ENSEMBLE CLASSIFIER AND RESAMPLING FOR IMBALANCED MULTICLASS LEARNING

Mohd Shamrie Sainin[1], Faudziah Ahmad[2] and Rayner Alfred[3]

[1,2]School of Computing, Universiti Utara Malaysia, shamrie@uum.edu.my
[3]Faculty of Computing and Informatics, Universiti Malaysia Sabah, ralfred@ums.edu.my

ABSTRACT. An ensemble classifier called DECIML has previously reported that the classifier is able to perform on benchmark data compared to several single classifiers and ensemble classifiers such as AdaBoost, Bagging and Random Forest. The implementation of the ensemble using sampling was carried out in order to investigate if there are any improvements in the classification performances of the DECIML. Random sampling with replacement (SWR) method is applied to minority class in the imbalanced multiclass data. Results show that the SWR is able to increase the average performance of the ensemble classifier.

Keywords: ensemble classifier, DECIML, imbalance, multiclass, data mining, machine learning, sampling

INTRODUCTION

A multiclass classification is a special case within statistical classification of assigning one of several class labels to an input object. Unlike the binary classification, learning multiclass problems is a complex task to exploit as each example can only be assigned from more than two class labels. Binary classifiers normally can be extended to solve the multiclass problem. However classifiers that are designed for binary classification are not effective to be used in imbalanced multiclass classification tasks (Lerteerawong & Athimethphat, 2011).

Data with multiclass labels has more than two classes and imbalance problem in this data occurs when one of the classes (the minority class) is heavily under-represented in comparison to the other class (the majority class) in training dataset (Fergani & Belkacem, 2014). With the existence of imbalance data in a multiclass classification task, traditional classification methods cannot be applied efficiently and effectively since they generally assume data are well distributed (Ding, 2011). The common issue in imbalanced data is standard machine learning methods assume that all examples have the same importance and tend to overlook the minority class examples to achieve high accuracy (Liu & Li, 2014).

Methods for imbalance problem can be categorized in two groups based on their approaches, namely data-level (also known as sampling) and algorithm-level. This categorization was first mentioned by Garcia, Sanchez, Mollineda, & Sotoca (2007). Both data-level and algorithm-level approach have their own advantages and drawbacks. While data-level methods (under-sampling and over-sampling) are able to produce balanced data for the classifier to work, however the methods could lead to duplicates or potential data loss, thus overfitting may occur (Liu & Li, 2014). Algorithm –level on the other hand Algorithm-level has two possible utilizations; either specific new algorithm construction or the existing algorithm

is tuned in order to produce high performance on imbalance problem. Ensemble methods recently have gained its recognition for improving the classifiers on imbalanced multiclass problem (Liu & Li, 2014; Sainin & Alfred, 2012).

In this research, the ensemble method incorporates several single classifiers with basic sampling known as sampling with replacement (SWR). SWR is applied to increase the number of samples in minority class so as to obtain balanced data. Finally, all classifiers in the ensemble will be combined through weighted voting. This paper aims to present the experimental results. In summary, the ensemble method with SWR produces better performance for imbalanced multiclass when average accuracy, f-measure and g-mean used as evaluation metric.

RELATED WORK

Data-Level Method for Imbalanced Multiclass

Data-level method is concerned about how the data are presented to the classifier to address the imbalance problem. There are two methods that are associated with data-level method which are row-based (e.g. sampling) and column-based (feature selection). The traditional re-sampling methods are known as over-sampling and under sampling. More specific sampling methods can be implemented as sampling with replacement, sampling without replacement, random-oversampling, improved sampling, and combination of such sampling methods. Sampling has the drawback in classification as mentioned in Wang & Yao (2012). Over-sampling tends to introduce more examples and causes over-fitting the minority class (indicated by the low recall and high precision or F-measure). In under-sampling, it is sensitive to the number of minority classes and can suffer from performance loss on majority classes.

In this paper, random sampling with replacement is investigated to identify if there is a significant effect of sampling to classification performance. Sampling with replacement is a method to create a finite population from a random sample may be selected until a desired number of samples are obtained, where each selected sample will be put back to the original random sample for the next selection. Recent research by Farid, Rahman, & Rahman (2011) shows that by using the sampling with replacement method, significant improvement achieved in their classification performance over various benchmark datasets. However, their benchmark data are not specifically addressing the issue of imbalanced multiclass problem.

Algorithm-Level Method for Imbalanced Multiclass

Due to the drawbacks of sampling (Lerteerawong & Athimethphat, 2011), methods that are specifically implemented using existing algorithms to solve the imbalance problem are another interesting alternative. Many of earlier machine learning algorithms are considered as a single classifier which was proposed to solve the binary and multiclass data classification problem (Ding, 2011). Some of the learning algorithms have been theoretically studied for their effectiveness in various application domains that they become standard machine learning topics. However, direct single classifiers have some disadvantages and shortcomings where none of them has been consistently performing well over various datasets (Zhang & Yang, 2008).

Combining classifiers, or known as an ensemble of classifiers, is defined as a method that consists of many individually trained classifiers whose decision are combined when classifying new example (Garcia et al., 2007). Although ensembles are a well-established research line (Garcia et al., 2007), they are only valid for binary class imbalance problem, while ensembles for the imbalanced multiclass problem are still maturing. Due to current development

in data mining and machine learning research for multiclass and imbalance learning, ensemble approaches have been given substantial attention and the most successful approaches are Boosting and Bagging (Guo, Yin, Dong, Yang, & Zhou, 2008).

Researchers are still unsure when dealing with the imbalance data. They could not decide whether to use ensemble algorithms only or combine both ensemble and sampling based on bagging or boosting. While using new distribution of data may introduce bias and different setting for learning, the performance such as classification accuracy and complexity of different learning methods may be affected. Thus in order to follow as closely to the standard benchmark data, two approaches will be used. The first approach is using direct ensemble implementation (without largely altering the original data distribution). The second approach is to run the sampling with replacement to create new data distribution and then implement the ensemble. Both approaches are tested on their classification performances.

Recent work by Sainin & Alfred (2012) which proposed the ensemble based classifier called a Direct Ensemble Classifier for Imbalanced Multiclass Learning (DECIML) was used to investigate single classifier performance. The researchers reported that the average accuracy using the DECIML on 16 imbalanced multiclass benchmark data was higher than the other tested single classifiers. According to the authors, the implementation of DECIML uses Naïve Bayes (NB), 1-Nearest Neighbor (1-NN) and k-Nearest Neighbor (kNN) due to drawbacks and possible gain through these single classifier algorithms as ensemble classifiers. The combination of those algorithms offers diverse ensemble construction where different type of learning method (NB is probability-based and Nearest Neighbor is distance-based) is applied to imbalanced multiclass problem. While previous implementation of DECIML was using the standard benchmark data (without altering the original data distribution), this research run the sampling with replacement to create the new data distribution and use the ensemble implementation to investigate their classification performance.

DECIML AND SAMPLING WITH REPLACEMENT

The implementation of DECIML with SWR carried out to investigate if there are any improvements in the classification performances of the ensembles. The DECIML consist of two ensembles namely NB1+1NN and NB+kNN.

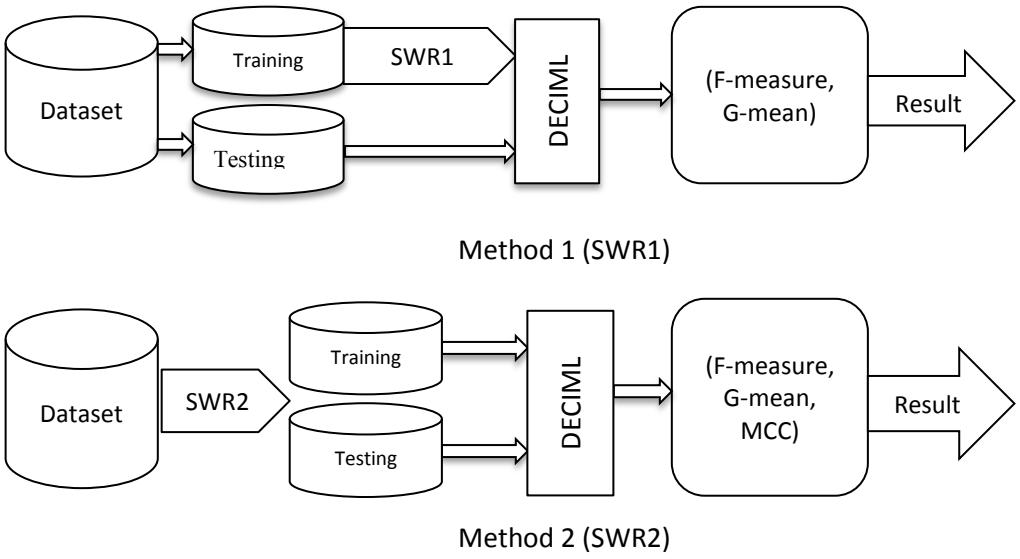

Figure 1. Experiment design for DECIML and sampling with replacement

Experiment Design

The process begins by identifying the dataset with imbalanced multiclass problem. Figure 1 depicts the overall process with two techniques for SWR1 and SWR2. In SWR1, original dataset is divided into two sets of training and testing, and then sampling is done on the training set only, where a new data created using SWR1 is added to the training data. Thus, there is an increase in the number of training samples when SWR is applied. The SWR2 method is to provide another perspective on testing the DECIML ensemble classification performance using a new proportion of training and test dataset. The proportion is constructed using original data in order to implement the SWR2. Then, examples are selected randomly using SWR, thus some of the selected examples may occur more than once. Finally, examples that are not selected during the selection with replacement will become as testing examples. In order to measure the performance of the DECIML with SWR, F-measure and G-mean are reported.

Dataset

Similar to previous study in Sainin & Alfred (2012), publicly available dataset repositories were examined such as the UCI (Asuncion & Newman, 2007), KEEL (Luengo, Derrac, Sánchez, & Herrera, 2011), UCR Time Series (Keogh Xi, X., Wei, L. and Ratanamahatana, C. A., 2006), and previously used dataset in multiclass imbalance publication (IEEE, ACM, Springer, Science Direct, etc.). The study used 11 selected datasets for multiclass imbalance data in 5 different domains, example size vary from 100 to 60,000 and feature size from less than 10 to 36. The imbalance ratio ranges from 1:19 to 1:4559. Datasets with imbalance ratio (i.e. exactly 1:19) were modified to meet the criteria for high imbalance in multiclass dataset. Out of the datasets used, five datasets have highly imbalanced multiclass data (i.e. more than 1:19 ratio). The detailed properties of the benchmark dataset used are summarized in Table 1.

Table 1. Benchmark dataset description

Data	#Examples	#Att	#Class	#Min	#Max	Ratio	Domain
Statlog(Landsat)	5865	36	6	56	1072	1:19	Physical
Glass	209	9	7	4	76	1:19	Physical
Car	1728	6	4	65	1210	1:19	Business
Thyroid	7200	21	3	351	6666	1:19	Life
New Thyroid	193	5	3	8	150	1:19	Life
Nursery	12857	8	4	227	4320	1:19	Social
Lympography	148	18	4	2	81	1:41	Life
Ecoli	336	7	8	2	143	1:72	Life
Yeast	1484	8	10	5	463	1:93	Life
PageBlocks	5473	10	5	28	4913	1:175	Computer
Statlog(Shuttle)	58000	9	7	10	45586	1:4559	Physical

Experiment Results

In order to compare the performance of the DECIML implementation without SWR and with SWR, Table 2, and 3 shows the comparison of the evaluation metric. The values in bold indicate that the metric is higher than other metric values among the three ensemble strategies for sampling. The 'No SWR' columns indicate the original implementation of DECIML ensembles without sampling; SWR1 and SWR2 are the results of sampling methods. Higher average values in F-measure and G-means for SWR1 (NB+kNN) shows that both evaluation metrics agree on the method SWR1. Through the observation, SWR1 and SWR2 slightly

improve the classification performance of DECIML on almost all of the dataset except for the Ecoli and Yeast datasets. None of the methods are well suited to every problem; however SWR1 with DECIML (NB+kNN) provides better overall classification performance where it wins five of the benchmark dataset with higher average F-measure rate. Thus, the sampling method using SWR is able to contribute to the average performance on some of the dataset using DECIML (NB+kNN).

Table 2. F-measure comparison

Data	No SWR		SWR1		SWR2	
	NB+1NN	NB+kNN	NB+1NN	NB+kNN	NB+1NN	NB+kNN
Statlog (Landsat)	0.960	0.950	0.967	0.967	**0.971**	**0.971**
Glass	0.690	0.708	0.844	0.844	**0.952**	**0.952**
Car	0.806	0.806	0.902	0.902	0.908	**0.949**
Thyroid	0.920	0.930	0.935	**0.941**	0.749	0.766
New Thyroid	1.000	1.000	1.000	1.000	1.000	1.000
Nursery	0.966	0.960	0.980	0.981	0.977	**0.995**
Ecoli	0.685	**0.703**	0.663	0.685	0.634	0.645
Yeast	**0.718**	0.699	0.646	0.649	0.691	0.695
Pageblocks	0.980	0.980	**0.985**	**0.985**	0.914	0.936
Statlog (Shuttle)	0.902	0.945	0.956	**0.993**	0.962	0.962
Lympography	0.709	0.740	0.952	**0.983**	0.981	0.981
Average	0.849	0.856	0.894	**0.903**	0.885	0.896
Win	1	1	1	**5**	2	4

Table 3. G-mean comparison

Data	No SWR		SWR1		SWR2	
	NB+1NN	NB+kNN	NB+1NN	NB+kNN	NB+1NN	NB+kNN
Statlog (Landsat)	0.964	0.965	0.976	0.976	**0.978**	**0.978**
Glass	0.748	0.775	0.839	0.839	**0.938**	**0.938**
Car	0.781	0.788	0.891	0.891	0.913	**0.947**
Thyroid	0.921	0.918	0.920	**0.922**	0.791	0.761
New Thyroid	1.000	1.000	1.000	1.000	1.000	1.000
Nursery	0.961	0.958	0.976	**0.977**	0.974	0.994
Ecoli	0.750	0.781	0.756	**0.798**	0.734	0.755
Yeast	0.651	0.666	0.700	0.689	0.691	**0.710**
Pageblocks	**0.975**	**0.975**	**0.975**	**0.975**	0.904	0.916
Statlog (Shuttle)	0.950	0.993	**0.999**	0.993	**0.999**	**0.999**
Lympography	0.803	0.848	0.930	**0.977**	0.972	0.972
Average	0.944	0.960	0.989	**0.997**	0.982	0.990
Win	1	1	2	**6**	3	5

CONCLUSION

In this paper, the performance of the DECIML (algorithm-level approach) was investigated based on two approaches, SWR1 and SWR2. Results show that sampling in some extent could improve performance and it is proven that with the SWR sampling, the predictive accuracy was improved as shown by the F-measure and G-mean. However, there are also some drawbacks; sampling on some of the datasets show degradation in the F-measure and this may be due to possible increase in data duplication. Further investigation can be conducted to study the effects of applying feature transformation on the DECIML classification performance. This feature transformation includes feature selection and construction methods.

REFERENCES

Asuncion, A., & Newman, D. J. (2007). *UCI Machine Learning Repository* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science. .

Ding, Z. (2011). Diversified Ensemble Classifiers for Highly Imbalanced Data Learning and Their Application in Bioinformatics. *Dissertation*. Computer Science Department, Georgia State University. Georgia State University. Retrieved from http://digitalarchive.gsu.edu/cs_diss/60

Farid, D. M., Rahman, M. Z., & Rahman, C. M. (2011). An Ensemble Approach to Classifier Construction based on Bootstrap Aggregation. *International Journal of Computer Applications*. *25*(5), 30–34.

Fergani, M., & Belkacem, B. A. (2014). A New Multi-Class WSVM Classification to Imbalanced Human Activity Dataset. *Journal of Computers*, *9*(7).

Garcia, V., Sanchez, J. S., Mollineda, R. A., & Sotoca, J. M. (2007). The class imbalance problem in pattern classification and learning. *Tamida 2007, Saragossa, Spain*, 283–291.

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the Class Imbalance Problem. In *Fourth International Conference on Natural Computation*, 4, 192–201.

Keogh Xi, X., Wei, L. and Ratanamahatana, C. A., E. (2006). The UCR Time Series Classification/Clustering Homepage: www.cs.ucr.edu/~eamonn/time_series_data/.

Lerteerawong, B., & Athimethphat, M. (2011). An Empirical Study of Multiclass Classification with Class Imbalance Problems. *International Conference on Business and Information, BAI2011*. Sapporo, Japan.

Liu, X.-Y., & Li, Q.-Q. (2014). Learning from Combination of Data Chunks for Multi-class Imbalanced Data. *International Joint Conference on Neural Networks (IJCNN)*, 1680–1687.

Luengo, J., Derrac, J., Sánchez, L., & Herrera, F. (2011). 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), 255–287.

NIPS. (2003). NIPS Feature Selection Challenge. Retrieved from http://www.nipsfsc.ecs.soton.ac.uk/datasets/

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligent Rev*, 33(1-2), 1–39.

Sainin, M. S., & Alfred, R. (2012). A Direct Ensemble Classifier for Imbalanced Multiclass Learning. In *Data Mining Optimization (DMO2012)*, 59–66. Langkawi.

Sun, Y., & Wang, Y. (2006). Boosting for Learning Multiple Classes with Imbalanced Class Distribution. In *Sixth International Conference on Data Mining*, 592–602.

Wang, S., & Yao, X. (2012). Multiclass Imbalance Problems: Analysis and Potential Solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *2*(4), 1119–1130.

Zhang, Z., & Yang, P. (2008). An Ensemble of Classifiers with Genetic Algorithm Based Feature Selection. *IEEE Intelligent Informatics Bulletin*, *9*(1), pp. 18–24.

Zhou, Z. H. (2008). Ensemble learning. *Encyclopedia of Biometrics*, 1–5.