Int. j. inf. tecnol. (February 2022) 14(1):539–545
https://doi.org/10.1007/s41870-018-0251-8

ORIGINAL RESEARCH

Issues and challenges of class imbalance problem in classification

Prabhjot Kaur[1] · **Anjana Gosain**[2]

Received: 22 May 2018 / Accepted: 8 October 2018 / Published online: 13 October 2018

Abstract Class imbalance problem is the problem of classification when we seek out exceptional cases using traditional classification algorithms. Traditional classification algorithms are designed to look for either bigger classes or classes with the similar size. These algorithms when used to identify smaller class from the data either fails to detect or gives erroneous results. Researchers have worked on this problem using various concepts, logics or by modifying existing classification algorithms. This paper discusses existing research trends used to solve class imbalance problem. It also highlights the issues and gaps related to this problem.

Keywords Classification · Class imbalance problem · Data level techniques · Ensemble methods · Algorithm level methods

1 Introduction

Classification is a data mining tool which identifies classes from the data based upon certain criteria. There are many real life scenarios where we look for exceptional cases from the whole data-set like looking for credit card frauds from the whole data set of credit card transactions, brain tumor images from the data-set of images, web spam detection from the data base of all e-mails etc., [22, 38, 51, 60]. When the traditional classification procedures were used with above mentioned scenarios, they did not give accurate results as the results were deviating towards the bigger class whereas the need was to sense the smaller class. This issue is interpreted as Class imbalance problem. We were using existing classification algorithms to detect classes from the unbalanced data whereas those algorithms were designed to identify classes from balanced data [22, 38, 51, 60].

Imbalanced data is a combination of classes with unequal size. In Class imbalance domain, we refer these classes as minority (Smaller) and majority (bigger) class and the purpose of proposed solutions is to accurately identify minority class. Researchers have suggested many ways to solve this issue. As per the existing proposed work by the researchers, we can divide the solutions into four categories. Data level, algorithm level, Feature based and hybrid (Data + Algorithm) algorithms. Data level algorithms basically pre-process the data and convert it to a balanced data-set so that existing classification algorithms can be used to handle this problem. Depending upon the logic suggested by the authors data-level algorithms are further divided into oversampling, undersampling and hybrid (Oversampling + Undersampling) sampling categories. In oversampling methods, data is balanced by increasing the size of smaller class either by copying the existing data or by using some other intelligent method. After balancing, the existing classification procedures are applied to classify the data [1, 2, 4, 7, 8, 23, 24, 28, 36, 39, 43, 45, 61]. Undersampling methods decrease the size of majority class either by randomly deleting or by using some other intelligent approach to remove the data from the class so as to balance the data-set before applying traditional classification algorithms [12, 20, 25,

✉ Prabhjot Kaur
thisisprabhjot@gmail.com

Anjana Gosain
anjana_gosain@hotmail.com

[1] Department of IT, MSIT, C4 Janakpuri, New Delhi, India

[2] USICT, GGSIP University, Sector 16C, Dwarka, New Delhi, India

41, 46, 48, 59]. In addition to algorithm and data level approaches, feature selection is another important aspect that can alone alleviate the class imbalance problem. Another study observed that instead of feature selection, interaction between different features is also important. Highly co-related feature can results into more accurate partitions [10, 17, 37, 53, 63]. Recently, the work is reported where the PCA technique is clubbed with the algorithm or data level procedures [14, 35] to solve this issue. Hybrid method uses the concept of undersampling and oversampling in combination to pre-process the data before classification [3, 32, 42]. In algorithm level approaches, authors either worked upon the internal structure of the traditional classification procedures in order to modify the sensitivity of the algorithm towards the bigger class or developed new method to aaliviate class imbalance situation [5, 11, 13, 15, 19, 27, 29, 30, 34, 40, 47, 49, 50, 54–58, 62, 64]. Hybrid category combines algorithm or data level methods with the ensemble approaches like bagging, boosting, random forest etc., [6, 9, 16, 18, 21, 26, 31, 33, 44, 52]. After analyzing the above methods from the year 1997 to 2016, we represented various research trends taken to solve this issue graphically in this paper. It will help the researchers to tackle this problem and face the challenges, which are coming in this domain, in a better manner and in the right direction.

2 Research trends

From the above study, we have recognized four categories which are further divided into nine categories as displayed (Fig. 1). All the techniques suggested in past to alleviate class imbalance problem have used 18 different approaches in their concept as listed in Table 1. Some of the techniques have used more than one approach to tackle the problem. Based upon above analysis, we have decided following parameters to know the research trends in class imbalance domain.

2.1 Publication trend category wise

Figure 2 shows the publication trend category wise for the four categories as data level, algorithm level, Feature based and hybrid level. The work done reported under algorithm level is highest followed by data level and Hybrid level, which has reported almost similar %age of techniques. Considering the sub-categorywise analysis (Fig. 3), we observed that maximum number of techniques (26.58%) are reported in cost-sensitive algorithm level. In data-level category, maximum publications are reported in oversampling (18.99%) and in case of hybrid approach, it is Boosting level (13.92%). Latest category that have been observed during survey is the Hybrid Level Rotation Forest category (1.27%). It is noticed that in the recent years Hybrid ensemble approaches are becoming very famous [55–64].

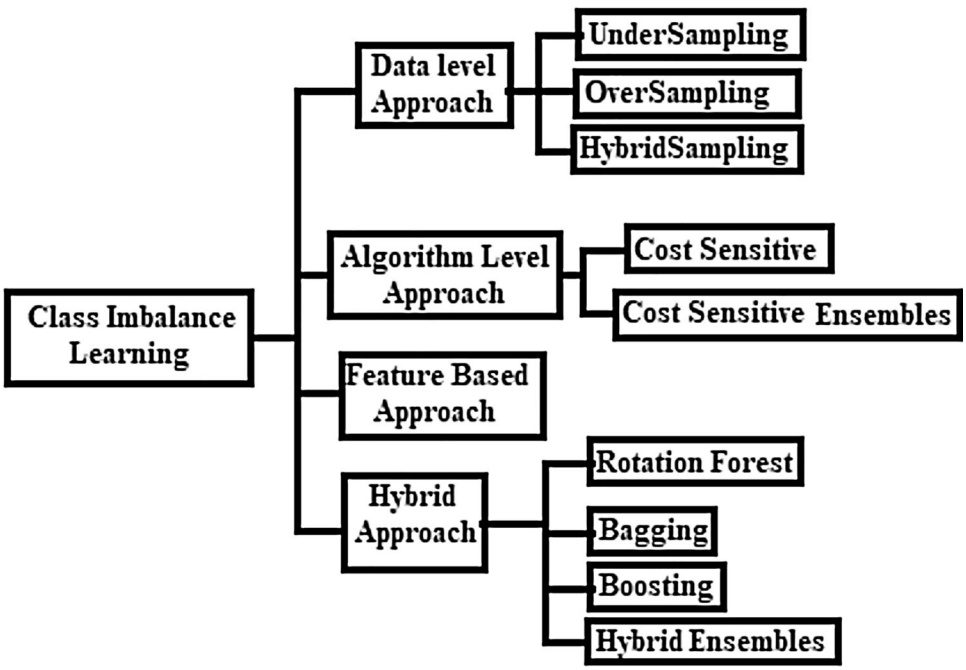


Fig. 1 Categories of class imbalance domain

Table 1 Approaches used in proposed techniques

S. no	Name of the approach	S. no	Name of the approach
1	Nearest neighbor	10	Rough sets
2	Random principle	11	Greedy divide and conquer
3	Genetics	12	Kernel function
4	Clustering	13	Fuzzy rule base
5	Neural networks	14	Bagging
6	PCA (principal component analysis)	15	Boosting
7	SVM (support vector machine)	16	Rotation forest
8	Noise filter	17	Geometric mean
9	Fuzzy logic	18	Immune networks

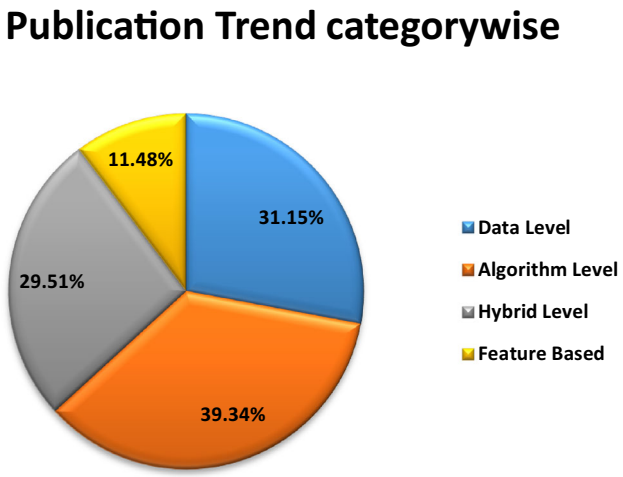


Fig. 2 Publication trend categorywise (color figure online)

2.2 Use of approaches by the techniques

To address the Class Imbalance Problem, authors have used various approaches to enhance the classifier's performance. Figure 4 recorded the trend of popularity in terms of usage of approaches in developing various techniques whereas Fig. 5 recorded it in terms of duration i.e., starting and recent year of the approach used in developing techniques. We observed that most popular approach in terms of usage is the nearest neighbor with 17.86% usage. Other closer approaches are SVM (16.43%), Boosting (15.71%) and Kernel function (14.29%). In terms of duration, the most popular approach is Nearest neighbor with 19 years duration (1997–2015).

SVM and Bagging are sharing popularity with 17 years duration (1999–2016). There are approaches which are used in the single technique only like noise filter (2014), Rough sets (2011), Geometric mean (2013), Rotation forest (2015) and Immune network (2015).

2.3 Tools used by the techniques

Tools are required by researchers for quick implementation and automatic analysis of their work. Different kinds of tools are used by the authors to develop techniques. Based on the availability of information in research papers WEKA (Waikato Environment for Knowledge Analysis), MATLAB and KEEL are the famous tools used by researchers for implementing and analyzing information

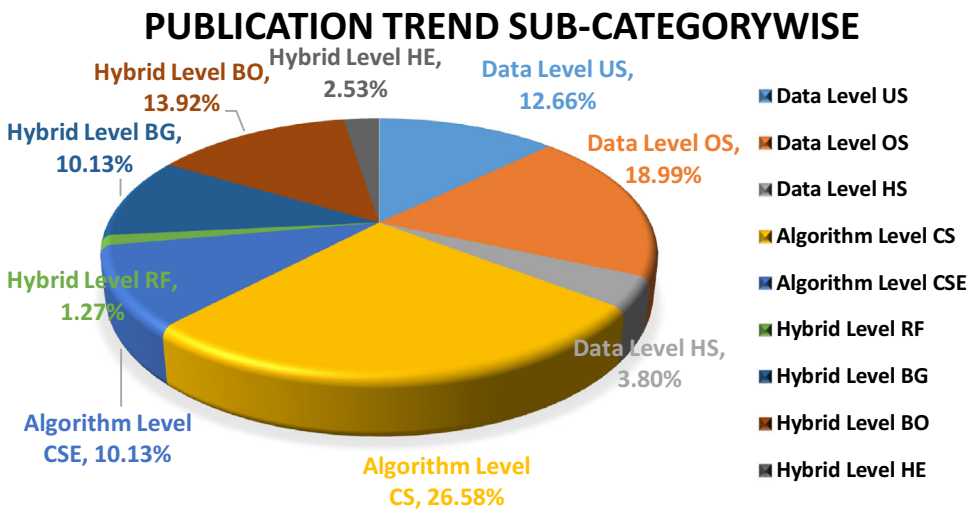
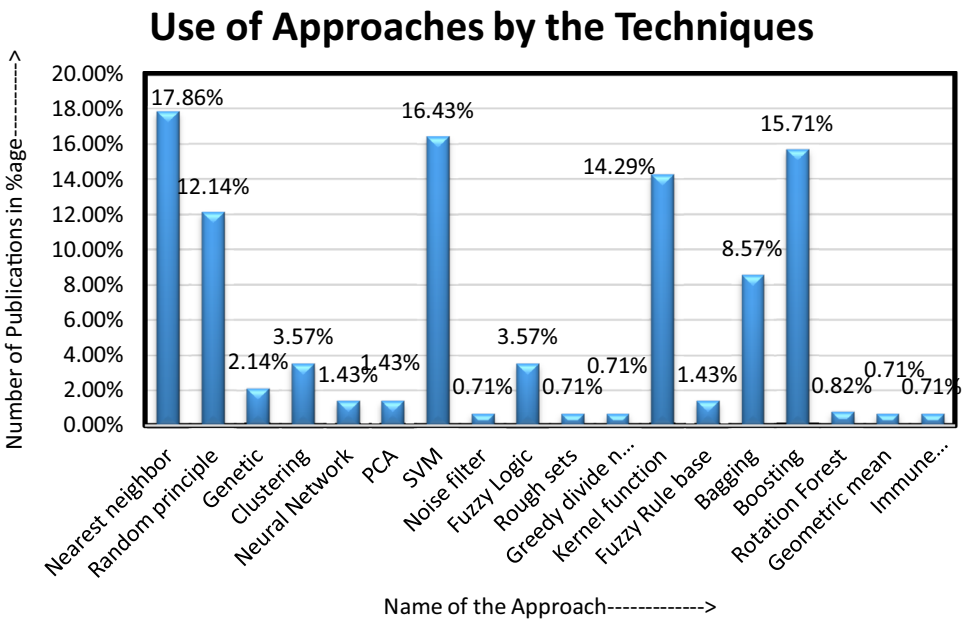


Fig. 3 Publication trend sub-categorywise. *US* undersampling, *OS* oversampling, *HS* hybridsampling, *CS* cost sensitive, *CSE* cost sensitive ensembles, *RF* random forest, *BG* bagging, *BO* boosting, *HE* hybrid ensembles (color figure online)



(Fig. 6). WEKA is the popular tool for analyses. Recently KEEL is used by authors wherein WEKA is already embedded in the tool itself.

2.4 Data set used

We observed from this study that majority of the techniques are evaluated with the data-sets available at UCI repository. Figure 7 shows that 56% techniques out of 79 have used data-set from UCI repository.

3 Issues and challenges related to class imbalance problem

This section discusses various issues that are recognized in class imbalance problem and can be taken as a research challenge to address this problem.

"What if the imbalance ratio is changing dynamically?" Imbalance Ratio (IR) is the ratio of instance count in the bigger class to the instance count in smaller class. IR value

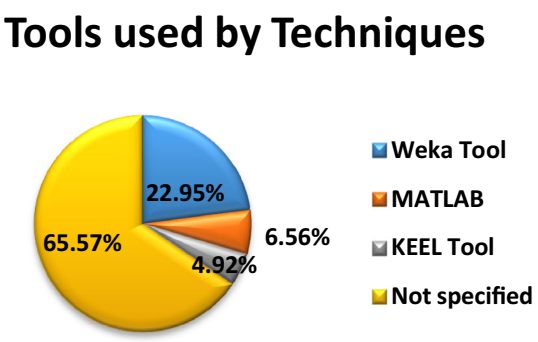


Fig. 6 Tools used by techniques (color figure online)

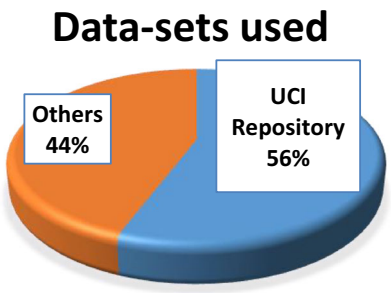


Fig. 7 Data-sets used in papers (color figure online)

may vary from > 1 to any number. The problem become more risky with the enhanced value of IR. No such technique in literature exists which can act dynamically by taking this factor into consideration. One technique may work efficiently for one specific value of IR [51].

"*Where is the best re-balance option?*" "Whether IR = 1 will achieve best results?" Another issue is that performance of techniques does not only depends upon the balancing of data otherwise at IR = 1, techniques will perform in the best manner. So, where is the best re-balance option and on which other factors it depends upon is another open question that can be looked into.

"*Is class imbalance the only problem with data?*" Majorly, the work done under this field is to remove class imbalance effect in the data-sets but if we consider the real situations, there are other data distribution complexities that play a major role in the degraded performance of classifiers. Very less literature is available which deals with the combine effects of CIP and other abnormalities like class overlapping, small disjuncts, class distribution within class etc.

"*Is data free from noise?*" Another important issue in real data-sets is noise, which is present in real data-sets of every possible field in one form or another. In some cases, we have missing values which acts as a noise. In medical data, there is the possibility of vague information in the data due to the acquisition process of images. In web data, there is a possibility of manipulated or changed information due to signal noise/impulse noise etc., very less work is recorded where the researchers have processed noise within the techniques. The techniques are developed either by neglecting the missing values or by assuming that data is cleaned before classification. An efficient technique is still to be developed which can handle such situation along with the other data distribution complexities.

"Which is the best performance metric to assess the techniques developed for CIP?" There are many performance metrics that are designed specifically to deal with Skewed Data Sets (SDS) like F-measure, ROC, AUC, Precision, G-Mean, PRC Curve, K-S Statistics, Recall, Specificity. The reason behind developing these metrics is that the accuracy performance metric used with traditional classifiers gives biased results towards the majority class. But, it is really an open question that which performance metric should be preferable in the specific situation and which metric is more relevant in one situation than the other.

"What if the class distribution of training set differs from the test set?" Class distribution is another important issue in developing an efficient technique as the distribution of test and training data may differ but the techniques are designed by assuming that the distribution of training and test data is same [51].

There is very less literature on Multiclass imbalance problem [38]. Major research is on binary classification. Although researchers have worked with multiple class data-sets but by reducing the multiclass to binary problem by joining majority and minority class separately. These kinds of problems do not work well when applied to the multiclass problem.

References

1. Ai X, Wu J, Sheng VS, Zhao P, Cui Z (2015) Immune centroids oversampling method for binary classification. Comput Intell Neurosci 2015:19
2. Akbani R, Kwek S, Japkowicz N (2004) Applying support vector machines to imbalanced datasets. European conference on machine learning. Springer, Berlin, pp 39–50
3. Al-Rifaie MM, Alhakbani HA (2016) Handling class imbalance in direct marketing dataset using a hybrid data and algorithmic level solutions. SAI Comput Conf (SAI) 2016:446–451
4. Barua S, Islam MM, Yao X, Murase K (2014) MWMOTE–majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans Knowl Data Eng 26(2):405–425
5. Batuwita R, Palade V (2010) FSVM-CIL: fuzzy support vector machines for class imbalance learning. IEEE Trans Fuzzy Syst 18(3):558–571
6. Breiman L (1999) Pasting small votes for classification in large databases and on-line. Mach Learn 36(1–2):85–103
7. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C (2009) Safe-level-smote: Safe-level-synthetic minority over-sampling

technique for handling the class imbalanced problem. Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, pp 475–482
8. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
9. Chawla NV, Lazarevic A, Hall LO, Bowyer KW (2003) SMOTEBoost: Improving prediction of the minority class in boosting. European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, pp 107–119
10. Chen XW, Wasikowski M (2008) Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 124–132
11. Chi Z, Yan H, Pham T (1996) Fuzzy algorithms: with applications to image processing and pattern recognition, vol 10. World Scientific, Singapore
12. Chyi YM (2003) Classification analysis techniques for skewed class distribution problems. Department of Information Management, National Sun Yat-Sen University, Taiwan
13. Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola JS (2002) On kernel-target alignment. Advances in neural information processing systems. MIT Press, Cambridge, pp 367–373
14. D'Addabbo A, Maglietta R (2015) Parallel selective sampling method for imbalanced and large data classification. Pattern Recogn Lett 62:61–67
15. Dai HL (2015) Class imbalance learning via a fuzzy total margin based support vector machine. Appl Soft Comput 31:172–184
16. Fattahi S, Othman Z, Othman ZA (2015) New approach with ensemble method to address class imbalance problem. J Theor Appl Inf Technol 72(1):23–33
17. Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3:1289–1305
18. Galar M, Fernández A, Barrenechea E, Herrera F (2013) EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. Pattern Recogn 46(12):3460–3471
19. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A review on ensembles for the class imbalance problem: bagging-, boosting- and hybrid-based approaches. IEEE Trans Syst Man Cybern C 42(4):463–484
20. García S, Herrera F (2009) Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. Evolut Comput 17(3):275–306
21. Guo H, Viktor HL (2004) Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. ACM SIGKDD Explor Newsl 6(1):30–39
22. Guo X, Yin Y, Dong C, Yang G, Zhou G (2008) On the class imbalance problem. In Natural Computation, 2008. ICNC'08. Fourth International Conference on vol. 4:192–201. IEEE
23. Han H, Wang WY, Mao BH (2005) Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. International Conference on Intelligent Computing. Springer, Berlin, pp 878–887
24. He H, Bai Y, Garcia EA, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on IEEE, 1322–1328
25. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
26. Hido S, Kashima H, Takahashi Y (2009) Roughly balanced bagging for imbalanced data. Stat Anal Data Min ASA Data Sci J 2(5–6):412–426
27. Hong X, Chen S, Harris CJ (2007) A kernel-based two-class classifier for imbalanced data sets. IEEE Trans Neural Netw 18(1):28–41
28. Hu S, Liang Y, Ma L, He Y (2009) MSMOTE: improving classification performance when training data is imbalanced. In Computer Science and Engineering. WCSE'09. Second International Workshop on, IEEE 2:13–17
29. Imam T, Ting KM, Kamruzzaman J (2006) z-SVM: an SVM for improved classification of imbalanced data. Australasian Joint Conference on Artificial Intelligence. Springer, Berlin, pp 264–273
30. Kandola JS, Shawe-Taylor J (2003) Refining kernels for regression and uneven classification problems. In: Proceedings of AISTATS
31. Kim MJ, Kang DK (2013) Geometric mean based boosting algorithm to resolve data imbalance problem. In: Proceedings of PACIS, 2013, pp 1–27
32. Li DC, Wu CS, Tsai TI, Lina YS (2007) Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. Comput Oper Res 34(4):966–982
33. Lin CF, Wang SD (2002) Fuzzy support vector machines. IEEE Trans Neural Netw 13(2):464–471
34. Mangasarian OL, Wild EW (2001) Proximal support vector machine classifiers. In: Proceedings of KDD-2001: knowledge discovery and data mining, 2001
35. Maruthi Padmaja T, Raju BS, Hota RN, Krishna PR (2014) Class imbalance and its effect on PCA preprocessing. Int J Knowl Eng Soft Data Paradig 4(3):272–294
36. Mi Y (2013) Imbalanced classification based on active learning SMOTE. Res J Appl Sci Eng Technol 5:944–949
37. Mladenic D, Grobelnik M (1999) Feature selection for unbalanced class distribution and naive bayes. ICML 99:258–267
38. Mollineda RA, Alejo R, Sotoca JM (2007) The class imbalance problem in pattern classification and learning. In II Congreso Español de Informática (CEDI 2007). ISBN, 978–84
39. Nakamura M, Kajiwara Y, Otsuka A, Kimura H (2013) Lvq-smote–learning vector quantization based synthetic minority over–sampling technique for biomedical data. BioData Min 6(1):16
40. Pang S, Zhu L, Chen G, Sarrafzadeh A, Ban T, Inoue D (2013) Dynamic class imbalance learning for incremental LPSVM. Neural Netw 44:87–100
41. Rahman MM, Davis D (2013) Cluster based under-sampling for unbalanced cardiovascular data. Proc World Congr Eng 3:3–5
42. Ramentol E, Caballero Y, Bello R, Herrera F (2012) SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl Inf Syst 33(2):245–265
43. Sáez JA, Luengo J, Stefanowski J, Herrera F (2014) Managing borderline and noisy examples in imbalanced classification by combining SMOTE with ensemble filtering. In International Conference on Intelligent Data Engineering and Automated Learning, Springer, Cham, 61–68
44. Salunkhe UR, Mali SN (2016) Classifier ensemble design for imbalanced data classification: a hybrid approach. Proc Comput Sci 85:725–732
45. Stefanowski J, Wilk S (2008) Selective pre-processing of imbalanced data for improving classification performance. In International Conference on data warehousing and knowledge discovery, Springer, Berlin, 283–292
46. Seiffert C, Khoshgoftaar TM, Van Hulse J, Napolitano A (2010) RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Trans Syst Man Cybern A Syst Humans 40(1):185–197
47. Shibulal B, Al-Bahry SN, Al-Wahaibi YM, Elshafie AE, Al-Bemani AS, Joshi SJ (2014) Microbial enhanced heavy oil

recovery by the aid of inhabitant spore-forming bacteria: an insight review. Sci World J 2014:1–12
48. Tang Y, Zhang YQ (2006) Granular SVM with repetitive undersampling for highly imbalanced protein homology prediction. In Granular Computing, IEEE International Conference on IEEE, 457–460
49. Ting KM (2002) An instance-weighting method to induce cost-sensitive trees. IEEE Trans Knowl Data Eng 14(3):659–665
50. Tomar D, Agarwal S (2016) Prediction of defective software modules using class imbalance learning. Appl Comput Intell Soft Comput 2016:6
51. Visa S, Ralescu A (2005) Issues in mining imbalanced data sets-a review paper. Proceedings of the sixteen midwest artificial intelligence and cognitive science conference 2005:67–73
52. Wang BX, Japkowicz N (2010) Boosting support vector machines for imbalanced data sets. Knowl Inf Syst 25(1):1–20
53. Wasikowski M, Chen XW (2010) Combating the small sample class imbalance problem using feature selection. IEEE Trans Knowl Data Eng 22(10):1388–1400
54. Wu G, Chang EY (2003) Adaptive feature-space conformal transformation for imbalanced-data learning. Proceedings of the 20th International Conference on Machine Learning (ICML-03) 816–823
55. Wu S, Amari SI (2002) Conformal transformation of kernel functions: a data-dependent way to improve support vector machine classifiers. Neural Process Lett 15(1):59–67
56. Wu G, Chang EY (2003) Class-boundary alignment for imbalanced dataset learning. In ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC 49–56
57. Wu G, Chang EY (2005) KBA: kernel boundary alignment considering imbalanced data distribution. IEEE Trans Knowl Data Eng 17(6):786–795
58. Yang CY, Yang JS, Wang JJ (2009) Margin calibration in SVM class-imbalanced learning. Neurocomputing 73(1–3):397–411
59. Yen SJ, Lee YS (2009) Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl 36(3):5718–5727
60. Yong Y (2012) The research of imbalanced data set of sample sampling method based on K-means cluster and genetic algorithm. Energy Proc 17:164–170
61. Zhang Y, Wang D (2013) A cost-sensitive ensemble method for class-imbalanced datasets. Abstract and applied analysis 2013:1–6
62. Zhao Z, Zhong P, Zhao Y (2011) Learning SVM with weighted maximum margin criterion for classification of imbalanced data. Math Comput Model 54(3–4):1093–1099
63. Zheng Z, Wu X, Srihari R (2004) Feature selection for text categorization on imbalanced data. ACM SIGKDD Explor Newsl 6(1):80–89
64. Zhuang D, Zhang B, Yang Q, Yan J, Chen Z, Chen Y (2005) Efficient text classification by weighted proximal SVM. In Data Mining, Fifth IEEE International Conference on IEEE, 8