International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue III, March 2018- Available at www.ijraset.com

Comparison Study: Classification Algorithms

Dr. Amita Goel[1], Saarthak Mahajan[2]
[1]*Associate Professor in Department of Informatin Technology, Maharaja Agrasen Institute of Technology)*
[2]*Department of Information Technology Maharaja Agrasen Institute of Technology Sector-22, Rohini, New Delhi-110086, India*

Abstract: Classification is one of the most important component of data mining that is used to identify a new observation into a unique category, on the basis of a training set. Specifically, the use of algorithms are the core engines behind the proper functioning of an Optical Character Recognition (OCR) used in various applications. An algorithm that implements classification is called a classifier. The classifier can be understood as a function that maps input data into a unique category. It is constructed from labeled training data that is used to predict the class label of a new data. In this paper, we discuss various classification algorithms and compare them with each other. We shall discuss the performance characteristics of each algorithm based on various parameters.
Keywords: Classification, Classifier, Category

I. INTRODUCTION

The ability to classify an unlabelled data into a new category is known as Classification [1]. In the context of Machine learning, classification is supervised learning as it is a learning where a training set of correctly identified observations are obtainable.
Unsupervised learning is known as clustering and involves putting data into a unique category based on some similarity or distance. An algorithm that maps input data to a category is a classifier. The terminology might vary according to the field. Classification in statistics is done with logistic regression, the observations are termed as explanatory variables and the categories to be predicted are outcomes. These are known as classes in Machine Learning and the observations are known as instances. It's been many decades since the research on how to incorporate this power into computers. This is done with the help of Machine Learning algorithms used for classification of characters and many other areas where an unlabelled data needs to be classified. This power is implemented with various concepts. OCR, deals with the recognition of characters that are optically processed. It has various applications in today's world – License Plate Detection, Pattern Recognition etc. It considerably helps to improve the interaction between humans and machines. It incorporates – Image Classification, labelling of objects or images into pre-defined categories. Although easy for humans, it is not so easy to implement Image classification in machines. It requires use of complex algorithms to correctly classify the unlabelled data. For past many years, huge amount of research has been done on printed recognition. A highly accurate system of any recognition system is dependent on proper functioning of classification algorithms.
Classification can be done on structured and unstructured data. The main goal of a classification problem is to identify the category the data belongs to. Ultimately, the classification algorithm leads to the formation of a classification model which is the function of the particular algorithm. There are many types of classification per se, Binary classification is the one with two linear outcomes. Multi classification is classification with more than two outcomes. Multi label classification is when each sample is mapped to a set of target labels.
In this paper, we discuss various algorithms to detect and classify unlabelled data. We also identify areas in these algorithms which could be improved for higher accuracy. In the end, performance parameters such as Level of Accuracy gives the final verdict on a particular algorithm.

II. CLASSIFICATION MODEL

The most important thing is to build a classification model [2]. The steps involved were to initialize the classifier to be used. The next step would be to train the classifier which uses a method to fit the model (training) for the given train data X and train data Y. After this we predict the target when an unlabelled observation returns a predicted label Y. Soon after the classifier model is evaluated. Classification predictive models are used to solve business problems involving non-numeric data. They predict categorical class labels; and prediction models predict continuous valued functions.
Some of the common classification models are linear models [3], decision trees, and naïve Bayes models. For very large datasets the linear models are relatively easier to scale. Decision tree is a powerful nonlinear technique that can be a little more difficult to scale

up and more computationally intensive to train, but delivers leading performance in many situations. Naïve Bayes models are more simple but are easy to train efficiently and parallelize (in fact, they require only one pass over the dataset).
They can also give reasonable performance in many cases when appropriate feature engineering is used. A naïve Bayes model also provides a good baseline model against which we can measure the performance of other models.

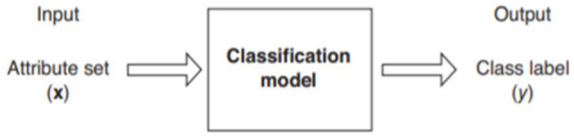

Figure 1: Classification as a function

A classification model is used to predict the class label of unknown records. From the above figure a classification model can be treated as a black box that automatically assigns a class label when presented with the attribute of an unknown record.

III. DIFFERENT ALGORITHMS

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces predicted instances. The prediction accuracy defines how "good" the algorithm is.

A. Linear Discriminant Analysis (LDA)

It is a method used in statistics [4], pattern recognition and machine learning to find a linear combination of features that distinguishes two or more classes or events. The combination could be used for linear classification or to reduce the dimensionality for later classification. In bankruptcy prediction based on accounting ratios and other financial variables, linear discriminant analysis was the first statistical method applied to systematically explain which firms entered bankruptcy vs. survived. It is also used in fields of face recognition, marketing etc.

B. Logistic Regression

Logistic regression was developed by statistician David Cox in 1958. The binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor. The model is a direct probability model and not a classifier. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences.

C. Naïve Bayes Classifier

It is a classification technique based on Bayes' theorem with an assumption of independence between predictors. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple.
Naive Bayesian model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

D. Support Vector Machines

In machine learning, support vector machines [5] (SVMs, also support vector networks) analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other. Our main goal is to find a line that uniquely divides the data into two regions. Such data which can be divided into two with a straight line (or hyperplanes in higher

dimensions) is called Linear Separable. Hence, SVM finds a straight line (or hyperplane) with largest minimum distance to the training samples.

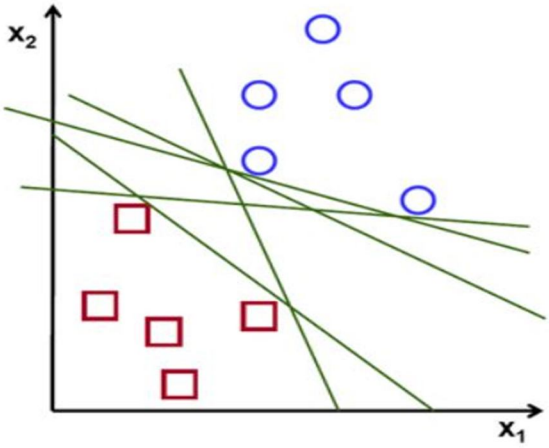

Figure 2: SVM

E. Kernel Classification: KNN

In K-nearest neighbors (KNN), [6] the input consists of k closest training samples in the feature space. The output depends on whether KNN is being used for regression or classification. It is a type of lazy learning where the function is only approximated locally and all computation is deferred until classification. It is quite sensitive to the local structure of the data.

F. Decision Trees

It is a type of supervised learning algorithm that is mostly used for classification problems. It works for both categorical and continuous dependent variables. In this algorithm [7], we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible. A decision tree is a simple representation for classifying examples. In these tree structures, leaves represent class label and the internal nodes or non-leaf node is labeled with an input feature,

IV. STUDY WITH DATA SET

The data set [8] that we've used to train on are the salaries of various employees with 2 classes – (>=50K and <=50K). There are 48,842 rows with 7 explanatory variables (Age, Capital Gains, Education, Hours per week, Marriage Status and Relationship, Occupation, Race and Sex). We have used the python language to run the above algorithms mentioned.

A. Logistic Regression

In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

B. Stochastic Gradient Descent

imple approach to fit linear models and is useful when sample is very large. But it requires a number of hyper-parameters and is sensitive to feature scaling.

C. Naïve Bayes

Based on Bayes theorem, it works on the assumption of independence between every pair of features. They work well in classifying Spam emails and documents.

D. Support Vector Machines

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap .New examples are then mapped into the space and predicted to belong to a category based on which side of the gap they fall.

E. Decision Trees

A decision tree produces rules for classifying unlabeled data. They can create complex trees which don't generalize well in case of the addition of new observed data.

F. Random Forest

Ranom forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

G. K-Nearest Neighbors

It is a type of lazy learning and only stores instances of trained data. The only drawback is that we need to determine the value of K and the computation cost is high as it needs to computer the distance of each instance to all the training samples.

To compare various algorithms we will form a comparison matrix which will have two parameters – Accuracy and F1-Score. Accuracy is a ratio of correctly predicted observation to the total number of observations. F1-Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. F1-Score is usually more useful than accuracy, especially if you have an uneven class distribution. After running the code along with the data set we get the following observations [9] –

Classification Algorithm	Accuracy	F1-Score
Logistic Regression	84.60%	0.6337
Naïve Bayes	80.11%	0.6005
Stochastic Gradient Descent	82.20%	0.5780
K-Nearest Neighbors	83.56%	0.5924
Decision Tree	84.23%	0.6308
Random Forest	84.33%	0.6275
Support Vector Machine	84.09%	0.6145

TABLE 1: Accuracy of Classifiers

V. CONCLUSION

The algorithms above don't show much deviation in terms of the accuracy measured against each other. The performance of the classifier and the classification model highly depends on the type of the data that is being tested. Determining a suitable classifier for a given problem is however still more an art than a science.

We also need to understand that in terms of classification there are of problems in the world that are being addressed. A suitable type of Algorithm is used to solve that particular classification problem.

VI. ACKNOWLEDGMENT

I would like to thank Dr. Amita Goel for her immense help and support, useful discussions and valuable recommendations.

REFERENCES

[1] https://en.wikipedia.org/wiki/Statistical_classification
[2] https://www-users.cs.umn.edu/~kumar001/dmbook/ch4.pdf
[3] https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-classification
[4] https://en.wikipedia.org/wiki/Fisher%27s_linear_discriminant
[5] https://docs.opencv.org/3.0beta/doc/py_tutorials/py_ml/py_svm/py_svm_opencv/py_svm_opencv.html- 94
[6] https://docs.opencv.org/3.0beta/doc/py_tutorials/py_ml/py_knn/py_knn_opencv/py_knn_opencv.html - 93.22
[7] Amit Gupta, Ali Syed, Azeem Mohammad, Malka.N. Halgamuge, "A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in DenverCity the USA", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 201
[8] http://www.census.gov/ftp/pub/DES/www/welcome.html
[9] https://analyticsindiamag.com/7-types-classification-algorithms/