International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques

Sayali D. Jadhav[1], H. P. Channe[2]

[1, 2]Department of Computer Engineering, Pune Institute of Computer Technology, Savitribai Phule Pune University, Pune, India

Abstract: ***Classification is a data mining technique used to predict group membership for data instances within a given dataset. It is used for classifying data into different classes by considering some constrains. The problem of data classification has many applications in various fields of data mining. This is because the problem aims at learning the relationship between a set of feature variables and a target variable of interest. Classification is considered as an example of supervised learning as training data associated with class labels is given as input. Classification algorithms have a wide range of applications like Customer Target Marketing, Medical Disease Diagnosis, Social Network Analysis, Credit Card Rating, Artificial Intelligence, and Document Categorization etc. Several major kinds of classification techniques are K-Nearest Neighbor classifier, Naive Bayes, and Decision Trees. This paper focuses on study of various classification techniques, their advantages and disadvantages.***

Keywords: Classification, Data Mining, Classification Techniques, K- NN classifier, Naive Bayes, Decision tree

1. Introduction

Data mining involves the use of complicated data analysis tools to discover previously unknown, interesting patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods [1]. Data mining is most important analysis step of knowledge discovery in database (KDD) process. The main goal of data mining is to extract the useful information from huge raw data and converting it to an understandable form for its effective and efficient use. In common, data mining tasks can be divided into two categories: descriptive and predictive classification techniques [2].

Data classification is the process of organizing data into categories/groups in such a way that data objects of same group are more similar and data objects from different groups are very dissimilar. Classification algorithm assigns each 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 [3].

Classification techniques can handle processing of largevolume of data. It can predict categorical class labels and classifies data based on model built by using training set and associated class labels and then can be used for classifying newly available test data. Thus, it is outlined as an integral part of data analysis and is gaining more popularity. Classification uses supervised learning approach. In supervised learning, a training dataset of records is available with associated classlabels [4].

Classification process is divided into two main steps. Thefirst is the training step where the classification model is built. The second is the classification itself, in which the trained model is applied to assign unknown data object to one out of a given set of class label [5]. This paper focuses on a survey of various classification techniques that are most commonly used in data mining. The comparative study between different algorithms (K-NN classifier, Bayesian network and Decision tree) is used to show the strength and accuracy of each classification algorithm in term of performance efficiency and time complexity. A comparative study would definitely bring out the advantages and disadvantages of one method over the other. This would provide the guideline for interesting research issues which in turn help other researchers in developing innovative algorithms for applications or requirements which are not available [5].

This paper is organized as follow: Section II covers K-NearestNeighbor algorithm, section III describes Naive Bayes classifier, while decision tree induction is discussed in section IV. Finally section V covers the comparative analysis of these algorithm followed by results of implementation and conclusions.

2. K- Nearest Neighbor Classification

The K-Nearest Neighbor Algorithm is the simplest of all machine learning algorithms. It is based on the principle that the samples that are similar, generally lies in close vicinity [6].K-Nearest Neighbor is instance based learning method. Instance based classifiers are also called lazy learners as they store all of the training samples and do not build a classifier until a new, unlabeled sample needs to be classified [7]. Lazy-learning algorithms require less computation time during the training phase than eager-learning algorithms (such as decision trees, neural networks and bayes networks) but more computation time during the classification process[8][9].

Nearest-neighbor classifiers are based on learning by resemblance, i.e. by comparing a given test sample with the available training samples which are similar to it. For a data sample X to be classified, its K-nearest neighbors are searched and then X is assigned to class label to which majority of its neighbors belongs to. The choice of k also affects the performance of k-nearest neighbor algorithm [5]. If the value of k is too small, then K-NNclassifier may be vulnerable to over fitting because of noise present in the training dataset. On the other hand, if k is too large, the nearest-neighbor classifier may misclassify the test sample

because its list of nearest neighbors may contain some data points that are located far away from its neighborhood.

K-NN fundamentally works on the belief that the data is connected in a feature space. Hence, all the points are considered in order, to find out the distance among the data points. Euclidian distanceor Hamming distance is used according to the data type of data classes used[10]. In this a single value of K is given which is used to find the total number of nearest neighbors that determine the class label for unknown sample. If the value of K=1, then it is called as nearest neighbor classification.

The K-NN classifier works as follows:
1. Initialize value of K.
2. Calculate distance between input sampleand training samples.
3. Sort the distances.
4. Take top K- nearest neighbors.
5. Apply simple majority.
6. Predict class label with more neighbors for input sample.

Following example shows that there are three classes X, Y and Z as shown in figure 1. Now, it is required to find out the class label for data sample P. Here, value of K=5 and the Euclidean distance is calculated for each sample pair and it is found that four nearest neighbor samples are falling in the class label X, while single tuple belongs to class label Z. So, the sample P is assigned to class X as it is the principal class for that sample.

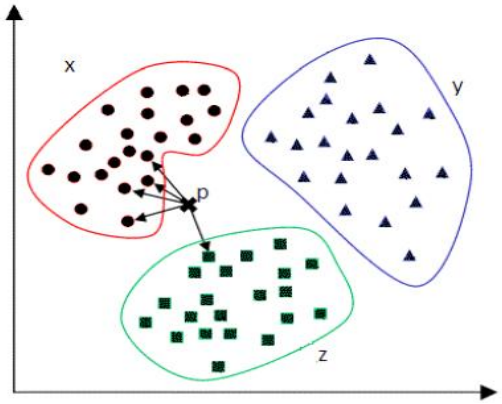

Figure 1: An example of K-NN classifier[2]

Advantages:
- Easy to understand and implement.
- Training is very fast.
- It is robust to noisy training data.
- It performs well on applications in which a sample can have many class labels[5].

Disadvantages:
- Lazy learners incur expensive computational costs when the number of potential neighbors whichto compare a given unlabeled sample is large [5].
- It is sensitive to the local structure of the data [1].
- Memory limitation.
- As it is supervised lazy learner, it runs slowly.

3. Naive Bayes Classification

Naive Bayes Classifier is the simple Statistical Bayesian Classifier [11]. It is called Naive as it assumes that all variables contribute towards classification and are mutually correlated. This assumption is called class conditional independence [12]. It is also called Idiot's Bayes, Simple Bayes, and Independence Bayes.They can predict class membership probabilities, such as the probability that a given data item belongs to a particular class label. 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.

The Naive Bayes Classifier technique is based on Bayesian Theorem and it is used when the dimensionality of the inputs is high [3]. Bayesian classification is based on Bayes Theorem and Bayes Theorem is stated as below: Let X is a data sample whose class label is not known and let H be some hypothesis, such that the data sample X may belong to a specified class C. Bayes theorem is used for calculating the posterior probability P(C|X), from P(C), P(X), and P(X|C). Where
P(C|X) is the posterior probability of target 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.

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)}$$

The Naive Bayes classifier [2] works as follows:
1. Let D be the training dataset associated with class labels.Each tuple is represented by n-dimensional element vector, X=(x1, x2, x3,.....,xn).
2. Consider that there are m classes C1, C2, C3...., Cm. Suppose that we want to classify an unknown tuple X, then the classifier will predict that X belongs to the class with higher posterior probability, conditioned on X. i.e., the Naive Bayesian classifier assigns an unknown tuple X to the class Ci if and only if P(Ci|X) > P(Cj|X) For $1 \leq j \leq m$, and $i \neq j$, above posterior probabilities are computed using Bayes Theorem.

Advantages
- It requires short computational time for training.
- It improves the classification performance by removing the irrelevant features.
- It has good performance.

Disadvantages:
- The Naive Bayes classifier requires a very large number of records to obtain good results.
- Less accurate as compared to other classifiers on some datasets.

4. Decision Tree Induction

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value [13]. Decision tree

algorithm is a data mining induction techniques that recursively partitions a dataset of records using depth-first greedy approach or breadth-first approach until all the data items belong to a particular class.

A decision tree structure is made of root, internal and leaf nodes. It is a flow chart like tree structure, where every internal node denotes a test condition on an attribute, each branch represents result of the test condition, and each leaf node (or terminal node) is assigned with a class label. The topmost node is the root node. Decision tree is constructed in a divide and conquer approach. Each path in decision tree forms a decision rule. Generally, it utilizes greedy approach from top to bottom.

Decision tree classification technique is performed in two phases: tree building and tree pruning [14]. Tree building is performed in top-down approach. During this phase, the tree is recursively partitioned till all the data items belong to the same class label. It is very computationally intensive as the training dataset is traversed repeatedly. Tree pruning is done in a bottom-up manner. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting problem of tree. Over-fitting problem in decision tree results in misclassification error.

There are many decision tree based algorithms like ID3, C4.5, C5.0, CART etc. These algorithms have the merits of high classifying speed, strong learning ability and simple construction [15]. Decision tree can be explained with an example as depicted below.

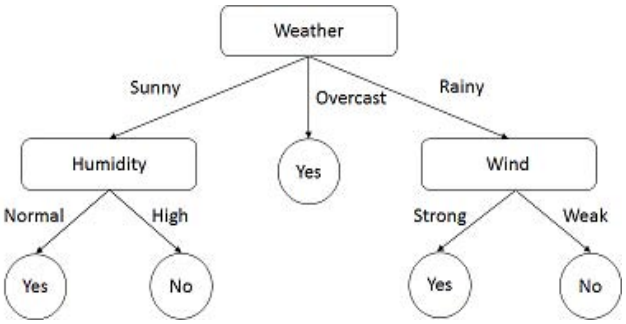

Figure 2: An example of decision tree induction

Example shows a weather forecasting process which deals with predicting whether the weather is sunny, overcast or rainy and the amount of humidity if it is sunny. This tree model can be applied to determine whether the atmosphere is suitable to play the tennis or not. So, a person can easily find the present climate and based on that decision can be made whether match can be possible or not.

Advantages:

- Decision Trees are very simple and fast.
- It produces the accurate result [1].
- Representation is easy to understand i.e. comprehensible.
- It supports incremental learning [5].
- It takes the less memory.
- It can also deal with noisy data.
- It uses different measures such as Entropy, Gini index, Information gain etc.to find best split attribute.

Disadvantages:

- It has long training time.
- Decision trees can have significantly more complex representation for some concepts due to replication problem [5].
- It has a problem of over fitting [15].

5. Comparison among K-NN, Naive Bayes and Decision Treetechniques

Table I shows the comparison between K-NN, Naive Bayes and Decision Tree Techniques

Table I

Parameter	*KNN*	*Naive Bayes*	*Decision Tree*
Deterministic/ Non-deterministic	Non-deterministic	Non-deterministic	Deterministic
Effectiveness on	Small data	Huge data	Large data
Speed	Slower for large data.	Faster than KNN.	Faster
Dataset	It can't deal with noisy data.	It can deal with noisy data.	It can deal with noisy data.
Accuracy	Provides high accuracy.	For obtaining good results it requires a very large number of records.	High accuracy

6. Results

Following tables shows summary of results of implementation of different classifiers using WEKA tool. Table II shows the results of accuracy of classifiers. Table III shows the results of time taken by classifiers for classifying given datasets.

Table 2: Results of Accuracy of Classifiers

Dataset	Size of Dataset	KNN	Naïve Bayes	Decision Tree
Weather Nominal	Small (14 instances)	100%	92.857%	100%
Segment Challenge	Medium (1500 instances)	100%	81.667%	99%
Supermarket	Large (4627 instances)	89.842%	63.713%	63.713%

Table 3: Results of Time taken for Classification

Dataset	Size of Dataset	Time	KNN	Naïve Bayes	Decision Tree
Weather Nominal	Small (14 instances)	To Build Model	0 sec	0 sec	0.02 sec
		To Test Model	0.02 sec	0 sec	0 sec
Segment Challenge	Medium (1500 instances)	To Build Model	0 sec	0.08 sec	0.16 sec
		To Test Model	0.42 sec	0.31 sec	0.06 sec
Super market	Large (4627 instances)	To Build Model	0.02 sec	0.06 sec	0.06 sec
		To Test Model	45.55 sec	0.28 sec	0.03 sec

7. Conclusion

From survey and analysis on comparison among data mining classification algorithms (Decision tree, KNN, Bayesian), it shows that all Decision Tree's algorithms are more accurate and they have less error rate and they are easier algorithms as compared to K-NN and Bayesian. The knowledge in decision tree is represented in the form of [IF-THEN] rules which is easier for humans to understand. The result of implementation in WEKA on the same dataset showed that Decision Tree outperforms and Bayesian classification having the same accuracy as of decision tree but other predictive method like K-NN does not giving good results. The comparative study has shown that each algorithm has its own set of advantages and disadvantages as well as its own area of implementation. None of the algorithm can satisfy all constrains and criteria. Depending on application and requirements, specific algorithm can be chosen.

References

[1] S.Archana and Dr. K.Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Vol. 2 Issue. 2, February 2014.
[2] BhaveshPatankar and Dr. Vijay Chavda, "A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 12, December 2014.
[3] Sagar S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms", Oriental Journal of Computer Science & Technology, Vol. 8, April 2015.
[4] Meenakshi and Geetika, "Survey on Classification Methods using WEKA", International Journal of Computer Applications, Vol. 86, No.18, January 2014.
[5] H. Bhavsar and A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, Issue. 4, September 2012
[6] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification", IEEE Transactions on Information Theory, vol. 13, No. 1, pp. 21-27, 1967.
[7] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Elevier, 2011.
[8] K. P. Soman, "Insight into Data Mining Theory and Practice", New Delhi: PHI, 2006.
[9] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica, vol. 31, pp. 249-268, 2007.
[10] M. Soundarya and R. Balakrishnan, "Survey on Classification Techniques in Data mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2014.
[11] R. Duda, and P. Hart, "Pattern Classification and Scene Analysis", John Wiley and Sons, New York, 1973.
[12] N. Friedman, D. Geiger, and Goldazmidt, "Bayesian Network Classifiers", Machine Learning, vol. 29, pp. 131-163, 1997.
[13] "Decision tree learning" pdf.
[14] Matthew N. Anyanwu and Sajjan G. Shiva, "Comparative Analysis of Serial Decision Tree ClassificationAlgorithms", Researchgate, January 2009.
[15] Brijain R. Patel and Kushik K.Rana, "A Survey on Decision Tree Algorithm for Classification", International Journal of Engineering Development and Research, 2014.