An Analysis on Ensemble Classifiers in Ensemble Classification Problems

P. Naveen Sundar Kumar, V. V. Nagendra Kumar, K. E. Naresh Kumar

Abstract*: Day to Day the amount of data was increasing rapidly. Due to analyzing the huge amount of data various technologies are also introduced. Traditional data mining approaches can be used to perform data analysis through classification algorithms. In data mining a single classifier can be used to perform data analysis. Sometimes, multiple or combined classifier can also be used to perform data analysis. But, the performance of ensemble classifier is better than single classifier. Based on improved accuracy the various number of ensemble classifiers are introduced. Now, this paper can reviews various ensemble classifiers based on their accuracy.

Index Terms*: Classification, Classifier Ensemble, Ensemble Classification.

I. INTRODUCTION

To perform data analysis, by learning the complex data and making intelligent decisions in easier way based on machine learning. In data mining, the machine learning is mainly used to automate the classification by constructing good models from given large amount of data/datasets. The major problem in machine learning is to perform the supervised learning. It is similar to classification by learning the training dataset and then constructing the effective data model to perform testing of data for deriving results.

Active learning is machine learning approach used in learning process. It optimizes the model quality by actively posing constraints and getting knowledge from users to labeling the data. In present generation the data analysis is a major problem due to having the large amount of data. To perform the classification by analyzing the data to extract models describing data classes. These models can define as classifiers.

The implementation of classification can be performed in two steps:

i. Learning Phase
ii. Classifying Phase

Learning phase involves the training dataset by constructing the model. In second step, the model can be referred for classifying process. Based on the entire process the classifier accuracy can be measured. To estimate the accuracy of classifier the training data can be used. Sometimes, the classifier tends to over fit the data. To over fit the data due to classifier causes a problem in classification approach. Therefore the major problem in classification problem is to maintain the accuracy.

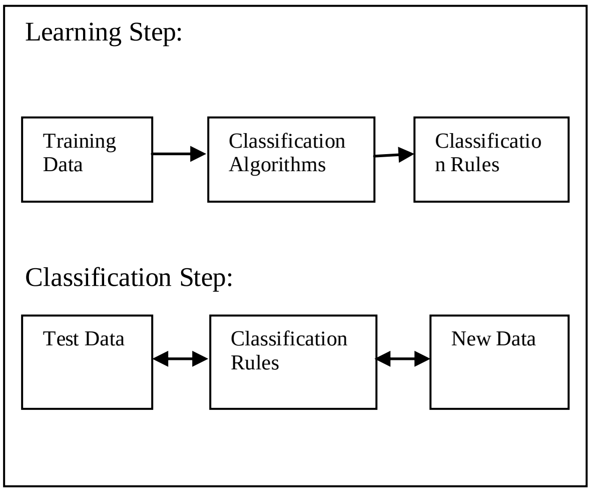

Figure 1: Classification Process

To increase the accuracy of classification then ensemble methods can be used. Ensemble methods can train more number of classifiers to solve the similar problem. To differentiate the ensemble from general classification algorithm by constructing a set of classifiers/learns and then combining them. Ensemble learning can be referred as committee based learning or learning multiple classifier system. The framework of ensemble classification consists of set of classifiers. The root of ensemble methods is an earlier than the introduction of Occam's razor. The idea of Ensemble is introduced by the Greek philosopher Epicurus (341-270 B.C), who derived the principle of multiple explanations. The ensemble methods are mainly implemented in various alternative ways as combining classifiers, ensembles of weak learners and mixture of experts (uses divide and conquer). The ensemble is better than the best single classifier by reducing error rate with observation of Hansen and salamon's [1990].

The construction of an ensemble is two-step process. i.e., generating classifiers and second step of combining them. When the base learner is accurate and diverse then ensemble is also to be good. The construction of single classifier computational cost is required for constructing an ensemble In general Ensemble methods can make use of single classification algorithms to generate homogeneous ensembles. But some other ensembles use multiple learning algorithms to produce heterogeneous learners. Mixture of Learners can also lead to improve the performance of classification.

Revised Manuscript Received on July 22, 2019.

P. Naveen Sundar Kumar, Assistant Professor, Dept. of CSE, RGMCET, Nandyal

V.V. Nagendra Kumar, Assistant Professor, Dept. of MCA, RGMCET, Nandyal

K.E. Naresh Kumar, Assistant Professor, Dept. of CSE, RGMCET, Nandyal

Retrieval Number: I30700789S319/2019©BEIESP
DOI: 10.35940/ijitee.I3070.0789S319

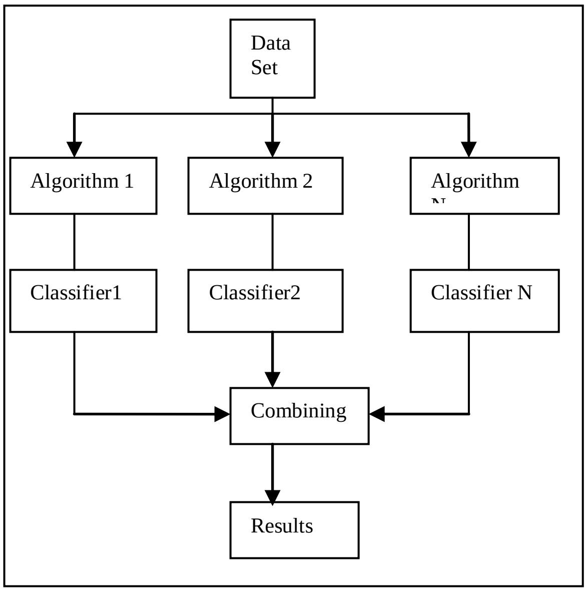

Figure 2: Ensemble Construction Process

II. POPULAR ENSEMBLE METHODS

The ensemble methods are combining classifiers derived from single base learning algorithm or different base learning algorithms. It allows the base learner (weak Learner) to boosting into strong learners.

The popular ensemble methods are:

a. Bagging
b. Boosting
c. Stacking
d. Random Forests

A. Bagging

It can derive as Bootstrap Aggregating. It performs bootstrap and Aggregation. It generates different classifiers or base learners based on adopting bootstrap distribution. Based on bootstrap sampling, it can provides various data subsets for training classifiers or base learners. For instance, consider an example of any training dataset consists of ‘m’ training examples. Each sample ‘m’ obtained by sampling with replacement. The subset size can be less than the original dataset; each subset applies to an individual classifier/model. Multiple subsets used for implementing predictions in parallel and not depending on each other. Finally, these predictions are combined together to make a final decision based on majority by voting for classification.

The algorithm to implement Bagging is:

B. Boosting

The boosting can generates a strong classifier based on various numbers of weak learners. It is a sequential process in which every model can be defined through the correction of errors of its previous model. It allows weights to each model. It mainly performs by creating a subset from the original dataset. Initially, all data points are assigned with equal weights based on base model. The Predictions can be made on the whole dataset using base model. Difference of actual and predicted values determines errors. The next model is created on the entire dataset by correcting the errors of previous model. Based on the process multiple models are created. Finally combining the predictions of all weak learners with weighted mean can be calculated Strong Learner.

C. Stacking

It can perform the predictions by combining predictions of multiple models. It can implement by splitting the training dataset into number of divisions. The base model is fitted on 1 to n-1 parts, the prediction can be made on ‘nth’ part. This process repeats for another base model to make predictions on training dataset and test dataset. These combined prediction results can be used to construct a new model to make a final prediction.

D. Random Forests

It is an extension to bagging. The base estimators used for random forests are decision trees. Initially it can create random subsets from original dataset using bootstrapping. At each node in decision tree, random set of features used to decide best split. The decision can be fitted for each subset. Then, the final prediction can be made by calculating average of predictions from all decision trees.

III. WHY ENSEMBLES BETTER THAN SINGLE CLASSIFIER

Ensembles can combine multiple models into single classifiers. These Ensemble methods are extensively used in various applications like medical diagnosis, Bio Medical imaging and Credit scoring etc. Due to exhibition of better performance of ensembles rather than single classifiers. These Ensembles are also used to reduce the model variance and bias also. The main conclusion to be derived as “Why Ensembles are better than single classifiers” by performing the following study of Comparison. The study of observation performs on four different data sets to comparing the classification accuracy of two ensemble classifiers Boosting Variant of Gradient Boosting or Random Forest and the other set of 2 Classifiers are Logistic Regression and Neural Networks. The comparison can only provide analysis of measuring the performance of four distinct classifiers including ensembles also.

First, let consider two single classifiers are implemented to construct their individual model to comparison. These single models can be statistical models (), Semi Parametric models (), Tree based models () etc. These models can perform various comparisons to perform the final classification with simple conclusion.

A. Ensemble Classifier

Ensemble methods can be defines as a set of multiple base classifiers are combined to predict the performance of a final classifier. Many theorems and Practical derivations are proved that combined models can performs increased classification accuracy. (Finlay, 2011; Paleologo, et al., 2010).

These Ensembles allows to create base classifiers either in dependent or independent manner. Suppose, Consider a bagging ensemble performs an independent model creation using Bootstrap subsets of sampling with replacement of given dataset. (Breiman, 1996). In another format of ensembles the Boosting approach is a dependent ensemble. It allows classifiers/models are trained at every iteration by correcting errors of previous models into new ensemble (Freund & Schapire, 1996). Various versions of boosting and bagging are introduced with several advantages. (Breiman, 2001; Friedman, 2002; Rodriguez, et al., 2006).

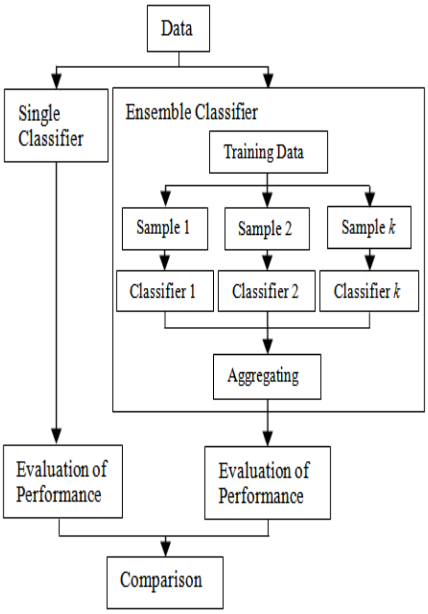

Figure 3: Utami, et al., 2014 reference diagram of single learner vs Ensemble.

IV. EXPERIMENTAL SET-UP

Consider four different datasets with partitioning of 70% training and the remaining 30% test dataset. The following Four datasets are used to perform the prediction of classification accuracy using a single classifier and ensemble classifiers.

Cardata: It consists of Auto dataset to determine whether to buy second hand car or not having 1,020,455 rows and 27 variables.

Time series data: It consists of time series data to predict stock price change index having 1,688,948 rows and 36 variables.

HMEQ: It describes Loan datasets to determine if customers will default or not having 5,960 rows and 13 variables.

Based on Keel, a data mining software to predict the performance of classifiers the following results are obtained by determining the misclassification rate of each classifier on given datasets.

Table 1: Results showing misclassification rates of all dataset

Car Data:

Random Forests	0.02
Neural Networks	0.105
Linear Regression	0.111
Gradient Boosting	0.131

Time Series Data:

Random Forests	0.258
Neural Networks	0.175
Linear Regression	0.179
Gradient Boosting	0.4

HMEQ Dataset:

Random Forests	0.09
Neural Networks	0.11
Linear Regression	0.12
Gradient Boosting	0.1

Based on the above results ensemble Classifiers (Random Forests and Gradient Boosting are producing better classification accuracies than single classifier (Neural Networks and Linear Regression).

V. CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

REFERENCES

1. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54, 627-635.
2. Breiman, L. (1996). Bagging predictors. Machine Learning, 24, 123-140
3. Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.
4. Freund, Y., & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. In L. Saitta (Ed.), Proc. of the 13th Intern. Conf. on Machine Learning (pp. 148-156). Bari, Italy: Morgan Kaufmann.
5. Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38, 367-378.
6. Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. European Journal of Operational Research, 210, 368-378.
7. Lessmann, S.,Baesens, B.,Seow, HV and Thomas, LC., (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247 (1), 124-136
8. Myoung-Jong, K., Sung-Hwan, M. and Ingoo, H.(2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index. Expert Systems with Applications 31(2):241-247
9. Paleologo, G., Elisseeff, A., & Antonini, G. (2010). Subagging for credit scoring models. European Journal of Operational Research, 201, 490-499.
10. Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28, 1619-1630.
11. Tsai, C.-F. , & Hsiao, Y.-C. (2011). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. Decision Support Systems, 50 ,258–269
12. Zhang,G., Patuwo, B.E., and Hu, M.Y.,(1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting 14 (1998) 35–62.