Indian Journal of Science and Technology, *Vol 7(9), 1369–1375, September 2014*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Attempt to Improve Classification Accuracy through Implementation of Bootstrap Aggregation with Sequential Minimal Optimization during Automated Evaluation of Descriptive Answers

C. Sunil Kumar[1*], R. J. Rama Sree[2]

[1]Research and Development Center, Bharathiar University, Coimbatore, India; sunil_sixsigma@yahoo.com
[2]Rashtriya Sanskrit Vidyapeetha, Tirupati, India; rjramasree@yahoo.com

Abstract

In this paper, Bootstrap Aggregation (Bagging) ensemble learning technique was implemented using Sequential Minimal Optimization (SMO) with polynomial kernel in order to improve the classification accuracy during automated evaluation of descriptive answers. The performances obtained through bagging were recorded on five datasets each with 900 training samples and with each of the datasets treated using Symmetric Uncertainty Feature Selection filter. The performances obtained with bagging implementation were quantitatively analyzed in comparison with performances obtained with a plain simple application of SMO – Polynomial kernel on the datasets. Accuracy, F Score, Kappa and Area under ROC curve were used as model evaluation metrics. Based on the results, a conclusion was derived that Bagging with SMO-polynomial kernel classifier did not yield better accuracies when compared with classification accuracies obtained from SMO - Polynomial kernel. It was observed that, with bagging better Area Under the ROC curves were obtained signifying that prediction confidence of the models were improved.

Keywords: Auto Evaluation, Bagging, Bootstrap Aggregation, Descriptive Answers, Ensembling, SMO

1. Introduction

Evaluation of answers and providing a score to each answer is a hard classification task (i.e., assigning a single category to each document) where in the human evaluator or the system is supposed to interpret the answer and classify the answer into one of the possible scores pre-allocated for the answer. Supervised learning method can be applied to classify the answers into appropriate score based on the likelihood suggested by training samples. The supervised learning process requires extracting various text features from the documents meant as training set and then train the classification models using one or more sophisticated machine learning algorithms.

In the previous experiments with various supervised machine learning classifiers, an average classification accuracy of 76% was obtained when tested across 5 datasets using 10 fold cross validation. Naive Bayes, Logistic Regression, Random Forests, Support Vector Machine (SVM), Decision Stump and Decision Trees were the various supervised machine learning algorithms considered and used during the experiments. From this experiment results, a conclusion was derived that the classification accuracy with Sequential Minimal Optimization (SMO) outperformed all other algorithms considered for experimentation. The accuracies obtained and the average accuracy are shown in Table 1. The 76% accuracy is the best accuracy obtained using Sequential Minimal Optimization with default parameters and polynomial kernel.

Ensemble learning involves learning various alternative delineations of a notion by using different

*Author for correspondence

Table 1. Classification accuracies with various classifiers

Classifier	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Average Accuracy
Naive Bayes	37%	72%	76%	84%	83%	70%
Logistic Regression	57%	61%	80%	94%	73%	73%
Support Vector Machines (SMO)	60%	64%	77%	87%	90%	76%
Random Forests	44%	66%	64%	85%	81%	68%
Decision Stump	50%	72%	61%	88%	67%	68%
Decision Tree - J48	53%	60%	58%	88%	67%	65%

training data or by using various learning algorithms. A consolidated decision is arrived at by using the learnings obtained. When combing multiple independent and diverse decisions each of which is at least more accurate than random guessing, random errors cancel each other out, and correct decisions are reinforced[1].

Ensemble learning can be achieved through various techniques such as Bootstrap Aggregation (Bagging), Boosting, and Stacking. Bagging involves having multiple learners learn from resampled training set derived with replacements from the original training set[2]. Then, a consolidated prediction is arrived at either through voting or through weighted measurement of all the learnings obtained by the learners. Boosting involves sequential learning of predictors. First classifier learns from the entire training data. The subsequent classifiers focus on subset of training data that were misclassified by the previous classifier. The process is repeated with multiple classifier as desired by the researcher. Each learning machine in the sequence specializes in correctly predicting some areas of the dataset[3]. In Stacking, multiple classifiers that belong to absolutely different classes of machine learning methods use all the training data and predict the classes. Voting method is then applied to determine the correct class[4,5].

For the purpose of this research covered under this paper, bagging ensemble learning technique alone is focused on and bagging is used to verify if it improves the classification accuracy during automated evaluation of descriptive answers.

The rest of this paper is organized as follows. Section 2 discusses the data used, experimental setup, the preliminaries of the tools and techniques used in this paper. Section 3 describes the models built and measurements made during the experiments. Finally, analysis of results, concluding remarks and further research plans are indicated in Section 4.

2. Experimental Setup

The set up in which the experiments are conducted for this research are specified in this section.

2.1 Data Collection

In June 2012, The William and Flora Hewlett Foundation (Hewlett) sponsored the Automated Student Assessment Prize (ASAP)[6] to machine learning specialists and data scientists to develop an automated scoring algorithm for student-written essays. As part of this competition, the competitors were provided with hand scored essays under 8 different prompts (questions). 5 of the 8 essays prompts were used for the purpose of this research.

2.2 Data Characteristics

All the graded essays from ASAP are according to specific data characteristics. All responses are written by students of Grade 10. On average, each essay is approximately 50 to 60 words in length. Some are more dependent upon source materials than others. The data contains ASCII formatted text for each essay followed by one or more human scores, and (where necessary) a final resolved human score. Where it is relevant, more than one human score exists, so as to signify the reliability of the human scorers[7]. For the purpose of evaluation of the performance of the models in this research, the score predicted by the models need to comply with the final resolved human score in training example.

The data used for training, validation and testing the models are answers written by students for 5 different questions. Set of answers for a question is considered as one unique dataset. So, there are a total of 5 datasets. The questions that students are asked to provide responses to are from diversified fields of Chemistry, English Language Arts and Biology. Table 2 shows additional details of the datasets

Table 2. Essay prompt descriptions

Dataset	Subject	Essay prompt briefing	Average length of response	Possible rubrics (Scores)
1	Chemistry	Students were asked to perform a lab experiment related to change of masses of various chemical substances - Marble, Limestone, Wood, and Plastic. Lab experiment shows reduction in mass of the chemical substances except wood. Post reading the group's procedure, Students need to describe what additional information they would need in order to replicate the experiment. They are expected to include at least three pieces of information.	50	0,1,2,3
2	English Language Arts	Students were asked to explain how pandas in China are similar to koalas in Australia and how they both are different from pythons. Students need to support their response with information from the 2 page article provided on this subject as part of the essay prompt.	50	0,1,2
3	English Language Arts	A 2 page essay about "Invasive species items" is provided to the students, post reading the article Students are asked to explain the significance of the word "invasive" to the rest of the article. They need to support their responses with information from the article.	50	0,1,2
4	Biology	Starting with mRNA leaving the nucleus, Students need to list and describe four major steps involved in protein synthesis.	60	0,1,2,3
5	Biology	Students need to list and describe three processes used by cells to control the movement of substances across the cell membrane.	50	0,1,2,3

used for the research covered in this paper. Further details on essay prompts can be found from kaggle.com[8].

In each of the 5 training datasets used for this research, the training set is 900 samples in size. The previous research for determining appropriate sample size for automated answers scoring using SMO revealed that using 900 samples for training proved to yield slightly better results than using other sample sizes therefore the decision to use 900 samples as the training sample size[9]. Also, a survey conducted in 10 educational institutes concluded that the educational institutions are comfortable to provide a maximum of 1000 training sample answers for a question therefore the explicit decision to consider 900 samples as the appropriate training sample size.

2.3 Weka Workbench

For the purpose of designing and evaluating the experiments, a machine learning workbench called Weka is used. Weka (Waikato Environment for Knowledge Analysis) is a free offering from University of Waikato, New Zealand. This workbench has a user-friendly interface and it incorporates numerous options to develop and evaluate machine learning models[10,11]. These models can be utilized for a variety of purposes, including automated essay scoring.

All experiments performed were executed on a Dell Latitude E5430 laptop. The laptop is configured with Intel Core i5 -3350M CPU @ 2.70 GHz and with 4 GB RAM however Weka workbench is configured to use a maximum of 1 GB. The laptop runs on Windows 7 64 bit operating system.

2.4 Statistical Feature Extraction

Below features are focused on from input training data set to build feature table –

a) Unigrams - An n-gram of size 1 is referred to as a "unigram".
b) Bigrams - An n-gram of size 2 is a "bigram" (or, less commonly, a "digram").
c) Trigrams - An n-gram of size 3 is a "trigram".
d) Stop words - The most common, short function words, such as the, is, at, which, and on.
e) Stemming – It is a process of reducing inflected (or sometimes derived) words to their stem, base or root form—generally a written word form. Porter stemmer is used for stemming purpose.
f) Punctuations - unigrams representing things like periods, commas, or quotation marks

Included features – Unigrams, Bigrams, Trigrams, and Stemming.

Excluded features – Stop words, Punctuations.

2.5 Symmetrical Uncertainty Attribute Evaluation based Feature Selection

Dimensionality i.e., too many features is a curse in text classification. To reduce the dimensionality and to ensure learning happens only through relevant and non-redundant features, a fast rank based attribute evaluation technique called Symmetrical Uncertainty Attribute Evaluation[12] was used during the experiment. This rank based feature selection method evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class.

Once the filter was applied, significant reduction in the number of features was observed. Table 3 shows the comparison between the initial set of features vs. the count of features post the application of feature selection.

2.6 Model Building using Sequential Minimal Optimization (SMO) and Bagging

For all the five datasets treated through feature selection filter, models were built using SMO with default parameters and Polynomial kernel. SMO is a Sequential Minimal Optimization principle based SVM method[13], introduced by Platt in 1997. The measurements obtained from SMO – Polynomial kernel were used as benchmark.

Bagging models were built using bagging meta-classifier and SMO–polynomial kernel as the classification algorithm. For each dataset, number of iterations in bagging classifier were set as 1, 5, 10, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, and 100. The number in the iteration signifies the number of models created and those models contribute to voting in a bagging iteration. For example, with a bagging iteration of 80, 80 training data subsets are created by randomly sampling the entire training dataset and with replacement. Then these 80 training data subsets are used to train 80 models. Once trained, each of the models predicts the classification of a test data items. Now, bagging meta-classifier enables voting amongst the 80 model predictions for each test data item and whatever prediction is the majority is stamped as the prediction for the test data item.

Table 3. Features reduction with symmetrical uncertainty filter

Dataset	Number of features with no attribute selection applied	Number of features with symmetrical uncertainty filter applied
1	25190	254
2	22847	126
3	29475	400
4	20915	378
5	19599	373

3. Tests and Measurements

The various models that were built during the experiment, the tests, the measurements obtained and various conclusions made through analysis of the measurements from the experiments are described in this section.

Models were built on Weka workbench, randomized 10-fold cross-validation with 10 iterations was adopted in order to test the performance the models.

Measurements were made under two broad categories namely calibration scores and discriminatory scores. Calibration scores measure whether the model assigns the correct class value to the test instances. Many of these scores can be computed solely from the confusion matrix obtained from the result of the classifications done by the model. Discriminatory scores measure how good can the prediction model separate instances with different classes are called discriminatory scores[14,15].

Under the calibration scores umbrella, Accuracy, F score and Cohen's Kappa were compared for the datasets. Area under the ROC curve is captured as part of discrimination of models. Though several measurements were captured, the primary focus of this experiment is to confirm if the accuracy increases through implementation of bagging ensemble technique.

3.1 Accuracy

Accuracy is measured by percentage of correctly predicted instances divided by the total number of instances[16]. TP, TN, FP, FN in the equation 1 below refers to True Positives, True Negatives, False Positives and False Negatives respectively.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \longrightarrow (1)$$

Table 4 shows the classification accuracy obtained with bagging iterations vs. the benchmark classification accuracy obtained with SMO – Polynomial kernel.

Table 4. Accuracies with bagging implementation vs SMO - Polynomial kernel

Dataset	1	2	3	4	5
Benchmark	60	64	77	87	90
Iteration 1	55.84	61.7	73.84	84.34	89.28
Iteration 5	58.78	62.99	76.03	85.24	89.6
Iteration 10	59.44	63.27	76.63	85.37	89.78
Iteration 15	59.49	63.49	76.72	85.77	90.01
Iteration 20	59.53	63.42	76.77	85.88	89.8
Iteration 25	59.67	63.44	77.11	86.07	89.9
Iteration 30	59.87	63.63	77	85.97	89.94
Iteration 35	59.98	63.63	76.91	86.01	90.03
Iteration 40	59.86	63.62	76.86	85.93	90.01
Iteration 45	60	63.54	77.04	85.99	90.12
Iteration 50	60.03	63.64	77.11	86.06	90.1
Iteration 60	60.04	63.66	77.16	86	90.09
Iteration 70	60.24	63.67	77.02	86	90.13
Iteration 80	60.29	63.66	77.02	86.1	90.07
Iteration 90	60.13	63.7	77.06	86.11	90.06
Iteration 100	60.21	63.84	77.07	86.12	90.1

3.2 F Score

The F score measures accuracy using the statistics precision p and recall r[17]. Precision is the ratio of True Positives (TP) to all predicted positives (TP + FP). Recall is the ratio of true positives to all actual positives (TP + FN). The F score is given by –

$$F\ Score = 2 * ((\text{Precision} * \text{Recall}) / (Precision + Recall)) \longrightarrow (2)$$

Table 5 shows the F Score obtained with bagging iterations vs. the benchmark F Score obtained with SMO – Polynomial kernel

3.3 Cohen's Kappa

Kappa statistic is used to measure the agreement between predicted and observed categorizations of a dataset, while correcting for an agreement that occurs by chance. However, like the plain success rate, it does not take costs into account. Better models will have Kappa closer to 1[18].

Table 6 shows the Kappas obtained with bagging iterations vs. the benchmark Kappa obtained with SMO – Polynomial kernel.

Table 5. F Scores with bagging implementation vs. SMO - polynomial kernel

Dataset	1	2	3	4	5
Benchmark	0.595	0.584	0.767	0.864	0.894
Iteration 1	0.73	0.28	0.75	0.92	0.96
Iteration 5	0.77	0.29	0.78	0.93	0.96
Iteration 10	0.77	0.3	0.78	0.93	0.96
Iteration 15	0.77	0.3	0.78	0.93	0.96
Iteration 20	0.78	0.3	0.78	0.93	0.96
Iteration 25	0.78	0.3	0.79	0.93	0.96
Iteration 30	0.78	0.3	0.79	0.93	0.96
Iteration 35	0.78	0.3	0.79	0.93	0.96
Iteration 40	0.78	0.31	0.79	0.93	0.96
Iteration 45	0.78	0.31	0.79	0.93	0.96
Iteration 50	0.78	0.31	0.79	0.93	0.96
Iteration 60	0.78	0.31	0.79	0.93	0.96
Iteration 70	0.78	0.31	0.79	0.93	0.96
Iteration 80	0.78	0.31	0.79	0.93	0.96
Iteration 90	0.78	0.31	0.79	0.93	0.96
Iteration 100	0.78	0.31	0.79	0.93	0.96

Table 6. Cohen's Kappa with bagging implementation vs SMO - polynomial kernel

Dataset	1	2	3	4	5
Benchmark	0.4613	0.2985	0.5862	0.6374	0.6266
Iteration 1	0.4	0.25	0.53	0.54	0.58
Iteration 5	0.44	0.27	0.56	0.57	0.59
Iteration 10	0.45	0.28	0.57	0.57	0.59
Iteration 15	0.45	0.28	0.58	0.59	0.6
Iteration 20	0.45	0.28	0.58	0.59	0.59
Iteration 25	0.45	0.28	0.58	0.59	0.6
Iteration 30	0.46	0.28	0.58	0.59	0.6
Iteration 35	0.46	0.28	0.58	0.59	0.61
Iteration 40	0.46	0.28	0.58	0.59	0.61
Iteration 45	0.46	0.28	0.58	0.59	0.61
Iteration 50	0.46	0.28	0.58	0.59	0.61
Iteration 60	0.46	0.28	0.58	0.59	0.61
Iteration 70	0.46	0.28	0.58	0.59	0.61
Iteration 80	0.46	0.28	0.58	0.6	0.61
Iteration 90	0.46	0.28	0.58	0.6	0.61
Iteration 100	0.46	0.29	0.58	0.6	0.61

3.4 Area under the Receiver Operating Characteristics Curve (AUC)

AUC is a single scalar that represents models performance based on two dimensional ROC representation. A perfect model will have an AUC value of 1 where as a random guessing model will have a value of 0.5[19].

Table 7 shows the AUC obtained with bagging iterations vs. the benchmark AUC obtained with SMO – Polynomial kernel.

4. Results, Discussion and Next Steps

The primary focus and sole objective for this research remains to be improving accuracy of classification through implementation of bagging. Going by the accuracies recorded with the application of bagging in several iterations, it is very evident that implementing bagging using SMO–Polynomial kernel did not yield better accuracies than that of the accuracies obtained with SMO–Polynomial kernel. This result was observed across the iterations.

Some studies available in the literature suggests that bagging does not work with linear classifiers[20]. Also, some studies suggest that bagging assists "instable" classifiers such as decision trees or concept learners[21]. SMO is a linear classifier. Unlike the instable classifiers referred in the studies, SMO is a stable classifier[22,23]. Therefore these evidences doubly justify the accuracy results obtained by application of bagging on datasets meant for implementation of automated evaluation of descriptive answers.

Table 7. AUC with bagging implementation vs. SMO - Polynomial kernel

Dataset	1	2	3	4	5
Benchmark	0.808	0.666	0.81	0.823	0.839
Iteration 1	0.92	0.67	0.83	0.79	0.84
Iteration 5	0.94	0.73	0.88	0.88	0.93
Iteration 10	0.94	0.74	0.89	0.9	0.92
Iteration 15	0.94	0.75	0.89	0.91	0.92
Iteration 20	0.94	0.75	0.89	0.91	0.94
Iteration 25	0.94	0.75	0.89	0.92	0.94
Iteration 30	0.94	0.75	0.89	0.92	0.93
Iteration 35	0.95	0.75	0.89	0.92	0.92
Iteration 40	0.95	0.75	0.89	0.92	0.92
Iteration 45	0.95	0.75	0.89	0.92	0.92
Iteration 50	0.95	0.76	0.9	0.93	0.94
Iteration 60	0.95	0.76	0.9	0.93	0.93
Iteration 70	0.95	0.76	0.9	0.93	0.95
Iteration 80	0.95	0.76	0.9	0.93	0.95
Iteration 90	0.94	0.76	0.9	0.93	0.95
Iteration 100	0.95	0.76	0.9	0.93	0.95

F score is a measure of accuracy using the precision and recall statistic. The F score weighs recall and precision equally. Maximizing both precision and recall simultaneously yields a good F score. Extremely good performance on one and poor performance on the other yield in poor F score. F score is of much value in measuring the accuracies of skewed datasets. From the results, it was observed that except for Dataset 2, the F score is significantly higher when using bagging. This result proves un-skewed statistical tests with bagging.

Kappa statistic which is the measure of amount of agreement, corrected for the agreement that would be expected by chance. From the results, it was observed that bagging did not have any better effect on kappa statistic than that of kappa statistic obtained with directly applying SMO–Polynomial kernel.

Across all the datasets, AUC recorded higher than that of the benchmark AUC. This essentially signifies that, with implementation of bagging, the confidence by which the predictions were made by the classifiers is higher. This in some sense also means that the models in bagging have become more discriminative.

In this paper, bagging ensemble learning technique was applied to improve the classification accuracy of automated evaluation of descriptive answers however with no success. Further research is required to apply other ensemble learning techniques such as boosting, stacking to improve the classification accuracy. Improving the classification accuracy with the application of extreme features engineering such as lemmatization, implementing spelling corrections etc., is one other area to explore. Further experimentation can be done with bagging in combination with an instable classifier such as decision trees to confirm if it yields any better accuracy than SMO–Polynomial kernel accuracy .

5. References

1. Raymond J. Mooney, Machine Learning: Ensembles, University of Texas at Austin. Available from: www.cs.utexas.edu/~mooney/cs391L/slides/ensembles.pdf
2. Breiman L. Bagging Predictors. Mach Learn. 1996; 24:2:123–40.

3. Freund Y, Schapire RE. Experiments with a New Boosting Algorithm, Machine Learning: Proceedings of the Thirteenth International Conference; 1996 Jul 3–6; Italy.
4. Wolpert DH. Stacked generalization. Neural Networks. 1992; 5:241–59.
5. Seewald AK. How to Make Stacking Better and Faster While Also Taking Care of an Unknown Weakness, Nineteenth International Conference on Machine Learning; 2002 Jul 8–12; Australia. 554–61.
6. Kaggle. 2012 Jun 25. Available from: http://www.kaggle.com/c/asap-sas
7. Evaluation. 2012 Jun 25. Available from: http://www.kaggle.com/c/asap-sas/details/evaluation
8. Data. Available from: http://www.kaggle.com/c/asap-sas/data
9. Sunil Kumar, Rama Sree RJ. Experiments towards determining best training sample size for automated evaluation of descriptive answers through sequential minimal optimization. ICTACT Journal on Soft Computing. 2014; 04 (02):710 –4.
10. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. SIGKDD Explorations. 2009; 11(1):10–18.
11. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn. San Francisco: Morgan Kaufmann; 2005.
12. Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Learning; 2003 Aug 21–23; Wahington DC. 856–63.
13. Platt JC. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. 1998. Available from: CiteSeerX:10.1.1.43.4376
14. Model evaluation: quantifying the quality of predictions. Scikit learn; Available from: http://scikit-learn.org/stable/modules/model_evaluation.html
15. Method scoring. Orange; Available from: http://orange.biolab.si/docs/latest/reference/rst/Orange.evaluation.scoring/
16. Evaluation of Classifier's Performance. Machine Learning Corner; 2013 Apr 30. Available from: http://mlcorner.wordpress.com/2013/04/30/evaluation-of-classifiers-performance
17. Kaggle. Mean F Score. Available from: http://www.kaggle.com/wiki/MeanFScore
18. What is Kappa coefficient (Cohen's Kappa). Available from: http://www.pmean.com/definitions/kappa.htm
19. Kaggle. Area under the curve. 2012 [updated 2012 Oct 21]. Availble from: https://www.kaggle.com/wiki/AreaUnderCurve
20. Why does bagging work so well for decision trees, but not for linear classifiers? Available from: http://www.quora.com/Why-does-bagging-work-so-well-for-decision-trees-but-not-for-linear-classifiers
21. Assayag I. Ensemble Learning, moodle-guest.idc.ac.il/mod/resource/view.php?id=64521
22. What are stable and unstable learning algorithms?. Research Gate. Available from: http://www.researchgate.net/post/What_are_stable_and_unstable_learning_algorithms, February 28, 2013
23. Buciu I, Kotropoulos C, Pitas I. Demonstrating the stability of support vector machines for classification. Signal Processing. 2006; 86(9):2364–80.