ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

One-Vs-All Binarization Technique in the Context of Random Forest

Md Nasim Adnan and Md Zahidul Islam

Centre for Research in Complex Systems (CRiCS)
School of Computing and Mathematics, Charles Sturt University
Bathurst, NSW 2795, Australia
madnan@csu.edu.au, zislam@csu.edu.au.

Abstract. Binarization techniques are widely used to solve multi-class classification problems. These techniques reduce the classification complexity of multi-class classification problems by dividing the original data set into two-class segments or replicas. Then a set of simpler classifiers are learnt from the two-class segments or replicas. The outputs from these classifiers are combined for final classification. Binarization can improve prediction accuracy when compared to a single classifier. However, to be declared as a superior technique, binarization techniques need to prove themselves in the context of ensemble classifiers such as Random Forest. Random Forest is a state-of-the-art popular decision forest building algorithm which focuses on generating diverse decision trees as the base classifiers. In this paper we evaluate one-vs-all binarization technique in the context of Random Forest. We present an elaborate experimental result involving ten widely used data sets from the UCI Machine Learning Repository. The experimental results exhibit the effectiveness of one-vs-all binarization technique in the context of Random Forest.

1 Introduction

Nowadays the amount of data is increasing at an astonishing pace; so much that 90% of the data in the world today have been generated in the previous two years [1]. The large volume of data amplifies the need for developing sophisticated means for automatic knowledge discovery. Data mining is the method of automatically discovering useful information from large data sets [2]. Classification and clustering are two popular data mining tasks that are applied for knowledge discovery and pattern recognition.

Classification aims to generate a function that maps the set of non-class attributes $\{A_1, A_2, ..., A_m\}$ to a predefined class attribute C [2] and the function is commonly known as the classifier. There are different types of classifiers that includes Decision Trees [3], [4], [5], Artificial Neural Networks [6], and Support Vector Machines [7]. Among these classifiers, decision trees are rigorously used in the real world scenario as they can be broken down to generate logic rules to infer valuable knowledge [8]. Due to their immense popularity, decision trees with better prediction accuracy can render huge impact on many sensitive application areas such as medical diagnosis.

In general, it is easier to induce a decision tree to distinguish between two class values (known as binary-class values) than more than two class values (known as multi-class values) since the decision boundaries for the binary-class values are simpler compared to that of multi-class values [9]. This is why binarization techniques are applied on the multi-class problems to decompose the original problem into easy to solve binary classification problem [10].

A lot of binarization techniques can be found in the literature [11]. Most of them can be categorized into two groups called "One-Vs-One"(OVO) and "One-Vs-All"(OVA) as described in the following:

1. OVO technique resolves the multi-class problem in such a way that the training data set is partitioned into several segments where each segment contains a distinct pair of class values. If there is C number of distinct class values then the total number of segments will be: $C(C-1)/2$.

2. OVA technique resolves the multi-class problem in such a way that the training data set is replicated into several data sets where each replica contains one original class vale and all other class values of the training data set are replaced as "O". Thus, if there is C number of distinct class values then the total number of replicas will be: C. The number of records in each replica is the same as the original training data set.

After binarization, for both OVO and OVA technique we get one classifier (decision tree for this paper) from each of the segments/replicas of the training data set. Thus, the outputs from more than one decision trees need to be combined for final classification.

Decision forest is essentially an ensemble of decision trees where an individual decision tree acts as the base classifier and the classification is performed by taking a vote based on the predictions made by each decision tree of the decision forest [2]. A decision forest overcomes some of the shortcomings of a decision tree. A decision tree is entirely formed from the training data set. This enables a decision tree to have remarkable classification performance on the examples (records) of the training data set. However, the classification performance on the seen examples does not necessarily get translated into predicting the class values of the unseen (unlabelled) records of the testing data set. Decision tree in particular, lacks in generalization performance. Nevertheless, different decision trees have different generalization errors. Thus the combination of several decision trees can help overcoming the generalization errors of a single decision tree.

We argue that after binarization the solution in effect becomes an ensemble of classifiers which has advantages over a single classifier. Thus, to determine whether binarization techniques are in general preferable or not we need to evaluate the techniques in the context of ensembles. In literature, Fürnkranz [10] compared the suitability of OVO strategies for decision trees and decision lists within Bagging and Boosting. In this paper, for the first time we evaluate OVA binarization technique in the context of state-of-the-art ensemble of decision trees namely - Random Forest [12].

2 OVA Binarization in the Context of Random Forest

For our experimentation we choose to generate 100 decision trees for Random Forest which is large enough to ensure convergence of ensemble effect [14]. Thus, when building 100-tree Random Forest without applying any binarization technique, we need to generate 100 different bootstrap samples. It is worthy to mention that decision tree is very sensitive to the training data set [2]. That is - if the training data set is slightly perturbed by adding or removing some records or attributes, the resultant decision tree can be very different. Thus, the bootstrap samples can render significant diversity among the trees which in turn increases the ensemble accuracy. To be fair in comparison, we have to generate around 100 trees for Random Forest even when OVO or OVA technique is applied. Let us assume that we have a data set with five (05) different class values. If we want to binarize a bootstrap sample of the data set using OVO technique we get approximately $5 \times (5 - 1)/2 = 10$ segments and thus 10 trees from a single bootstrap sample. As a result, only 10 different bootstrap samples of the original training data set can be used which can significantly trim down diversity. Moreover, OVO binarization generates data segments from the bootstrap samples in such a way that there is only a pair of distinct class values is present in each data segment. The records containing other class values are excluded. This phenomenon significantly reduces the size of the data segment compared to the bootstrap sample whose size is equal to that of the original training data set. On the other hand, in our example when OVA is applied on a bootstrap sample we get exactly 5 replicas and thus 5 trees from a single bootstrap sample. Thus, we can use 20 different bootstrap samples. We admit that in this case also diversity can be degraded yet compared to OVO it would be significantly higher. More importantly, OVA generates replicas with the same size of a bootstrap sample where each replica contains one original class value and all other class values of the training data set are replaced as “O”. For the reasons stated above, we find OVA is more applicable for one to one comparison in the context of Random Forest.

3 Experimental Results

We conduct an elaborated experimentation on ten (10) data sets with more than two class (multi-class) values covering almost every well known multi-class data sets from the UCI Machine Learning Repository [15]. The data sets used in the experimentation are listed in Table 1. We generate 100 trees for both Regular Random Forest (R_RF) and One-Vs-All Random Forest (OVA_RF). Majority voting is used to aggregate the ensemble results. All the results reported in this paper are obtained using 10-fold-cross-validation (10-CV) for every data set. The best results are emphasized through **bold-face**.

Ensemble accuracy is one of the most important performance indicators for any decision forest algorithm. From Table 2, we see that the OVA_RF outperforms the R_RF for eight (08) out of ten data sets considered (including one

Table 1: Description of the data sets

Data Set Name	Attributes	Records	Distinct Class Values
Balance Scale	04	625	3
Car Evaluation	06	1728	4
Dermatology	34	366	6
Ecoli	08	336	8
Glass Identification	10	214	7
Hayes-Roth	05	160	3
Iris	04	150	3
Lenses	04	24	3
Soybean (Small)	35	47	4
Statlog (Vehicle)	18	946	4

Table 2: Ensemble Accuracies (in percentage)

Data Set Name	R_RF	OVA_RF
Balance Scale	**83.8670**	82.5640
Car Evaluation	74.2340	**81.6600**
Dermatology	85.5890	**92.6780**
Ecoli	**84.5240**	**84.5240**
Glass Identification	66.0940	**73.6390**
Hayes-Roth	63.3840	**70.3080**
Iris	**95.3330**	94.6660
Lenses	71.6670	**78.3330**
Soybean (Small)	99.0910	**100.0000**
Statlog (Vehicle)	71.1760	**73.5320**
Average	79.4959	**83.1904**

tie). On an average OVA_RF achieve a significant improvement in prediction accuracy (**83.1904%**) over that of the R_RF (79.4959%). Another important observation to note from Table 3 is that OVA_RF performs far better with the data sets having more distinct class values (more than 3 in this case).

We already know the decision boundaries to distinguish the binary-class values are simpler compared to that of multi-class values. Thus decision trees generated following binarization should be simpler and generated faster. From Table 3 we find that trees generated from OVA_RF have comparatively less nodes and depth. This implies that the rules generated from OVA_RF are more concise, and thus more preferable [16]. As expected, trees generated from OVA_RF takes comparatively less time than R_RF (see Table 4).

Table 3: Tree Structure

	Average Tree Nodes		Average Tree Depth	
Data Set Name	R_RF	OVA_RF	R_RF	OVA_RF
Balance Scale	103.5000	**66.1717**	3.0160	**2.0111**
Car Evaluation	129.5560	**42.0550**	4.8650	**2.1050**
Dermatology	17.6960	**10.0304**	3.5190	**1.9441**
Ecoli	43.4420	**10.4885**	7.9280	**3.3712**
Glass Identification	46.3860	**13.9000**	8.6480	**4.0284**
Hayes-Roth	14.8810	**14.1778**	2.6040	**2.4626**
Iris	10.0400	**7.1859**	3.8080	**2.5778**
Lenses	7.1740	**5.7788**	2.4010	**2.0374**
Soybean (Small)	6.4750	**5.1700**	1.8480	**1.4020**
Statlog (Vehicle)	150.5020	**65.4100**	14.3900	**10.7180**
Average	52.9652	**24.0368**	5.3027	**3.2658**

4 Conclusion

The main contribution of this paper is to evaluate the OVA randomization technique in the context of Random Forest. From our study, we find that OVA randomization has a great potential in the context of Random Forest. In future, we plan to apply OVA randomization methods on some of the latest forest building algorithms such as Rotation Forest [17].

References

[1] IBM Co.: Bringing big data to the Enterprise. http://www-01.ibm.com/software/au/data/ bigdata/, (Last Accessed: 25 Jul 2013)

[2] Tan, P., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Education, Boston (2006)

[3] Breiman, L., Friedman J., Olshen R., Stone C.: Classification and Regression Trees. Wadsworth International Group, CA, U.S.A. (1985)

[4] Quinlan, J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)

[5] Quinlan, J. R.: Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, vol. 4, pp.77-90 (1996)

[6] Zhang, G. P.: Neural Networks for Classification: A Survey. IEEE Transactions on Systems, Man, and Cybernetics, vol. 30, pp. 451-462 (2000)

[7] Burges, C. J. C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, vol. 2, pp. 121-167 (1998)

Table 4: Average Tree Generation Time (in seconds)

Data Set Name	R_RF	OVA_RF
Balance Scale	0.8895	**0.4094**
Car Evaluation	0.7728	**0.2913**
Dermatology	0.3356	**0.1530**
Ecoli	0.9340	**0.1101**
Glass Identification	1.1467	**0.2234**
Hayes-Roth	0.1331	**0.1073**
Iris	0.0787	**0.0438**
Lenses	0.0440	**0.0321**
Soybean (Small)	0.0932	**0.0394**
Statlog (Vehicle)	5.5128	**1.5633**
Average	0.9940	**0.2973**

[8] Murthy, S. K.: Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Mining and Knowledge Discovery, vol. 2, pp. 345-389, (1998)

[9] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: An verview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognition, vol. 44, pp. 1761-1776, (2011)

[10] Furnkranz, J.: Round robin classification. Journal of Machine Learning Research, vol. 2, pp. 721-747 (2002)

[11] Lorena, A. C., Carvalho, A. C., Gama, J. M.: A review on the combination of binary classifiers in multiclass problems. Artificial Intelligence Review, vol. 30, pp. 19-37, (2008)

[12] Breiman, L.: Random Forests. Machine Learning, vol. 45, pp. 5-32 (2001)

[13] Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2nd ed. Morgan Kaufmann, San Francisco, U.S.A. (2006)

[14] Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning, vol. 63, pp. 3-42 (2006)

[15] UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets.html (Last Accessed: 3 June 2014)

[16] Geng, L., Hamilton, H. J.: Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys, vol. 38 (2006)

[17] Rodriguez, J. J., Kuncheva, L. I., Alonso, C. J.: Rotation Forest: A New Classifier Ensemble Method. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 28, pp. 1619-1630, (2006)