The Effect of Ensemble Learning Models on Turkish Text Classification

Deniz Kılınç*

Department of Software Engineering, Faculty of Technology, Celal Bayar University,
Manisa, Turkey
deniz.kilinc@cbu.edu.tr
*Corresponding author / İletişimden sorumlu yazar

Recieved / Geliş: 6th June (Haziran) 2016
Accepted / Kabul: 1st August (Ağustos) 2016
DOI: http://dx.doi.org/10.18466/cbujos.04526

Abstract

Due to rapid development of the Internet and related technologies, the amount of text-based content generated through Internet applications is increasing from day to day. Since text-based content is unstructured, accessing and managing this data is almost impossible. Consequently, there is a need for automatic text classification process. Text mining is a discipline in the Data Mining field and offers algorithms in order to perform text classification. The main objective of text classification is forming a learning model by using a training data set with pre-defined categories and placing data with unknown categories into correct categories. Different text classification algorithms such as decision trees, Bayesian classifiers, rule-based classifiers, neural networks, k-nearest neighbor classifier, support vector machines and ensemble learning methods exist in the literature. In this study, the effect of ensemble learning models on Turkish text classification was evaluated. A publicly available data set named TTC-3600 which consists of 3600 news collected from 6 news portals was selected. Text classification process was performed on TTC-3600 data set by using 4 base classification algorithms Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, J48 Decision tree and their Boosting, Bagging and Rotation Forest ensemble learning models. The experimental results shows that ensemble learning models generally give more accurate results by increasing the success of base classifiers.

Keywords — Text Classification, Data Mining, Ensemble Learning Model, Machine Learning, Turkish Text Mining.

1 Introduction

The rapid development in the Internet and network technologies has led to intensive use of Internet-based applications and the amount of text-based content generated through these applications is also increased [1]. Blogs, social media sites and web-based news sites are some of the sources that produce this type of contents. Since unstructured text-based content cannot be manageable and accessible manually, there is a need for automatic text classification process. Text mining, which is a subfield of Data Mining, is a discipline that offers algorithms and methods to perform text classification process [2]. Creating a model by using a training data set with pre-defined categories and placing data with unknown categories into correct categories is the basic idea of text classification [3]. In the literature, there are various text classification algorithms such as decision trees, Bayesian classifiers, rule-based classifiers, neural networks, k-nearest neighbor classifier, support vector machines and ensemble learning methods [4].

Considering the related literature, although there are many text classification researches in other languages, the number of researches conducted in Turkish is relatively less. In a study conducted by Torunoglu et al. [5], the importance of pre-processing in text classification studies was investigated. In another study conducted by Güran et al. [6], Naïve Bayes (NB), Decision Tree (J48) and K-Nearest Neighbor (K-NN) classification algorithms were tried on Turkish data sets and the best classification results were obtained from Decision Tree classifier. Amasyalı and Beken [7] proposed a different approach to classification and divided Turkish words into semantic categories. They achieved the best classification results by using Linear Regression Classification Algorithm. In another study, Amasyalı and Diri [8] tried to prove that n-gram-based approaches produce more accurate results. They evaluated NB, Support Vector Machine (SVM), J48 and Random Forrest classification algorithms. Çataltepe et al. [9] investigated the effect of length of the word roots on the classi-fication and concluded that Centroid classification using shortened roots was more successful. Tüfekçi and Uzun [10] investigated the effect of different term weighting methods on texts author detection. According to the experimental results, the best results were obtained from SVM algorithm.

In this study, the effect of ensemble learning models on Turkish text classification was evaluated. TTC-3600 [11] was used as the data set, which consists of news collected from 6 news portals and agencies that are very well known in Turkey. TTC-3600 is also publicly available in order to be used in the experimental work of other researchers. Text classification process was per-formed on TTC-3600 data set by using 4 base classification algorithms NB, SVM, K-NN, J48 and their Boosting, Bagging and Rotation Forest ensemble learning models. Base classifiers were selected from different classification categories. The experimental results indicates that ensemble learning models generally give more accurate results by increasing the success of base classifiers.

The rest of the paper organized as follows: In section 2, the methods and data set used are introduced. Section 3 presents and discusses the experimental results obtained. Finally, section 4 concludes the paper with some future directions.

2 Text Classification

Text classification is used to assign text documents into pre-defined categories depending on the content of these documents [4]. Considering the related literature, the most widely used classification algorithms are NB, J48, K-NN and SVM and these algorithms are also selected as base classifiers in this study.

2.1 Naïve Bayes (NB)

Naive Bayes (NB) classifier is a well-known and high-performance classification algorithm used to find the category of the data with a statistical approach. It has a rule-based approach. According to a basic rule that creates, the possibility of being member of a category is calculated for the data [12]. The equation of NB classifier is given below.

$$P(c_k|i) = P(i|c_k)\,P(c_k)/P(i) \qquad (2.1.1)$$

Where, $P(c_k|i)$ and $P(i|c_k)$ denote the probability of instance i being in class c_j and the probability of generating instance d given class c_k, respectively. $P(c_k)$ is the probability of presence of class c_k and $P(i)$ is the probability of instance i occurring.

2.2 J48 - Decision Tree

Decision tree is a fundamental method used for classification of the data sets in data mining. The aim of decision tree is to construct a model that estimates the value of a target variable by using different input variables. It consists of four basic steps. The rules are created by creating the training data set. The root node, internal node and sheets are determined with selected features. Information gain is calculated using the probability of occurrence of unexpected events and the uncertainty for each selected character. The highest rate of information gain is determined as the root [13]. J48 classifier is a version of the C4.5 algorithm [14], which uses divide-and-conquer approach.

2.3 K-Nearest Neighbor (K-NN)

Another base classifier used to determine classes is K-Nearest Neighbor (K-NN) algorithm, which doesn't require any training phase. The closest member of the class is identified by using Euclidean distance method in order to identify the data from an unknown category. All data are placed in an n-dimensional vector space model to perform classification [15].

2.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is developed from the theory of Structural Risk Minimization which minimizes the probabilistic upper bound of the error on the test set [16]. SVM finds an ideal hyperplane in the feature space and utilizes a regularization parameter named C. The generalization characteristic of the SVM is the main advantage in comparison with other classifiers. SVM gives very accurate results in binary classifi-cation problems and it is one of the important algorithms used in text classification field in recent years.

2.5 Ensemble Learning Models

In recent years, the emergence of Ensemble learning models is one of the most important improvements made in the machine learning and classification areas. The main objective of Ensemble learning approach is generating more accurate results by bringing classification values that are obtained by using different base classifiers together. As seen in Figure 1, $H_1(x)$ and $H_\mathrm{M}(x)$ indicate the base classifiers and the output classification decision is given by $H(X)$.

In this process, calculations are made by giving particular weight scores to other classifiers. Combining different classification algorithms and deciding about weight scores is one of the major challenges. The biggest advantage of using Ensemble classification algorithm is probability of having more accurate values because this algorithm uses a combination of data of other methods. The most widely known and used ensemble learning algorithms are Bagging, Boosting and Rotation Forest.

Bagging [17] is an ensemble learning method that aims to train the base classifier again by deriving new training sets from an existing training set. In Bagging, a new training set with n samples generated from an existing training set by using the random selection method. In this case, some of the training examples are not included in the new training set while some samples are included multiple times. Each base classifier is trained by training sets including different samples and their results are combined by the majority rule.

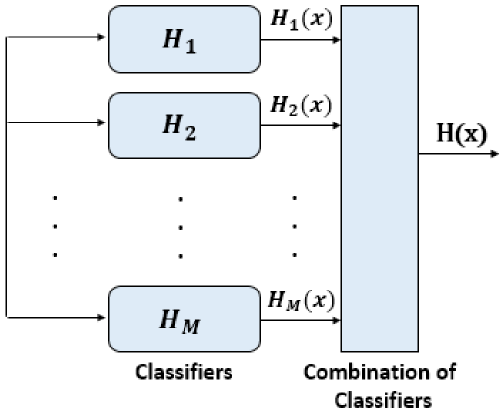

Figure 1. Ensemble learning

Boosting is another important ensemble learning method [18]. In this method, the data that couldn't be determined by the previous classifier is used. Each sample has a weight in the learning data set. After each learning process, weights of the samples are updated by taking classification error of each classifier into account. The weighted average based on the accuracy of each classifier is selected to classify a new sample and classification process is performed. The most commonly used algorithm for Boosting is Adaboost ensemble algorithm.

Rotation Forest is a new generation ensemble learning algorithm proposed to improve the performance of base classifiers [19]. Bootstrap algorithm is used as a trainer in Rotation Forest by utilizing multiple trees. The data set to be used in each decision tree is determined by Principle Component Analysis (PCA). During the training phase of decision trees in the forest by using the Rotation Forest algorithm, training data set is randomly divided into subsets and feature extraction is performed by applying principal component analysis on each subset.

2.6 TTC-3600 Data set

In this study, the data set named TTC-3600 [11], which has compatible formats with data mining tools and prepared to be used in Turkish text mining researches, was used. This data set can be accessed via Internet. The dataset consists of a total of 3600 documents including 600 news/texts from 6 categories like economy, culture-arts, health, politics, sports and technology are obtained from 6 well-known news portals and agencies (Hürriyet, Posta, İha, HaberTürk, Radikal and Zaman).

Three additional dataset versions were created on TTC-3600 by implementing different stemming methods [20] in order to deepen the scientific work of researchers. Considering the results of previous studies, the TTC-3600 data set version with most accurate results were used. This version has 3600 documents and 5693 features. Stop-words filtering and stemming process was also done by using a mainspring tool. [21].

2.7 Evaluation Measure

There are many performance measures to evaluate classification algorithms. All measures are basely generated from a confusion matrix which contains actual and predicted classification information. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are the four different prediction outcomes within the matrix. In this study, we utilized the mostly used evaluation measure in text classification named Accuracy (ACC). ACC is the ratio of the total number of correctly classified samples which is calculated using Equation 2.4.1.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.4.1)$$

3 Experimental Results

In the experimental study, 4 base classification algorithms with their corresponding Bagging, Boosting and Rotation Forest ensembles were selected, which results 12 different classification models. All experimental evaluations were performed using WEKA data mining tool [22]. Each classifier model was tested with 10-fold cross validation, which is a well-known strategy for performance estimation. In this strategy, each dataset is split into 10-blocks. One single block is retained as the validation data for testing the model, and the remaining k − 1 blocks are used as training data. The cross-validation process is then repeated 10 times [23].

Table 1 shows the classification accuracy results of both base classifiers and their corresponding ensembles. The column named Base shows ACC results that were obtained by directly using classification algorithms without using any ensemble model. Rot.Forest, Bagging and Boosting columns were obtained as a result of performing ensemble learning models that are used to improve results of these base classifiers.

Table 1. Accuracy results of base classifiers and their corresponding ensembles

Accuracy (ACC)				
Classifier	**Base**	**Rot. Forest**	**Bagging**	**Boosting**
NB	72.08%	73.36%	73.91%	75.97%
J48	77.13%	81.77%	81.69%	85.52%
SVM	82.38%	82.65%	81.80%	81.75%
K-NN	55.11%	61.83%	54.55%	55.61%

Considering the results of base classifiers, NB, J48, SVM and K-NN algorithms gave the following accuracy values; 72.08%, 77.13%, 82.38% and 55.11%, respectively. As seen in Figure 2, K-NN algorithm has the worst performance value, where as the best ACC value (82.83%) was obtained from SVM algorithm. The source of success of SVM is high dimensional feature space of this text classification algorithm. SVM is independent from the dimensionality of the feature space and is able to learn with small amount of data.

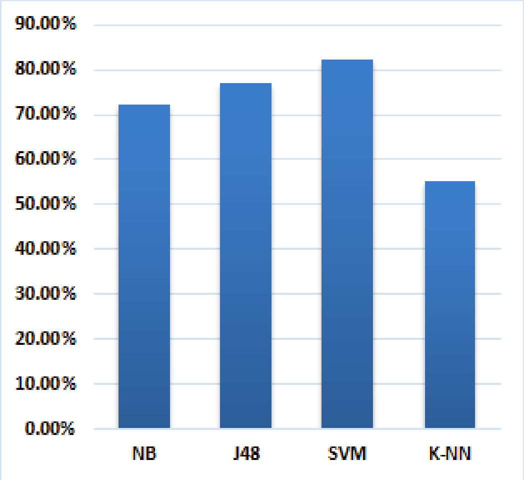

Figure 2. The accuracy results of base classifiers

As seen in Figure 3, the new results obtained by applying ensemble learning models Bagging, Boost-ing and Rotation Forest on base classifiers are more accurate or close the base results. More accurate results were obtained with the help of all ensemble learning models performed on NB and J48 classifiers compared to base results.

For example, J48 base classification algorithm gives an accuracy rate of 77.13%; however, when Rotation Forest, Bagging and Boosting ensemble models are used, base accuracy value is increased by 4%, 4% and 8%, respectively. In addition, there is a 6% improvement in the ACC value from 55.11% to 61.83% when Rotation Forest ensemble model is performed in the K-NN algorithm. Bagging and Boosting ensemble models didn't have a significant effect on K-NN base classifier.

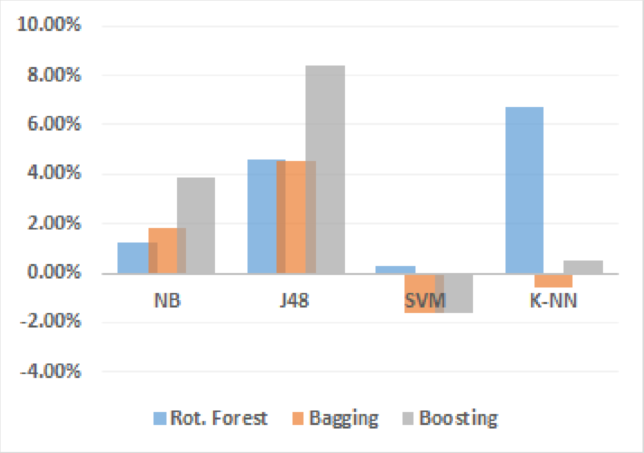

Figure 3. The accuracy difference of ensemble models

It is also seen from the Figure 3 that ensemble models had not a positive effect on the SVM classifier. Since SVM is a strong classifier, ensemble model cannot improve the accuracy of SVM in general [24].

4 Conclusion

In this study, the effect of ensemble learning models on Turkish text classification was investigated. Four well-known base classifiers (NB, K-NN, SVM, and J48) and three ensemble models (Bag-ging, Boosting, and Rotation Forest) were experimentally evaluated on TTC-3600 data set which includes a total of 3600 documents in the categories of economy, culture, arts, health, politics, sports and technology. The experimental results showed that the ensemble learning models mostly improved the classification accuracy of base classifiers in the task of Turkish text classification. Base classifiers with Rotation Forest and Boosting ensemble models had the highest classification accuracies.

In the future works, other ensemble learning models can be empirically evaluated with the combination of different feature selection methods. It is also planned to test the n-gram based representation of the TTC-3600 dataset.

5 References

[1] Fan, W.; Bifet, A. Mining big data: current status, and forecast to the future. ACM sIGKDD Explorations Newsletter. 2013; 14(2), 1-5.

[2] Sebastiani, F. Text categorization. In: Text Mining and Its Applications, UK: WIT Press. 2005; pp. 109-129.

[3] Azzalini, A.; Scarpa, B.; Walton, G. Data Analysis and Data Mining: An Introduction, New York: Oxford University Press, 2012.

[4] Sebastiani, F. Machine learning in automated text categorization. ACM Comput. Surv. 2002; 34(1), 1-47.

[5] Torunoğlu, D.; Çakırman, E.; Ganiz, M.C. et al. Analysis of preprocessing methods on classification of Turkish texts. In: Proceedings of International Symposium on Innovations in Intelligent Systems and Applications. 2011; pp. 112-118.

[6] Guran, A.; Akyokus, S.; Guler, N.; Gurbuz, Z. Turkish text categorization using n-gram words. In: Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA). 2009; pp. 369-373.

[7] Amasyalı, M.F.; Beken, A. Measurement of Turkish word semantic similarity and text categorization application. In: Proceedings of IEEE Signal Processing and Communications Applications Conference, Newyork: IEEE. 2009; pp. 1-4.

[8] Amasyali, M.F.; Diri, B. Automatic Turkish text categorization in terms of author, genre and gender. In: Natural Language Processing and Information Systems, Berlin: Springer. 2006; pp. 221-226.

[9] Çataltepe, Z.; Turan, Y.; Kesgin, F. Turkish document classification using shorter roots. In: Proceedings of IEEE Signal Processing and Communications Applications Conference (SIU), Newyork: IEEE, Eskisehir, Turkey. 2007; pp. 1-4.

[10] Tufekci, P.; Uzun, E. Author detection by using different term weighting schemes. In: Proceedings of IEEE Signal Processing and Communications Applications Conference (SIU), Newyork: IEEE, Trabzon, Turkey. 2013; pp. 1-4.

[11] Kılınç, D.; Özçift, A.; Bozyigit, F.; Yıldırım, P.; Yücalar, F. and Borandag, E. TTC-3600: A new benchmark dataset for Turkish text categorization. Journal of Information Science, p.0165551515620551. 2015.

[12] John G.H.; Langley P. Estimating continuous distributions in Bayesian classifiers. In: Proc. 11th Conf. Uncertainty in Artificial Intelligence. 1995; pp. 338-345.

[13] Cha, S.H.; Tappert, C.C. A Genetic Algorithm for Constructing Compact Binary Decision Trees. Journal of Pattern Recognition Research. 2009; 4(1), 1-13.

[14] Quinlan, J.R. C4.5: Programs for Machine Learning. Machine Learning. 1993; 16(3), 235-240.

[15] Aha, D.W.; Kibler, D.; Albert, M.K. Instance-based Learning Algorithms. Machine Learning. 1991; 6(1), 37-66.

[16] Cortes, C.; Vapnik, V. Support-vector network. Machine Learning. 1995; vol. 20, pp. 273–297.

[17] Breiman, L. Bagging predictors. Machine Learning, 1996; 24(2), 123–140.

[18] Freund, Y.; Schapire, R.E. Experiments with a new boosting algorithm. In: Proceedings of 13th International Conference on Machine Learning, San Francisco: Morgan Kaufman. 1996; pp. 148–156.

[19] Rodriguez, J.J.; Kuncheva, L.I.; Alonso, C.J. Rotation

forest: a new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. 2006; vol. 28, pp. 1619–1630.

[20] Tunali, V.; Bilgin, T.T. Examining the impact of stemming on clustering Turkish texts. In: Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium. 2012; pp. 1-4.

[21] Akin, A.A.; Akin, M.D. Zemberek, an open source NLP framework for Turkic Languages. 2007.

[22] Witten, I.H.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques, second ed., San Fransisco: Morgan Kaufman, 2005.

[23] McLachlan, Geoffrey J.; Do, Kim-Anh; Ambroise, Christophe. Analyzing microarray gene expression data. ISBN: 978-0-471-22616-1, Wiley, 2014.

[24] Dong, Y.S.; Han, K.S. Boosting SVM classifiers by ensemble. In: Special interest tracks and posters of the 14th international conference on World Wide Web. 2005; pp. 1072-1073.