ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (8-14), Month: April - June 2014, Available at: www.researchpublish.com

A REVIEW ON DATA MINING ALGORITHMS

Maninder Singh
Dept. of CSE, Guru Nanak Dev University, Amritsar (Punjab) India.

Abstract: **This paper presents different machine learning algorithms. Machine learning algorithms are becoming popular day by day in real time applications like intrusion detection system, diabetes mining, e-mail spam classification etc. The idea behind the data mining is very straight it is same like a human being become intelligent from examples and experience. In data mining; rules are developed by taking the behaviour of given system (data set). Then these rules are used to evaluate the behaviour/ outcome for the given circumstances. The overall objective of this paper is to classify some well-known data mining algorithms.**

Keywords: **Machine learning, Data mining, Decision trees, clustering.**

I. Introduction

Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behaviour of their customers and potential customers. It discovers information within the data that queries and reports can't effectively reveal.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

II. What Can Data Mining Do?

Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, health care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed.

For businesses, data mining is used to discover patterns and relationships in the data in order to help make better business decisions. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty. Specific uses of data mining include:

- Market segmentation - Identify the common characteristics of customers who buy the same products from your company.
- Customer churn - Predict which customers are likely to leave your company and go to a competitor.
- Fraud detection - Identify which transactions are most likely to be fraudulent.
- Direct marketing - Identify which prospects should be included in a mailing list to obtain the highest response rate.
- Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.
- Market basket analysis - Understand what products or services are commonly purchased together; e.g., beer and diapers.
- Trend analysis - Reveal the difference between a typical customer this month and last.

III. Data Mining Techniques

Several core techniques that are used in data mining describe the type of mining and data recovery operation. Unfortunately, the different companies and solutions do not always share terms, which can add to the confusion and apparent complexity.

Let's look at some key techniques and examples of how to use different tools to build the data mining.

A. Association Rules:

An association rule is a rule which implies certain association relationships among a set of objects (such as “occur together” or “one implies the other”) in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: “30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items”. Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

Association (or relation) is probably the better known and most familiar and straightforward data mining technique. Here, you make a simple correlation between two or more items, often of the same type to identify patterns. For example, when tracking people's buying habits, you might identify that a customer always buys cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream.

Building association or relation-based data mining tools can be achieved simply with different tools. For example, within Info Sphere Warehouse a wizard provides configurations of an information flow that is used in association by examining your database input source, decision basis, and output information. Figure 2 shows an example from the sample database.

B. Classification:

You can use classification to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, you can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, you might apply it into a particular class by comparing the attributes with our known definition. You can apply the same principles to customers, for example by classifying them by age and social group.

Additionally, you can use classification as a feeder to, or the result of, other techniques. For example, you can use decision trees to determine a classification. Clustering allows you to use common attributes in different classifications to identify clusters.

C. Clustering:

By examining one or more attributes or classes, you can group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is useful to identify different information because it correlates with other examples so you can see where the similarities and ranges agree.

Clustering can work both ways. You can assume that there is a cluster at a certain point and then use our identification criteria to see if you are correct. The graph in Figure shows a good example. In this example, a sample of sales data compares the age of the customer to the size of the sale. It is not unreasonable to expect that people in their twenties (before marriage and kids), fifties, and sixties (when the children have left home), have more disposable income.

D. Prediction:

Any prediction can be thought of as classification or estimation. The difference is one of emphasis. When data mining is used to classify a phone line as primarily used for internet access or a credit card transaction as fraudulent, we do not expect to be able to go back later to see if the classification was correct. Our Classification may be correct or incorrect, but the uncertainty is due to incomplete knowledge only: out in the real world, the relevant actions have already taken place. The phone is or is not used primarily to dial the local ISP. The credit card transaction is or is not fraudulent. With enough efforts, it is possible to check.

Predictive tasks feel different because the records are classified according to some predicted future behavior or estimated future value. With prediction, the only way to check the accuracy of the classification is to wait and see. Examples of prediction tasks include:

- Predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer
- Predicting which customers will leave within next six months
- Predicting which telephone subscribers will order a value–added service such as three-way calling or voice mail.

Any of the techniques used for classification and estimation can be adopted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior.

E. Sequential patterns:

Often used over longer-term data, sequential patterns are a useful method for identifying trends, or regular occurrences of similar events. For example, with customer data you can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history.

F. Decision trees:

Related to most of the other techniques (primarily classification and prediction), the decision tree can be used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each answer leads to a further question to help classify or identify the data so that it can be categorized, or so that a prediction can be made based on each answer.

Decision trees are often used with classification systems to attribute type information, and with predictive systems, where different predictions might be based on past historical experience that helps drive the structure of the decision tree and the output.

G. Combinations:

In practice, it's very rare that you would use one of these exclusively. Classification and clustering are similar techniques. By using clustering to identify nearest neighbors, you can further refine your classifications. Often, we use decision trees to help build and identify classifications that we can track for a longer period to identify sequences and patterns.

H. Long-term (memory) processing:

Within all of the core methods, there is often reason to record and learn from the information. In some techniques, it is entirely obvious. For example, with sequential patterns and predictive learning you look back at data from multiple sources and instances of information to build a pattern.

In others, the process might be more explicit. Decision trees are rarely built one time and are never forgotten. As new information, events, and data points are identified, it might be necessary to build more branches, or even entirely new trees, to cope with the additional information.

You can automate some of this process. For example, building a predictive model for identifying credit card fraud is about building probabilities that you can use for the current transaction, and then updating that model with the new (approved) transaction. This information is then recorded so that the decision can be made quickly the next time.

I. Statistics:

The problem of abstracting knowledge from data has been tackled by statisticians, long before the first artificial intelligence papers were published. For example, correlation analysis applies statistical tools for analyzing the correlation between two or more variables. Cluster analysis offers methods for discovering clusters in large set of objects described by vector of values. Factor analysis tries to point the most important variables describing clusters. Some of the popular techniques that are used for supervised classification tasks are Linear Discriminants, Quadratic Discriminants, K-nearest Neighbor, Naïve Bays, Logistic Regression and CART.

J. Machine Learning:

Statistical methods have difficulty incorporating subjective, non quantifiable information in their models. They also have to assume various distributions of parameters and independence of attributes. Various studies have concluded that machine learning produces comparable (and often better) predictive accuracy. Its good performance as compared to statistical methods can be attributed to the fact that it is free from parametric and structural assumptions that underlie statistical methods. Another weakness of statistical approaches to data analysis is the problem of interpreting the results. Some of the machines learning techniques are mentioned below.

K. Neural Networks:

Artificial neural networks are computational models composed of many non linear processing elements arranged in a pattern similar to biological neuron networks. A typical neural network has an activation value associated with each node and a weight value associated with each connection. An activation function governs the firing of nodes and the propagation of data through network connections in massive parallelism. The network can also be trained with examples through connection weight adjustments.

L. Genetic Algorithms

Genetic algorithms are search algorithms based on mechanics of natural selection and natural genetics. They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search. In every generation, a new set of strings is created using bits and pieces of the fittest of the old; an occasional new part is tried for good measure. While randomized, genetic algorithms are no simple random walk. They efficiently exploit historical information to speculate on new search points with expected improved performance. A simple GA that yields good result, is composed of three operators namely reproduction, crossover and mutation. GAs differs from more normal optimization and search procedures in four ways:

- GAs work with coding of parameter set, not the parameter themselves.
- GAs search from a population of points, not a single point.
- GAs use objective function information, not derivatives or other auxiliary knowledge.
- GAs use probabilistic transitional rules, not deterministic rules.

M. Support Vector Machines:

SVMs are the learning machines that can perform binary classification and regression estimation tasks. They are becoming increasingly popular as a new paradigm of classification and learning because of two important factors. First, unlike the other classification techniques, SVMs minimize the expected error rather than minimizing the classification error. Second, SVMs employ the duality theory of mathematical programming to get a dual problem that admits efficient computational methods.

N. Fuzzy Logic:

Fuzzy logic, which may be viewed as an extension of classical logical systems, provides an effective conceptual framework for dealing with the problem of knowledge representation in an environment of uncertainty and imprecision [Zad89]. Some of the essential characteristics of fuzzy logic relate to the following:

- In fuzzy logic, exact reasoning is viewed as a limiting case of approximate reasoning.
- In fuzzy logic everything is a matter of degree.
- Any logical system can be fuzzified.
- In fuzzy logic, knowledge is interpreted as a collection of elastic or equivalently, fuzzy constraint on a collection of variables.

Summary of basic concepts and techniques underlying the application of fuzzy logic to knowledge representation and description of number of examples relating to its use as a computational system is provided in [Zad89]. Fuzzy logic in its pure form is not a technique for classification but it has been a very useful concept in many hybrid techniques for classification.

O. Rough Sets Techniques:

RS theory deals with approximation of sets or concepts by means of binary relations constructed from empirical data based on the notion of indiscernibility and the inability to distinguish between objects. Such approximations can be said to form models of our target concepts, and hence in its typical use, falls in under the bottom up approach to model construction. Rough set applications to data mining generally proceed along the following directions:

- Decision rule induction from attribute value table
- Data filtration by template generation - This mainly involves extracting elementary blocks from data based on equivalence relation. Genetic algorithms are also sometimes used in this stage for searching.

IV. Related Work

Jindal and Liu (2007) [1] has proposed mining of opinions from product reviews, forum posts and blogs as an important research topic with many applications. Existing research has been focused on extraction, classification and summarization of opinions from these sources. The issue in the context of product reviews has been studied. There is still no published study on this topic, although Web page spam and email spam have been investigated extensively. Review spam is quite different from Web page spam and email spam, and thus requires different detection techniques.

Ma et al. (2009) [2] has proposed that detecting and filtering are still the most feasible ways of fighting spam emails. There are many reasonably successful spam email filters in operation. Proactively catching new strains of spam emails, where no previous knowledge is available, is still a major challenge. Negative selection is a branch of artificial immune systems. It has a strong temporal nature and is especially suitable for discovering unknown temporal patterns. This nature makes it a good candidate in quickly discovering and detecting new strains of spam emails.

Wang and Liu (2010) [3] has presented various approaches to solve the spreading spam problem. Most of these approaches cannot flexibly and dynamically adapt to spam. A novel approach to counter spam based on trusted behavior recognition during transfer sessions has been proposed. A behavior recognition of email transfer patterns which enables normal servers to detect malicious connections before email body has been delivered. An integrated Anti-Spam framework combining the trusted behavior recognition with Bayesian Analysis has been designed. The effectiveness of both the trusted Behavior recognition and the integrated filter have been evaluated.

Algur et al. (2010) [4] has discussed that the estimation mining from product reviews, forum posts and blogs as an important research topic today with many applications. Existing research has focused more towards classification and summarization of these online opinions. An important issue related to the trustworthiness of online opinions has been neglected most often. There is no reported study on assessing the trustworthiness of reviews, which is crucial for all opinion based applications, although web spam and email spam have been investigated extensively.The trust worthiness of the reviews has been accessed as spam or a non spam review which includes both duplicate and near duplicate reviews classified as spam reviews, and partially related and unique reviews classified as non spam reviews .A novel and effective technique, namely, Conceptual level similarity measure used for detecting spam reviews based on the product features that have been commented in the reviews has been proposed.

Chen et al. (2010) [5] has discussed Email as a kind of semi-structured document, and using spam-specific features could improve the email classification results. The decision tree data mining technique to dig out the potential association rules among these attributes of email, and then to identify unknown email’s category based on these rules has been applied. The efficiency of the method is not lower than that of other existing methods of checking whole email content text.

Salama et al. (2012) [6] has presented a comparison among the different classifiers decision tree (J48), Multi-LayerPerception (MLP), Naive Bayes (NB), Sequential Minimal Optimization (SMO), and Instance Based for K-Nearest neighbor (IBK) on three different databases of breast cancer (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) and Wisconsin Prognosis Breast Cancer (WPBC)) by using classification accuracy and confusion matrix based on 10-fold cross validation method. A fusion at classification level between these classifiers to get the most suitable multi-classifier approach for each data set has been introduced. The experimental results have shown that in the classification using fusion of MLP and J48 with the PCA is superior to the other classifiers using WBC data set. The PCA has been used in WBC dataset as a features reduction transformation method which combines a set of correlated features. All experiments have been conducted in WEKA data mining tool.

Liu and Yang (2012) [7] has proposed that the spam messages in mobile phone were flooded and the management to it was not effective. The characteristics of spam messages, its forming reason and its harm has been analysed. It has been discussed that the classification method of filtering spam messages, and points out it is the key work of the researchers to develop the more effective classification method.

Thornton and Burman (2012) [8] have described that how a novel application of Data Mining techniques can be used to provide the engine for a tool which can be used to identify email correspondence which may be an early indication of virtual bullying or harassment. The approach that makes use of linear discriminant approaches to classify normal, and non-normal, style of email correspondence for each sender has been taken. The change in email style could be used to provide an early indication of virtual harassment/bullying. This approach has great potential for use in large organization where it often appears to be hard to identify unacceptable information transmission between two colleagues. By identifying indicative behavior it has been made possible to start company anti bullying processes in a more timely manner.

V. Conclusion and Future work

This paper has presented different algorithms of data mining. The overall goal is to evaluate the use of the data mining algorithms for different kind of systems. Each algorithm has different goal and objective to classify the data set in different manner. Like clustering based algorithms has ability to develop the cluster of homogeneous data element from a given data set like Sex attribute divide data set into Males and Females. Also apriori for evaluating the relationships among attributes like bonus attribute has great influence on the montly income of a person.

In near future we will use different data mining algorithms to efficiently detect the email spams from a given data set of emails. The implementation of the given behaviour will be done using some popular data mining tool like Weka, MATLAB, Web-miner etc.

References

[1] Jindal, Nitin, and Bing Liu. "Analyzing and detecting review spam." In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, pp. 547-552. IEEE, 2007.

[2] Ma, Wanli, Dat Tran, and Dharmendra Sharma. "A novel spam email detection system based on negative selection." In Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. Fourth International Conference on, pp. 987-992. IEEE, 2009.

[3] Wang, Cong, and Jianyi Liu. "Trusted Behavior Based Spam Filtering." In Web Information Systems and Mining (WISM), 2010 International Conference on, vol. 2, pp. 96-100. IEEE, 2010.

[4] Algur, Siddu P., Amit P. Patil, P. S. Hiremath, and S. Shivashankar. "Conceptual level similarity measure based review spam detection." In Signal and Image Processing (ICSIP), 2010 International Conference on, pp. 416-423. IEEE, 2010.

[5] Chen, Hao, Yan Zhan, and Yan Li. "The application of decision tree in Chinese email classification." In Machine Learning and Cybernetics (ICMLC), 2010 International Conference on, vol. 1, pp. 305-308. IEEE, 2010.

[6] Salama, Gouda I., M. B. Abdelhalim, and Magdy Abd-elghany Zeid. "Experimental comparison of classifiers for breast cancer diagnosis." In Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on, pp. 180-185. IEEE, 2012.

[7] Liu, Guoxiang, and Fengxia Yang. "The application of data mining in the classification of spam messages." In Computer Science and Information Processing (CSIP), 2012 International Conference on, pp. 1315-1317. IEEE, 2012.

[8] Burn-Thornton, K., and T. Burman. "The Use of Data Mining to Indicate Virtual (Email) Bullying." In Intelligent Systems (GCIS), 2012 Third Global Congress on, pp. 253-256. IEEE, 2012.

[9] SHI, Lei, Q. Wang, and X. M. Ma. "Spam Email Classification Using Decision Tree Ensemble." Journal of Computational Information Systems 8, no. 3 (2012): 949-956.