Survey of expert opinion: which machine learning method may be used for which task? Special issue on machine learning of International Journal of HCI, 1996.
2. Fayyad U.M., Piatetsky-Shapiro G., Smyth P and Uthurusamy R. .Advances in Knowledge Discovery and Data Mining. Cambridge MA : AAAI Press/MIT Press, 1996.
3. Agrawal R., Mannila H., Srikant R., Toivonen H. and Verkamo A. Fast discovery association rules. In Advances in Knowledge Discovery and Data Mining. Cambridge MA: AAAI Press/ MIT Press.
4. Chester M. . Neural Networks: A Tutorial. Englewood Cliffs NJ: Prentice-Hall, 1993.
5. Booker L.B., Goldberg D.E. and Holland J.H. . Classifier Systems and Genetic Algorithms, 1989.
6. Clark P. and Niblett T. . The CN2 induction algorithm. Machine Learning, 3, 261-83, 1996.
7. Adriaans P. and Zantinge D., " Data Mining", Addison-Wesley,1996.
8. Agarwal, R., & Srikant, R. Mining sequential patterns. In Proceedings of the eleventh international conference on data engineering, Taipei, Taiwan (pp. 3–14), 2005.

9. Aleksandar Lazarevic, Levent ErtÄoz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network in-trusion detection. In *SIAM Conference on Data Mining (SDM)*, 2003.
10. Anoop Singhal and Sushil Jajodia. Data mining for intrusion detection. In *Data Mining and Knowledge Discovery Handbook*, pages 1225{1237. Springer, 2005.
11. Chen, H., Chung, W., Xu Jennifer, J., Wang, G., Qin, Y., Chau, M., "Crime Data Mining: A General Framework and Some Examples". Technical Report, Published by the IEEE Computer Society, 0018-9162/04, pp 50-56, April 2004.
12. Feldman, R., & Sanger, J. The text mining handbook. Cambridge University Press 2006.
13. Gyorgy Simon, Hui Xiong, Eric Eilertson, and Vipin Kumar. Scan detection: A data mining approach. Technical Report AHPCRC 038, University of Minnesota, Twin Cities, 2005.
14. Gyorgy Simon, Hui Xiong, Eric Eilertson, and Vipin Kumar. Scan detection: A data mining approach. In *Proceedings of SIAM Conference on Data Mining (SDM)*, 2006.
15. Jadhav, S. R., and Kumbargoudar, P., "Multimedia Data Mining in Digital Libraries: Standards and Features READIT, pp 54-59, 2007.
16. Kirkos, E., Spathis, C., and Manolopoulos., Y., "Data Mining techniques for the detection of fraudulent financial statements." Expert Systems with Applications 32(4), 995-1003, 2007.
17. Smith, L., Lipscomb, B., and Simkins, A., "Data Mining in Sports: Predicting Cy Young Award Winners". Journal of Computer Science, Vol. 22, Page No. 115-121,April 2007.
18. Spence, R. Information visualization. Addison-Wesley 2001.
19. Vipin Kumar, Jaideep Srivastava, and Aleksander Lazarevic, editors. *Managing Cyber Threats{Issues, Approaches and Challenges*. Springer Verlag, May 2005.
20. Varun Chandola, Eric Eilertson, Levent ErtÄoz, GyÄorgy Simon and Vipin Kumar, Data Mining for Cyber Security, Book chapter in *Data Warehousing and Data Mining Techniques for Computer Security*, Springer, 2006.