Solving Imbalanced Problem of Muticlass Data Set with Class Balancer and Synthetic Minority Over-sampling Technique

Pumitara Ruangthong[1], Pradit Songsangyos[2], Soontaree Kankaew[3]

[1]Computer Science Division, Rajamangala University of Technology Suvarnabhumi
Phra Nakhon Si Ayutthaya, Thailand
pumitara@gmail.com
[2]Computer Science Division, Rajamangala University of Technology Suvarnabhumi
Phra Nakhon Si Ayutthaya, Thailand
pradit.s@rmutsb.ac.th
[3]Computer Science Division, Rajamangala University of Technology Suvarnabhumi
Phra Nakhon Si Ayutthaya, Thailand
soontaree.k@rmutsb.ac.th

บทคัดย่อ— ปัญหาการจำแนกประเภทของข้อมูลที่มีลักษณะแบบหลายคลาสในข้อมูลชุดเดียวกัน มักให้ผลลัพธ์ที่ไม่ดีเท่าที่ควร ดังนั้นในงานวิจัยนี้ได้มุ่งเน้นประเด็นการแก้ปัญหาความไม่สมดุลของข้อมูลที่มีลักษณะแบบหลายคลาส จึงได้ทำการเปรียบเทียบข้อมูลที่ผ่านการแก้ปัญหาความไม่สมดุลของคลาสกับข้อมูลปกติที่ยังไม่ผ่านการแก้ปัญหาความไม่สมดุลเพื่อหารูปแบบการทำนายผลที่เหมาะสมกับข้อมูลแบบหลายคลาสที่มีลักษณะความไม่สมดุลของข้อมูล เนื่องจากข้อมูลเป็นแบบหลายคลาส ดังนั้นงานวิจัยนี้จึงได้ให้ความสำคัญกับผลลัพธ์ของแต่ละคลาสอย่างเท่าเทียมกัน การพิจารณาผลลัพธ์จะดูจากการกระจายของความถูกต้องของการทำนายในแต่ละคลาส

คำสำคัญ-ข้อมูลแบบหลายคลาส ปัญหาความไม่สมดุลของข้อมูล, การจำแนกประเภท ปรับความสมดุลของคลาส

Abstract— Classifying multiclass data set frequently leads to poor results. Therefore, this research tends to solve imbalanced multiclass data set. We compare the data undergone class imbalance problem solving process with the unsolved data to look for predictive modelling most suitable for imbalanced multiclass data set. As it contains multiple classes, we treat each class equitably. Dispersion of accurate prediction for each class is of result consideration.

Keywords-multiclass; imbalanced data; classification; class balancer

I. INTRODUCTION

In the present time there are many researches concerning multiclass problem solving. The difficulties occur when there are more than 2 classes because the results typically incline to the class with greater numbers. For example, classifying a data set with more than 3 classes, if it comes from diverse existent specimen, it becomes more complicated since the data set contains high imbalance. We cannot expect a higher result. For that reason, it is essential that we use a suitable algorithm to prepare the data set before classification process.

In 2007, L. Bing dealt with imbalanced multiclass problem using hierarchical classification [1]. A. Astha handled the same problem using SMOTE and cluster based undersampling in 2015 [2]. In, 2010 S. Zhuoli, K. Kyunghee and S. Tadashi conducted a research on multiple classes with self-learning approach using multiple dimensional quasi Gaussian [3]. In 2016, P. Ruangthong and P. Songsangyos solved the imbalanced problem by creating class balancer [5] to predict tendency of getting higher education of students.

II. METHODOLOGY

A. Class balancer

The data set used in this research is svmguide2 [11] data set which contains integer and decimal data type. There are 391 instances and 3 classes. We applied class imbalance algorithm by reassigning weight in the data set in order that all the classes are even. The total amount of weights among the instances remains the same. The weights in the first group of data given by this filter are solely modified so that it is allowed to be used with the Filtered Classifier. The following is the example of reweighting Shown in Table I and II.

TABLE I. WEIGHTS OF THE INSTANCES BEFORE REWEIGHTING

Class	Number of Instances	Weight
1	221	221.00
2	117	117.00
3	53	53.00

TABLE II. WEIGHTS OF THE INSTANCES AFTER REWEIGHTING

Class	Number of Instances	Weight
1	221	130.33
2	117	130.33
3	53	130.33

Reweighting formula

$$w = \frac{x_i+\ldots+x_n}{c} \qquad (1)$$

W = weight of each class

C = number of classes in data set

x_i = number of instances of each class

We also used Synthetic Minority Over-sampling Technique (SMOTE) [5]. It is an algorithm for adjusting minor class to become closer to the other one. This way classification can produce more accurate results.

B. Measure

Classification of class imbalanced data set needs measures that can indicate results of prediction in each aspect as shown in Table III.

TABLE III. CONFUSION MATRIX

Real	Classifiers	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Precision is the possibility that a received document (chosen at random) is appropriate [6][8][9][10].

$$precision = \frac{TP}{TP+FP} \qquad (2)$$

Recall is the possibility that an appropriate document (chosen at random) is received in a search.

$$recall = \frac{TP}{TP+FN} \qquad (3)$$

The harmonic mean of precision and recall is the measure that merges precision and recall, which is F-measure.

$$F - Measure = \frac{2.Precision.recall}{(Precision+recall)} \qquad (4)$$

III. PREPARE EXPERIMENTAL RESULTS

According to experimental results, in the original data, model logistic regression gave training:80, testing:20 section at the highest at precision 0.86, recall 0.84 and f-measure 0.84. Then in class balancer [5] algorithm using reweighting, Random Forest Model gave training:80, testing:20 section at the highest at precision 0.83 recall 0.82 and f-measure 0.82. As to solve class imbalance problem with SMOTE algorithm [12], this gave training:70, testing:30 section at the highest by Multilayer Perceptron Model at precision 0.83. As for training:80, testing:20 section, the highest was Random Forest Model at precision 0.83, while recall 0.82 and f-measure 0.82 at the highest was Multilayer Perceptron Model as shown in Table IV. According to the line graph in Figure 1, it displays the comparison of results from each model and each predictive indicator.

Since this experiment used multiclass data set, predictive results of each class have dispersion. As shown in the table below, the original data still has high predictive results when focusing only on the average but with thorough consideration we can see that classification makes a great difference. For example, Class 1 includes more instances than Class 2, so Class 1 delivers larger results than Class 2. Conversely, after undergone reweighting or balancing by SMOTE algorithm, the values of each class appear to be proximate and still have high results of classification.

IV. CONCLUSIONS

Multiclass data set can always be problematic in many different ways. There are also many solutions for class imbalance problem besides reweighting with class balancer or using SMOTE algorithm to balance each class. Researchers can modify algorithm for balancing data or experimentally conduct classification with other model than the one used in this research.

REFERENCES

[1] L. Bing and Z. Yun, "Hierarchical Classification for Imbalanced Multiple Classes in Machine Vision Inspection," Image and Graphics, 2007, pp.536-541.

[2] A. Astha Agrawal, L. Herna and P. Eric, "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling," 2015 7th International Joint Conference on Knowledge Discovery, 2015, pp.226-234.

[3] S. Zhuoli, K. Kyunghee and S. Tadashi, "A self-learning multiple-class classifier using multi-dimensional quasi-Gaussian analog circuits," Proceedings of 2010 IEEE International Symposium on Circuits and Systems, 2010, pp.2330-2333.

[4] R. Pumitara, P. Songsangyos, S. Kankaew and K. Thaksina, "Solving Data Imbalance Problem by Creating Class Balancer Using Data Mining to Predict Decision-Making for Higher Education of School Students" The 10th National Conference and 2016-2 International Conference on Applied Computer Technology and Information Systems and 2016-2 and National Conference on Business Administration, 2016, pp. 25-28.

[5] Data Mining Software from The University of Waikato.

[6] Y. Baeza, Ricardo, N. Ribeiro and Berthier, Modern Information Retrieval. New York, NY: ACM Press, Addison-Wesley, 1999.

[7] B. Hjørland, foundation of the concept of relevance, Journal of the American Society for Information Science and Technology, pp. 217-237.

[8] Makhoul, John; Kubal Francis; Schwartz, Richard; and Weischedel, Ralph, Performance measures for information extraction, in Proceedings of DARPA Broadcast News Workshop, Herndon, VA, February 1999.

[9] "Machine literature searching X. Machine language; factors underlying its design and development".

[10] van Rijsbergen, Cornelis Joost Information Retrieval, London, GB; Boston, MA: Butterworth, 2nd Edition, 1979.

[11] UC Irvine Machine Learning Repository.

[12] N. V. Chawla, W. K. Bowyer, O. L. Hall and W. P. Kegelmeyer, "SMOTE:Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, 2002, pp. 321-357.

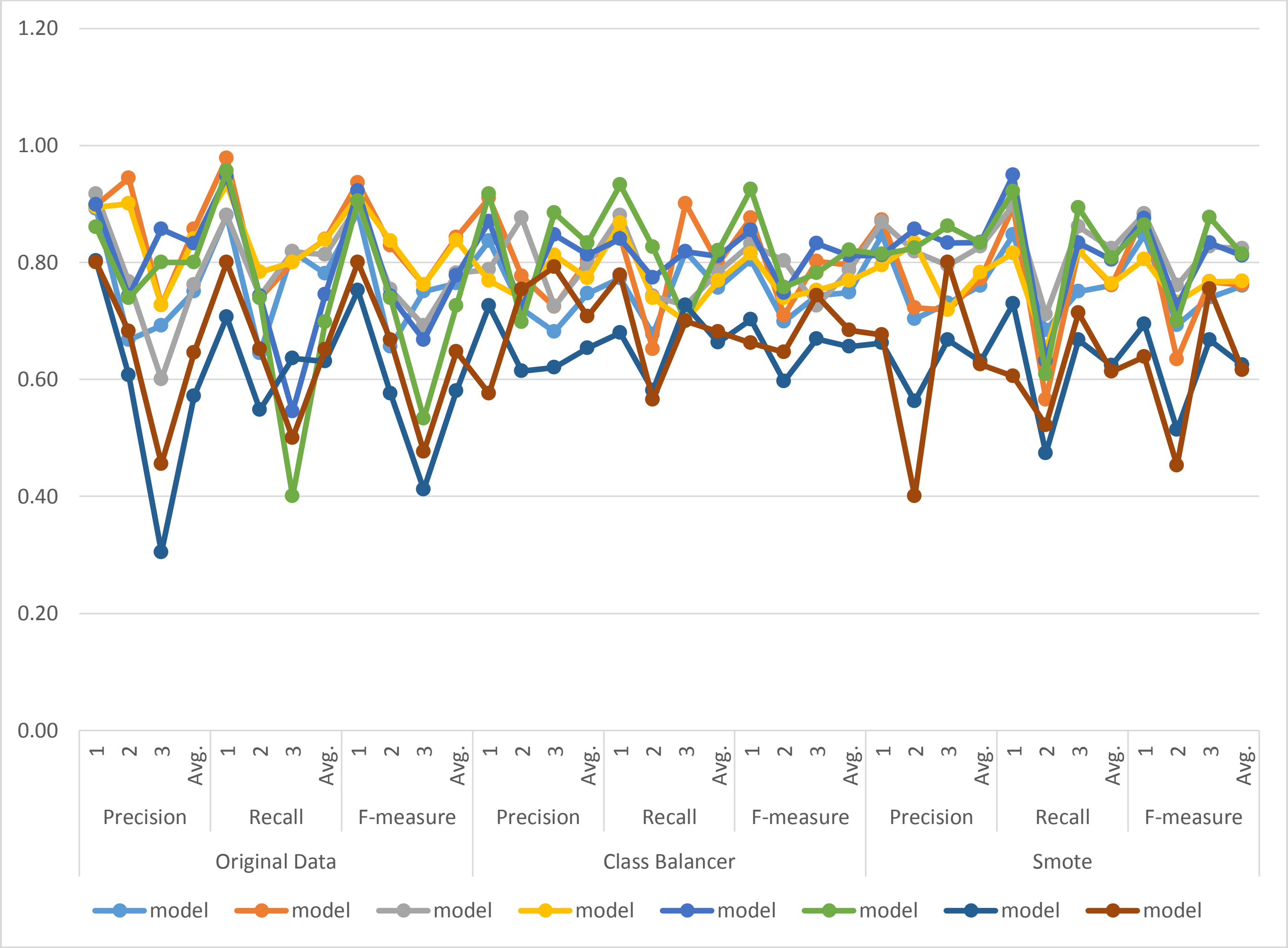

Figure 1. The line graph showing comparison of prediction results for each model

TABLE IV. Results from classification using Class Balancer and SMOTE

Data			model							
			logistic regression		Multilayer Perceptron		Random Forest		Random Tree	
Training/Testing (%)			70	80	70	80	70	80	70	80
Original Data	Precision	1	0.89	0.90	0.92	0.89	0.90	0.86	0.80	0.80
		2	0.67	0.94	0.77	0.90	0.74	0.74	0.61	0.68
		3	0.69	0.73	0.60	0.73	0.86	0.80	0.30	0.46
		Avg.	0.75	0.86	0.76	0.84	0.83	0.80	0.57	0.65
	Recall	1	0.88	0.98	0.88	0.93	0.95	0.96	0.71	0.80
		2	0.65	0.74	0.74	0.78	0.74	0.74	0.55	0.65
		3	0.82	0.80	0.82	0.80	0.55	0.40	0.64	0.50
		Avg.	0.78	0.84	0.81	0.84	0.74	0.70	0.63	0.65
	F-measure	1	0.89	0.94	0.90	0.91	0.92	0.91	0.75	0.80
		2	0.66	0.83	0.75	0.84	0.74	0.74	0.58	0.67
		3	0.75	0.76	0.69	0.76	0.67	0.53	0.41	0.48
		Avg.	0.76	0.84	0.78	0.84	0.78	0.73	0.58	0.65
Class Balancer	Precision	1	0.84	0.91	0.79	0.77	0.87	0.92	0.73	0.58
		2	0.72	0.78	0.88	0.74	0.72	0.70	0.61	0.75
		3	0.68	0.72	0.73	0.81	0.85	0.89	0.62	0.79
		Avg.	0.75	0.80	0.80	0.77	0.81	0.83	0.65	0.71
	Recall	1	0.77	0.84	0.88	0.87	0.84	0.93	0.68	0.78
		2	0.68	0.65	0.74	0.74	0.77	0.83	0.58	0.57
		3	0.82	0.90	0.73	0.70	0.82	0.70	0.73	0.70
		Avg.	0.76	0.80	0.78	0.77	0.81	0.82	0.66	0.68
	F-measure	1	0.80	0.88	0.83	0.82	0.86	0.93	0.70	0.66
		2	0.70	0.71	0.80	0.74	0.75	0.76	0.60	0.65
		3	0.74	0.80	0.73	0.75	0.83	0.78	0.67	0.74
		Avg.	0.75	0.80	0.79	0.77	0.81	0.82	0.66	0.68
Smote	Precision	1	0.85	0.87	0.87	0.80	0.81	0.81	0.66	0.68
		2	0.70	0.72	0.82	0.83	0.86	0.82	0.56	0.40
		3	0.73	0.72	0.80	0.72	0.83	0.86	0.67	0.80
		Avg.	0.76	0.77	0.83	0.78	0.83	0.83	0.63	0.63
	Recall	1	0.85	0.90	0.90	0.82	0.95	0.92	0.73	0.61
		2	0.68	0.57	0.71	0.65	0.63	0.61	0.47	0.52
		3	0.75	0.82	0.86	0.82	0.83	0.89	0.67	0.71
		Avg.	0.76	0.76	0.82	0.76	0.80	0.81	0.62	0.61
	F-measure	1	0.85	0.88	0.88	0.81	0.88	0.86	0.69	0.64
		2	0.69	0.63	0.76	0.73	0.73	0.70	0.51	0.45
		3	0.74	0.77	0.83	0.77	0.83	0.88	0.67	0.76
		Avg.	0.76	0.76	0.82	0.77	0.81	0.81	0.63	0.62