International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

General Article

A Comparative Study of Different Data Mining Algorithms

Shrey Bavisi[Á*], Jash Mehta [Á] and Lynette Lopes[Á]

[A]Computer Department, DJSCOE, Vile Parle (W), Mumbai – 400056, India

Accepted 02 Sept 2014, Available online 01 Oct 2014, **Vol.4, No.5 (Oct 2014)**

Abstract

Data Mining is used extensively in many sectors today, viz., business, health, security, informatics etc. The successful application of data mining algorithms can be seen in marketing, retail, and other sectors of the industry. The aim of this paper is to present the readers with the various data mining algorithms which have wide applications. This paper focuses on four data mining algorithms K-NN, Naïve Bayes Classifier, Decision tree and C4.5. An attempt has been made to do a comparative study on these four algorithms on the basis of theory, its advantages and disadvantages, and its applications. After studying all these algorithms in detail, we came to a conclusion that the accuracy of these techniques depend on various characteristics such as: type of problem, dataset and performance matrix.

Keywords: *Data mining, k-NN, Naïve Bayes classifier, Decision Tree, C4.5, classification.*

1. Introduction

Data mining is a process of exploring huge data, typically business related data which is also called as big data. This process is performed to find hidden patterns and relationship present in the data. The overall objective of the data mining process is to extract information from a large data set and transform it into a comprehensible structure for further use. Generally, the tasks of data mining are two types:

1) Descriptive data mining: In descriptive data mining, the data set is summarized in a concise manner and presents interesting properties of the data.
2) Predictive data mining. The ultimate goal is prediction, which is the most common application of data mining, in this behavior of future data sets is predicted.

The process of data mining consists of two stages:

1) Exploration Stage: This stage consists of pre-processing data, i.e., before applying data mining algorithms on the data, the data sets must be assembled from all disparate sources. In other words the data must be extracted from all sources such that the disparity is eliminated. A common source of such data is a data warehouse of a company, hospital, retail chain etc. In this stage the data is cleansed and transformed so that the noise and missing values are dealt with.
2) Data mining Stage: This stage takes place after performing exploration.

- Class description: Class description provides a concise summarization of data. This is also called as characterization of data.
- Association: Association is discovery of dependencies or correlations in the data sets. An association rule expressed as X=>Y means that, database tuples that satisfy X are likely to satisfy Y
- Classification: Classification analyses a set of training data and based on the features of the training data the classification rules are generated and models are constructed which can be used in future for testing data
- Clustering: Clustering analysis is grouping the similar data in the data sets. Similarity can be expressed in terms of distance functions.

There are large numbers of data mining algorithms which are used in the field of Engineering, Meteorology, Informatics, Corporate Business, Sales Forecasting, Business Forecasting Domains, Neurophysiology, Finance, Medicine and many more. But, in this paper we will focus mainly on commonly used mining algorithms such as:

1) k-NN (k-Nearest Neighbours): KNN is a simple classification and regression algorithm.
2) Naïve Bayes classifier: Naïve Bayes classifier is a supervised learning algorithm which is used for data classification using statistical method.
3) Decision trees: Decision trees are powerful and popular tools for classification and prediction.
4) C4.5: C4.5 is an algorithm that was developed by Ross Quinlan. This algorithm generates Decision trees which can further be used for problems related to classification.

*Corresponding author: **Shrey Bavisi**

2. Data Mining Algorithms

2.1 k-NN (k-Nearest Neighbors)

KNN is a simple classification and regression algorithm that stores all the available cases and classifies new incoming cases based on a certain similarity measure. Conceptually, KNN is a simple algorithm; however it is still able to solve complex problems. *k*-NN algorithm is a type of instance-based learning or lazy learning, wherein the function is approximated only locally. All computation is ceased until classification.

KNN can be used to weight the contributions of the neighbors, in both classification and regression, so that the contribution of nearer neighbors is more towards the average and less towards the distant ones. The training examples in KNN are vectors which are in a multidimensional feature space, each having a class label. Storing of the feature vectors along-with class labels of the training samples are a part of the training phase.

In the classification phase, the user-defined constant is k, and an unlabeled vector is classified by assigning the label which is most frequent among the *k* training samples nearest to that query point. A distance metric used for continuous variables is Euclidean distance.

Advantages

1) *K-NN* methods have a simplicity and lack of parametric assumptions.
2) In the presence of a large enough training set, these methods perform shockingly well, especially in cases when each class is characterized by multiple combinations of predictor values. For instance, in real estate databases there are likely to be multiple combinations of {home type, asking price, number of rooms, neighborhood, etc.} that characterize homes that sell quickly against those that remain unsold for a long period in the market.
3) It is robust with regards to the search space.
4) Classes need not be linearly separable.
5) Classifier can be updated online and that to at very little cost given the fact that new instances with known classes are presented.
6) There are only a handful of parameters to tune: distance metric and *k*.
7) Zero cost of the learning process.
8) Local approximation can help learn complex concepts by using simple procedures.

Disadvantages

1) Although no time is required to estimate parameters from the training data, the time to find the nearest neighbors in a large training set can be prohibitive.
2) Expensive testing of each instance, as we need to compute its distance to all known instances. Specialized algorithms and heuristics exist for specific problems and distance functions, which can mitigate this issue. This is problematic for datasets with a large number of attributes.
3) Sensitiveness to noisy or irrelevant attributes, which can result in less meaningful distance numbers. Scaling and/or feature selection are typically used in combination with KNN to mitigate this issue.
4) Performance depends on the number of dimensions that we have.

Applications

1) Classification and Interpretation in the fields of banking, legal, medical, news, etc.
2) Problem-solving in pronunciation, planning etc.
3) Function learning in dynamic control.
4) Classification of program as intrusive or normal.
5) Efficient statistical classification of satellite measurements.
6) Melting point prediction.
7) Optical character recognition.

2.2 Naïve Bayes Classifier

Naïve Bayes classifier is a supervised learning algorithm which is used for data classification using statistical method. It is a probabilistic classifier that helps to classify the given input over a set of classes using probability distribution. This method goes by the name Naïve because it naively assumes independence of the attributes given the class. Classification is then completed by applying Bayes rule to work out the probability of the correct class given the particular attributes of a scenario. The following equation (1) is used in Naïve Bayes classifier:

$$P(C \mid F_1.....F_n) = \frac{P(F_1.....F_n \mid C) * P(C)}{P(F_1.....F_n)} \quad (1)$$

Where, C: No. of classes, F_1-F_n: Evidence.
Equation (1) can also be written as:

$$\text{Posterior Probability} = \frac{\text{Prior Probability} * \text{Likelihood}}{\text{Evidence}} \quad (2)$$

For example, consider following equation for Credit Card Fraud detection:

$$P(Fraud \mid Evidences) = \frac{P(Evidences \mid Fraud) * P(Fraud)}{P(Evidences)} \quad (3)$$

The terms used in Naïve Bayes Classifier are as follows:

Posterior Probability: In Bayesian statistics, the posterior probability of a random event or an uncertain proposition is the conditional probability that is assigned after the relevant evidence is taken into account. In equation (3), P (Fraud | Evidences) is the posterior probability; the probability of the hypothesis (the transaction being fraudulent) after considering the effect of the evidences (the attribute values based on training examples).

Prior Probability: In Bayesian statistical inference, a prior probability distribution, often called simply

the prior, of an uncertain quantity *p* is the probability distribution that would express one's uncertainty about *p* before some evidence is taken into account. In equation (3), P (fraud) is the prior probability; the probability of the hypothesis given only past experiences while ignoring any of the attribute values.

Maximum Likelihood: Maximum likelihood is a technique used to estimate the parameter that can be used in Naïve Bayes classifier. This method selects a parameter that will maximize the likelihood function. In equation (3), P (Evidences | Fraud) is called the likelihood.

Advantages

1) This algorithm is simple to implement.
2) It has great computational efficiency and classification performance.
3) The amount of data required for this supervised learning algorithm is less, whereas other sophisticated algorithms require huge amount of data for learning process.
4) It gives accurate results for most of the classification and prediction problems.

Disadvantages

1) Any classification algorithm requires large amount of data for better accuracy. Same is the case with Naïve Bayes classifier in which, the precision of algorithm decreases if the amount of data is less.
2) If there is no occurrence of the class label and attribute value, then this algorithm will consider the probability of that attribute to be zero. This problem is called as the 'Zero Problem'.

Applications

1) It can be used in Credit Card Fraud detection, Heart Disease prediction, predicting the Sex of an individual, rain prediction etc.
2) It can also be used for spam filtering.
3) Additionally, network intrusion detection is supported by this algorithm.

2.3 Decision Tree

Decision trees are powerful and popular tools for classification and prediction. The attractiveness of this algorithm is due to the fact that, decision trees represent only rules, in contrast to neural networks. The rules can be readily expressed so that humans can understand them or even directly use them in database access language like SQL so that records falling into a particular category may be retrieved. A decision tree allows the calculation of forward and backward and because of that correct decision will be made automatically. Decision tree is a classifier in the form of tree structure where each node is either:

1) Leaf node: Indicates the value of the target attribute
2) Decision node: Specifies some test to be carried out on a single attribute-value with one branch and sub-tree for each possible outcome of the test.

Representation

Example: Decision tree representation for weekend plans.

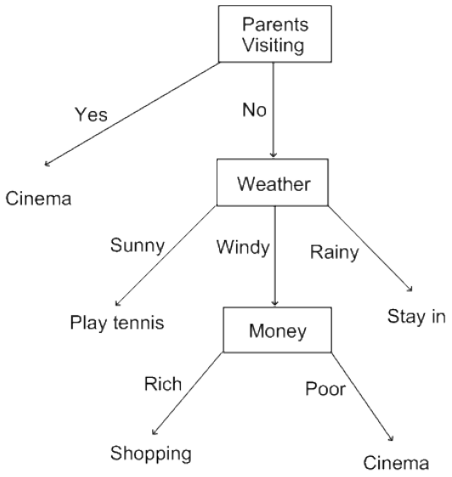

Fig.1 Decision Tree Representation

Logical representation:

For shopping: (Parents Visiting=No Λ Weather=Windy Λ Money=Rich)

Similarly we can deduce logical representation for other results.

If –then Rules:

If Parents Visiting=Yes THEN go to Cinema.

If Parents Visiting=No Λ Weather=Sunny THEN Play tennis.

This way we can represent a decision tree with If-then rules.

Advantages

1) The rules generated by a decision tree are easy for the humans to understand.
2) Less computation is required for classification in decision trees.
3) Decision trees are not affected continuous variables, noisy data and outliers.
4) Decision trees provide a clear suggestion of which fields are most critical for classification or prediction.
5) They are fast and scalable.

Disadvantages

1) Performs poorly with many class and small data.
2) Computationally expensive to train.
3) It does not function well with categorical variables having multiple levels.
4) Over fitting: Too many branches may result in poor accuracy (Pre-Pruning and Post-Pruning can be used to handle this type of problem).

Applications

1) Decision trees are commonly used decision analysis, which helps to reach a target by identifying a strategy.
2) It is used for calculating conditional probabilities.
3) They are used in health care systems.
4) Strategic Business decision making is possible with help of decision trees.
5) They are also used in the field of finance and philosophy.

2.4 C-4.5

C4.5 is an algorithm that was developed by Ross Quinlan. This algorithm generates Decision trees which can further be used for problems related to classification. C4.5 algorithm further make changes into the previous ID3 algorithm and deals with both discrete and continuous attributes, pruning trees after construction and also missing values. C4.5 is also known to as a statistical classifier. C4.5 algorithm was made further more memory efficient by creating its successor which is called C5.0. C5.0 is used to make even smaller decision trees.

A set of training examples are required where-in each example can be seen as a pair: an input object and its corresponding output value or class. The algorithm builds a classifier by analyzing the training set, such that it correctly classifies both test and training examples. A test example is an input object and the output value must be correctly predicted.

C4.5 algorithm makes use of information gain as splitting criteria. It has the provision to accept values that have numerical or categorical values. To deal with the continuous values, it generates a threshold and divides the attributes that have values above the threshold values or equal to threshold values or below the threshold. Missing values can be easily handled by C4.5 algorithm because of the fact that missing attribute values do not come in use in the gain calculations by C4.5.

There are a few base cases that this algorithm must hold:

1) The samples in the list must all belong to the same class. At the time when this happens, simply a leaf node is created for decision tree saying to choose that particular class.
2) If information gain is not provided by any of the classes then, in that case, C4.5 creates a decision node higher up the tree by making use of the expected value of the class.
3) If an instance of a previously-unseen class has been encountered, C4.5 again creates a decision node which is higher up the tree by making use of the expected value.

Advantages

1) It builds models that can be interpreted easily.
2) It is easy to implement.
3) Can make use of both categorical and continuous values.
4) It can handle noisy data.

Disadvantages

1) Even a small change in data can cause different decision trees to be built. This happens more in the case where the variables are very close to each other in value.
2) In case of a small training set, the C4.5 algorithm does not work very well(less accurate/efficient).

Applications

1) Financial distress prediction model.
2) Screening of patients.
3) Web-Based Learning Assessment System.

Conclusion

The No Free Lunch Theorem or Conservation Law is a theoretical result in which no single learning algorithm can perform better than any other when the expected generalization accuracy is the performance measure. It assumes that all possible targets are equally likely. However, this assumption is clearly wrong when in practice because for a given domain, it is often found that not all concept are equally probable. It has been found that the algorithm which performs the best depends on the type of problem that is being considered, the performance matrix used and the dataset characteristics. For instance, Naïve Bayes performs better in medical domains whereas the decision tree algorithms tend to perform better in sequential domains. So the performance of an algorithm varies from domain to domain. The algorithms discussed in this review literature include C4.5, KNN, Naïve Bayes and Decision tree. The Naïve Bayes model is simple, elegant and extremely robust making it way more appealing. It is also one of the oldest classification algorithms which in-spite of being simple provides effective results. KNN, on the other hand is an easily understood and easily implemented classification technique. C4.5 algorithm is also used in classification problems where it is used to build decision trees. C4.5 deals with both numeric attributes as well as missing values, making it suitable for dealing with real life problems. Thus, all the algorithms are used effectively in various domains to overcome real world difficulties making data mining a boon for people living in an era of technology.

Acknowledgement

We would like to thank our honorable principal Dr. Hari Vasudevan of D. J. Sanghvi College of Engineering and Dr. Narendra Shekhokar, Head of Department of Computer Engineering, for giving us the facilities and providing us with a propitious environment for working in college. We would like to thank Prof. Lynette Lopes for guiding us in this paper. We would also like to thank S.V.K.M. for encouraging us in such co-curricular activities.

References

Zaki .S. Towfik, Sabiha Fathil Jawad (2012), Application of Decision Tree As a Data Mining Tool in a health care, *Journal*

of Kufa for Mathematics and Computer, Vol. 1, No. 5 102-109.

k-Nearest Neighbors Javier B_ejar 2012, http://www.lsi.upc.edu/~bejar/apren/docum/trans/03d-algind-knn-eng.pdf

Zhu Xiaoliang, Wang Jian, Yan Hongcan, Wu Shangzhuo (2009), Research and application of the improved algorithm C4.5 on Decision tree, *IEEE*, Vol. 2 184-187 Peng Liu, Lei Lei, Junjie Yin, Zheng Yao (2005), R-C4.5 decision tree model and its applications to health care dataset,*IEEE*,Vol. 2 1099-1103

Decision Trees, http://octaviansima.wordpress.com/2011/03/25/decision-trees-c4-5/

k-Nearest Neighbors 2005 http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm >

Decision Tree 2003 http://en.wikipedia.org/wiki/Decision_tree C4.5 algorithm 2005

http://en.wikipedia.org/wiki/C4.5_ algorithm Naïve Bayes classifier 2002

http://en.wikipedia.org/wiki/Naive_Bayes_classifier,Data Mining

http://en.wikipedia.org/wiki/Data_mining, Data Mining Techniques

http://www.statsoft.com/Textbook/Data-Mining Techniques#mining