Original Research Paper

Volume-8 | Issue-6 | June-2018 | PRINT ISSN No 2249-555X

Computer Science

IMPROVING ACCURACY OF CLASSIFICATION SYSTEM USING ENSEMBLE TECHNIQUE

Nikhil Jadhav	BE 4th Year, Kothrud, Pune
Vaibhav Joshi*	BE 4th Year, Kothrud, Pune *Corresponding Author
Bhushan Bhavsar	BE 4th Year, Kothrud, Pune
Vishwajeet Bokan	BE 4th Year, Bavdhan, Pune
Amrut A Patil	Assistant Professor, Bavdhan,Pune

ABSTRACT Ensemble techniques creates base classifiers and different individual output, after combining individual output one combined output is prepared usually by majority voting. To get better results by classifiers we use the technique called as Ensemble classification or Ensemble learning. Using Ensemble techniques is one of the general strategies to improve the accuracy and prediction. Ensemble Learning is simple, useful effective method that will combine prediction from multiple output of individual classifiers. Weather forecasting is a vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century. Ensemble classifiers in general targets to improve the accuracy and prediction multiple classifiers output to a single combined output with the help of majority voting. Boosting and bagging are the two techniques in the Ensemble Learning Boosting is the technique of combining group of individual output to strong output wherein Bagging is learning algorithm where multiple prediction are generated and subsequently aggregated by averaging.Email users are increasing day by day and they are facing the problem due to spam Emails. To filter the spam email there are using the Data mining techniques for classifying the Email data. They have used the ensemble of C5.0, SVM and ANN which are outperforming with high accuracy of 94.26%.

KEYWORDS : Classifiers, Datasets, Ensemble Learning, Machine Learning, Supervised Learning, Data Stream Classification.

INTRODUCTION

The climate prediction is not stable by which we have to predict the climate which is one of the major challenge in the day-to-day life hence to predict the proper climate we have to build an efficient system which will be helping in predicting the weather condition in proper format. One of the fastest uprising technology is by using ensemble classifiers. In this Ensemble technique it is the mixture of many individual classifiers by generating a single output with the help of majority voting and predicting the weather. In ensemble classification technique the individual decision from each of the classifiers is combined with majority voting and new prediction is generated. Supervised learning often uses Ensemble classification. In our system, we are going to provide input as datasets to classifiers. Output generated by each of these classifiers will be an individual prediction. The final prediction of the system will be selected upon majority voting of these predictions. Typical weather forecasting system uses an algorithm for prediction. But the accuracy may differ from algorithm to algorithm. There are many advantage of Ensemble system one of them is the ability of statistical learning from huge amount of data. The application in which weather prediction can be done and the climatic condition can be predicted are in the fields of sports, banking, meteorological department, etc., and because of this, scientists, mathematicians and researchers have come up with a wide range of algorithms for finding solution.

RELATED WORK

Weather prediction is the most vital challenge in today's days therefore to get rid of these problems it is necessary to study and design an efficient system that will predict the weather also in uncertain conditions. As discussed Ensemble technique is the appropriate option. Ensemble have been defined as consisting of a set of individual classifiers whose decision are combined when classifying new instances, the simplest thing included in our system is the voting system. [1][2]An ensemble of classifiers for confidence-rated classification of NDE signal in this they have proposed the advantage of a boosting algorithm while avoiding the problem of over fitting. Performance Analysis for visual data mining classification technique of decision tree in this they have proposed that visual data mining application for enhancing decision making. [3]Popular Ensemble method in this they have proposed the empirical study and evaluation of Bagging and Boosting strategies for decision making. [4]New application of ensembles of classifiers in this they have proposed that ensemble technique have managed several problem in the application of data mining. [5]Improving classification accuracy using ensemble learning technique in this they have proposed that classification accuracy has improved with the help of ensemble technique and also proposed that in future the technique can be established by using different algorithms.

A survey of machine learning techniques for spam filtering has recently published in which author has explored many data mining techniques like k-nearest neighbour, Artificial Neural Network, Naive Bayes etc. (Omar Saad , et al., 2012). Authors suggested that machine learning techniques may be one of the best techniques for anti-spam filtering. [6]They made an attempt to develop ensemble model for filtering email data which will distinguish between spam and non-spam emails. A framework for classification of spam e-mail data is viewed as two folds: the first stage is data preparation and model development; second stage is model testing with reduced number of features and selection of final model as anti-spam classifier.Various data mining techniques can be used as ensemble model to achieve higher accuracy by combining individual data mining techniques and models. Model are tested and found satisfactory testing accuracy of 94.26%. Feature reduction technique is further applied to best two models, and it is examined that models are outperforming with 34 features, hence ensemble models of C5.0, ANN and SVM could be recommended for classification of spam e-mail data.

A data set which is taken from the UCI repository is given as the input to the system. [7]Each agent consists of three classification algorithms. The coordinator, then, partitions the data set and sends each partition to all the agents. The best classifier i.e. the most accurately classified data set is then selected. This is called cross validation. Each agent contains three distinct classifiers. Each algorithm in each one of the mining agents performs the classification with some accuracy.Data can be divided into partitions that are processed by all the agents in the systems and the results from the partitions are then merged to improve the efficiency. A graph has been plotted for multiple data sets by comparing multi agent system with single agent system. It can achieve much higher accuracy for larger datasets and can be used for many applications. Ensemble of C5.0 and SVM and Ensemble of C5.0, ANN and SVM are best as compare to other ensemble models. Ensemble of C5.0, ANN and SVM is better than all other models. Classification accuracy in this case is highest i.e. 94.26%.

Still there is much need to improve the use and performance of the classifiers. One of the ways is sampling, which plays a major role for improving the quality of ensemble classifier. Sampling is the process of extracting the subset of samples from the original dataset. Samples

are gathered in a process which gives all the individuals in the population of equal chances, such that, sampling bias is removed. Its' prediction is better than the individual classifier. [8]Dependent framework model is boosting and independent framework is Random Forest. Sample is the input for the ensemble classifier to build the prediction model. Sampling is the process of taking the subset of sample from original dataset. Ensemble of classifier means combine the prediction models of more than one classification algorithms into single prediction model to improves the performance of classifiers. It combine the outputs of classification algorithms into single prediction model based on combine strategy such as voting, weighting, etc. Finally combined prediction model is applied on testing dataset to perform prediction.Existing sampling technique fails to extract proper samples for both balanced dataset and unbalanced dataset. So it is necessary to improve a new sampling technique to extract the quality samples for all datasets.

With respect to supervised classification, clustering is inherently an ill-posed problem. Given a set of data samples, each clustering solution is equally plausible with no prior knowledge about the underlying probability distribution of the data. [9]The clustering algorithms assume some model to describe the data, which, in effect, is reflected in the corresponding clustering results.Clustering algorithms require either an implicit or explicit initial estimation of the inherent cluster parameters (e.g., mean and variance). Cluster ensemble techniques can be useful to highlight the common cluster subspace information to be adopted for the entire data set by using some supervised classification strategies. The application of ensemble-based clustering techniques is relatively new in remote sensing.K-means with different cluster centroid initializations are grouped together using the concept of cluster alignment. The effectiveness of the proposed unsupervised land-cover classification technique has been analyzed on two data sets. The land-cover classification accuracy of the proposed self training-based unsupervised classifier has been compared with the ones produced by the clustering methods used for the ensemble independently.The outcome of the cluster ensemble is further used to initialize the class wise statistical parameters for a self-learning-based ML classifier. Initially, the image is clustered using three diverse clustering methods. A set of highly reliable samples per cluster is selected to initialize the cluster parameters for an EM-based parameter retraining method considering the image as a mixture of Gaussian functions. Experimental results prove that the proposed framework is invariant to the underlying clustering techniques and can correspond correct clusters given that the clustering algorithms are efficient in detecting the clusters substantially.

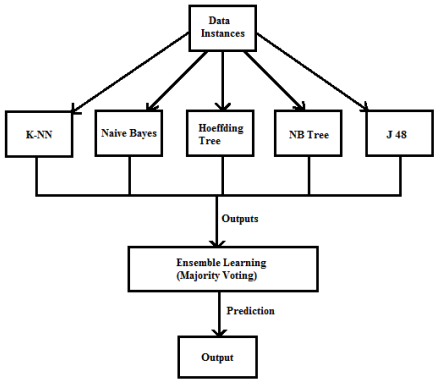
In recent years, with the rapid growth in the amount of data, the traditional algorithm cannot deal with the challenges of mass data. In order to cope with the challenges of big data, data mining based on big data technology has become a hot research, some scholars have proposed the classification algorithms for imbalanced data sets based on MapReduce, such as random forest algorithm based on MapReduce.[10]Over sampling method is a non-heuristic sampling method which increases minority class samples by random replication, is easy to cause the over decision boundary; a heuristic sampling method, which is represented by SMOTE algorithm , balances the category distribution of original data sets by adding some virtual samples. In recent years, many improved algorithms are proposed base on SMOTE, such as SMOTE-RSB algorithm combined with the theory of RST, which alter the samples from the sampling result when their similarity is greater than the given threshold; SMOTE-IPF algorithm uses multi noise alter to resample synthetic sample data; SMOTE-FRST algorithm is also combined with the theory of RST. Sampling method is also divided into non-heuristic method and heuristic method. The ensemble random forest algorithm has higher operation efficiency than most of the strong classification algorithms, and it is suitable for imbalanced classification model. And the strong classifiers failed to find the decision boundary due with the imbalance distribution of user features, the SMOTE sampling pre-processing is useful but it will spend a longer time to build the correct model on the large number of virtual samples, the ensemble random forest take fewer running time to get more accurate results because the sampling batch is parallel processed by executors, and it's also performed well than MapReduce approaches due to the memory cache mechanism of Spark. Algorithms like SVM and LR etc. is useless in the classification of imbalance distribution feature dataset, the ERF algorithm has a better performance than other algorithms when the number of features in a reasonable range.

A large amount of psychophysiological data can be readily measured, effective data reduction, analysis and interpretation methods are required for quantitatively assessing the CTL level. It was found that several issues received less attention when using psychophysiological measures and pattern recognition methods for CTL assessment. [11]However, the use of the same number of target classes lead to considerable misclassification especially during the period of the transition of task-load conditions. Therefore, an objective target class determination technique is required to reflect the individual difference.In reality it is often difficult to determine which optimal functions should be used for each subject. To overcome this difficulty, a possible solution is to build an ensemble of independent member classifiers to make a final classification decision based on the outputs of individual classifiers.The salient EEG and HR features are extracted and reduced to recognize the operator's CTL levels. A subject-average classification rate of 0.7667 is achieved for five- and four-level classification of CTL. Although it is a bit time-consuming to train the proposed SVM ensemble, the CTL level can be estimated in 1 s by invoking and applying the trained ensemble SVM-based classifier. It was expected that the ensemble classifier is more suited to cross-subject classification. The multi-class CTL classification was carried out via offline data analysis only, thus online classification experiments must be performed in the future to further substantiate the effectiveness and reliability of the system.

PROPOSED SYSTEM

In our system, we are going to provide input as datasets to classifiers. Output generated by each of these classifiers will be an individual prediction. The final prediction of the system will be selected upon majority voting of these predictions.

Figure 1: System Architecture



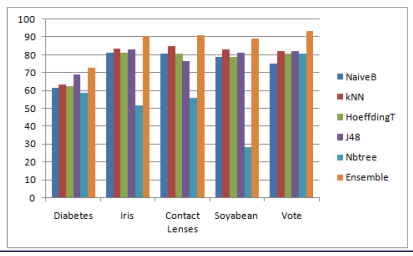
RESULTS

As various classifiers are ensembled to achieve the efficient prediction for the problem like Weather Forecast, Business, Medical decisions, etc. This technique outperforms the base classifiers output to the ensemble classifiers. Following table shows the efficiency comparison of base classifiers to ensemble which shows that it overtakes the efficiency of every classifier after the ensembling into one. This table shows the output of individual output of classifier and ensemble classifiers.

Table. 1Results of Proposed System

	NaiveB	kNN	HoeffdingT	J48	Nbtree	Ensemble
Diabetes	61.3021	63.2552	62.474	69.1146	58.5677	72.747
Iris	81	83.6667	81	83	51.6667	90.6666
Contact Lenses	80.8333	85	80.8333	76.6667	55.8333	91.1443
Soyabean	78.7042	82.9502	78.7042	81.3397	27.9649	89.092
Vote	75.3448	82.2414	80.6322	82.2414	80.6322	93.5578

Graph 1



Comparison of Individual vs Ensemble classifier

The above graph shows the difference in their efficiency for various applications like Diabetes, Iris, etc. These individual classifiers output compared with ensemble classifier is not efficient as ensemble classifier. It states that the ensemble of classifier make the system efficient as the graph shows the outperforming the base classifiers. So, this enhances and improves the classification technique with the help of ensemble techniques

CONCLUSION

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to appropriate forecast the target class for each case in the data. A typical classification system uses a base classifier to classify data instances using a relevant data set as a reference. The system that we have proposed uses ensemble learning technique. Same input is provided to multiple classifiers and a majority voting is taken to finalize the output. The proposed system will increase the accuracy in prediction by 5-10 percent. If accuracy of base classifier is 85-90 percent, our system will try to increase it up to 90-95 percent.

REFERENCES

[1] Portia Banerjee, SeyedSafdarnejad, LalitUdpa and SatishUdpa," An Ensemble of Classifiers for Confidence-rated Classification of NDE Signal" published in AIP conference, 1706, 2016.

[2] CM Velu and Krishnan Kashwan,"Performance Analysis for visual data mining classification technique of decision tree" published in IJCA, Vol 57- No.22 in Nov 2012.

[3] David Opitz and Richard Maclin,"Popular Ensemble Methods: An Empirical Study" published in Journal of Artificial Intelligence Research, 8/99 in 2011.

[4] R. Barandela, J. S. Sanchez, R. M. Valdovinos,"New Applications of Ensembles of Classifiers" published by Springer –Verlag London in April 2003.

[5] BhaveshPatankar and Dr.VijayaChavada," Improving classification accuracy using ensemble learning technique" published by IJSRCSEIT, Volume 1, 2016.

[6] H.S. Hota, S.K. Singhai and Akhilesh Kumar Shrivas,"Data Mining Techniques and its Ensemble Model Applied for Classification of E-Mail Data" published by ICIRT Nov 2012.

[7] NimishaPeddakam, SreevidyaSusarla, AnnepallyShivakesh Reddy,"Ensemble Classification Using Agents", published in IJARCSE Vol 5, Dec 2015.

[8] M. Balamurugan and S. Kannan,"ANALYSE THE PERFORMANCE OF ENSEMBLE CLASSIFIERS USING SAMPLING TECHNIQUES" published by ICTACT Journal of Soft Computing, Vol 6, July 2016.

[9] Biplab Banerjee, Francesca Bovolo, Avik Bhattacharya, Lorenzo Bruzzone, SubhasisChaudhuri and B. Krishna Mohan,"A New Self-Training-Based Unsupervised Satellite Image Classification Technique Using Cluster Ensemble Strategy", published by IEEE April 2015.

[10] WEIWEI LIN, ZIMING WU1, LONGXIN LIN, ANGZHAN WEN, AND JIN LI,"An Ensemble Random Forest Algorithm for Insurance Big Data Analysis" published by IEEE Aug 2017.

[11] Jianhua Zhang, Zhong Yin, and Rubin Wang," Pattern Classification of Instantaneous Cognitive Task load task GMM clustering, LaplacianEigenmap, and Ensemble SVMs" published by IEEE, Aug 2017.