SATURATION AND PERFORMANCE ANALYSES ON AN EDUCATIONAL DATA SET WITH MACHINE LEARNING ALGORITHMS

Tolga Demirhan
Trakya Üniversitesi Tunca Vocational College, Edirne Turkey
tolgademirhan@trakya.edu.tr

Özlem Uçar
Trakya Üniversitesi Engineering Faculty, Edirne Turkey
oucar@trakya.edu.tr

Abstract

In the present study, saturation and performance analyses were carried out on a data set which consists of educational instances. As a result of the study, the most successful results were obtained by the Random Forest algorithm in the data set created through BiasToUniform:1 and NoReplacement:false customizations in Resampling filter, which was also observed to be saturated. When the confusion matrix of the Random Forest algorithm was investigated, it was seen that 351 of 366 Successful instances and 358 of 366 Unsuccessful instances were predicted accurately by the algorithm. Performance values for the Random Forest algorithm were 97% for Accuracy, 97% for F-Measure and 94% for Kappa.

Keywords: Machine Learning, Educational Data Set, Saturation, Performance Analysis

INTRODUCTİON

Classification is a data mining process which is used for the extraction of information and making future predictions from data sets. In order to obtain successful results in classification, noise cleaning and conversion of record attributes are essential along with the consideration of the number of instances in the data set.

Classification algorithms provide information in the process of learning through the instances in the data set and are trained by means of this information. For that reason, the classification of data within data sets with and insufficient number of instances is highly arbitrary. Also referred to as saturation, this issue is crucial since it provides information on the adequacy of the number of instances in an educational data set [1]. Whether an algorithm is adequately saturated or not can be observed in learning curves, also referred to as Happy Graph, which are obtained using the subsets taken randomly from the data set in varying sizes. The accuracy of a classification algorithm can be evaluated using these learning curves. [2]

A learning curve is generated by plotting the training set size on the x-axis and the classification accuracy on the y-axis in a coordinate system. The learning curve can be defined as the relationship between the classification performance and the size of the training set. It is also used to help determine the most appropriate training data when it comes to the costs related to the provision of samples. [3]

If a learning curve follows a straight line or a downward trend in the y axis after a certain number of instances, or in other words, if the accuracy value does not increase despite the increase in the number of instances, it can be said that the highest possible learning success rate has been achieved and learning saturation has been reached with the related number of instances. [1,3]

In the present study, j48, Random Forest (RF) and Naive Bayes (NB) algorithms of Weka software were used to produce learning curves on the educational data set. [3] In order to take samples in varying sizes from the valid data set, Resampling (biasToUniformClass=0/1, noReplacement = true/false, SampleSizePercent = 20, 40, 60, 80, 100) which is a Weka filter and Cross-Validation (fold = 10) as the test type were used. Through the study, special emphasis was put on saturation in data sets and the factors affecting the obtaining of the data sets and the production of learning curves were investigated.

MATERIALS AND METHODS

The records in the data set which were used in the study comprises of the results of the questionnaires and chapter tests completed by volunteer primary school pupils. Upon the collection of the test and questionnaire data from the pupils by means of optical form, the data set was cleaned of problematic records. For the sake of a positive effect on the classification process, the number of classes in the classify value (sontest) attribute was converted into two classes.

In order to convert the multi-class structure of the Sontest attribute into two classes as Successful and Unsuccessful, averaging, which is also used to find the threshold value in the computation of the bell curve, was used. The average value of the sontest attribute in the available data set was 70. The scores below this value were converted into Unsuccessful and those above the average value were converted into Successful. Thanks to the preprocessing of the data set, a data set with the classify value (sontest) attribute with two classes and 732 instances was obtained.

Weka 3.8.1, which is one of the popular tools in data mining, was used for sampling and classification. Weka was developed by Waikato University in Australia as a Java-based software and it is an open-source data mining software with a number of users across the world. [4] The data set was converted into .arff format which Weka could understand.

```
@relation Basari_Analizi-weka.filters.unsupervised.attribute.Remo

@attribute probleme_dayali {PDO,GELENEKSEL,bos}
@attribute gelir_duzeyi {kotu,orta,iyi,cok_iyi,bos}
@attribute cinsiyet {kiz,erkek,bos}
@attribute yas {10,11,12,13,bos}
@attribute nerede_yasiyorsun {ailemle,yurtta,akrabada,hicbiri,bos
@attribute kiminle_yasiyorsun {annemle,babamla,ailemle,diger,bos}
@attribute odan_varmi {evet,hayir,bos}
@attribute annen_yasiyormu {evet,hayir,bos}
@attribute annen_calisiyormu {evet,hayir,bos}
@attribute annenin_egitim_durumu {ilkokul,ortaokul,lise,universit
@attribute baban_yasiyormu {evet,hayir,bos}
@attribute baban_calisiyormu {evet,hayir,bos}
@attribute babanin_egitim_durumu {ilkokul,ortaokul,lise,universit
@attribute kac_kardessiniz {tekim,iki,uc,dort_ve_daha_fazla,bos}
@attribute engelli_akraba {annem,babam,kardesim,hicbiri,bos}
@attribute odevlerine_kim_yardimci_oluyor {annem,babam,anne_ve_ba
@attribute telefon_numarasini_biliyormusun {evet,hayir,bos}
@attribute adresini_biliyormusun {evet,hayir,bos}
@attribute ulasim {servis,araba,yuruyerek,minibus,bos}
@attribute okulda_vakit_gecirme {hicbirzaman,bazen,siksik,herzama
@attribute en_coksevdigin_ders {fenveteknoloji,matematik,turkce,h
```

Fig 1: *.Arff File*

Moreover, as the classification algorithms to be run on the data set, Naive Bayes, J48 and Random Forest algorithms in the Weka library, which are frequently used in learning curve experiments, were used. In the study, customizations which would affect model success were avoided and the algorithms were used in their default values. Cross-Validation 10 was preferred as the test method.

A. Resampling:

Resampling: Weka includes three types of filters for performing sampling in different sizes on an existing data set. In the present study, Resample filter, which is within the Supervised / Instances category, was used. New data sets with an equal number of class instances can be formed by setting the BiasToUniform field in this filter to 0 or 1. If the BiasToUniform field is set to 0, the class instance distribution among the records in the data set is taken into account and the same field can be set to 1 if this is not necessary. Furthermore, after setting the NoReplacement field to false and a value above 100% is input into the SampleSizePercent field, the sample size may be increased above the number of instances in the existing data set. [5]

B. Classification Algorithms Used:

Naive Bayes: Bayesian classifiers are based on the Naive Bayes theory and assume that each attribute is independent of the other attributes. The conditional probability of a class label is predicted and assumptions are made on the model in order to make this possibility a product of the conditional probabilities. [6]

J48: All instances in a decision tree begin in the root node. The attribute which produces the best discrimination is used in the root node and branches to the inner nodes that are based on the division attribute. The process continues until each one of the instances belong to the same class or as long as there is no more attribute. Among the most useful features of decision trees are their understandability and easy interpretation in the form of rules. In the decision tree, the assumption is that the instances which belong to different classes have at least one attribute with a different value. J48 is a decision tree algorithm based on the popular C4.5 algorithm. [6]

Random Forest (RF): Random Forest is a classification algorithm used in decision trees, which includes a voting method and was developed by Leo Breiman and Adele Cutler. [7]

The purpose of the algorithm is to combine the decisions of a number of multivariate trees, each one of which are trained in different training sets, instead of producing a single

decision tree. In the identification of attributes in each level, the attribute is firstly identified through computations in all trees and then the most frequently used attribute is selected by combining the attributes in all the trees. The selected attribute is included in the tree and the processes are reiterated in the other levels. [8]

EVALUATION METRICS

Confusion Matrix is used for the evaluation of the classification models with respect to their levels of success. A dual-class confusion matrix is as follows.

Table 1. *Confusion Matrix*

		Predict	
		Successful	Unsuccessful
Actual	Successful	TP	FN
	Unsuccessful	FP	TN

TP (True Positive) - FN (False Negative) - FP (False Positive) - TN (True Negative)

Values which aid the interpretation of the performance of an algorithm, such as Accuracy, TP Rate, FP Rate, Sensitivity, Specificity, Precision, Recall, F-Measure, Kappa, ROC, MCC etc., can be obtained using the confusion matrix. In the present study, Accuracy, F-measure, Kappa values were taken into account for the comparison of classification algorithms run on the data sets.

Accuracy: Accuracy is the most popular and simple method in the evaluation of model success. Correct classification rate is the value which gives the rate of correct classifications an algorithm makes using the instances in the data set as both training and test data. [9] The accuracy value provides the classification performance related to whole test data independent of classes. [10]

$$\textbf{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

$$\textbf{Accuracy} = \frac{\text{Sum of the number of accurate predictions}}{\text{The number of instances}} \quad (2)$$

The accuracy of a classification algorithm can be evaluated by means of the learning curves formed using accuracy values. [2]

However, since the classify value used in the present study has a dual-class structure, it is considered to be inaccurate to evaluate classification success only by checking the accuracy value. [11]

Kappa: Kappa Statistic (KS) is a measure which is used to express the level of agreement between predicted and observed classifications in a data set quantitatively. KS value is between -1 and 1. -1 indicates total disagreement or a negative relationship. 1 indicates perfect agreement. KS values above 0.4 indicate that the agreement is acceptable beyond chance. [12] Kappa Statistic value is calculated with the formula below.

$$\boldsymbol{K} = \frac{(P_a - P_c)}{(1 - P_{c)}} \quad (3)$$

P(a) shows the accuracy of the classifier and P(c) is the expected accuracy of the classifier which makes random predictions on the same data set. [13]

F-Measure: F-Measure is expressed as the harmonic mean of precision and recall values. F-Measure is especially used to find out the performance of the classifier during the preparation of training data and to determine if the classes are sufficient for diagnosis. The acceptable F-Measure value is generally taken as 0,5. [12] This value is used as 0,5 in the present study.

$$\boldsymbol{F - Measure} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

Precision: Precision is the ratio of the number of True Positive (TP) instances predicted as class 1 to the total number of instances (TP + FP) predicted as class 1.

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \quad (5)$$

Recall: Recall is the ratio of correctly classified True Positive (TP) instances to actual total positive (TP + FN) instances.

$$\boldsymbol{Recall} = \frac{TP}{TP+FN} \quad (6)$$

Learning Curve: Learning Curve is one of the tools used to visually present the performance of classification. In the Learning Curve, classification performance is given on the Y axis while the number of training instances is given on the X axis. Learning Curves show that sufficient data has been obtained and no more learning will occur. Furthermore, Learning

Curve can also show the performance of different classification algorithms on different sets of data. [14]

The accuracy of the classification algorithm can be evaluated using the Learning Curve. In the learning curve given below, the accuracy value is observed to increase when the size of the data set increases. The curves in this view can also be given as a Happy Graph. [2]

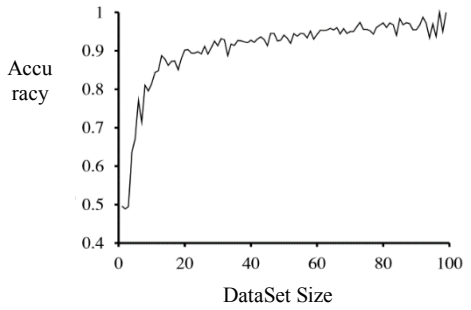

Fig 2: *Learning Curve – Happy Graph*

EXPERIMENTAL STUDY AND RESULTS

Experiments were performed using Weka 3.8.1. Naive Bayes, C4.5 decision tree, Random Forest (RF) algorithms were used as the basic classifiers. J48 in Weka was used as C4.5. To obtain samples of different sizes from the valid data set, Resampling filter

(biasToUniformClas=0/1,noReplacement= true/ false, SampleSizePercent=20%, 40%, 60%, 80%, 100%) from Weka filters was used and a total of 20 data sets of different sizes were generated from the 732-sample data set. These data sets are given below. These datasets were grouped into types I, II, III and IV due to the customizations made in the Resampling filter to facilitate tracking them.

Table 2: Resampling: BiasToUniform:0 - NoReplacement:false

	%	Class		Total Instance
		Successful	Unsuccessful	
I.TYPE	20	123	23	146
	40	246	46	292
	60	369	70	439
	80	492	93	585
	100	615	117	732

Table 3: Resampling: BiasToUniform:0 - NoReplacement:true

	%	Class		Total Instance
		Successful	Unsuccessful	
II.TYPE	20	123	23	146
	40	246	46	292
	60	369	70	439
	80	492	93	585
	100	615	117	732

Table 4: Resampling: BiasToUniform:1 - NoReplacement:false

	%	Class		Total Instance
		Successful	Unsuccessful	
III.TYPE	20	73	73	146
	40	146	146	292
	60	219	219	438
	80	292	292	584
	100	366	366	732

Table 5: Resampling: BiasToUniform:1 - NoReplacement:true

	%	Class		Total Instance
		Successful	Unsuccessful	
IV.TYPE	20	63	83	146
	40	146	117	263
	60	219	117	336
	80	292	117	409
	100	366	117	483

Among the classification results obtained by taking into account the accuracy, F-measure and Kappa values from the classification algorithms run on the generated data sets, the most successful results were obtained from the Type III data sets and the Random Forest algorithm. The analysis results obtained from these data sets by the Random Forest algorithm are given below.

Table 6. Resampling: BiasToUniform:1- NoReplacement: false

%	Class's Instances		Total Instance	Analysis Results		
	Successful	Unsuccessful		Accuracy	F-Measure	Kappa
20	73	73	146	0,82	0,82	0,64
40	146	146	292	0,88	0,88	0,76
60	219	219	438	0,93	0,93	0,86
80	292	292	584	0,97	0,97	0,95
100	366	366	732	0,97	0,97	0,94

The Accuracy, F-Measure and Kappa values obtained by the Random Forest algorithm from the Type III data sets are shown below in graphical form.

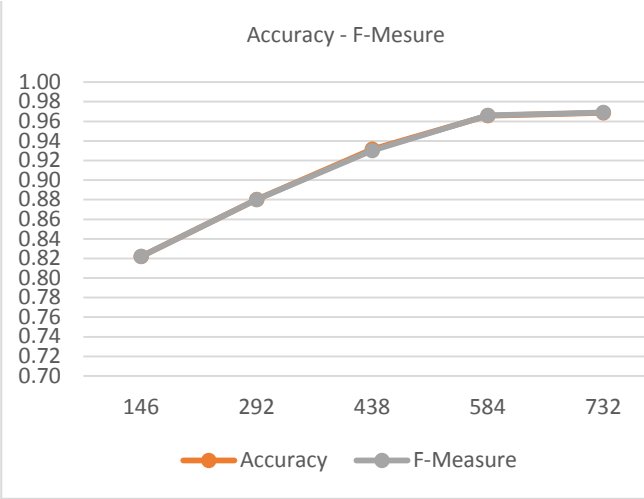

Fig 3: *Learning Curve*

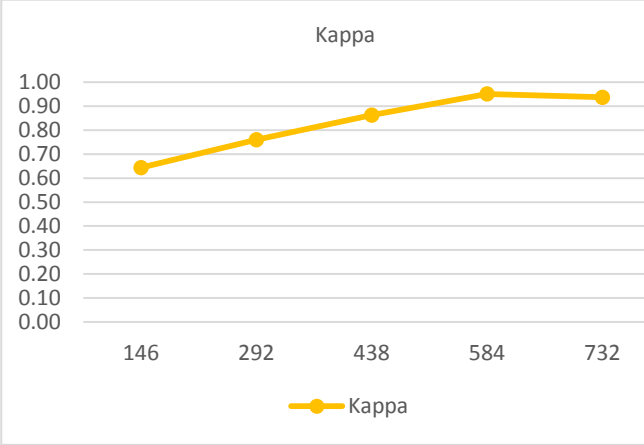

Fig 4: *Kappa Curve*

The results obtained by the Random Forest algorithm on the Type III data set with 732 instances are given below with the Weka Output screen.

```
Correctly Classified Instances         709               96.8579 %
Kappa statistic                          0.9372
Mean absolute error                      0.0872
Root mean squared error                  0.1667
Relative absolute error                 17.4325 %
Root relative squared error             33.3341 %
Total Number of Instances              732

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,959    0,022    0,978      0,959    0,968      0,937    0,997     0,996     BASARILI
                 0,978    0,041    0,960      0,978    0,969      0,937    0,997     0,997     BASARISIZ
Weighted Avg.    0,969    0,031    0,969      0,969    0,969      0,937    0,997     0,997

=== Confusion Matrix ===

   a   b   <-- classified as
 351  15 |   a = BASARILI
   8 358 |   b = BASARISIZ
```

Fig 5: *Weka Output*

CONCLUSION

In the present study, Random Forest, J48 and Naive Bayes algorithms were run on 20 data sets generated by using the Resampling filter on an educational data set. Saturation and learning curve behavior were investigated with respect to the related data sets and classification algorithms using the results obtained.

When the experimental results were scrutinized, it was seen that the data sets with which the most successful results had been obtained were among the Type III data sets which were generated by means of the BiasToUniform:1 and NoReplacement:false customizations on the Resampling filter.

When the classification algorithm experiments conducted on the Type III data sets were investigated in terms of Kappa values, it was seen that the obtained classification success was not arbitrary. The same experiments suggested that the classes were adequate for diagnosis when they were investigated in terms of F-Measure and Accuracy.

Rank ordering the success levels of the classification algorithms with Type III data sets showed that the most successful algorithm was Random Forest. In this order, J48 ranked the second and Naive Bayes ranked the third. When the confusion matrix obtained by the Random Forest algorithm, which produced the most successful results in the Type III data set with 732 instances, was examined, it was seen that the algorithm accurately predicted 351 of 366 Successful instances and 358 of 366 Unsuccessful instances. This value produces an Accuracy of 97%, an F-Measure of 97% and a Kappa value of 94%.

When the Type III data sets were investigated, it was seen that the instance count of two classes belonging to the classify value (sontest) attribute were equal-balanced. Looking at the criteria values obtained, it can be said that balanced data sets have an effect on classification success.

When the learning curves in Fig 5 were examined, it was seen that the accuracy values reached the highest possible success level in the experiments with a certain number of data sets and increasing the number of instances after this point had no effect on accuracy success. This shows that the model reached its learning saturation and a sufficient number of instances were used in the experiments. [2]

REFERENCES

[1] Eminağaoğlu, M., Özdevimli Öğrenme Yaklaşımı ile Bilgi Güvenliği Risklerinin Nitel Değerlendirilmesine Yönelik Bir Model,Doktora Tezi, 2011.

[2] Russell, S., and Norvig, P., Artificial Intelligence A Modern Approach Third Edition, ISBN-13: 978-0-13-604259-4,p:702-703, 2010.

[3] Weiss, G. M., Provost, F., "Learning when training data are costly: the effect of class distribution on tree induction". Journal of Artificial Intelligence Research, 19, s. 315-354, 2003.

[4] Dener, M., Dörterler, M., Orman, A., "Açık kaynak kodlu veri madenciliği programları: WEKA'da örnek uygulama", Akademik Bilisim'09 - XI. Akademik Bilişim Konferansı, Şanlıurfa, 787-796, 2009.

[5] https://weka.wikispaces.com/Classifying+large+datasets,2017.

[6] Mishra, T., Kumar, D., Gupta, S., Students' Employability Prediction Model through Data Mining, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 4, p:2279, 2016.

[7] Korkem,E., Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest Ve Naïve Bayes Sınıflama Yöntemleri Yaklaşımı, Yüksek Lisans Tezi, Hacettepe Üniversitesi, Sağlık Bilimleri Enstitüsü, 2013.

[8] http://bolubeyi.net/karar-agaclari-ve-algoritma-turleri.html,Mayıs 2017.

[9] Özhan E., Yapay Zeka Yöntemleri İle Dünya Depremlerinin Modellene Bilirliği, Isıtes2014 Karabük – Turkey,1188, 2014.

[10] Klement, W., Wilk, S., Michaowski, W., Matwin, S.: Dealing with Severely Imbalanced Data. In: ICEC 2009 Workshop, PAKDD, p:7, 2009.

[11] Bulut,F., Performance Analysis of Ensemble Methods on Imbalanced Datasets, Bilişim Teknolojileri Dergisi, Cilt: 9, Sayı: 2, Mayıs 2016 - DOI: 10.17671/Btd.81137.

[12] Aydın F., Kalp Ritim Bozukluğu Olan Hastaların Tedavi Süreçlerini Desteklemek Amaçlı Makine Öğrenmesine Dayalı Bir Sistemin Geliştirilmesi, Yüksek Lisans Tezi,p:39-41,2011.

[13] Kılıçaslan Y., Güner E. S., Yıldırım S., Learning-based pronoun resolution for Turkish with a comparative evaluation, Computer Speech & Language Volume 23, Issue 3, July 2009, Pages 311-331

[14] Remco R. Bouckaert, Efficient AUC Learning Curve Calculation19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006.