SENTIMENT ANALYSIS OF E-COMMERCE WEBSITE USING ENSEMBLE FEATURE SELECTION

R.Preethi, S. Sheeja*

Abstract

In any case, the content is normally large throughout nature due to the convenience of products in big amount with their information, offering by vendors, crowned with buyers render comments in the form of measure. Valuation become reflections connected with user approval based on a new scale [1].

Frauds are generally an overbearing analysis because they are engineered to stop detection[2].The performance has been analyzed with 5 classification algorithms of SVM ,RF, NB, Gradient Boosting Classifier and Ridge classifier. The overall performance with accuracy has been evaluated with each classifier and feature importance.

Keyword: SVM, Randomforest,NB,Gradient boosting Classifier and Ridge classifier,E-commerce, Ensembles, homogeneous, heterogeneous .

I INTRODUCTION

The feature score, recursive feature selection and elastic net feature selection has been used. If embedded feature selection is needed, a wrapper approach such as feature importance has been considered for the aggregation of the feature selection on the ensemble model. .The feature importance has been created for each feature on the data set. The performance has been analyse with 5 classification algorithms of SVM RF, NB, Gradient Boosting Classifier and Ridge classifier. The overall performance with accuracy has been evaluated with each classifier and feature importance. Ensemble highlight sets and AI for estimation characterization[3]. Two kinds of capabilities specifically "POS based highlights" and "the world-connection based capabilities" has been planned. These element determination techniques were ensemble with NB, ME, and SVM utilizing 3 strategies viz. fixed blend, weighted mix and meta-classifier blend. Word connection based weighted classifier yielded precision of 87.7% and normal 85.15%

1.1 ENSEMBLE FEATURE SELECTION METHODS

Opinion mining systems are highly domain dependant. The results can vary importantly from a domain to another which make the opinion mining a very stimulating and ambitious task. Prior studies has shown that many works in opinion mining exist on the product domain using single classifier [4,5].

Opinion mining is the process which is used to extract the information or knowledge automatically from the others opinion related with the particular topic.[6]

Ensemble feature selection can be a result for the before mentioned problem since, by union of the output of several feature selectors.The performance can be usually developed and the user is free from having to choose a single method. The goal of the paper is to offer a wide review of ensemble learning in the field of feature selection. Ensembles for feature selection can be classified into two types they are homogeneous selection is based on the same base feature selector and heterogeneous is based on different feature selectors .

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

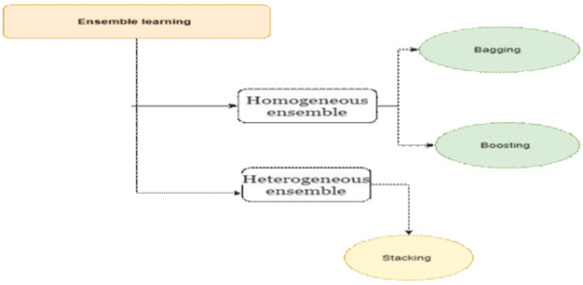

Figure 1 Types of Ensemble Learning

The single outputs produced by each feature selector should be carried out. These outputs can be in the form of subsets of features or ranking of features, and specific collection strategies are needed consequently. Opinion processing is not essential to a customer but criticalto an organization[7].

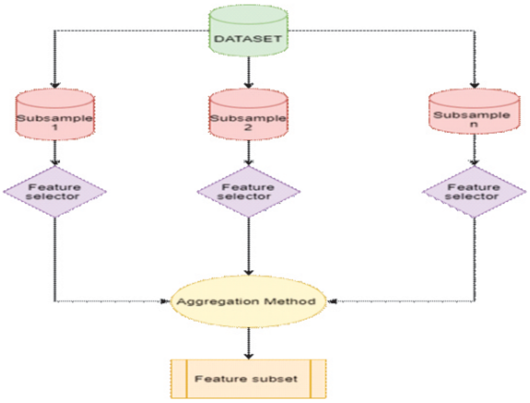

Figure 2 Homogeneous Feature Selection Ensembles

Heterogeneous feature selection ensembles are more popular than homogeneous, Some new works have also explored different designs transaction the order of the collection and withholding steps when ranker are used as base feature selectors.

II ENSEMBLE FEATURE SELECTION OF FEATURE SCORE, ELASTIC NET, RECURSIVE FEATURE ELIMINATION

The optimal machine learning problem approach is to take a dataset, perform extensive EDA on it, and understand many to most of the important properties of the predictors before getting as far as seriously training models on these variables. However, this is not always possible. Sometimes the dataset has too many variables. Datasets may easily have hundreds or even thousands of variables, quickly out running human understanding. While the number of features is small or you have time to sit down and consider them all feature selection is mainly a hand-driven process. In scenarios where the number of variables are overtaken, or time is limited, automated or semi-automated feature selection can speed things up.

2.1 Feature score

The feature score provides a way to rank drivers based on the features that a driver supports. The feature score supplements the identifier score, making it possible for driver writers to more easily and precisely distinguish between different drivers for a device that is based on well-defined criteria.

And even when you do have the incentive to hand-roll and accurate your features, automated feature selection provides some useful early directions for exploration during the exploratory process. The f-feature score based feature selection is that allows to select features from a dataset using a scoring function. It supports selecting columns in one of a few different configurations. k for when you want a specific number of columns, percentile for when you want to a percentage of the total number of columns, and so on.

2.2 Elastic net

Elastic net linear regression uses the control from both the lasso and ridge techniques to order regression models. The method combines both the lasso and ridge regression methods by erudition from their shortcomings to improve the condition of statistical models.

2.3 Recursive feature elimination

RFE is favourite because it is simple to assemble and use, because it is impressive at selecting those features (columns) in a training dataset that are more or most related in predicting the target variable.

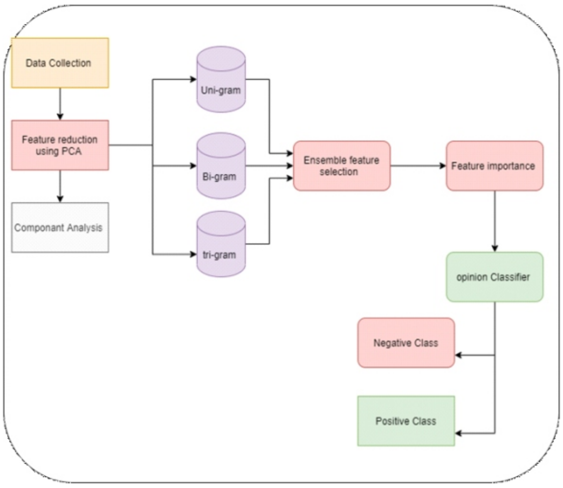

Figure 3 Proposed Ensemble feature selection

This is achieved by proper machine learning algorithm used in the core of the model, ranking features by value,the least important features, and re-fitting the model. This process is recurrent until a specified number of features remains.

2.4 Feature importance

Working with selected features instead of all the features reduces the risk .of fitting,develop accuracy, and decreases the training time. Another important distinction is global vs. local feature importance. The first applicable, , when we want to explain why a specific news has been fake from the news dataset.

III PERFORMANCE METRICS

3.1Confusion Matrix

A confusion matrix is null but a table with two dimensions . "Actual" and "Predicted" and moreover, both the dimensions have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", "False Negatives (FN)". True Positives (TP) − It is the case when both actual class & predicted class of data point is 1.

True Negatives (TN) − The case when both actual class & predicted class of data point is 0.

False Positives (FP) − The case when actual class of data point is 0 & predicted class of data point is 1.

False Negatives (FN) − The case when actual class of data point is 1 & predicted class of data point is 0.

3.2 Classification Accuracy

It is most common demonstration metric for classification algorithms. It may defined as the definite quantity of exact predictions made as a ratio of all predictions successful.confusion matrix can be easily calculated with the help of following formula.

Accuracy=TP+TNTP+FP+FN+TN(1)

3.3 Classification Report

The report consists of the scores of Precision, Recall, F1 and Support. They are explained as follows

Precision

Precision, used in document recovery, may be defined as the number of correct documents returned by ML model. confusion matrix can be calculated with the help of following formula .

Precision = TPTP+FP(2)

3.4 Recall or Sensitivity

Recall be defined as the number of positives returned by ML model. confusion matrix can be calculated with the help of following formula .

Recall = TPTP+FN(3)

Specificity

Specificity, in contrast to recall, may be defined as the number of negatives returned by ML model. confusion matrix can be calculated with the help of following formula .

Specificity = TNTN+FP(4)

3.5 Support

Mathematically, F1 score is the weighted average of the precision and recall. The best value of F1 would be 1 and worst would be 0. F1 score can be calculated with the help of following formula .

Lee [8] spearheaded in applying AI viz. NB, Maximum Entropy (ME), and SVM for parallel assumption grouping of film surveys[8].

This will be a trade-off between precision (higher with higher threshold) and recall (lower with higher threshold).The performance of the proposed model has been efficiently measured by using 5 different classifier model of SVM ,RF, NB, Gradient Boosting Classifier and Ridge classifier.

Framed a half breed strategy for Particle Swarm Improvement (PSO) and SVM for slant order of film surveys[9].

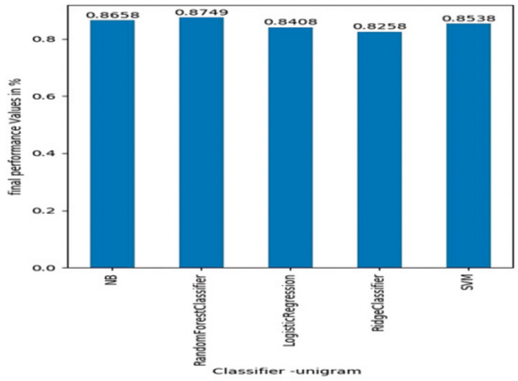

Figure 4 Comparison of algorithm

The performance was evaluated using SVM RF, NB Gradient Boosting Classifier, and Ridge classifiers. The overall performance in terms of accuracy was evaluated for each classifier and feature value. The result shows that the overall process has been higher of 87.49% for random forest with the proposed ensemble feature selection

IV BAGGING

The main idea is to estimate each member of the ensemble from a training dataset, and to guess the collection by uniform averaging over class . A bootstrap sample of S items is selected uniformly at random with variation. This means each classifier is trained on a sample of examples taken with a replacement from the training set, and each sample size is equal to the size of the training set. The presentation of three well known gathering techniques viz. sacking, boosting, and arbitrary subspace dependent on five base students specifically NB, ME, DT, K-Nearest Neighbor (KNN), and SVM for notion arrangement.[10]

4.1 Bayesian boosting

Boosting is an repetitive process, has become one of the option framework for classifier design, unitedly with the more accepted classifier like Bayesian classifier. NB classifier is used as interior classifier and the number of repeat repetition to combine the classifier .

Classifier		Positive Class				Negative Class		
	Accuracy	Precision	Recall	F-score	Precision	Accuracy	Recall	F-score
NB	0.86	0.86	0.732	0.818	0.866	0.80	0.507	0.640
SVM	0.85	0.77	0.780	0.870	0.821	0.800	0.836	0.880
Randomforest	0.87	0.84	0.886	0.866	0.888	0.818	0.814	0.854
Ridge classifier	0.82	0.83	0.883	0.872	0.844	0.876	0.836	0.886
LR	0.840	0.83	0.865	0.876	0.834	0.835	0.823	0.801
Bagging (majority voting)	0.88	0.862	0.856	0.843	0.848	0.856	0.832	0.823
Boosted SVM (NB+SVM)	0.91	0.871	0.868	0.868	0.878	0.881	0.882	0.876

Table1 Comparison of classifier algorithm

V CONCULSION

For neutral class, best precision and f-score is observed using ME, while the best recall is obtained using boosted SVM. The general performance of classifiers are used in this experiment,weighted average precision, recall, and f-score values were obtained in which weight of individual classes positive (P) and negative (N) is considered. The best weighted average precision, recall, and f-score is obtained using bagging method,followed by boosted SVM (NB+SVM) as the next performer on training dataset.

REFERNCES

[1] Dr.S.Sheeja, R.preethi. Psychology and Education, Latent dirichlet allocation based e-commerce recommendation system using deep neural network, 2021, 58(2) Pg.953-959

[2] Wang, S., Liu, C., Gao, X., Qu, H., & Xu, W. (2017). Session-Based Fraud Detection in Online E-Commerce Transactions Using Recurrent Neural Networks. Lecture Notes in Computer Science, 241–252. doi:10.1007/978-3-319-71273-4_20

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

[4] S. Poria, A. Gelbukh, A. Hussain, N. Howard, D. Das, S. Bandyopadhyay, Enhanced SenticNet with Affective Labels for ConceptBased Opinion Mining, Knowledge-Based Approaches to Concept-Level Sentiment Analysis, IEEE Intelligent Systems, (2013) 1.

[5] Maks, P. Vossen, A lexicon model for deep sentiment analysis and opinion mining applications, Decision support systems 53 (2012)

[6] Dr.S.Sheeja, R.Preethi, Online product safety surveilanceanalysis from online reviews using LDSA, Journal of Advanced research in dynamical and control systems, Special Issue on Recent Trends in Engineering and Managerial Excellence, May 2017

[7] Dr.S.Sheeja, R.Preethi, Current trends and tools of sentiment analysis and opinion mining: international journal of future generation, Vol. 13 No. 3 (2020)

[8] Lee, Predicting the helpfulness of online reviews using multilayer perceptron neural networks, Expert Systems with Applications 41 (2014) 3041–3046.

[9] Abd. S.H. Basari, B. Hussin, I.G.P. Ananta, J. Zeniarja, Opinion Mining of Movie Review using Hybrid Method of Support Vector.

[10]G. Wang et al., Sentiment classification: The contribution of ensemble learning, Decision Support Systems (2013)