Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution

A. Fernández[1,*], S. García[1], and F. Herrera[2]

[1] Dept. of Computer Science, University of Jaén
Tel.:+34-953-213016; Fax:+34-953-212472
{alberto.fernandez,sglopez}@ujaen.es
[2] Dept. of Computer Science and A.I., University of Granada,
herrera@decsai.ugr.es

Abstract. Classifier learning with datasets which suffer from imbalanced class distributions is an important problem in data mining. This issue occurs when the number of examples representing one class is much lower than the ones of the other classes. Its presence in many real-world applications has brought along a growth of attention from researchers.

The aim of this work is to shortly review the main issues of this problem and to describe two common approaches for dealing with imbalance, namely sampling and cost sensitive learning. Additionally, we will pay special attention to some open problems, in particular we will carry out a discussion on the data intrinsic characteristics of the imbalanced classification problem which will help to follow new paths that can lead to the improvement of current models, namely size of the dataset, small disjuncts, the overlapping between the classes and the data fracture between training and test distribution.

Keywords: Imbalanced Datasets, Sampling, Cost Sensitive Learning, Small Disjuncts, Overlapping, Dataset Shift.

1 Introduction

In many applications, there exists a significant difference between the class prior rates, that is the probability a particular example belongs to a particular class. This situation is known as the class imbalance problem [1,2] and it is dominant in a high number of real problems including, but not limited to, telecommunications, WWW, finances, ecology, biology, medicine and so on; for which it is considered as one of the top problems in data mining [3]. Furthermore, it is worth to point out that the positive or minority class is usually the one that has the highest interest from the learning point of view and it also implies a great cost when it is not well classified [4].

The hitch with imbalanced datasets is that standard classification learning algorithms are often biased towards the majority classes and therefore there is a

[*] Corresponding author.

E. Corchado, M. Kurzyński, M. Woźniak (Eds.): HAIS 2011, Part I, LNAI 6678, pp. 1–10, 2011.

Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution

A. Fernández[1,*], S. García[1], and F. Herrera[2]

[1] Dept. of Computer Science, University of Jaén
Tel.:+34-953-213016; Fax:+34-953-212472
{alberto.fernandez,sglopez}@ujaen.es
[2] Dept. of Computer Science and A.I., University of Granada,
herrera@decsai.ugr.es

Abstract. Classifier learning with datasets which suffer from imbalanced class distributions is an important problem in data mining. This issue occurs when the number of examples representing one class is much lower than the ones of the other classes. Its presence in many real-world applications has brought along a growth of attention from researchers.

The aim of this work is to shortly review the main issues of this problem and to describe two common approaches for dealing with imbalance, namely sampling and cost sensitive learning. Additionally, we will pay special attention to some open problems, in particular we will carry out a discussion on the data intrinsic characteristics of the imbalanced classification problem which will help to follow new paths that can lead to the improvement of current models, namely size of the dataset, small disjuncts, the overlapping between the classes and the data fracture between training and test distribution.

Keywords: Imbalanced Datasets, Sampling, Cost Sensitive Learning, Small Disjuncts, Overlapping, Dataset Shift.

1 Introduction

In many applications, there exists a significant difference between the class prior rates, that is the probability a particular example belongs to a particular class. This situation is known as the class imbalance problem [1,2] and it is dominant in a high number of real problems including, but not limited to, telecommunications, WWW, finances, ecology, biology, medicine and so on; for which it is considered as one of the top problems in data mining [3]. Furthermore, it is worth to point out that the positive or minority class is usually the one that has the highest interest from the learning point of view and it also implies a great cost when it is not well classified [4].

The hitch with imbalanced datasets is that standard classification learning algorithms are often biased towards the majority classes and therefore there is a higher misclassification rate in the minority class instances. Therefore, throughout the last year, many solutions have been proposed to deal with this problem, which can be categorised into two major groups:

[*] Corresponding author.

E. Corchado, M. Kurzyński, M. Woźniak (Eds.): HAIS 2011, Part I, LNAI 6678, pp. 1–10, 2011.

1. **Data sampling:** in which the training instances are modified in such a way as to produce a balanced class distribution that allow classifiers to perform in a similar manner to standard classification [5,6].
2. **Algorithmic modification:** this procedure is oriented towards the adaptation of base learning methods to be more attuned to class imbalance issues [7]. We must also stress in this case the use of cost-sensitive learning solutions, which basically assume higher misclassification costs with samples in the rare class and seek to minimise the high cost errors [8,9].

In this contribution, our initial goal is to develop a review on this type of methodologies, analysing the different approaches that have been traditionally applied and describing the main features of each one of them.

Additionally, most of the studies on the behavior of several standard classifiers in imbalance domains have shown that significant loss of performance is mainly due to skew of class distributions. However, there are also several investigations which also suggest that there are other factors that contribute to such performance degradation, for example, size of the dataset, small disjuncts, overlapping between classes and dataset shift among others [10,11,12,13,14].

According to the previous issues, we aim to carry out a discussion about the source where the difficulties for imbalanced classification emerge, focusing on the analysis of significant data intrinsic characteristics such as the ones previously mentioned. We must point out that some of these topics have recent studies associated, but that they still need to be addressed in detail in order to have models of quality in this classification scenario.

In order to do so, this contribution is organised as follows. First, Section 2 presents the problem of imbalanced datasets, introducing its features and the metrics employed in this context. Next, Section 3 describes the preprocessing and cost sensitive methodologies that have been proposed to deal with this problem. Section 4 is devoted to analyse and discuss some open problems on the topic. Finally, Section 5 summarises and concludes the work.

2 Imbalanced Datasets in Classification

In the classification problem field, the scenario of imbalanced datasets appears with high frequency. The main property of this type of classification problem is that the examples of one class outnumbers examples of the other one [1,2]. The minority classes are usually the most important concepts to be learnt, since they represent rare cases [15] or because the data acquisition of these examples is costly [16].

Since most of the standard learning algorithms consider a balanced training set, this situation may cause the obtention of suboptimal classification models, i.e. a good coverage of the majority examples whereas the minority ones are misclassified more frequently; therefore, those algorithms which obtains a good behavior in the framework of standard classification do not necessarily achieves the best performance for imbalanced data-sets [17]. There are several reasons behind this behaviour which are enumerated below:

1. The use of global performance measures for guiding the search process, such as standard accuracy rate, may benefit the covering of the majority examples.
2. Classification rules that predict the positive class are often highly specialised and thus their coverage is very low, hence they are discarded in favour of more general rules, i.e. those that predict the negative class.
3. It is always not easy to distinguish between noise examples and minority class examples and they can be completely ignored by the classifier.

In recent years, the imbalanced learning problem has received a high attention in the machine learning community. Specifically, regarding real world domains the importance of the imbalance learning problem is growing, since it is a recurring problem in many applications. As a few examples, we may find very high resolution airbourne imagery [18], remote-sensing [19], face recognition [20] and especially medical diagnosis [21,22].

We must also point out that in imbalanced domains the evaluation of the classifiers' performance must be carried out using specific metrics to take into account the class distribution. Since in this classification scenario we intend to achieve good quality results for both classes, one way to combine the individual measures of both the positive and negative classes, and to produce an evaluation criteria, is to use the Receiver Operating Characteristic (ROC) graphic [23]. This graphic allows to visualise the trade-off between the benefits (TP_{rate}) and costs (FP_{rate}), thus it evidences that any classifier cannot increase the number of true positives without also increasing the false positives. The Area Under the ROC Curve (AUC) [24] provides a single measure of a classifier's performance for evaluating which model is better on average. Other metrics of interest to be stressed in this area are the geometric mean of the true rates [25] and the F-measure [26].

3 Addressing Classification with Imbalanced Data: Preprocessing and Cost Sensitive Learning

A large number of approaches have been previously proposed to deal with the class-imbalance problem. These approaches can be categorised in two groups: the internal approaches that create new algorithms or modify existing ones to take the class-imbalance problem into consideration [25,27] and external approaches that preprocess the data in order to diminish the effect of their class imbalance [5,28]. Furthermore, cost-sensitive learning solutions incorporating both the data and algorithmic level approaches assume higher misclassification costs with samples in the minority class and seek to minimise the high cost errors [8,29].

Regarding this, in this section we first introduce the main features of preprocessing techniques and next, we describe the cost-sensitive learning approach.

3.1 Preprocessing Imbalanced Datasets: Resampling Techniques

In the specialised literature, we can find some papers about resampling techniques studying the effect of changing class distribution to deal with imbalanced datasets where it has been empirically proved that, applying a preprocessing step in order to balance the class distribution, is usually a positive solution [5,30,31]. The main advantage of these techniques is that they are independent of the underlying classifier.

Resampling techniques can be categorised into three groups: *undersampling methods*, which create a subset of the original dataset by eliminating instances (usually majority class instances); *oversampling methods*, which create a superset of the original dataset by replicating some instances or creating new instances from existing ones; and finally, *hybrids methods*, that combine both sampling approaches.

Among these categories, there exist several different proposals, from which the most simple ones are non heuristic methods such as random undersampling and random oversampling. In the first case, the major drawback is that it can discard potentially useful data, that could be important for the induction process. For random oversampling, several authors agree that this method can increase the likelihood of occurring overfitting, since it makes exact copies of existing instances.

According to the previous facts, more sophisticated methods have been proposed. Among them, the "Synthetic Minority Oversampling Technique" (SMOTE) [6] have become one of the most significant approaches in this area. In brief, its main idea is to create new minority class examples by interpolating several minority class instances that lie together for oversampling the training set.

Regarding undersampling techniques, the application of genetic algorithms for the correct identification of the most significant instance have shown to achieve very positive results [32,33]. Also, a training set selection can be carried out for enhancing the learning stage of several classification algorithms in the area of imbalanced data-sets [34].

Finally, some combination of preprocessing of instances with data cleaning techniques could lead to diminish the overlapping that is introduced from sampling methods. Some representative works in this area [5] include the condensed nearest neighbour rule and Tomek Links integration method, the neighbourhood cleaning rule based on the edited nearest neighbour (ENN) rulewhich removes examples that differ from two of its three nearest neighbours, and the integrations of SMOTE with ENN and SMOTE with Tomek links.

3.2 Cost-Sensitive Learning

Cost-sensitive learning takes into account the variable cost of a misclassification of the different classes [8,9]. In this case, a cost matrix codifies the penalties of classifying examples of one class as a different one. However, given a dataset, this matrix is not usually given [1,35]. Specifically, when dealing with imbalanced problems it is usually of most interest to recognise the positive instances rather than the negative ones and therefore, the cost when mistaking a positive instance is higher than the cost of mistaking a negative one.

Three main general approaches have been proposed to deal with cost-sensitive problems:

1. Methods based on modifying the training data set. The most popular technique lies in resampling the original class distribution of the training data set according to the cost decision matrix by means of undersampling/ oversampling, modifying decision thresholds or assigning instance weights [9,29].
2. Other methods change the learning process in order to build a cost-sensitive classifiers, for example, in the context of decision tree induction, the tree-building strategies are adapted to minimise the misclassification costs [36].
3. Methods based on the Bayes decision theory that assign instances to the class with minimum expected cost [7,8].

4 Analysing Open Problems Related to Intrinsic Data Characteristics in Imbalanced Classification

As it was stated in the introduction of this work, skewed class distribution does not hinder the learning task by itself [1,2], but usually a series of difficulties related with this problem turn up.

In this section, we aim to develop a discussion on the nature of the problem itself, emphasising several data intrinsic characteristics that do have a strong influence on imbalanced classification, in order to be able to address this problem in a more reasonable way.

First, we present the issues about the size of the dataset and the difficulties related to the apparition of small disjuncts in the data. Next, we focus on the class overlap and how it is extremely significant on imbalanced domains. Finally, we define the dataset shift problem and its relationship to imbalanced datasets classification.

4.1 Small Sample Size and Small Disjuncts

One problem that can arise in classification is a small sample size [37]. This issue is related to the “lack of information” where induction algorithms do not have enough data to make generalisations about the distribution of samples. This problem is increased in the presence of high dimensional data, i.e a large number of features.

The combination of imbalanced data and the small sample size problem presents a new challenge to the research community [38]. In this scenario, the minority class can be poorly represented and the knowledge model to learn this data space become too specific, leading to overfitting. Therefore, two datasets can not be considered to present the complexity with the same imbalance ratio (the ratio between the positive and negative instances [39]) but it is also significant how good do the training data represents the minority instances.

On the other hand, the existence of the imbalanced classes is closely related to the problem of small disjuncts. This situation occurs when the concepts are represented within small clusters, which arise as a direct result of underrepresented subconcepts [11,40]. Although those small disjuncts are implicit in most of the problems, the existence of these small disjuncts highly increases the complexity of the problem in the case of imbalance because it becomes hard to know whether these examples represent an actual subconcept or are merely attributed to noise [41].

4.2 Overlapping or Class Separability

The problem of overlapping between classes appears when a region of the data space contains a similar quantity of training data from each class. This lead to develop an inference with almost the same a priori probabilities in this overlapping area, which makes very hard or even impossible the distinction between the two classes. Indeed, any “linearly separable” problem can be solved by any simple classifier regardless of the class distribution.

There are several works which aim to study the relationship between overlapping and class imbalance. Specifically, in [12] we may find a study where the authors propose several experiments with synthetic datasets varying the imbalance ratio and the overlap existing between the two classes. Their conclusions stated that it is not the class probabilities the main responsible for the hinder in the classification performance, but instead the degree of overlapping between the classes.

Also, in [42] the authors developed a similar study with several algorithms in different situations of imbalance and overlap focusing in the the k-NN algorithm. In this case, the authors proposed two different frameworks: on the one hand, they try to find the relation when the imbalance ratio in the overlap region is similar to the overall imbalance ratio whereas on the other hand, they search for the relation when the imbalance ratio in the overlap region is inverse to the overall one (the positive class is locally denser than the negative class in the overlap region). They shown that when the overlapped data is not balanced, the IR in overlapping can be more important than the overlapping size. In addition, classifiers based on more global learning attain greater TP rates whereas more local learning models obtain better TN rates than the former.

More recent works [13] have extracted empirically some interesting findings with real world datasets. Specifically, the authors depicted the performance of the different datasets ordered according to different data complexity measures (including the IR) in order to search for some regions of interesting good or bad behaviour. They could not characterize any interesting behaviour according IR, but they do for example according the so called metric $F1$ or *maximum Fishers discriminant ratio* [43], which measures the overlap of individual feature values.

A closely related issue is the how deep is the impact of noisy and borderline examples from the minority class on the classifier performance [44], and also its relationship with the use of focused re-sampling methods with respect to the simplest random and cluster oversampling.

As a final remark, a positive approach at the algorithm-level could consist in working with different granular levels, in a way that more general submodels of knowledge could cover the largest part of the problem space, whereas in more difficult areas, that is, boundary zones with a high degree of overlapping, we could use more specific discrimination functions in different paradigms of learning algorithms. In [45] the authors introduced a fuzzy system with hierarchical fuzzy partitions for managing specific regions, i.e. the most difficult areas.

4.3 Dataset Shift

The problem of dataset shift [46] is defined as the case where training and test data follow different distributions. This is a common problem that can affect all kind of classification problems, and it often appears due to sample selection bias issues. A mild degree of dataset shift is present in most real-world problems, but general classifiers are often capable of handling it without a severe performance loss.

However, the dataset shift issue is specially relevant when dealing with imbalanced classification, because in highly imbalanced domains, the minority class is particularly sensitive to singular classification errors, due to the typically low number of examples it presents [14]. In the most extreme cases, a single misclassified example of the minority class can create a significant drop in performance. There are two different potential approaches in the study of the effect and solution of dataset shift in imbalanced domains:

1. The first one focuses on intrinsic dataset shift, that is, the data of interest includes some degree of shift that is producing a relevant drop in performance. In this case, we may develop techniques to discover and measure the presence of dataset shift [47], but adapting them to focus on the minority class. Furthermore, we may design algorithms that are capable of working under dataset shift conditions, either by means of preprocessing techniques [48] or with ad hoc algorithms that are [49]. In both cases, we are not aware of any proposals in the literature that focus on the problem of imbalanced classification in the presence of dataset shift.
2. The second branch in terms of dataset shift in imbalanced classification is related to induced dataset shift. Most current state of the art research is validated through stratified cross-validation techniques, which are another potential source of shift in the machine learning process. A more suitable validation technique needs to be developed in order to avoid introducing dataset shift issues artificially.

5 Concluding Remarks

In this contribution we have reviewed the topic of classification with imbalanced datasets, focusing on two main issues: (1) to present the main approaches for dealing with this problem, namely preprocessing of instances and cost-sensitive

learning, and (2) to develop a throughout discussion on the data intrinsic characteristics of this scenario of data mining.

Specifically, we have pointed out that the imbalance ratio have a significant effect on the classifiers' performance, but that there other issues that must be taken into account such as small sample size, small disjuncts, class overlapping and dataset shift. Overcoming these problems can be the key for developing new approaches that improve the correct identification of both the minority and majority classes, and therefore we have stressed them as future trends of research for imbalanced datasets.

Acknowledgment

This work had been supported by the Spanish Ministry of Science and Technology under Project TIN2008-06681-C06-01.

References

1. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 23(4), 687–719 (2009)
2. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284 (2009)
3. Yang, Q., Wu, X.: 10 challenging problems in data mining research. International Journal of Information Technology and Decision Making 5(4), 597–604 (2006)
4. Elkan, C.: The foundations of cost–sensitive learning. In: Proceedings of the 17th IEEE International Joint Conference on Artificial Intelligence (IJCAI 2001), pp. 973–978 (2001)
5. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behaviour of several methods for balancing machine learning training data. SIGKDD Explorations 6(1), 20–29 (2004)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over–sampling technique. Journal of Artificial Intelligent Research 16, 321–357 (2002)
7. Zadrozny, B., Elkan, C.: Learning and making decisions when costs and probabilities are both unknown. In: Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD 2001), pp. 204–213 (2001)
8. Domingos, P.: Metacost: A general method for making classifiers cost–sensitive. In: KDD 1999, pp. 155–164 (1999)
9. Zadrozny, B., Langford, J., Abe, N.: Cost–sensitive learning by cost–proportionate example weighting. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), pp. 435–442 (2003)
10. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intelligent Data Analysis Journal 6(5), 429–450 (2002)
11. Weiss, G.M., Provost, F.J.: Learning when training data are costly: The effect of class distribution on tree induction. Journal of Artificial Intelligence Research 19, 315–354 (2003)

12. Prati, R.C., Batista, G.E.A.P.A., Monard, M.C.: Class imbalances *versus* class overlapping: An analysis of a learning system behavior. In: Monroy, R., Arroyo-Figueroa, G., Sucar, L.E., Sossa, H. (eds.) MICAI 2004. LNCS (LNAI), vol. 2972, pp. 312–321. Springer, Heidelberg (2004)
13. Luengo, J., Fernández, A., García, S., Herrera, F.: Addressing data complexity for imbalanced data sets: analysis of SMOTE–based oversampling and evolutionary undersampling. In: Soft Computing (in press 2011), doi:10.1007/s00500–010–0625–8
14. Moreno-Torres, J.G., Herrera, F.: A preliminary study on overlapping and data fracture in imbalanced domains by means of genetic programming–based feature extraction. In: 10th International Conference on Intelligent Systems Design and Applications (ISDA 2010), pp. 501–506 (2010)
15. Weiss, G.M.: Mining with rarity: a unifying framework. SIGKDD Explorations 6(1), 7–19 (2004)
16. Weiss, G.M., Tian, Y.: Maximizing classifier utility when there are data acquisition and modeling costs. Data Mining and Knowledge Discovery 17(2), 253–282 (2008)
17. Fernandez, A., García, S., Luengo, J., Bernadó-Mansilla, E., Herrera, F.: Genetics-based machine learning for rule induction: State of the art, taxonomy and comparative study. IEEE Transactions on Evolutionary Computation 14(6), 913–941 (2010)
18. Chen, X., Fang, T., Huo, H., Li, D.: Graph–based feature selection for object–oriented classification in VHR airborne imagery. IEEE Transactions on Geoscience and Remote Sensing 49(1), 353–365 (2011)
19. Williams, D., Myers, V., Silvious, M.: Mine classification with imbalanced data. IEEE Geoscience and Remote Sensing Letters 6(3), 528–532 (2009)
20. Kwak, N.: Feature extraction for classification problems and its application to face recognition. Pattern Recognition 41(5), 1718–1734 (2008)
21. Mazurowski, M.A., Habas, P.A., Zurada, J.M., Lo, J.Y., Baker, J.A., Tourassi, G.D.: Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks 21(2-3) (2008)
22. Peng, X., King, I.: Robust BMPM training based on second–order cone programming and its application in medical diagnosis. Neural Networks 21(2-3), 450–457 (2008)
23. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition 30(7), 1145–1159 (1997)
24. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering 17(3), 299–310 (2005)
25. Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition 36(3), 849–851 (2003)
26. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)
27. Ducange, P., Lazzerini, B., Marcelloni, F.: Multi–objective genetic fuzzy classifiers for imbalanced and cost–sensitive datasets. Soft Computing 14(7), 713–728 (2010)
28. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalanced data sets. Computational Intelligence 20(1), 18–36 (2004)
29. Zhou, Z.H., Liu, X.Y.: Training cost–sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering 18(1), 63–77 (2006)
30. Fernández, A., García, S., del Jesus, M.J., Herrera, F.: A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data–sets. Fuzzy Sets and Systems 159(18), 2378–2398 (2008)

31. Fernández, A., del Jesus, M.J., Herrera, F.: On the 2–tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data–sets. Information Sciences 180(8), 1268–1291 (2010)
32. García, S., Cano, J., Herrera, F.: A memetic algorithm for evolutionary prototype selection: a scaling up approach. Pattern Recognition 41(8), 2693–2709 (2008)
33. García, S., Herrera, F.: Evolutionary under-sampling for classification with imbalanced data sets: Proposals and taxonomy. Evolutionary Computation 17(3), 275–306 (2009)
34. García, S., Fernández, A., Herrera, F.: Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. Applied Soft Computing 9, 1304–1314 (2009)
35. Sun, Y., Kamel, M.S., Wong, A.K.C., Wang, Y.: Cost–sensitive boosting for classification of imbalanced data. Pattern Recognition 40(12), 3358–3378 (2007)
36. Ling, C.X., Yang, Q., Wang, J., Zhang, S.: Decision trees with minimal costs. In: ICML (2004)
37. Raudys, S., Jain, A.: Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(3), 252–264 (1991)
38. Wasikowski, M., Chen, X.W.: Combating the small sample class imbalance problem using feature selection. IEEE Transactions on Knowledge and Data Engineering 22(10), 1388–1400 (2010)
39. Orriols-Puig, A., Bernadó-Mansilla, E.: Evolutionary rule–based systems for imbalanced datasets. Soft Computing 13(3), 213–225 (2009)
40. Orriols-Puig, A., Bernadó-Mansilla, E., Goldberg, D.E., Sastry, K., Lanzi, P.L.: Facetwise analysis of XCS for problems with class imbalances. IEEE Transactions on Evolutionary Computation 13, 260–283 (2009)
41. Jo, T., Japkowicz, N.: Class imbalances versus small disjuncts. ACM SIGKDD 6(1), 40–49 (2004)
42. García, V., Mollineda, R., Sánchez, J.S.: On the k–NN performance in a challenging scenario of imbalance and overlapping. Pattern Analysis Applications 11(3-4), 269–280 (2008)
43. Ho, T., Basu, M.: Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 289–300 (2002)
44. Napierała, K., Stefanowski, J., Wilk, S.: Learning from imbalanced data in presence of noisy and borderline examples. In: Szczuka, M., Kryszkiewicz, M., Ramanna, S., Jensen, R., Hu, Q. (eds.) RSCTC 2010. LNCS, vol. 6086, pp. 158–167. Springer, Heidelberg (2010)
45. Fernández, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. International Journal of Approximate Reasoning 50(3), 561–577 (2009)
46. Quiñonero Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset Shift in Machine Learning. The MIT Press, Cambridge (2009)
47. Cieslak, D.A., Chawla, N.V.: A framework for monitoring classifiers performance: when and why failure occurs? Knowledge and Information Systems 18(1), 83–108 (2009)
48. Moreno-Torres, J.G., Llorà, X., Goldberg, D.E., Bhargava, R.: Repairing Fractures between Data using Genetic Programming-based Feature Extraction: A Case Study in Cancer Diagnosis. Information Sciences, doi:10.1016/j.ins.2010.09.018 (in press)
49. Bickel, S., Brückner, M., Scheffer, T.: Discriminative learning under covariate shift. Journal of Machine Learning Research 10, 2137–2155 (2009)