Survey on Classification Techniques for Data Mining

Vivek Agarwal
Dept. of Comp. Engg. Smt.
Kashibai Navale College Of
Engineering, Pune

Saket Thakare
Dept. of Comp. Engg. Smt.
Kashibai Navale College Of
Engineering, Pune

Akshay Jaiswal
Dept. of Comp. Engg. Smt.
Kashibai Navale College Of
Engineering, Pune

ABSTRACT

This paper focuses on the various techniques that can be implemented for classification of observations that are initially uncategorized. Our objective is to compare the different classification methods and classifiers that can be used for this purpose. In this paper, we study and demonstrate the different accuracies and usefulness of classifiers and the circumstances in which they should be implemented.

General Terms

Classification, Sentiment, Review, Accuracy, Positive, Negative, Neutral.

Keywords

Classifiers, Naïve Bayes, SMO, Decision Tree, SVM, Sentiment, Analysis.

1. INTRODUCTION

In the present world scenario, we are floating with tremendous amount of data around us. In order to extract useful information from this data, we need to process this data and classify it using various classifiers. In this work, we aim to study the methods of classifying an observation using machine learning techniques and the different classifiers involved. We shall compare and contrast, the advantages and disadvantages associated with the classifiers under consideration and study their efficiency and accuracy. We consider various research works based on these classification techniques and draw conclusions based on their end-results.

2. CLASSIFICATION

Classification is the process of categorizing an observation, by using training set of data containing observations whose category is known beforehand. There are various classification algorithms used for this purpose by different classifiers. The classifiers are basically associated with a mathematical function, which helps in mapping a test observation to its category.

2.1 Naïve Bayes.

Naïve Bayes is a probability based classifier used for learning a categorized set of documents. It is based on the Baye's theorem. Geetika Gautam [1] has used this classifier for both training and classification stages. Geetika has used this classifier to extract the sentiments related to a twitter user's review regarding a product.

$$\boldsymbol{P_{NB}(c/d)} = \frac{\boldsymbol{(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)})}}{\boldsymbol{P(d)}}$$

$C^{*}=argmac_{c}\ P_{NB}(c|d)$

Here, the class c* is assigned to a tweet d, where, f represents a feature and n_i(d) represents the count of feature f_i found in tweet d. Total number of features are m and the parameters P(c) and P(f|c) are obtained using maximum likelihood estimates.

In order to maximize the entropy defined on conditional probability distribution, we can calculate the maximum entropy.

$$\boldsymbol{P_{ME}(c|d)} = \frac{\mathbf{exp}[\sum_i \leftthreetimes_i f_i(c|d)]}{\sum_c \mathbf{exp}[\sum_i \leftthreetimes_i f_i(c,d)}$$

Where c is the class, d is the tweet and $\leftthreetimes$ is a weight vector. The weight vector decides the significance of a vector in classification.

The Naïve Bayes classifier assumes all features to be conditionally independent [2]. Even though this classifier yields good results in [3], it hasn't shown superior results as compared to some other classifiers.

The performance measure of Naïve Bayes found by [1] is as follows.

Table 1 Naive Bayesian Classification Measurements

Positive Recall	91.2%
Negative Recall	85.4%
Positive Precision	49.3%
Negative Precision	39.3%

2.2 Support Vector Machines (SVMs)

The algorithms used by SVMs are based on kernel substitution method. These can be defined as systems which uses hypothesis space of linear functions and in a high dimensional feature space. Using SVMs, we can construct non-linear classification without being stuck in local minima [4]. The SVMs defines decision boundaries and then kernels are used for computation. The input data in [1] is two sets of vectors of size m each. The task would be to find a margin between two classes that is far from any document. SVMs also support concepts of classification and regression. These are helpful in statistical learning and recognizing factors precisely[5].

The performance measure of SVMs found by [1] is as follows.

Table 2 Naive Bayesian Classification Measurements

Positive Recall	88.3%
Negative Recall	83.5%
Positive Precision	43.8%
Negative Precision	35.7%

2.3 Sequential Mining Optimization (SMO)

The SMO efficiently solves the optimization problems when training support vector machines. It uses Iterative approach to solve optimization problem and breaks it into series of smaller sub-problems and solve it analytically. [1] demonstrates that each such sub-problem involves two language multipliers. The constraints can be reduced to the following equations 0≤α1, α2≤C. y1α1+y2α2=k which is solved analytically. The purpose of the algorithm is to find a multiplier that violates KTT conditions and picks second multiplier that optimizes the pair.

2.4 Decision Trees

Decision tree is a predictive modeling based technique developed by Rose Quinlan .It makes use of recursive tree structure [6] and is a sequential classifier. There are three kinds of nodes in the decision tree. The node from which the tree is directed and has no incoming edge is called the root node. A node with outgoing edge is called internal or test node whereas all the other nodes are called leaves (also known as terminal or decision node) [7]. The data set in decision tree is analyzed by developing a branch like structure with appropriate decision tree algorithm. Each internal node of tree splits into branches based on the splitting criteria. Each test node denotes a class. Each terminal node represents the decision. They can work on both continuous and categorical attributes [8].

A decision tree is a flow chart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The decision tree classifier has two phases

i) Growth phase or Build phase.

The gini index is used to measure the impurity of data partition as follows.

$$\boldsymbol{Gini(t) = 1 - \sum_{j} \quad [p(j|t)]^2}$$

Where p(j | t) is the relative frequency of class j at node t.

Entropy is an information-theoretic measure of the “uncertainty‟ contained in a training set, due to the presence of more than one possible classification.

$$\boldsymbol{Entropy(t) = -\sum_j p(j|t) log_2 p(j|t)}$$

Where p (j | t) is the relative frequency of class j at node t.

Information gain is the change in information entropy from prior state to a state that takes some information. It measures reduction in entropy achieved because of split. It is given by

$$\boldsymbol{Gain_{split} = Entropy(p) - (\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)\,)}$$

where Parent Node, p is split into k partitions and ni is number of records in partition i.

Gain Ratio is the variant introduced by the Australian academician Ross Quinlan in his influential system C4.5 in order to reduce the effect of the bias resulting from the use of information gain. Gain Ratio adjusts the information gain for each attribute to allow for the breadth and uniformity of the attribute values.

$$\boldsymbol{GainRatio_{split} = \frac{Gain_{split}}{SplitInfo}}$$

ii) Pruning phase

This phase avoids model over fitting [9]. The issue of over fitting arises due to random errors, noise in data or the coincidental patterns, which can lead to strong performance degradation.

The work of D.Lavanya [10] is based on Evaluating Performance of Decision Tree Classifiers on the basis of Medical Datasets. They aim to compare the classification accuracies of different algorithms as follows.

Table 3 Classifiers Accuracy

Data Set	ID3 Accuracy (%)	C4.5 Accuracy (%)	CART Accuracy (%)
Diabetes	57.5	73.8	75.1
Heart Statlog	61.4	76.6	78.5
Thyroid	65.60	67.92	69.16
Breast Cancer	90.41	94.56	94.84
Arrhythmia	42.69	64.38	70.57

2.5 K-Nearest Neighbor Classifier

The KNN approach is used when the data is present in the feature space. The data can be scalars or maybe multidimensional vectors. As, these point are present in feature space, they have a notion of distance associated with them. This need not be the Euclidian distance, although it is more commonly used with it. The training data consists of a set of vectors and a class label associated with each vector. In simplest cases it may be either positive or negative, for example extracting a positive or negative sentiment from a review. KNN can also work with multiple classes. The k in KNN represents the total number of neighbors that influence the classification. Choice of k is very critical, a small value of k means that noise will have a higher influence on the result. A large value makes it computationally expensive and kind of violates the basic philosophy behind KNN. It is a type of instance base learning or lazy learning. It is very simple but its accuracy can be affected by noisy or irrelevant features. It is thus, a little less preferred over the other classification methodologies.

3. COMPARISON

A comparison of the various classifiers and their accuracies has been performed. The results obtained are as follows

Table 4 Accuracy Comparison of Classifiers.

Classifiers	Accuracy
Naïve Bayes	85.5
Support Vector Machines	86.2
SMO	84.9
Decision Trees	85.3

The Fig 1. Depicts graphically, the comparison of the various classifiers under survey.

4. CONCLUSION

In this paper, we have surveyed the behavior of multiple classifiers and the mathematical functions that their algorithms follow. We have encountered some advantages and disadvantages associated with each classifier and compiled a comparative study of their accuracies. We have concluded that Naïve Bayes’ classifier is one of the simplest and quicker algorithms around. The convergence of Naïve Bayes classifier is relatively quicker and requires lesser training data. Its main disadvantage is that it doesn’t learn interaction between features. SVMs have high accuracy, handle overfitting efficiently and work well with appropriate kernel, even if data is not linearly separable. The disadvantages of SVMs are that it is highly memory intensive and hard to interpret. Decision Trees have good accuracy, easy to interpret, handles feature interactions and they are non-parametric. The disadvantage with Decision Trees is that you have to rebuild it when new examples come on and can easily overfit. SMO can handle optimization problems when used with SVMs using Iterative approach and breaking the problem into sub-problems.

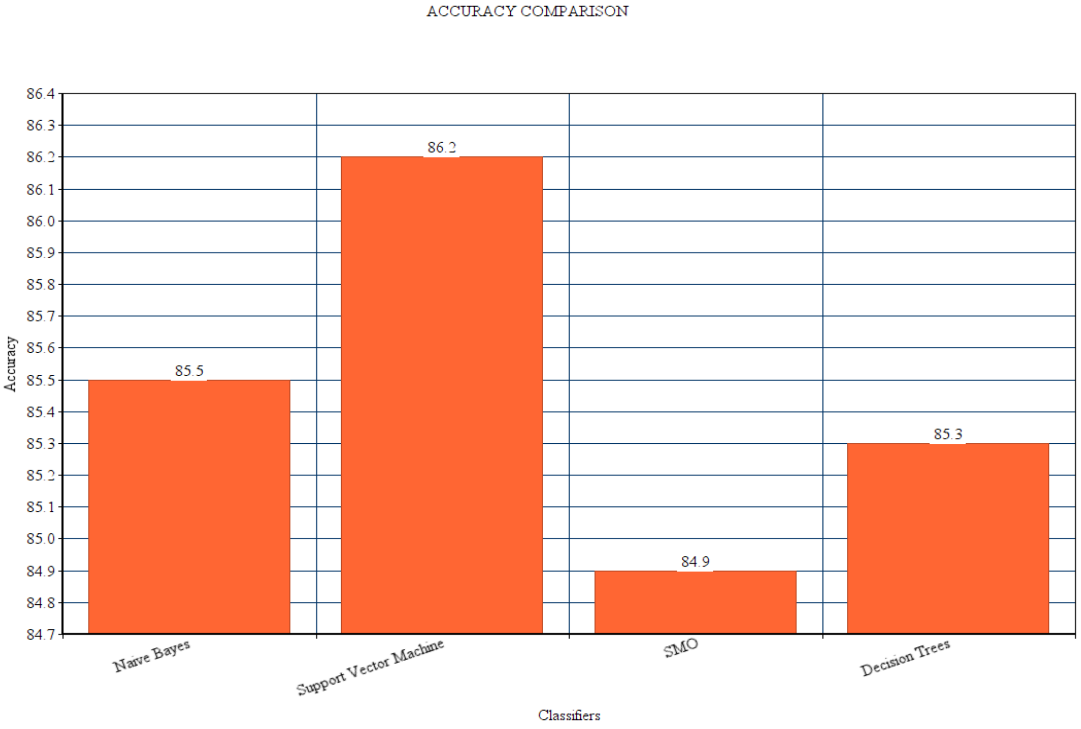

Fig 1: A comparison of the accuracies of some of the classifiers.

5. ACKNOWLEDGMENTS

We would like to thank all the experts and those who have researched the various classifiers for helping me compile this paper. We are grateful to all our professors for believing in us and our college for giving us this opportunity. We thank all the authors and data scientists who have contributed into the formulations of all these techniques and helped data mining.

6. REFERENCES

[1] Geetika Gautam, Divakar yadav, “Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis”, IEEE 2014 International Conference, pp 437-442.

[2] K.P.Murphy. (2006). Naive Bayes classifiers. [Online].Available:http://www.cs.ubc.calmurphyk/TeachinglCS340-Fall06/readingiNB. pdf.

[3] A. Pak and P.Paroubek, "Twitter as a Corpus for Sentiment Analysisand Opinion Mining", in Proc. 7 thconference on International Language Resources and Evaluation LREC'lO ,May 2010.

[4] K.P.Bennet and C.Campbell. "Support Vector Machines: Hype or Hallelujah?" in Proc. SIGKDD Explorations, 2000, vol. 2, no. 2, pp 1-13.

[5] M.A. Hearst,"Support vector machines," IEEE Intelligent Systems, pp.18-28, 1998.

[6] Quinlan, J. R. (1987). "Generating production rules from decision trees". Proceedings of the 10th international joint conference on Artificial intelligence , pp. 304-307.

[7] Maimon, O., & Rokach, L. (2010). "Data Mining and Knowledge Discovery Handbook". (2nd, Ed.) Springer

[8] Han, J., & Kamber, M. (2006). "Data Mining: Concepts and Techniques" (2nd ed.). Morgan Kaufmann Publishers.

[9] Barros, R. C., Basgalupp, M. P., Carvalho, A. C., &Freitas, A. A. (2010, Jan). "A Survey of Evolutionary Algorithms for DecisionTree Induction". IEEE Transactions on Systems,Mans and Cybernetics, Vol. 10,No. 10,pp. 1-22.

[10] D.Lavanya, Dr. K.Usha Rani," Performance Evaluation of Decision Tree Classifiers on Medical Datasets", IJCA Vol. 26, July 2011.