2nd International Conference on Sensor Network and Computer Engineering (ICSNCE 2018)

An Ensemble Learning Method for Text Classification Based on Heterogeneous Classifiers

Fan Huimin
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, 710021, China
e-mail: 492896361@ qq.com

Zhao Yingze
School of School of Marxism
Xi'an Jiaotong University
Xi’an, China
e-mail: yingze1013@163.com

Li Pengpeng
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, 710021, China
e-mail: m18295715879@163.com

Li Danyang
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, 710021, China
e-mail: 821563942@qq.com

Abstract—**Ensemble learning can improve the accuracy of the classification algorithm and it has been widely used. Traditional ensemble learning methods include bagging, boosting and other methods, both of which are ensemble learning methods based on homogenous base classifiers, and obtain a diversity of base classifiers only through sample perturbation. However, heterogenous base classifiers tend to be more diverse, and multi-angle disturbances tend to obtain a variety of base classifiers. This paper presents a text classification ensemble learning method based on multi-angle perturbation heterogeneous base classifier, and validates the effectiveness of the algorithm through experiments.**

Keywords-Machine Learning; Ensemble Learning; Text Classification

I. INTRODUCTION

The main idea of ensemble learning is to generate multiple learners through certain rules and then adopt some integrated strategy to make the final decision[1]. In general, multiple learners in the so-called ensemble learning are all homogenous “weak learners”. Based on these weak learners, multiple learners are generated through sample set perturbation, and a strong learner is obtained after integration. With the deepening of integrated learning, its broad definition gradually accepted by scholars. It refers to a collection of multiple classifiers using learning methods, without distinction between the nature of the classifier。 However, the research of ensemble learning with homogenous classifiers is still the most common, and it is usually only perturbed by a single angle such as algorithm training set[2][3]. The random forest algorithm adds the perturbation of the classification attribute to the traditional bagging algorithm, and thus obtains a better classification effect[4]. This shows that the multi-angle perturbation can produce a larger difference base learner, and the ensemble learning model has higher classification accuracy. In addition, the research shows that the diversity of base learners based on the heterogeneous base classifier is stronger, so the classification model has stronger classification accuracy and generalization performance[5][6]. Therefore, this paper combines the above two factors and designs a text classification ensemble learning method based on multi-angle perturbation heterogeneous base classifier.

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

II. Ensemble Learning

"Weak Learning Is Equivalent to Strong Learning" is a theoretical issue raised by Kearns and Valiant in 1989. The Boosting algorithm arises from the proof of this issue. Then the Boosting algorithm derived a number of variants, including Gradient Boosting, LPBoosting and so on. Because of the characteristics of boosting that training classifiers serially, the training process takes up more resources and has lower efficiency. Therefore, whether it is possible to use a few classifiers and obtain the same performance is a matter of concern to researchers. Zhou Zhihua and others on the "selective ensemble"[7][8] of boosting algorithm helped to overcome this problem. "Selective ensemble" only used the classifier with has good classification results to integrate the classifiers. This idea can finish the construction of ensembled model more efficiently without changing the original algorithm that training base classifiers. In recent years, a method of selective integration based on clustering, selection, optimization and other methods has also been developed.

The theoretical basis of ensemble learning shows that strongly learner and weak learner are equivalent, so we can find ways to convert weaker learners into strongly learners, without having to look for hard-to-find Learner. Currently there is a representative ensemble learning method boosting, bagging. The traditional Bagging algorithm and Boosting algorithm as well as many derived algorithms of the two algorithms are ensemble learning based on homogenous base classifier. And diversity is only obtained through sample disturbances, while multi-angle disturbances and heterogeneous classifiers can improve model classification accuracy. This paper first trains and integrates homogenous base classifiers, compares and analyzes changes in the accuracy of base classifiers and integrated models, and then integrates k-nearest neighbor classifiers, Bayesian classifiers, and logistic regression classifiers in text classifiers. The integration model of the heterogeneous base classifier compares the diversity with the base classifier homogenous Bagging algorithm to measure the KW value and accuracy.

III. Ensemble Learning Model Based on Heterogeneous Base Classifier

In order to obtain an integrated learning model with higher accuracy, more base classifiers with more diversity and good classification results should be obtained as much as possible. From the perspective of diversity, we can try to select a combination of many "attributes" from the variable factors in the classification process. Here, "attribute" refers to everything that causes the change of the algorithm classification result. From the general process of text classification analysis, feature selection, feature dimension, classifier selection and classifier parameters can be used as a basis for the diversity of the classifier.

For each classification model, its algorithm parameters, feature selection algorithm, feature dimension are disturbed. In this paper, many kinds of classifiers are integrated, and an integrated learning model based on multi-angle perturbation heterogeneous basis classifiers is designed. Inputs in the process of model training are feature selection algorithm set S, feature dimension set N, classifier set C, adjustable parameter set A and parameter optional value set (dictionary) V. Training steps are as follows:

Step 1: Pre-process the sample set.

Step 2: Select an algorithm for each feature, make a feature selection for each feature dimension, and add the feature selection result to the feature selection result list L.

Step 3: Perform Step 4 for each classifier.

Step 4: Train and save to the classifier list C-output for each parameter of the classifier in combination with eachresult in the L list.

The output of the model is the classifier list C-output. The testing process of the model is as follows: After the pre-processing and the vectorization of the sample to be tested, a series of classification models are used to predict the samples to obtain a plurality of classification results. The majority of voting integration strategies lead to the final classification result.

The feature selection algorithm, feature dimension, and classifier all serve as a source for the diversity of the base classifiers. In this paper, feature selection algorithm can use chi-square statistics, information gain and mutual information algorithm. Classifier perturbation can be trained by Bayesian classifier, k-nearest neighbor classifier and logistic regression classifier. Since the parameters of the classifier are also variables, they can also be used as disturbance variables.

IV. Experiment Analysis

The experiment uses Sogou Labs' entire network news dataset, and randomly selects 600 news documents from five categories of financial, education, automotive, entertainment and women, and uses the body part and its category markers as the experimental text data Set (balanced data set). The experiment will use 80% data as the training set and the rest as test sets.

A. *The impact of changes in featuredimensions.*

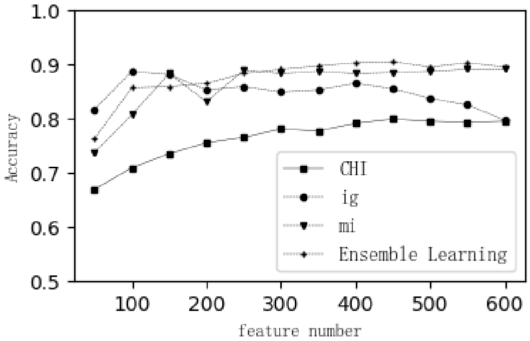

Figure 1. Experiment on the variation of feature dimension between integrated model and single classifier model

With the increase of feature dimensions, the accuracy of each model is on the rise. When the number of features is small, the accuracy of the integrated model is only lower than that of the information gain algorithm. When the number of features exceeds 300, the integrated model performs best. It can be seen that the classification effect of the integrated model is not always better than that of a single classifier. When the feature dimension is small, the accuracy of the integrated model is lower than that of the information gain algorithm model. In the experimental results obtained from experimental data in this paper, when the feature dimension exceeds 400 dimensions, the accuracy of the model tends to be stable, and the accuracy of ensemble learning model is always higher than that of a single classifier model.

B. *The effect of feature selection algorithm and classifiers*

TABLE I. Experiment ofthe perturbation of feature selection algorithm

Type	feature selection algorithm	classifier	accuracy/%
Base classifier	CHI	KNN	79.6
	IG		83.8
	MI		88.8
Ensembleclassifier	Above three kinds		89.6
Base classifier	CHI	Bayesian	77.8
	IG		90.2
	MI		82.4
Ensembleclassifier	Above three kinds		86.2
Base classifier	CHI	Logistic regression	79.4
	IG		91.4
	MI		87.6
Ensembleclassifier	Above three kinds		89.2

It can be seen from the experimental results that under the same conditions, the classification results of multiple classifiers combined with multiple feature selection algorithms are quite different. That is to say, the diversity between the base classifiers obtained by the perturbation feature selection algorithm is strong. Therefore, a variety of feature selection algorithms can be used as one of the sources of the base classifiers.

As can be seen from Table 1, when the feature selection algorithm is chi-square statistics, information gain or mutual information algorithm, the classification accuracy of the single classifier is lower, and the accuracy of the integrated classifier classification is higher than that of any single classifier. The disturbance of classifier makes the algorithm vary greatly in accuracy, so the disturbance of classifier can also be used as one of the sources of the diversity of classifier.

C. *Effect of classifier parameters*

Due to the different settings of the base classifier parameters will lead to some differences between the training model, this paper designed experiments to further examine the accuracy of the basic learning model in the disturbance of classifier parameters. The experimental results are shown in Table 2-4.

TABLE II. PARAMETER PERTURBATION EXPERIMENT OF K NEAREST NEIGHBOR CLASSIFIER

Type	K	classifier	Accuracy/%
Base classifier	5	KNN	85.2
	10		85.6
	15		84.6
	20		81.2
	25		78.6
	30		78
ensemble classifier	-	-	81.6

TABLE III. PERTURBATION EXPERIMENT OF BAYESIAN CLASSIFIER PARAMETER

type	Type of classifier	classifier	Accuracy/%
Base classifier	Polynomial	Bayesian classifier	89.6
	Gaussian		91
	Bernoulli		84.8
ensemble classifier	-	-	93.2

TABLE IV. PERTURBATION EXPERIMENT OF LOGISTIC REGRESSION CLASSIFIER PARAMETER

Type	The way of classification	Loss function optimization method	classifier	Accuracy/%
Base classifier	One to many	liblinear	Logistic regression	90.6
		newton-cg		90.6
		lbfgs		90.6
		sag		90.8
	multi-category	newton-cg		90.8
		lbfgs		90.8
		sag		90.8
ensemble classifier	-	-	-	90.6

From the data in Table 2-4 found:

Compared with the above three groups of experiments, the K-nearest neighbor classifier has a strong diversity among the classifiers in the selection of "K value" and the Bayesian classifier perturbation of the "classifier type" parameters. Therefore, Base classifiers with higher classification accuracy are candidates. However, the logistic regression classifier is insensitive to the two parameters of "classification method" and "loss function optimization method". The accuracy of the base classifier is almost constant and the diversity is lower. In the multi-angle perturbation integrated model, only one of the classifiers can be selected.

D. *Multi-angle disturbance*

Through the above three groups of experiments, we have screened the selected parameters of the base classifier with strong diversity. From the experimental data obtained from the above three experiments, the KW diversity measure between homogeneity classifiers that make up each classifier can be calculated as shown in Table 5.

TABLE V. BASE CLASSIFIER DIVERSITY MEASURE KW VALUE

Ensemble learning model	Disturb variable	KW value
KNN	Feature Selection Algorithm	0.06
Bayesian classifier		0.05
Logistic regression classifier		0.04
CHI	classifier	0.07
IG		0.03
MI		0.05
KNN	K value	0.04
Bayesian classifier	Classifier type	0.04
Logistic regression classifier	Classification and optimization methods	0
Multiangle perturbation heterogeneous basis classifier	Multi-angle disturbance	0.07

The range of KW values is [0,1]. When KW is 0 or 1, the base classifiers are the same, and there is no diversity among base classifiers. When KW is 0.25, the base classifier has the highest diversity. As can be seen from table 5, the integrated models with the most diversity of base classifiers in table 5 are all based on heterogeneous base classifiers. The KW value of this model is better than that based on the rest of the integrated learning models.

Using the integrated method, the above feature selection algorithm, feature dimension, classifier and its parameters are taken as input to integrate all the base classifiers, and an integrated model based on multi-angle perturbation heterogeneous base classifiers is obtained. The multi-angle disturbance integrated learning model parameters are summarized in Table 6.

TABLE VI. MODEL PARAMETERS

variable	Value / classifier	Classifier property value
Feature Selection Algorithm	CHI、IG、MI	-
Characteristic dimension	400、450、500	-
Classifier	Bayesian classifier	Type: Gaussian, Bernoulli, Polynomial
Classifier	KNN	K=5、10、15
Classifier	Logistic regression classifier	Classification: one to many; optimization methods: sag

The parameters shown in Table 6 are used as inputs to the model to train the integrated learning model designed in this paper. Compare this model with the Bagging text classification model with only sample perturbation. The experimental results are shown in Table 7.

TABLE VII. THE COMPARISON BETWEEN THE MODEL AND BAGGING MODEL

model type	variable	classifier	KW value	accuracy/%
Bagging	Sample disturbance	KNN	0.10	83
Bagging	Sample disturbance	Bayesian	0.03	85.4
Bagging	Sample disturbance	Logistic regression	0.06	83
Heterogeneous classifier model	Multi-angle disturbance	-	0.07	92

The experimental results show that the Bagging algorithm based on K-nearest neighbor classifier has higher KW value, that is to say, the classifier has strong diversity but low accuracy. Bagging algorithm based on Bayesian classifier and logistic regression classifier has low KW value and accuracy, that is, the base classifier has less diversity and low accuracy. The integrated learning model based on multi-angle disturbance heterogeneous basis classifier designed in this paper has the highest classification accuracy and the strong diversity of base classifiers.

V. CONCLUSION

This paper analyzes the algorithmic process of Bagging and Boosting, and finds that both of them are integrated learning strategies based on homogeneity classifier. At present, the research on heterogeneous base classifier integrated learning is less. In this paper, we design a learning model of multi-angle perturbation heterogeneous basis classifier. Multi-angle perturbation of heterogeneous classifiers, and try to integrate them. The experimental results show that the integrated learning model based on multi-angle perturbation-based heterogeneous base classifiers proposed and designed in this paper has higher classification accuracy and rich base classifier diversity. This will provide an important basis for further research on heterogeneous classifier integration.

REFERENCES

[1] Lai J H. Ensemble Learning for Text Classification[J]. 2017.

[2] Wang G, Sun J, Ma J, et al. Sentiment classification: The contribution of ensemble learning[J]. Decision support systems, 2014, 57: 77-93.

[3] Xia R, Zong C, Li S. Ensemble of feature sets and classification algorithms for sentiment classification[J]. Information Sciences, 2011, 181(6): 1138-1152.

[4] Jia J, Liu Z, Xiao X, et al. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach[J]. Journal of theoretical biology, 2016, 394: 223-230.

[5] Rodriguez J J, Kuncheva L I, Alonso C J. Rotation forest: A new classifier ensemble method[J]. IEEE transactions on pattern analysis and machine intelligence, 2006, 28(10): 1619-1630.

[6] Wu Z, Lin W, Zhang Z, et al. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis[C]//Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on. IEEE, 2017, 1: 531-536.

[7] Li N, Jiang Y, Zhou Z H. Multi-label Selective Ensemble[C]//International Workshop on Multiple Classifier Systems. Springer, Cham, 2015: 76-88.

[8] Qian C, Yu Y, Zhou Z H. Pareto Ensemble Pruning[C]//AAAI. 2015: 2935-2941.