Local Clustering Conformal Predictor for Imbalanced Data Classification

Huazhen Wang[1], Yewang Chen[1], Zhigang Chen[2], and Fan Yang[2]

[1] School of Computer Science and Technology, Huaqiao University, Xiamen, 361021, China
[2] Department of Automation, Xiamen University, Xiamen, 361005, China

Abstract. The recently developed Conformal Predictor (CP) can provide calibrated confidence for prediction which is out of the traditional predictors' capacity. However, CP works for balanced data and fails in the case of imbalanced data. To handle this problem, Local Clustering Conformal Predictor (LCCP) which plugs a two-level partition into the framework of CP is proposed. In the first-level partition, the whole imbalanced training dataset is partitioned into some *class-taxonomy* data subsets. Secondly, the majority class examples proceed to be partitioned into some *cluster-taxonomy* data subsets by clustering method. To predict a new instance, LCCP selects the nearest cluster, incorporated with the minority class examples, to build a re-balanced training data. The designed LCCP model aims to not only provide valid confidence for prediction, but significantly improve the prediction efficiency as well. The experimental results show that LCCP model presents superiority than CP model for imbalanced data classification.

Keywords: conformal predictor, imbalanced data, local clustering.

1 Introduction

Traditional pattern classification methods focus on the improvement of the accuracy on the test set while neglects confidence analysis of the results [1,2]. Suffered from this weakness, the machine learning methods are constantly criticized by traditional statisticians and shows inapplicable in many realistic practice. Moreover, traditional pattern recognition algorithms are generally designed based on the balanced distributed data, and thus always deteriorate terribly on imbalanced datasets. In other words, traditional pattern recognition algorithms tend to prefer the majority class examples to the minority class examples, regardless of the fact that the minority class examples might be important for the users [3, 4]. Thus, the cross-study in these two areas seems changeable and significant.

The recently developed Conformal Predictor (CP) can provide confidence analysis of results and output reliable prediction [5]. However, the CP model works for the evenly distributed dataset and cannot effectively solve the problem of imbalanced data learning. It is worth noting, in order to address the cost sensitive learning, a modified model named MCP (Mondrian Conformal Predictor) can provide label-conditional valid confidence [6]. It shows that CP can be incorporated with multi-partition technology to fit a variety of particular learning settings. This encourages our

H. Papadopoulos et al. (Eds.): AIAI 2013, IFIP AICT 412, pp. 421–431, 2013.

exploring on the possibility of multi-partition technology in imbalanced data learning to improve the feasibility of CP.

In this paper, we introduce a two-level partition method into the framework of the CP model, and then build a modified model named Local Clustering Conformal Predictor (LCCP) for classification of imbalanced data. LCCP model adopts cluster technology to explore the local construction and selects the nearest cluster to be the representative of majority class examples without potential loss. And then constitutes a re-balanced train dataset for confidence prediction. The designed LCCP model aims to not only provide valid confidence for prediction, but significantly improve the prediction efficiency as well.

2 Related Work

2.1 Conformal Predictor(CP)

To address the problem of reliable prediction, Professor Vovk proposed the Conformal Predictor (CP) which can output prediction tailed by valid confidence[7]. According to the CP, the i.i.d assumption is equivalent to the Kolmogorov algorithmic randomness statistic test[8]. When carried on a new test instance, CP applies transductive inference learning to incorporate the train data with the test instance and thus establish a *test data sequence*. Then the algorithmic randomness test is carried out in the test data sequence and subsequently the p value of it is applied to response the prediction. CP model has aroused increased interest in the literature of machine learning and has been applied successfully in the classification of medical data[9], image data[10], and so on. Besides classification, CP has been extended to regression[11-12],feature selection[13], and so on.

At the aspect of the framework of CP, some modified models have been proposed to improve the flexibility of CP. In order to improve the computation efficiency of CP model, Papadopoulos proposed ICP (Inductive CP) for large data sets [14]. In our previous work, the HCCP (Hybrid-Compression CP) not only improves the computational efficiency but preserves the prediction efficiency as well [15]. In order to address the cost sensitive learning, Vovk proposed MCM (Mondrian Confidence Machine) model, which is renamed MCP (Mondrian Conformal Predictor) nowadays [6] and has applied interesting implementations on the gene expression data [16] and breast cancer data[17]. According to all the modifications, Vovk proposed OCM (On-line Compression Model), which is a universal framework that can regulate all the existing CP-related models[6].

2.2 Classification of Imbalanced Data

To address the particular problem of imbalanced data, the solutions can be divided into two categories: data-level methods and algorithms-level methods. The former applies over-sampling or under-sampling to build a re-balanced training data. The SMOTE model is one of the typical approaches [18]. On the other hand, some particular algorithms have been designed based on the assumptions of imbalanced data distribution, such as cost-sensitive learning, active learning and so on[19].

Here we give a brief review of cluster-based under-sampling methods for imbalanced data, because it shows more related to our work[20-23]. These algorithms differ on whether the clustering is done on the whole training data or inside each category. The first one clusters the whole imbalanced dataset into several groups, and then selects some representatives from each group to rebuild the majority class examples[20]. The latter performs clustering inside each class and then tapes pseudo-class labels for those sub-groups. After that, they expanded the two-class learning problem in the multi-class classification setting [21-23].

3 Local Clustering Conformal Predictor for Imbalanced Data

3.1 The Framework of CP

It is necessary to present the framework of CP model because our LCCP model is derived from it. The reality outputs the training data sequence $Z^{(n-1)} = (Z_1, Z_2, ..., Z_{n-1})$, and now a new instance x_n is given to be recognized. CP exhausts all the labels $y \in Y = \{1,2,...,C\}$ (C is the number of classes) to be the candidate label for x_n, and thus forms the corresponding *test example* $Z_n^y = (x_n, y)$. Next, CP incorporates each Z_n^y with $Z^{(n-1)}$ to construct the *test data sequence*. Consequently, there are C test data sequences, such as:

$$z^{(n)y} = \{(z_1, z_2, ..., z_{n-1}, z_n^y), y = 1, 2, ..., C\} \tag{1}$$

Subsequently, CP designs a function $\Lambda : Z^{(n)y} \to \alpha^{(n)y}$,which maps each example Z_i to a single nonconformity point α_i, and thus conforms a one-dimension *nonconformity measurement sequence*:

$$\alpha^{(n)y} = \{(\alpha_1, \alpha_2, ..., \alpha_{n-1}, \alpha_n^y), y = 1, 2, ..., C\} \tag{2}$$

where α_i measures the degree of the nonconformity between Z_i an $Z^{(n)y}$. Based on $\alpha^{(n)y}$, the p value which serves as the probability of y being the true label y_n is computed as follows:

$$p_n^y = \frac{\left|\{i = 1, 2, ..., n-1 : \alpha_i \geq \alpha_n^y\}\right| + 1}{n} \tag{3}$$

Given a significance level ε, CP outputs the prediction for x_n as follows:

$$\tau_n^\varepsilon = \{y : p_n^y > \varepsilon, y = 1,2,..., C\} \tag{4}$$

where τ_n^ε is a region prediction rather than a point prediction. An error occurs when the prediction set τ_n^ε does not contain the true label y_n. Thus, CP has been proven , in the online setting, the error rate is not greater than the significance level ε ,i.e.,

$$P\{p_n^y(z_1, z_2, ..., z_{n-1}, z_n^y) \leq \varepsilon\} \leq \varepsilon \tag{5}$$

The above equation (5) is known as the *validity theorem* which has been theoretically proved [6] and tested in practice.

3.2 Local Clustering Conformal Predictor (LCCP)

a) Binary Classification Setting

Consider a typical binary classification setting, i.e. $y \in \{1,2\}$. Moreover, $y = 1$ refers to the minority class, $y = 2$ corresponds the majority class. Thus the process of LCCP can be designed as follows：

1) *first-level partition*: dividing the whole training data sequence $z^{(n-1)}$ into two *class-taxonomy* data subsets, and then produces the *minority-class examples sequence* $Z^{(n_1)} = (Z_1, Z_2, ..., Z_{n_1})$ and the *majority-class examples sequence* $Z^{(n_2)} = (Z_1, Z_2, ..., Z_{n_2})$.
2) *second-level partition*: clustering $Z^{(n_2)}$ into J *cluster-taxonomy* data subsets, i.e., $Z^{(n_{2j})} = \{(Z_1, Z_2, ..., Z_{n_{2j}}), j = 1,2,...,J\}$. J is set to be same as the ratio of the majority examples to the minority examples; Subsequently, the J class centroids $Cen_j\, j = 1,2,...,J$ can be carried out.
3) *re-imbalanced training data building*: computing the distance $D_j\, j = 1,2,...,J$ between x_n and $Cen_j\, j = 1,2,...,J$; extracting $Z^{(n_{2k})}$ corresponding to the minimum distance D_k ; combining the $Z^{(n_{2k})}$ with $Z^{(n_1)}$ to build the re-imbalanced training data $Z^{(n_1+n_{2k})} = (Z^{(n_1)} \bigcup Z^{(n_{2k})})$
4) *reliable prediction*: executes reliable prediction for x_n based on $Z^{(n_1+n_{2k})}$.

b) Multi-class Classification Setting

Next, we further discuss the three-class classification setting. If the learning setting is made of two minority classes and one majority class, only the majority examples should be imposed to the clustering algorithm. Subsequently, one of *cluster-taxonomy* data subsets is selected to build the re-imbalanced training data.

If the learning setting is made of one minority class and two majority classes, both of the majority examples should be clustered. Afterwards, one of cluster-taxonomy data subsets is selected to build the re-imbalanced training data. According to the above scheme, we can extend the LCCP algorithm to the multi-class classification setting in general.

4 Experimental Setup

4.1 The Selection of Clustering Algorithm

In order to contribute the merits of LCCP to the compact clusters, we set up the clustering process in reverse form, i.e., we assemble some compact clusters to get the majority examples. At the second stage that provides a prediction based on the re-balanced training dataset, we apply Random Forest to design the nonconformity measurement, the detail process can be seen in our previous work [24].

4.2 Datasets

a) Synthetic Imbalanced Dataset: Simplex

The base Simplex dataset is depicted in Fig. 1, which shows the good separability distribution. The number of classes is 4, and the number of examples per class is 2500. One of the class examples are used to be minority examples while the rest to be the majority examples, i.e., two imbalanced ratios of 1:2 and 1:3 are available .

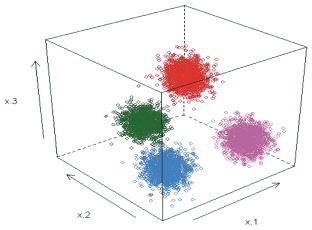

Fig. 1. The scatter distribution of Simplex

b) Real Imbalanced Dataset: TEP

Tennessee Eastman Process (TEP) was a government sponsored program for the evaluation of fault detection in the large scale industrial process[25]. There are 52 monitors (i.e. , features) to diagnose 21 faults in the process. In our experiments, total 17600 points with 800 points per fault and an additional 800 points as the normal class are sampled. In the experiment, we select the normal form as the minority class, and then compile the remainder of the examples to be the majority class, which creates a series of different imbalanced ratios, such as 1:2,1:3,…, 1:13.

5 Experimental Results

5.1 The Validity of Confidence of LCCP

Compared with the classical CP , the experimental results for Simplex is illustrated in Fig. 2, and the experimental results for TEP is demonstrated in Fig. 3.

In the upper zone of Fig.2 and Fig.3, with the significance level 5% (the corresponding confidence level 0.95), the x-axis represents the size of test data and the y-axis represents the number of errors. It can be seen that with the expansion of the size of test data, the errors increase accordingly. But the slope of the curve (i.e., *error calibration line*) is constant and close to the significance level 5%, which reveals the validity theorem seen in formula (5). Notable is, the slope of LCCP is apparently smaller than that of CP on TEP dataset, i.e., about 3.8% for LCCP while 5% of CP. It shows that the error rate of LCCP is less than the significance level, which is also applicable in formula (5).

In the lower zone of Fig.2 and Fig.3, the x-axis represents the confidence and the y-axis represents accuracy rate. The diagonal line with legend '*base calibration*'

exhibits the optimal relationship between the accuracy rate and the confidence. As clearly shown in Fig.2, for the Simplex data, the *accuracy calibration lin*e of LCCP and CP both closely attached to the "*base calibration line*", which reveals that the accuracy of LCCP can be calibrated by the confidence. In addition, as shown in Fig.4, the *accuracy calibration line* of LCCP is slightly higher than the *base calibration line*, which implies that the accuracy rates of LCCP are always greater than the corresponding confidence, which also corresponds with the formula (5) to demonstrate the validity of LCCP.

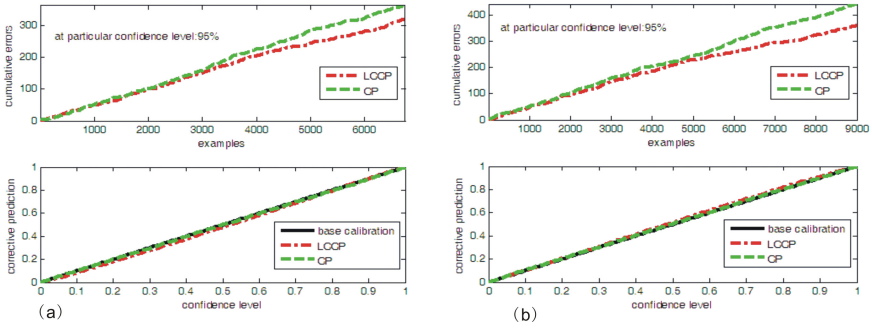

Fig. 2. The comparison of calibration on Simplex dataset at different imbalanced ratio (a) 1:2 (b) 1: 3

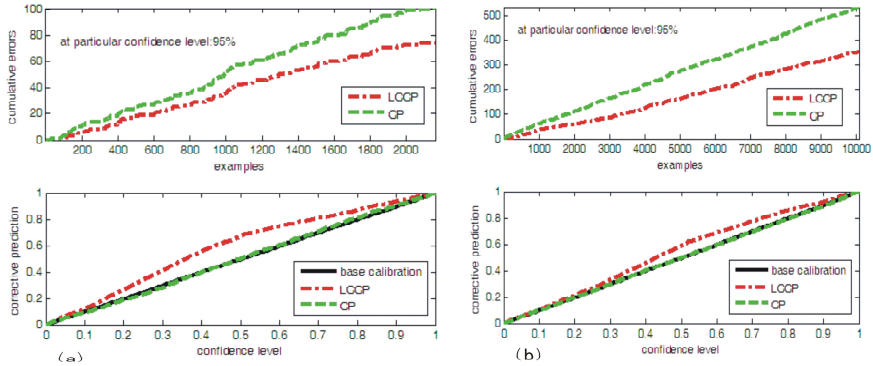

Fig. 3. The comparison of calibration on TEP dataset at different imbalanced ratio (a) 1:2 (b) 1: 3

5.2 The Prediction Efficiency of LCCP

The *favorite prediction* which contains not only one label but also being the true label has been recognized as a key index to exhibits the prediction efficiency. The performance of LCCP is illustrated in Fig.4 and Fig.5.

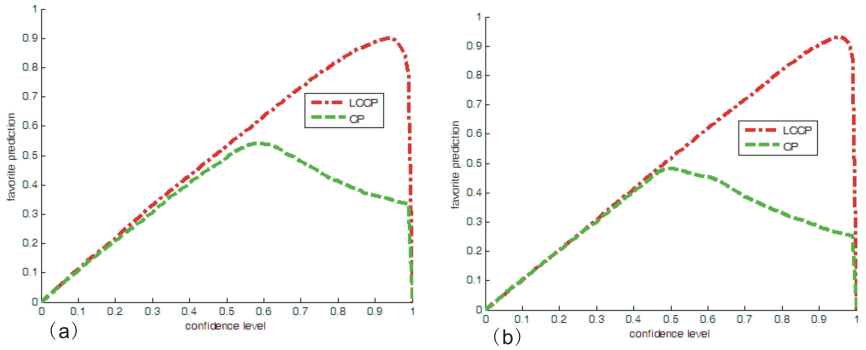

Fig. 4. The comparison of favorite prediction on Simplex dataset at different imbalanced ratio (a)1:2 (b)1: 3

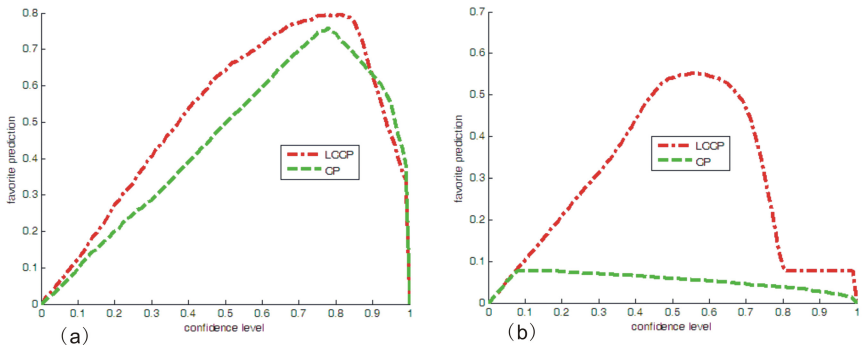

Fig. 5. The comparison of favorite prediction on TEP dataset at different imbalanced ratios (a)1:2 (b)1:13

As can be seen from Fig.4, the favorite prediction of LCCP performs significantly higher than the CP with the confidence level within [0.5 1] which is preferred in practice. On the other hand, we can find that , on TEP dataset, the gap of favorite prediction ratio becomes higher and higher along with the increasing of the imbalanced ratio. This means a large gap between the randomness levels of the C test data sequences. Furthermore, the corresponding label for the higher one must always be the true label. The superiority of LCCP comes from the mechanism that LCCP selects the nearest cluster for the test instance.

Next, Given the confidence level 0.85、0.95、0.99 which are more interested in real practice, all the favorite prediction rates at the different imbalanced ratios on TEP dataset are shown in table 1.

Table 1. The comparison of the predictive efficiency

confidence level	imbalanced ratio							
	1:2		**1:3**		**1:4**		**1:5**	
	LCCP	**CP**	**LCCP**	**CP**	**LCCP**	**CP**	**LCCP**	**CP**
0.85	0.77	0.68	0.67	0.57	0.79	0.50	0.24	0.77
0.95	0.46	0.54	0.47	0.45	0.49	0.38	0.16	0.46
0.99	0.34	0.38	0.26	0.29	0.22	0.27	0.16	0.34
	1:6		**1:7**		**1:8**		**1:9**	
	LCCP	**CP**	**LCCP**	**CP**	**LCCP**	**CP**	**LCCP**	**CP**
0.85	0.14	0.06	0.12	0.06	0.11	0.05	0.10	0.04
0.95	0.14	0.04	0.12	0.04	0.11	0.03	0.09	0.03
0.99	0.14	0.02	0.12	0.02	0.11	0.01	0.09	0.01
	1:10		**1:11**		**1:12**		**1:13**	
	LCCP	**CP**	**LCCP**	**CP**	**LCCP**	**CP**	**LCCP**	**CP**
0.85	0.16	0.03	0.09	0.03	0.07	0.03	0.07	0.03
0.95	0.08	0.02	0.08	0.02	0.07	0.02	0.07	0.02
0.99	0.08	0.01	0.08	0.01	0.07	0.01	0.07	0.01

Table 1 illustrates the comparison of *favorite prediction* between LCCP and CP. It is clear that LCCP performs distinctly higher favorite prediction than CP under all of the imbalanced ratios. This highlights again that LCCP can deeply dig the local distribution structure of the majority examples through *second-level clustering*. LCCP guarantees the quality of the re-balanced training data, and thus promote significantly prediction efficiency in the subsequent learning. On the contrary, the performance of CP decline dramatically with the high imbalanced ratio, because it has to execute the prediction in the whole imbalanced dataset.

5.3 The Performance under the Domain-Related Indices

Considering the particular imbalanced learning problem, It is essential to evaluate LCCP by some domain-related indices. That is, given the TP (the number of correct predictions among the minority examples), TN (the number of correct predictions among the majority examples), FN (the number of error predictions among the minority examples) and FP (the number of error predictions among the majority examples), some specific indices, such as *Recall*, *Precision*, *F*, *G-man's* and A*UC* value, are designed to demonstrate the power of classification for minority class or the integrated ability of classification [19].

Nonetheless, the indices above can only be set based on the point prediction setting, which seems incompatible with the region prediction with the LCCP model. Thus, the prediction method of LCCP has to be changed to be the point prediction mode, which selects a single label corresponding the maximum *p value* based on formula (3). In such setting, the performance of LCCP is illustrated in table 2.

Table 2. The performance of LCCP under domain-specified indices

dataset	Imbalanced ratio	Recall	Precision	F	G-means	AUC
Simplex	**1:2**	1.00	1.00	1.00	1.00	1.00
	1:3	1.00	0.97	0.99	0.99	0.99
TEP	**1:2**	1.00	0.99	1.00	1.00	1.00
	1:3	1.00	0.97	0.98	0.99	0.99
	1:4	1.00	0.92	0.96	0.99	0.99
	1:5	1.00	0.52	0.69	0.903	0.91
	1:6	1.00	0.38	0.55	0.85	0.86
	1:7	1.00	0.30	0.46	0.814	0.83
	1:8	1.00	0.31	0.47	0.851	0.86
	1:9	1.00	0.29	0.45	0.86	0.87
	1:10	1.00	0.31	0.47	0.88	0.89
	1:11	1.00	0.26	0.42	0.864	0.87
	1:12	1.00	0.21	0.34	0.826	0.84
	1:13	1.00	0.22	0.36	0.85	0.86

As shown in table 2, on the Simplex dataset, all of the indices are approximately to be 1, which indicates that LCCP can successfully address the problem of the imbalanced data. On the contrary, the performance of LCCP on the TEP data set fluctuates across these indices. The values of *Recall* index show very high to be around 1, but the performance of the *Precision* index gradually decreases verse the increasing of imbalanced ratio. However, the three indices, *F*, *G-means*, *AUC value*, demonstrate high values all over 0.8.

It is clearly that LCCP can recognize the minority examples well and performs quite well on the whole dataset. But LCCP seems poor in the classification of majority examples, especially in the case of the higher imbalanced ratio. The underlying cause of the situation lies in the high overlap of the based TEP dataset[26]. It indicates that the clustering algorithm plays a significant influence on the *second-level partition*. Proper selection, the clustering algorithm can widen the divergence among the clusters and thus boost the ability of classification for the majority examples.

6 Conclusions

In this paper we propose Local Clustering Conformal Predictor (LCCP) to provide valid prediction for the imbalanced data. The experimental results show that LCCP not only provide valid confidence for prediction, but significantly improve the prediction efficiency as well. Furthermore, the LCCP model seems virtually a general framework to deeply dig the local distribution structure of the dataset and thus can promote the prediction efficiency in other application.

Acknowledgements. The work is supported by the Natural Science Foundation of Fujian Province, under Grant No 2012J01274; the Research Grant Council of Huaqiao University under Grant No 09BS515; National Natural Science Fundation of China under Grant No 61202144; Natural Science Foundation of Fujian Province under Grant No 2012J05125.

References

1. Li, H.R.: Reliability and Validity in Qualitative Research. PhD thesis, Harbin Engineering University (2009)
2. Melluish, T., Saunders, C., Nouretdinov, I., Vovk, V.: Comparing the Bayes and Typicalness Frameworks. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 360–371. Springer, Heidelberg (2001)
3. Elazmeh, W., Japkowicz, N., Matwin, S.: Evaluating misclassifications in imbalanced data. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) ECML 2006. LNCS (LNAI), vol. 4212, pp. 126–137. Springer, Heidelberg (2006)
4. Li, F., Mi, H., Yang, F.: Exploring the stability of feature selection for imbalanced intrusion detection data. In: 9th IEEE International Conference on Control and Automation, Santiago, pp. 750–754 (2011)
5. Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning Research 9, 371–421 (2005)
6. Vovk, V., Gammerman, A., Shafer, A.G.: Algorithmic Learning in a Random World. Springer, New York (2005)
7. Saunders, C., Gammerman, A., Vovk, V.: Transduction with confidence and credibility. In: 16th International Joint Conference on Artificial Intelligence, Stockholm, pp. 722–726 (1999)
8. Gammerman, A., Vovk, V.: Kolmogorov complexity: Sources, theory and applications. The Computer Journal 42(4), 252–255 (1999)
9. Bellotti, T., Luo, Z., Gammerman, A.: Qualified predictions for microarray and proteomics pattern diagnosis with confidence machines. International Journal of Neural Systems 15(4), 247–258 (2005)
10. Vega, J., Murari, A., Pereira, A.: Accurate and reliable image classification by using conformal predictors in the TJ-II Thomson scattering. Review of Scientific Instruments 81, 10–18 (2010)
11. Papadopoulos, H., Vovk, V., Gammerman, A.: Regression Conformal Prediction with Nearest Neighbours. J. Artif. Intell. Res (JAIR) 40, 815–840 (2011)
12. Papadopoulos, H., Haralambous, H.: Reliable Prediction Intervals with Regression Neural Networks. Neural Networks 24(8), 842–851 (2011)
13. Li, F., Kosecka, J., Wechsler, H.: Strangeness based feature selection for part based recognition. In: Computer Vision and Pattern Recognition Workshop, p. 22 (2006)
14. Papadopoulos, H.: Inductive Conformal Prediction: Theory and Application to Neural Networks. In: Tools in Artificial Intelligence, ch.18, pp. 315–330. I-Tech, Vienna (2008)
15. Huazhen, W., Chengde, L., Fan, Y., Jinfa, Z.: An online Algorithm with confidence for Real-Time Fault Detection. Journal of Information and Computational Science 6(1), 305–313 (2009)
16. Fan, Y., Huazhen, W., Hong, M., Weiwen, C.: Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. Bmc Bioinformatics10(1), S22, 14-18 (2009)

17. Devetyarov, D., Nouretdinov, I., Burford, B.: Conformal predictors in early diagnostics of ovarian and breast cancers. In: Progress in Artificial Intelligence, pp. 1–13 (2012)
18. Chawla, N.V., Bowyer, K.W., Hall, L.O.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16(1), 321–357 (2002)
19. Grzymala, J.W., Stefanowski, J.: A comparison of two approaches to data mining from imbalanced data. Journal of Intelligent Manufacturing 16(6), 565–573 (2005)
20. Yen, S.-J., Lee, Y.-S.: Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst. Appl. 36(3), 5718–5727 (2009)
21. Ji, H., Zhang, H.X.: Classification with Local Clustering in Imbalanced Data Sets. Advanced Materials Research 219, 151–155 (2011)
22. Wu, J., Xiong, H., Chen, J.: COG.: Local decomposition for rare class analysis. Data Mining and Knowledge Discovery 20(2), 191–220 (2010)
23. Prachuabsupakij, W., Soonthornphisaj, N.: Clustering and combined sampling approaches for multi-class imbalanced data classification. In: Zeng, D. (ed.) Advances in Information Technology and Industry Applications. LNEE, vol. 136, pp. 717–724. Springer, Heidelberg (2012)
24. HuaZhen, W., ChengDe, L., Fan, Y., XueQin, H.: Hedged predictions for traditional Chinese chronic gastritis diagnosis with confidence machine. Computers in Biology and Medicine 39(5), 425–432 (2009)
25. Lyman, P., Georgakis, C.: Plant-wide control of the Tennessee Eastman problem. Computers and Chemical Engineering 19(3), 321–331 (1995)
26. Kulkarni, A., Jayaraman, V., Kulkarni, B.: Knowledge incorporated support vector machines to detect faults in tennessee eastman process. Computers and Chemical Engineering 29(10), 2128–2133 (2005)