The use of combined methods to classify data with difficult distributions (including unbalanced)

Romeo Rego

Wroclaw University of Science and Technology, wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: Classification · Imbalanced Data · Classifier · Classifier Ensemble

1 Related works

As imbalanced data is common in such fields as fraud detection or medical diagnosis, interest in this field increased in recent years. Because of that there are many interesting works, many of which employ ensembles of classifiers. Although research on learning from imbalanced data is conducted for more than two decades, there are still unsolved problems which need to be addressed. Krawczyk [1] provided a great overview of these issues. In his work, he discusses different types of machine learning tasks (binary/multi-class classification, regression, etc.) and problems related to them, as well as proposes some methods of solving them. Particularly he acknowledges importance and efficency of ensemble classifiers. He also warns that these methods are not without flaws and there is a need of studying them more deeply. According to him, most important things are lack of understanding of diversity in imbalanced learning, lack of clear indicators on how large ensembles should be constructed and methods of combining decisions of individual classifiers other than majority voting.

Some of these problems were addressed by Ksieniewicz [2] who proposed the *Undersampled Majority Class Ensemble*. He divided a imbalanced dataset into smaller balanced datasets by splitting samples from majority class into equal subsets. Each of these datasets were given to a member of an ensemble. In an extended version he also introduced an additional classifier trained on a whole oversampled dataset. He also proposed a novel method of *response pruning*, which bases on an analysis of statistical dependencies of the classifiers response on the testing set. His result show that this approach, especially with added classifier trained on oversampled dataset and *response pruning* significantly improves the results of an ensemble.

Ksieniewicz and Woźniak [3] propose a completely new tool, a so called *exposer*. It is an ensemble of subspace projections, spanned on combined features of data. It adjusts itself to tested objects, allowing for dynamical feature extraction. Exposer also makes use of *random subspace* method and randomly selects attributes. One of the big advantages of this tool is that all the attributes are fully interpretable but invertible. It also shows robustness to *curse of dimensionality* which makes it useful for big data problems. As experiments show, *Exposer Classifier Ensemble (ECE)* in most cases outperforms all of the reference methods with a difference even as high as 10%.

Uma R. Salunkhea and Suresh N. Malib [4] proposed a hybrid approach for ensemble classifier construction. In their work they use both re-sampling of the data as well as different ensembles. In data preprocessing phase they used over-sampling (SMOTE) and random under-sampling of the majority class having previously identified necessary data samples. In classifier formation phase they used two methods. First was to construct a bagging ensemble using different training data subsests and J48 algorithm. Second method was to construct a StackingC ensemble and to use different algorithms, namely J48, Logistic Regression and Bagging. Their experiments were conducted on eight datasets from KEEL repository. Results show that on all used datasets a high score was achieved, additionally StackingC ensemble significantly outperforms the other method.

Ksieniewicz and Burduk [5] created a new and robust method of integrating decisions of base classifiers in an ensemble. The proposed algorithm uses a weighted scoring function but it calculates the weight for a recognized objects instead of base classifiers. A K-Means algorithm is used for every class algorithm separately. Then the weighted scoring function takes the distance of classified object from decision boundary and the centroids of clusters into account. The advantage of this method is that it is insensitive to sample count because centroids can be calculated irrespectively of the number of objects in a cluster. Comprehensive experiments have shown that this algorithm works better than other algorithms in the context of statistical tests, especially in the case of balanced accuracy and G-mean.

Park and Ghosh [6] introduced two decision tree ensembles in which they used a new decision criterion based on α-divergence which generalises several known splitting criteria such as those used in C4.5 and CART. Their experiments show that by changing the value of α one can obtain less correlated decision trees which is beneficial in an ensemble of classifiers. This is proven by a better scores in terms of AUROC metric. In one of proposed ensembles a lift-aware stopping criterion was used. Because of that the ensemble produces a set of interpretable rules that provide higher lift values for a given coverage.

2 Experiment design

2.1 Hypothesis

Based on the related works, described in section 1 it can be seen that ensemble classifiers are popular in imbalanced data learning. Moreover they seem to perform very well in this field. One can draw a conclusion that combining decisions of multiple models into one final decision, reduces (or completely removes) biases of specific models and improves overall score. Yet it has to be proven that it is really so. The goal of this work will be to verify if combining multiple models into an ensemble allows for achieving significantly better results in comparison to individual models that make up the ensemble.

2.2 Experiment plan

When building an ensemble, there are multiple ways of constructing its members. There are four most common scenarios, which are:

1. Using different training sets: every classifier is trained with different subset of the training data. Algorithm stays the same.
2. Using different feature subsets: every classifier is trained with different subset of features.
3. Using different algorithms: every member of an ensemble is using different algorithm. Training data is the same for all members.
4. Combination of methods presented above.

In this work, only the third scenario is covered. Three algorithms were chosen for construction of the ensemble, namely: Gaussian Process Classifier, Support Vector Machine and a Feedforward Neural Network. Gaussian Process Classifier (GPC) [10] works by placing a Gaussian Process prior on latent function *f*. This function is then squashed through a link function and thus a probabilistic classification is obtained. This work uses *scikit-learn* GPC implementation. Support Vector Machine [11] is a widely used algorithm, which has proven to be very versatile and robust, while not being as complicated as neural networks. This algorithm uses so called *support vectors*, which are simply co-ordinates of individual observations, to find a hyperplane which separates classes best. The last of algorithms is the most basic type of neural network called Feedforward Neural Network (FFNN) or Multi Layer Perceptron (MLP). It consists of many hidden layers of neurons, connected sequentially. Each neuron (Perceptron) calculates a weighted sum of its inputs and gives it as an input to the activation function. This function then calculates an output of a neuron, which is either 0 or 1.

Performance of constructed ensemble is verified on twenty datasets from KEEL repository [7] and five generated datasets. Number of datasets is determined by the requirements of the statistical test, which needs at least 25 samples to achieve meaningful results. KEEL repository provides a large collection of benchmark datasets for imbalanced data learning problem. To enrich the experiment with even more different imbalance ratios, five synthetic datasets were generated. They are listed in table 1 as *generated*{*1-5*}. These datasets cover imbalance ratio span not present in KEEL datasets. All datasets used in experiments are described in table 1.

In order to evaluate robustness of the constructed ensemble, each of the individual models used in it was also trained separately. Results of these model were then compared to results of the ensemble. This allows for evaluation whether the

Name	Attributes	Examples	Imbalance ratio
glass1	9	214	1.82
wisconsin	9	683	1.86
iris0	4	150	2
haberman	3	306	2.78
vehicle1	18	846	2.9
vehicle0	18	846	3.25
new-thyroid1	5	215	5.14
segment0	19	2308	6.02
glass6	9	214	6.38
yeast3	8	1484	8.1
ecoli3	7	336	8.6
page-blocks0	10	5472	8.79
yeast-2_vs_4	8	514	9.08
yeast-0-5-6-7-9_vs_4	8	528	9.35
glass-0-1-6_vs_2	9	192	10.29
shuttle-c0-vs-c4	9	1829	13.87
abalone9-18	8	731	16.4
shuttle-c2-vs-c4	9	129	20.5
yeast-2_vs_8	8	482	23.1
generated1	11	1000	30.0
generated2	12	1000	50.0
generated3	13	1000	70.0
generated4	14	1000	90.0
generated5	15	1000	110.0
abalone19	8	4174	129.44

Table 1. Datasets used in experiments.

ensemble indeed performs better than singular models. Metric used for assessing the performances of the models is Balanced Accuracy [8]. It is an average of Specificty and Sensitivity. Equation 1 shows how this metric is calculated.

$$Balanced\ accuracy = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}) \tag{1}$$

Unlike classic accuracy, it is a metric that is well suited for evaluating classifiers dealing with imbalanced data. Area Under Receiver Operating Statistic (AUROC) was also considered, but in the end rejected as there is a risk of misleading results given by this metric [9].

2.3 Statistical tests

Although models performances were evaluated using appropriate metric, it is still unknown whether these results are statistically significant. To prove that these results did not happen by chance, a statistical tests were performed. A method chosen in this work is Wilcoxon Signed Rank Test. It is a paired test, which means that it has to be performed for each pair of models. Used on four models, it gives total number of 12 tests.

References

1. Krawczyk, B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5, 221–232 (2016). https://doi.org/10.1007/s13748-016-0094-0
2. Ksieniewicz, P.. (2018). Undersampled Majority Class Ensemble for highly imbalanced binary classification. Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications, in PMLR 94:82-94
3. Ksieniewicz, P. & Woźniak, M.. (2017). Dealing with the task of imbalanced, multidimensional data classification using ensembles of exposers. Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, in PMLR 74:164-175
4. Uma R. Salunkhea, Suresh N. Malib. Classifier Ensemble Design for Imbalanced Data Classification: A Hybrid Approach (2016). https://doi.org/10.1016/j.procs.2016.05.259
5. Ksieniewicz, Paweł & Burduk, Robert. (2020). Clustering and Weighted Scoring in Geometric Space Support Vector Machine Ensemble for Highly Imbalanced Data Classification. 10.1007/978-3-030-50423-6_10.
6. Park, Yubin & Ghosh, Joydeep. (2014). Ensembles of α-Trees for Imbalanced Classification Problems. Knowledge and Data Engineering, IEEE Transactions on. 26. 131-143. 10.1109/TKDE.2012.255.
7. J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3 (2011) 255-287.
8. K. H. Brodersen, C. S. Ong, K. E. Stephan and J. M. Buhmann, ”The Balanced Accuracy and Its Posterior Distribution,” 2010 20th International Conference on Pattern Recognition, Istanbul, 2010, pp. 3121-3124, doi: 10.1109/ICPR.2010.764.
9. Bekkar, Mohamed & Djema, Hassiba & Alitouche, T.A.. (2013). Evaluation measures for models assessment over imbalanced data sets. Journal of Information Engineering and Applications. 3. 27-38.
10. Carl Eduard Rasmussen and Christopher K.I. Williams, “Gaussian Processes for Machine Learning”, MIT Press 2006
11. Noble, William Stafford. “A biologist ’ s introduction to support vector machines.” (2006).