A Survey on Application of Machine Learning Algorithms on Data Mining

Tapas Ranjan Baitharu, Subhendu Kumar Pani

Abstract: In the context of data mining the feature size is very large and it is believed that it needs a bigger population. Hence, this translates directly into higher computational load.Data and information have become major assets for most of the organizations. The success of any organisation depends largely on the extent to which the data acquired from business operations is utilized. classification is an important task in KDD (knowledge discovery in databases) process. It has several potential applications. The primary objective of this paper is to review the data mining and study of machine learning algorithm .The performance of classifiers is strongly dependent on the data set used for learning..

Keywords: Data Mining, Machine learning Algorithm, Knowledge Discovery Databases.

I. INTRODUCTION.

1.1 Data mining .

In different kinds of information databases, such as scientific data, medical data, financial data, and marketing transaction data; analysis and finding critical hidden information has been a focused area for researchers of data mining[1][2][4]. How to effectively analyze and apply these data and find the critical hidden information from these databases, data mining technique has been the most widely discussed and frequently applied tool from recent decades. Although the data mining has been successfully applied in the areas of scientific analysis, business application, and medical research and its computational efficiency and accuracy are also improving, still manual works are required to complete the process of extraction.

Data mining is considered to be an emerging technology that has made revolutionary change in the information world. The term `data mining' (often called as knowledge discovery) refers to the process of analysing data from different perspectives and summarizing it into useful information by means of a number of analytical tools and techniques, which in turn may be useful to increase the performance of a system[3].

Technically, "data mining is the process of finding correlations or patterns among dozens of fields in large relational databases". Therefore, data mining consists of major functional elements that transform data onto data warehouse, manage data in a multidimensional database, facilitates data access to information professionals or analysts, analyse data using application tools and techniques, and meaningfully presents data to provide useful information.

Manuscript Received December, 2013.

Tapas Ranjan Baitharu, Associate Prof. Dept. of CSE, Orissa Engineering College, Odisha, India.

Subhendu Kumar Pani, Associate Prof. Dept. of CSE, Orissa Engineering College, Odisha, India.

1.2 Data Mining Process

Data Mining is an iterative process consists of the following list of stages:

i. Data cleaning
ii. Data integration
iii. Data selection
iv. Data transformation
v. Data mining
vi. Pattern evaluation
vii. Knowledge presentation

Data cleaning: This task handles missing and redundant data in the source file. The real world data can be incomplete, inconsistent and corrupted. In this process, missing values can be filled or removed, noise values are smoothed, outliers are identified and each of these deficiencies are handled by different techniques.

Data integration: Data integration process combines data from various sources. The source data can be multiple distinct databases having different data definitions. In this case, data integration process inserts data into a single coherent data store from these multiple data sources.

In the data selection process, the relevant data from data source are retrieved for data mining purposes.

Data transformation: This process converts source data into proper format for data mining. Data transformation includes basic data management tasks such as smoothing, aggregation, generalization, normalization and attributes construction.

Data mining: In Data mining process, intelligent methods are applied in order to extract data patterns. Pattern evaluation is the task of discovering interesting patterns among extracted pattern set. Knowledge representation includes visualization techniques, which are used to interpret discovered knowledge to the user.

Pattern Evaluation: During data mining, a large number of patterns may be discovered. However, all those patterns may not be useful in a particular context. It is highly required to assess the usefulness of the discovered patterns based on some criteria, so that truly useful and interesting patterns representing knowledge can be identified.

Knowledge Presentation: Finally, the mined knowledge has to be presented to the decision-maker using suitable techniques of knowledge representation and visualization.

II. TECHNIQUES AND ALGORITHMS

Researchers identify two fundamental goals of data mining: prediction and description. Prediction makes use of existing variables in the database in order to predict unknown or future values of interest, while description focuses on finding patterns describing the data the subsequent presentation for user interpretation.

The relative emphasis of both prediction and description differs with respect to the underlying application and technique. There are several data mining techniques fulfilling these objectives. Some of these are classification, clustering, association and pattern discovery.

- 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. 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. Some well-known classification models are:

a) Classification by decision tree induction
b) Bayesian Classification
c) Neural Networks
d) Support Vector Machines (SVM)

- Clustering: Clustering is a technique for 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. Some commonly used clustering methods are:

a) Partitioning Methods
b) Hierarchical Agglomerative (divisive) methods
c) Density based methods
d) Grid-based methods
e) Model-based methods

- 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. Some association rule types are:

a) Multilevel association rule
b) Multidimensional association rule
c) Quantitative association rule

III. MACHINE LEARNING ALGORITHMS

We select five commonly used classifiers for prediction classification in data mining qualitative performance. These classifiers are described in this section and their WEKA names are given in Table-3.1.

- K-Nearest Neighbour: This classifier is considered as a statistical learning algorithm and it is extremely simple to implement and leaves itself open to a wide variety of variations. In brief, the training portion of nearest-neighbour does little more than store the data points presented to it. When asked to make a prediction about an unknown point, the nearest-neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training point according to some distance metric. The distance metric used in nearest neighbour methods for numerical attributes can be simple Euclidean distance.
- Decision Tree: A decision tree partitions the input space of a dataset into mutually exclusive regions, each of which is assigned a label, a value or an action to characterize its data points. The decision tree mechanism is transparent and we can follow a tree structure easily to see how the decision is made. A decision tree is a tree structure consisting of internal and external nodes connected by branches. An internal node is a decision making unit that evaluates a decision function to determine which child node to visit next. The external node, on the other hand, has no child nodes and is associated with a label or value that characterizes the given data that leads to its being visited. However, many decision tree construction algorithms involve a two - step process. First, a very large decision tree is grown. Then, to reduce large size and over-fitting the data, in the second step, the given tree is pruned. The pruned decision tree that is used for classification purposes is called the classification tree. A popular decision tree algorithm is C4.5. It can help not only to make accurate predictions from the data but also to explain the patterns in it. It deals with the problems of the numeric attributes, missing values, pruning, estimating error rates, complexity of decision tree induction, and generating rules from trees [18]. In terms of predictive accuracy, C4.5 performs slightly better than CART and ID3 [17]. C4.5's successor, C5.0, shows marginal improvements to decision tree induction but not enough to justify its use. The learning and classification steps of C4.5 are generally fast [19]. However, scalability and efficiency problems, such as the substantial decrease in performance and poor use of available system resources, can occur when C4.5 is applied to large data sets.
- Bayesian Networks: This classifier is a powerful probabilistic representation, and its use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute Ai given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of A1…..An and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes. In particular, the Naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. Although the naive Bayesian (NB) algorithm is simple, it is very effective in many real world datasets because it can give better predictive accuracy than well-known methods like C4.5 and BP [20],[21] and is extremely efficient in that it learns in a linear fashion using ensemble mechanisms,

- such as bagging and boosting, to combine classifier predictions [22]. However, when attributes are redundant and not normally distributed, the predictive accuracy is reduced [23].
- Neural Network: Back-Propagation (BP) Neural Networks can process a very large number of instances; have a high tolerance to noisy data; and has the ability to classify patterns which they have not been trained [19]. They are an appropriate choice if the results of the model are more important than understanding how it works [24]. However, the BP algorithm requires long training times and extensive testing and retraining of parameters, such as the number of hidden neurons, learning rate and momentum, to determine the best performance [25].
- Support Vector Machine: Support vector machines exist in different forms, linear and non-linear. A support vector machine is a supervised classifier. What is usual in this context, two different datasets are involved with SVM, training and a test set. In the ideal situation the classes are linearly separable. In such situation a line can be found, which splits the two classes perfectly. However not only one line splits the dataset perfectly, but a whole bunch of lines do. From these lines the best is selected as the "separating line". The best line is found by maximizing the distance to the nearest points of both classes in the training set. The maximization of this distance can be converted to an equivalent minimization problem, which is easier to solve. The data points on the maximal margin lines are called the support vectors. Most often datasets are not nicely distributed such that the classes can be separated by a line or higher order function. Real datasets contain random errors or noise which creates a less clean dataset. Although it is possible to create a model that perfectly separates the data, it is not desirable, because such models are over-fitting on the training data. Over-fitting is caused by incorporating the random errors or noise in the model. Therefore the model is not generic, and makes significantly more errors on other datasets. Creating simpler models keeps the model from over-fitting. The complexity of the model has to be balanced between fitting on the training data and being generic. This can be achieved by allowing models which can make errors. A SVM can make some errors to avoid over-fitting. It tries to minimize the number of errors that will be made. Support vector machines classifiers are applied in many applications. They are very popular in recent research. This popularity is due to the good overall empirical performance. Comparing the naive Bayes and the SVM classifier, the SVM has been applied the most.

Table-3.1: WEKA names of selected classifiers

Generic Name	WEKA Name
Bayesian Network	Naïve Bayes (NB)
Neural Network (NN)	Multilayer Perceptron
Support Vector Machine	SMO
C4.5 Decision Tree	J48
K-Nearest Neighbour	1Bk

IV. APPLICATION AREAS

There are several applications of data mining. Some common used applications of data mining are given below:

a) Fraud or non-compliance anomaly detection: Data mining isolates the factors that lead to fraud, waste and abuse. The process of compliance monitoring for anomaly detection (CMAD) involves a primary monitoring system comparing some predetermined conditions of acceptance with the actual data or event. If any variance is detected (an anomaly) by the primary monitoring system then an exception report or alert is produced, identifying the specific variance. For instance credit card fraud detection monitoring, privacy compliance monitoring, and target auditing or investigative efforts can be done more effectively [5].

b) Intrusion detection: It is a passive approach to security as it monitors information systems and raises alarms when security violations are detected. This process monitors and analyzes the events occurring in a computer system in order to detect signs of security problems. Intrusion detection systems (IDSs) may be either host based or network based, according to the kind of input information they analyze [6]. Over the last few years, increasing number of research projects (MADAM-ID, ADAM, Clustering project, etc.) have been applied data mining approaches (either host based or network based) to various problems (construction of operational IDSs, clustering audit log records, etc.) of intrusion detection [13].

c) Lie detection (SAS Text Miner): SAS institute introduced lie-detecting software, called SAS Text Miner. Using intelligence of this tool, managers can be able to detect automatically when email or web information contains lies. Here data mining can be applied successfully to identify uncertainty in a deal or angry customers and also have many other potential applications [14]. Many other market mining tools are also available in real practice viz. Clementine, IBM's Intelligent Miner, SGI's MineSet, SAS's Enterprise Miner, but all pretty much the same set of tools.

d) Market basket analysis (MBA): Basically it applies data mining technique in understanding what items are likely to be purchased together according to association rules, primarily with the aim of identifying cross-selling opportunities. Sometimes it is also referred to as product affinity analysis. MBA gives clues as to what a customer might have bought if an idea had occurred to them. So, it can be used in deciding the location and promotion of goods by means of combo-package and also can be applied to the areas like analysis of telephone calling patterns, identification of fraudulent medical insurance claims, etc. [15].

e) Aid to marketing or retailing: Data mining could help direct marketers by providing useful and accurate trends on purchasing behavior of their customers and also help them in predicting which products their customers may be interested in buying. In addition, trends explored by data mining help retail-store managers to arrange shelves, stock certain items, or provide a certain discount that will attract their customers. In fact data mining allows companies to identify their best customers, attract customers, aware customers via mail marketing, and maximize profitability by means of identifying profitable customers [16].

f) Customer segmentation and targeted marketing: Data mining can be used in grouping or clustering customers based on the behaviors (like payment history, etc.),

which in turn helps in customer relationship management (epiphany) and performs targeted marketing. Usually it becomes useful to define similar customers in a cluster, holding on good customers, weeding out bad customers, identify likely responders for business promotions.

g) Phenomena of "`beer and baby diapers"': This story of using data mining to find a relation between beer and diapers is told, retold and added to like any other legend. The explanation goes that when fathers are sent out on an errand to buy diapers, they often purchase a six-pack of their favorite beer as a reward. An article in The Financial Times of London (Feb. 7, 1996) stated, "The oft-quoted example of what data mining can achieve is the case of a large US supermarket chain which discovered a strong association for many customers between a brand of babies nappies (diapers) and a brand of beer [17].

h) Financial, banking and credit or risk scoring: Data mining can assist financial institutions in various ways, such as credit reporting, credit rating, loan or credit card approval by predicting good customers, risk on sanctioning loan, mode of service delivery and customer retention (i.e. build profiles of customers likely to use which services), and many others. A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. In addition, data mining can also assist credit card issuers in detecting potentially fraudulent credit card transaction. In general, data mining methods such as neural networks and decision trees can be a useful addition to the techniques available to the financial analyst [18].

i) Medicare and health care: Applying data mining techniques, it is possible to find relationship between diseases, effectiveness of treatments, to identify new drugs, market activities in drug delivery services, etc. However, a pharmaceutical company can analyze its recent sales to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health care systems. Such dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sale situation.

V. CONCLUSION

To overview, evolution, parameter and the applications of GA and PSO are presented in a simple way. Although PSO has been used mainly to solve unconstrained, single objective optimization problems, PSO algorithms have been developed mainly to solve constrained problems, multi objective optimization problems and problems with dynamically changing landscapes and to find multiple solutions.

REFERENCES

1. Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.
2. Provost, F., & Fawcett, T., Robust Classification for Imprecise Environments. Machine Learning, Vol. 42, No.3, pp.203-231, 2001.
3. Larose D T, Discovering knowledge in data: an introduction to data mining, John Wiley, New York, 2005.
4. Kantardzic M, Data mining: concepts, models, methods, and algorithms, John Wiley, New Jersey, 2003.
5. Goldschmidt P S, Compliance monitoring for anomaly detection, Patent no. US 6983266 B1, issue date January 3, 2006, Available at: www.freepatentsonline.com/6983266.html
6. Bace R, Intrusion Detection, Macmillan Technical Publishing, 2000.
7. J. Kennedy, R. C. Eberhart, and Y. Shi, Swarm Intelligence, Morgan Kaufmann, San Francisco, CA, 2001.
8. J. Kennedy, Small worlds and mega-minds: Effects of neighborhood topology on particle swarm performance, In Proceeding of the 1999 Conference on Evolutionary Computation, pp. 1931-1938, 1999.
9. J. Kennedy and R. Mendes, Population structure and particle swarm performance, Proceeding of the 2002 Congress on Evolutionary Computation, Honolulu, Hawaii, May 2002.
10. J. Kennedy and R. C. Eberhart, A discrete binary version of the particle swarm algorithm, in Proceeding of the 1997 Conference on Systems, Man, and Cybernetics, pp. 4104-4109, 1997.
11. C. K. Mohan and B. Al-kazemi, Discrete particle swarm optimization, Proceedings of the Workshop on Particle Swarm Optimization, Indianapolis, IN, 2001.
12. D. K. Agrafiotis and W. Cedeño, Feature selection for structure-activity correlation using binary particle swarms, Journal of Medicinal Chemistry, Vol. 45, pp. 1098-1107, 2002
13. Smyth P, Breaking out of the Black-Box: research challenges in data mining, Paper presented at the Sixth Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD-2001), held on May 20 (2001), Santra Barbara, California, USA.
14. SAS Institute Inc., Lie detector software: SAS Text Miner (product announcement), Information Age Magazine, [London, UK], February 10 (2002), Available at: http://www.sas.com/solutions/fraud/index.html.
15. Berry M J A and Linoff G S, Data mining techniques: for marketing, sales, and relationship management, 2 nd edn (John Wiley; New York), 2004.
16. Delmater R and Hancock M, Data mining explained: a manager's guide to customer-centric business intelligence, (Digital Press, Boston), 2002.
17. Fuchs G, Data Mining: if only it really were about Beer and Diapers, Information Management Online, July 1, (2004), Available at: http://www.information-management.com/ news/1006133-1.html.
18. Langdell S, Use of data mining in financial applications, (Data Analysis and Visualization Group at NAG Ltd.), Available at: http://www.nag.co.uk/IndustryArticles/ DMinFinancialApps.pdf