ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Techno Research Publishers, Bhopal, India.**
www.computerscijournal.org

ISSN: 0974-6471
April 2015,
Vol. 8, No. (1):
Pgs. 13-19

A Comparative Study of Classification Techniques in Data Mining Algorithms

SAGAR S. NIKAM

Department of Computer Science, K.K.Wagh College of Agriculture, Nashik, India.

(Received: February 16, 2015; Accepted: April 10, 2015)

ABSTRACT

Classification is used to find out in which group each data instance is related within a given dataset. It is used for classifying data into different classes according to some constrains. Several major kinds of classification algorithms including C4.5, ID3, k-nearest neighbor classifier, Naive Bayes, SVM, and ANN are used for classification. Generally a classification technique follows three approaches Statistical, Machine Learning and Neural Network for classification. While considering these approaches this paper provides an inclusive survey of different classification algorithms and their features and limitations.

Key words: C4.5, ID3, ANN, Naive Bayes, SVM, k-nearest neighbour.

INTRODUCTION

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data.The term could cover any context in which some decision or forecast is made on the basis of presently available information. Classification procedure is recognized method for repeatedly making such decisions in new situations. Here if we assume that problem is a concern with the construction of a procedure that will be applied to a continuing sequence of cases in which each new case must be assigned to one of a set of pre defined classes on the basis of observed features of data.Creation of a classification procedure from a set of data for which the exact classes are known in advance is termed as pattern recognition or supervised learning. Contexts in which a classification task is fundamental include, for example, assigning individuals to credit status on the basis of financial and other personal information, and the initial diagnosis of a patient's disease in order to select immediate treatment while awaiting perfect test results. Some of the most critical problems arising in science, industry and commerce can be called as classification or decision problems.Three main historical strands of research can be identified: statistical, machine *learning* and *neural network*. All groups have some objectives in common. They have all attempted to develop procedures that would be able to handle

a wide variety of problems and to be extremely general used in practical settings with proven success.

Statistical procedure based approach

Two main phases of work on classification can be identified within the statistical community. The first "classical" phase concentrated on extension of Fisher's early work on linear discrimination. The second, "modern" phase concentrated on more flexible classes of models many of which attempt to provide an estimate of the joint distribution of the features within each class which can in turn provide a classification rule [11]. Statistical procedures are generally characterised by having an precise fundamental probability model which provides a probability of being in each class instead of just a classification. Also it is usually assumed that the techniques will be used by statisticians and hence some human involvement is assumed with regard to variable selection and transformation and overall structuring of the problem.

Machine learning based approach

Machine Learning is generally covers automatic computing procedures based on logical or binary operations that learn a task from a series of examples. Here we are just concentrating on classification and so attention has focussed on decision-tree approaches in which classification results from a sequence of logical steps. These classification results are capable of representing the most complex problem given sufficient data. Other techniques such as genetic algorithms and inductive logic procedures (ILP) are currently under active improvement and its principle would allow us to deal with more general types of data including cases where the number and type of attributes may vary. Machine Learning approach aims to generate classifying expressions simple enough to be understood easily by the human and must mimic human reasoning sufficiently to provide insight into the decision process [11].Like statistical approaches background knowledge may be used in development but operation is assumed without human interference.

Neural network

The field of Neural Networks has arisen from diverse sources ranging from understanding and emulating the human brain to broader issues of copying human abilities such as speech and can be use in various fields such as banking, legal, medical, news, in classification program to categorise data as intrusive or normal. Generally neural networks consist of layers of interconnected nodes where each node producing a non-linear function of its input and input to a node may come from other nodes or directly from the input data. Also, some nodes are identified with the output of the network.

On the basis of this example there are different applications for neural networks that involve recognizing patterns and making simple decisions about them. In airplanes we can use a neural network as a basic autopilot where input units reads signals from the various cockpit instruments and output units modifying the plane's controls appropriately to keep it safely on course. Inside a factory we can use a neural network for quality control.

Classifaication algorithms

Classification is one of the Data Mining techniques that is mainly used to analyze a given dataset and takes each instance of it and assigns this instance to a particular class such that classification error will be least. It is used to extract models that accurately define important data classes within the given dataset. Classification is a two step process. During first step the model is created by applying classification algorithm on training data set then in second step the extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy. So classification is the process to assign class label from dataset whose class label is unknown.

ID3 Algorithm

Id3 calculation starts with the original set as the root hub. On every cycle of the algorithm it emphasizes through every unused attribute of the set and figures the entropy (or data pick up IG(A)) of that attribute. At that point chooses the attribute which has the smallest entropy (or biggest data gain) value. The set is S then split by the selected attribute (e.g. marks < 50, marks < 100, marks >= 100) to produce subsets of the information. The algorithm proceeds to recurse on each and every

item in subset and considering only items never selected before. Recursion on a subset may bring to a halt in one of these cases:

- Every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labelled with the class of the examples
- If there are no more attributes to be selected but the examples still do not belong to the same class (some are + and some are -) then the node is turned into a leaf and labelled with the most common class of the examples in that subset.
- If there are no examples in the subset, then this happens when parent set found to be matching a specific value of the selected attribute. For example if there was no example matching with marks >=100 then a leaf is created and is labelled with the most common class of the examples in the parent set.

Working steps of algorithm is as follows,

- Calculate the entropy for each attribute using the data set S.
- Split the set S into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum)
- Construct a decision tree node containing that attribute in a dataset.
- Recurse on each member of subsets using remaining attributes.

C4.5 Algorithm

C4.5 is an algorithm used to produce a decision tree which is an expansion of prior ID3 calculation. It enhances the ID3 algorithm by managing both continuous and discrete properties, missing values and pruning trees after construction. The decision trees created by C4.5 can be used for grouping and often referred to as a statistical classifier. C4.5 creats decision trees from a set of training data same way as Id3 algorithm. As it is a supervised learning algorithm it requires a set of training examples which can be seen as a pair: input object and a desired output value (class). The algorithm analyzes the training set and builds a classifier that must have the capacity to accurately arrange both training and test cases. A test example is an input object and the algorithm must predict an output value. Consider the sample training data set $S=S_1,S_2,...Sn$ which is already classified. Each sample S_i consist of feature vector ($x_{1,i}$, $x_{2,i}$, ..., $x_{n,i}$) where x_j represent attributes or features of the sample and the class in which S_i falls. At each node of the tree C4.5 selects one attribute of the data that most efficiently splits its set of samples into subsets such that it results in one class or the other. The splitting condition is the normalized information gain (difference in entropy) which is a non-symmetric measure of the difference between two probability distributions *P* and *Q*. The attribute with the highest information gain is chosen to make the decision. General working steps of algorithm is as follows,

- Assume all the samples in the list belong to the same class. If it is true, it simply creates a leaf node for the decision tree so that particular class will be selected.
- None of the features provide any information gain. If it is true, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Then, C4.5 creates a decision node higher up the tree using the expected value.

K Nearest Neighbors Algorithm

The closest neighbor (NN) rule distinguishes the classification of unknown data point on the basis of its closest neighbor whose class is already known. M. Cover and P. E. Hart purpose k nearest neighbour (KNN) in which nearest neighbor is computed on the basis of estimation of k that indicates how many nearest neighbors are to be considered to characterize class of a sample data point. It makes utilization of the more than one closest neighbor to determine the class in which the given data point belongs to and and consequently it is called as KNN. These data samples are needed to be in the memory at the run time and hence they are referred to as memory-based technique. T. Bailey and A. K. Jain enhance KNN which is focused on weights. The training points are assigned weights according to their distances from sample data

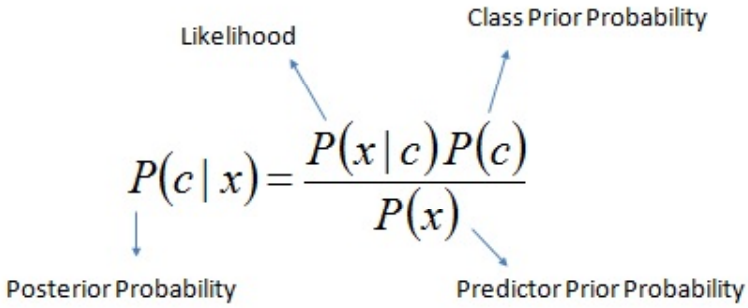
point. But at the same time the computational complexity and memory requirements remain the primary concern dependably. To overcome memory limitation size of data set is reduced. For this the repeated patterns which don't include additional data are also eliminated from training data set. To further enhance the information focuses which don't influence the result are additionally eliminated from training data set. The NN training data set can be organized utilizing different systems to enhance over memory limit of KNN. The KNN implementation can be done using ball tree, k-d tree, nearest feature line (NFL), principal axis search tree and orthogonal search tree. The tree structured training data is further divided into nodes and techniques like NFL and tunable metric divide the training data set according to planes. Using these algorithms we can expand the speed of basic KNN algorithm. Consider that an object is sampled with a set of different attributes. Assuming its group can be determined from its attributes; different algorithms can be used to automate the classification process. In pseudo code k-nearest neighbor classification algorithm can be expressed,

K →number of nearest neighbors
For each object *X*in the test set **do**
calculate the distance D(*X*, *Y*) between *X* and every object *Y* in the training set
neighborhood ! the *k* neighbors in the training set closest to *X*
X.class → SelectClass (neighborhood)
End for

Naïve Bayes Algorithm

The Naive Bayes Classifier technique is based on Bayesian theorem and is particularly used when the dimensionality of the inputs is high. The Bayesian Classifier is capable of calculating the most possible output based on the input. It is also possible to add new raw data at runtime and have a better probabilistic classifier. A naive Bayes classifier considers that the presence (or absence) of a particular feature(attribute) of a class is unrelated to the presence (or absence) of any other feature when the class variable is given. For example, a fruit may be considered to be an apple if it is red, round. Even if these features depend on each other or upon the existence of other features of a class, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. Algorithm works as follows,

Bayes theorem provides a way of calculating the posterior probability, *P*(*c*|*x*), from *P*(*c*), *P*(*x*), and *P*(*x*|*c*). Naive Bayes classifier considers that the effect of the value of a predictor (*x*) on a given



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute) of class.
- P(c) is called the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor of given class.
- P(x) is the prior probability of predictor of class.

class (*c*) is independent of the values of other predictors.

SVM Algorithm

SVM have attracted a great deal of attention in the last decade and actively applied to various domains applications. SVMs are typically used for learning classification, regression or ranking function. SVM are based on statistical learning theory and structural risk minimization principal and have the aim of determining the location of decision boundaries also known as hyperplane that produce the optimal separation of classes [1][2][3].Maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error [8]. Efficiency of SVM based classification is not directly depend on the dimension of classified entities. Though SVM is the most robust and accurate classification technique, there are several problems. The data analysis in SVM is based on convex quadratic programming, and it is computationally expensive, as solving quadratic programming methods require large matrix operations as well as time consuming numerical computations [4]. Training time for SVM scales quadratically in the number of examples, so researches strive all the time for more efficient training algorithm[5], resulting in several variant based algorithm.

SVM can also be extended to learn non-linear decision functions by first projecting the input data onto a high-dimensional feature space using kernel functions and formulating a linear classification problem in that feature space [4]. The resulting feature space is much larger than the size of dataset which are not possible to store in popular computers. Investigation on this issues leads to several decomposition based algorithms. The basic idea of decomposition method is to split the variables into two parts: set of free variables called as working set, which can be updated in each iteration and set of fixed variables, which are fixed at a particular value temporarily. This procedure is repeated until the termination conditions are met[5]. Originally, the SVM was developed for binary classification, and it is not simple to extend it for multi-class classification problem. The basic idea to apply multi classification to SVM is to decompose the multi class problems into several two class problems that can be addressed directly using several SVMs [6].

ANN Algorithm

Artificial neural networks (ANNs) are types of computer architecture inspired by biological neural networks (Nervous systems of the brain) and are used to approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are presented as systems of interconnected “neurons” which can compute values from inputs and are capable of machine learning as well as pattern recognition due their adaptive nature.

An artificial neural network operates by creating connections between many different processing elements each corresponding to a single neuron in a biological brain. These neurons may be actually constructed or simulated by a digital computer system. Each neuron takes many input signals then based on an internal weighting produces a single output signal that is sent as input to another neuron. The neurons are strongly interconnected and organized into different layers. The input layer receives the input and the output layer produces the final output. In general one or more hidden layers are sandwiched in between the two. This structure makes it impossible to forecast or know the exact flow of data.

Artificial neural networks typically start out with randomized weights for all their neurons. This means that initially they must be trained to solve the particular problem for which they are proposed. A back-propagation ANN is trained by humans to perform specific tasks. During the training period, we can evaluate whether the ANN’s output is correct by observing pattern. If it’s correct the neural weightings that produced that output are reinforced; if the output is incorrect, those weightings responsible can be diminished.

Implemented on a single computer, an artificial neural network is normally slower than more traditional solutions of algorithms. The ANN’s

Advantages and Disadvantages of Classification Algorithm

Sr. No	Algorithm	Features	Limitations
1	C4.5 Algorithm	-Build Models can be easily interpreted. -Easy to implement -Can use both discrete and continuous values -Deals with noise.	-Small variation in data can lead to different decision trees.· -Does not work very well on a small training data set. -Overfitting.
2	ID3 Algorithm	-It produces the more accuracy result than the C4.5 algorithm. -Detection rate is increase and space consumption is reduced.	-Requires large searching time. -Sometimes it may generate very long rules which are very hard to prune. -Requires large amount of memory to store tree.
3	K-Nearestneighbor Algorithm	-Classes need not be linearly separable. -Zero cost of the learning process. -Sometimes it is Robust with regard to noisy training data. -Well suited for multimodal classes.	-Time to find the nearest Neighbours in a large training data set can be excessive. -It is Sensitive to noisy or irrelevant attributes. -Performance of algorithm depends on the number of dimensions used.
4	Naive Bayes Algorithm	-Simple to implement. -Great Computational efficiency and classification rate· -It predicts accurate results for most of the classification and prediction problems.	-The precision of algorithm decreases if the amount of data is less. -For obtaining good results it requires a very large number of records.
5	Support vector-machine Algorithm	-High accuracy. -Work well even if data is not linearly separable in the base feature space.	-Speed and size requirement both in training and testing is more. -High complexity and extensive memory requirements for classification in many cases.
6	Artificial Neural Network Algorithm	-It is easy to use, with few parameters to adjust. -A neural network learns and reprogramming is not needed. -Easy to implement. -Applicable to a wide range of problems in real life.	-Requires high processing time if neural network is large. -Difficult to know how many neurons and layers are necessary. -Learning can be slow.

parallel nature allows it to be built using multiple processors giving it a great speed advantage at very little development cost. The parallel architecture allows ANNs to process very large amounts of data very efficiently in less time. When dealing with large continuous streams of information such as speech recognition or machine sensor data ANNs can operate considerably faster as compare to other algorithms. An artificial neural network is useful in a variety of real-world applications such as visual pattern recognition and speech recognition that deal with complex often incomplete data. In addition, recent programs for text-to-speech have utilized ANNs. Many handwriting analysis programs (such as those used in popular PDAs) are currently using ANNs.

CONCLUSION

This paper focuses on various classification techniques (statistical and machine learning based) used in data mining and a study on each of them. Data mining can be used in a wide area that integrates techniques from various fields including machine learning, Network intrusion detection, spam filtering, artificial intelligence, statistics and pattern recognition for analysis of large volumes of data. Classification methods are typically strong in modeling communications. Each of these methods can be used in various situations as needed where one tends to be useful while the other may not and vice-versa. These classification algorithms can be implemented on different types of data sets like share market data, data of patients, financial data,etc. Hence these classification techniques show how a data can be determined and grouped when a new set of data is available. Each technique has got its own feature and limitations as given in the paper. Based on the Conditions, corresponding performance and feature each one as needed can be selected.

REFERENCES

1. J. Han and M. Kamber, “Data Mining Concepts and Techniques”, Elevier, 2011.
2. V. Vapnik and C. Cortes , “Support Vector Network,” *Machine Learning,* **20**; 273-297, (1995).
3. C. J. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Mining and Knowledge Discovery,* **2**; (1998).
4. H. Bhavsar, A. Ganatra, “A Comparative Study of Training Algorithms for Supervised Machine Learning”, *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231 -2307, **2**(4); (2012)
5. G. Wang, “A Survey on Training Algorithms for Support Vector Machine Classifiers”, Fourth International Conference on Networked Computing and Advanced Information Management, 2008,IEEE.
6. G Madzarov, D. Gjorgievikj and I. Chorbev, “ A Multi-class SVM Classifier Utilizing Binary Decision Tree”, *Informatica,* pp. 233-241 (2009).
7. M . Aly, “Survey on Multiclass Classification Methods”, November (2005).
8. V. Vapnik, “Statistical Learning Theory”, Wiley, New York, (1998).
9. T.Joachims, “Making large-scale support vector machine learning practical”, In Advances in Kernel Methods: Support Vector Machines, (1999).
10. J.Platt, “Fast training of SVMs using sequential minimal optimization”, In B. Sch¨olkopf, C.Burges and A.Smola (ed.), Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, 1999, 185-208.
11. D. Michie, D.J. Spiegelhalter, C.C. Taylor “Machine Learning, Neural and Statistical Classification”, February 17, (1994).
12. Delveen Luqman Abd Al.Nabi, Shereen Shukri Ahmed, *“Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)”* (ISSN 2222-2863)**4**(8); (2013)
13. Riaan Smit” An Overview of Support Vector Machines, 30 March 2011.