An Approach to Imbalanced Data Classification Based on Instance Selection and Over-Sampling

Ireneusz Czarnowski(✉) and Piotr Jędrzejowicz

Department of Information Systems, Gdynia Maritime University,
Morska 83, 81-225 Gdynia, Poland
{i.czarnowski,p.jedrzejowicz}@umg.edu.pl

Abstract. The paper referees to a problem of learning from class-imbalanced data. The class imbalance problem arises when the number of instances from different classes differs substantially. Instance selection aims at deciding which instances from the training set should be retained and used during the learning process. Over-sampling is an approach dedicated to duplicate minority class instances. In the paper, a hybrid approach for the imbalanced data learning using the over-sampling and instance selection techniques is proposed. Instances are selected to reduce the number of instances belonging to the majority class, while the number of instances belonging to the minority class is expanded. The process of instance selection is based on clustering, where the authors' approach to clustering and instance selection using an agent-based population learning algorithm is applied. As a result a more balanced distribution of instances belonging to different classes is obtained and a dataset size is reduced. The proposed approach is validated experimentally using several benchmark datasets.

Keywords: Instance selection · Clustering · Imbalanced data · Over-sampling · Under-sampling

1 Introduction

In the real world, there are many problems for which available data are class imbalanced. Class imbalance means that the number of instances in classes is not equally distributed. In the case of the imbalanced data, *the distribution of the data in the feature space is usually skewed in class imbalanced datasets* [1]. The imbalanced data have to be dealt with in numerous practical applications, like for example text classification, medical diagnosing, financial decision making, discovering network intrusions, and others.

When the available data set has a skewed distribution one faces a data sample overlap, small sample size, and small disjoints. From the point of view of the traditional machine learning process, examples in the majority class will have a greater influence on the induced classifier, causing its classification weight to be in favor of the majority class. Conventional machine learning algorithms which are focused on optimizing the overall classification accuracy, in the class imbalanced situations, tend to achieve poor classification performance, especially for the minority class, which might be of special

© Springer Nature Switzerland AG 2019
N. T. Nguyen et al. (Eds.): ICCCI 2019, LNAI 11683, pp. 601–610, 2019.
https://doi.org/10.1007/978-3-030-28377-3_50

interest to the user [2, 3]. This is due to the fact, that classifiers try, in the first place, to correctly predict the majority class labels. The minority class in such a process might be ignored in favors of the majority one. However, in many real-life problems, the correct minority predictions are crucial, like for example in case of the medical diagnostics or software fault prediction. In the machine learning literature, class imbalance is recognized as one of the most challenging problems in data mining [4].

Research work in the field of imbalanced data classification has resulted in the development of numerous approaches and algorithms. Among them the most successful are the bagging ensemble methods, cost-sensitive methods, approaches based on sampling techniques [6] and methods based on the so-called, data level [7]. The data level methods aim to transform original data into a more balanced dataset, thus reducing the imbalance ratio between the majority and minority classes.

In general, data level techniques transform a dataset with a view to reducing the imbalance ratio between the majority and minority classes [9]. Such transformation takes place at the pre-processing stage and the resulting dataset can be further processed using standard machine learning methods. In the literature, three main types of the data level approaches have been, so far, proposed. These are over-sampling, under-sampling and a hybrid of over and under-sampling [5].

The aim of the under-sampling technique is balancing the distribution of data classes. In practice, under-sampling techniques just remove instances from the majority class. The strength of this approach depends, however, on what kind of rules for instance removal have been implemented [9]. Many methods belonging to this group base on clustering and instance selection. They are referred to as the cluster-based sampling group (see for example [19, 20]). These methods group instances from a majority class into clusters. Next, some instances, representing clusters are selected from each cluster. Main drawbacks of the cluster-based sampling include difficulty to decide on the optimal number of clusters and the lack of rules telling which instances should be selected as cluster representatives [9].

An improved approach for under-sampling based on clustering has been proposed in [12]. The algorithm reduces the number of instances in the majority classes by grouping instances and next selecting only the representative instances. Finally, the majority class set of instances is reduced and the required balance between the minority and the majority class instances is achieved. The clustering algorithm was based on the similarity coefficient calculations, originally proposed in [13]. The instance selection has been carried out using an agent-based population learning implementation. The main feature of the approach is that the number of clusters from which the instances are selected and the process of instance selection are carried out automatically. In this approach, the earlier described drawbacks of the cluster-based sampling reduced or, at least partly, eliminated.

Over-sampling techniques are dealing with expanding the number of minority class instances to balance the classes' distribution. The approach includes techniques providing for the artificial synthesis of the required number of new examples in the minority class. The most simple of the discussed approaches involve simple duplication of the minority class instances. More advanced approaches duplicate or synthesize instances from areas deemed as most promising [10].

Random over-sampling algorithms duplicate minority class instances until the balance between minority and majority class instances will be achieved. Their simplicity makes such an approach practicable. There is, however, a drawback. It is not clear which instances from the minority class should be duplicated. This makes looking for a more effective way of duplicating the minority class instances and, at the same time, avoiding problems with the random duplication, worth a research effort.

SMOTE is the most popular over-sampling method, proposed originally to improve random over-sampling [11]. SMOTE interpolates existing instances to generate new instances. However, the algorithm is not free from weaknesses. One of them is the assumption that all minority class instances are of equal importance. From a practical point of view, it means that each one instance of the minority class can be chosen to over-sample with uniform probability and the duplication may include instances which do not provide any useful information for identification of boundaries between classes.

To eliminate SMOTE's disadvantages numerous extensions of the basic algorithm have been, so far, proposed. These algorithms aim to emphasize certain minority class regions, others intend to reduce the within-class imbalance or attempt to avoid the generation of noise [10]. Example approaches include borderline-SMOTE [14], self-level-SMOTE [18], cluster-SMOTE [16], CURE-SMOTE [15], k-means SMOTE [10] and others (see for example [10] and [17]).

In this paper, a hybrid algorithm for the imbalanced data learning is proposed. The main idea of the proposed algorithm is based on balancing of the minority and majority classes by over-sampling and instance selection. The instance selection is carried out in the majority class. The process is based on clustering using the similarity coefficient as the criterion for grouping instances. The process of clustering is carried-out independently for instances from all classes. Next, the prototypes are selected from the induced clusters. The process of instance selection is integrated with the learning phase executed by the team of agents. In some cases to achieve the balanced distribution of instances between different classes requires an over-sampling. However, the process of over-sampling is run only when the instance selection does not assure the required balance.

The paper is organized as follows. Section 3 contains problem formulation and a detailed description of the proposed method. Section 4 provides details on the computational experiment setup and discusses experiment results. Conclusions and suggestions for future research are included in the final section.

2 An Approach to Imbalanced Learning

In this section, the imbalanced data classification problem is formulated and the details of the proposed approach are discussed.

2.1 Problem Formulation

The aim of learning from data is to output the hypothesis $h \in \mathrm{H}$ optimizing performance criterion F using dataset D, where D is the multiclass data set $D = D_1 \cup D_2 \ldots \cup D_d$ and d is the number of different classes.

In case of the imbalanced training set, $D_{minority}$ is the subset of D which contains the minority class dataset. It is assumed that the cardinality of $D_{minority}$ is definitely smaller than the cardinality of each of the remaining subsets of D representing the remaining classes. Among these remaining subsets, there is the majority class subset containing the majority class instances. Data level methods aim at transforming an imbalanced dataset into a better-balanced one by reducing the imbalance ration between the majority and minority classes. The reduction can be carried out by over-sampling or under-sampling.

The data level approach involves two stages. First, cardinalities of all classes including the minority class are identified. Next, the instance selection process aiming at reducing the cardinality of all datasets representing classes other than the minority one is carried out. Ideally, the reduction process should produce datasets with cardinalities not exceeding cardinality of the dataset representing the minority class. Formally, the number of instances from each subset $\forall_{i\in\{1,\ldots,d\}\setminus\{minority\}}\ D_i$ is reduced and the resulting subsets are denoted as $\forall_{i\in\{1,\ldots,d\}\setminus\{minority\}}\ S_i$, that also means that $\forall_{i\in\{1,\ldots,d\}\setminus\{minority\}}\ S_i \subset \mathrm{D}_i$.

In case when the reduction process cannot guarantee the required balance, i.e. $\exists_i |S_i| > \left|D_{minority}\right|$, then the second stage with the over-sampling process on $D_{minority}$ has to be entered. The process produces dataset obtained by duplication of instances from $D_{minority}$ and denoted as $S_{minority}$. In the ideal case the following holds:

$$\forall_{i\in\{1,\ldots,d\}\setminus\{minority\}} |S_i| \cong \left|S_{minority}\right| \text{ and} \tag{1}$$

$$\forall_{i\in\{1,\ldots,d\}} |S_i| < |D_i| \text{ and} \tag{2}$$

$$\mathrm{U}_{i=1}^{d} |S_i| < |D| \tag{3}$$

In case of the imbalanced data, when the over and under-sampling processes have been carried out, the task of the learner L is to output the hypothesis $h \in H$ optimizing performance criterion F using datasets $S_1, \ldots, S_d$, which are subsets of D containing instances obtained by the over and under-sampling processes, and where the condition (1) is satisfied.

2.2 The Proposed Approach

After the instance selection has been carried out with respect to subsets representing all classes except the minority one, we suggest the following procedure: :

– when the number of considered classes in D is equal to 2 (i.e. $d = 2$), then the over-sampling process is run on the minority class subset,

– when the number of considered classes in D is greater than 2 (i.e. $d > 2$), then at first, the reduced subset of instances containing the maximum number of instances is identified, and on all remaining subsets, the over-sampling procedure is run.

In all cases, under-sampling is based on the instance selection approach where instances are selected from clusters grouping similar instances, that is carried-out under umbrella of the instance selection procedure. The process of clustering is carried-out independently for each of the considered classes without the minority class. It is also assumed that only a single instance, as a reference instance, is selected from each cluster. Thus, the number of clusters produced at the clustering stage has a direct influence on the size of the reduced dataset. Reference instances are selected from the clusters during the learning process executed by the team of agents, as described in a detailed manner in [8], forming the reduced dataset.

The similarity-based clustering algorithm (SCA) produces clusters, where the similarity coefficients are calculated as shown in [8]. The SCA induces clusters with an identical similarity coefficient, and the number of clusters is determined by the value of this coefficient across all instances belonging to the considered class. Clusters are initialized automatically and without any user intervention.

In the paper, the population-based metaheuristics known as the population-learning algorithm (PLA) originally proposed in [21] has been applied for instance selection. The population-learning algorithm is an implementation of the set of agents and different optimization procedures executed by the agents within the asynchronous team of agents (A-Team). These agents cooperate and exchange information. Agents working in the A-Team achieve implicit cooperation by sharing the population of current solutions to the problem to be solved. The A-Team can be also defined as a set of agents and a set of memories, forming a network in which every agent remains in a closed loop [22]. The framework for the agent-based instance selection has been adopted from earlier papers of the authors including [12, 23–25] and [13]. In the [13] the cluster-based instance selection, as a tool for under-sampling has been proposed.

In case when the instance reduction in majority class datasets does not guarantee the required balance, that is $\exists_i |S_i| > |D_{minority}|$, then the over-sampling procedure is activated on all subsets of instances representing classes other than the majority one. The over-sampling procedure starts with identifying for each two closest clusters from the same class their neighbors. The closeness of neighbors is measured using the Euclidean measure. The number of neighbors k is a parameter of the approach and should be set by the user.

The pseudo-code explaining how a new (artificial) instance for the minority class is generated is shown as Algorithm 1. Algorithm 2 shows the pseudo-code of the agent-based population learning algorithm (PLA) where individuals (solutions) represent the selected instances. The pseudo-code covering the proposed method of the imbalanced data classification is shown as Algorithm 3.

Algorithm 1 Generation of an artificial instance (GAI)

Input: x_1, x_2 – reference instances for the minority class; S – a subset of instances; k - number of neighbors;
Output: x_a – an artificial instance;

Begin
For x_1 and x_2 find its k-nearest neighbor instances, which belong to S and where N contains the neighbor instances;
Generate randomly an artificial instance x_a located between instances from N;
Return x_a;
End

Algorithm 2 Agent-based population learning algorithm

Generate initial population P of individuals randomly;
Activate optimizing agents;
While (*stopping criterion is not met*) **do** {*in parallel*}
Read the individual from the common memory;
Execute the instance selection procedure;
Send an improved individual back to the common memory;
Evaluate the fitness of the newly arriving individual and update the common memory;
End while

Algorithm 3 Generation of the balanced instances set

Input: D - training set; k - number of neighbors; d - the number of classes;
Output: $S=S_1 \cup S_2 \ldots \cup S_d$ - sets of balanced instances forming a training set.

Begin
Set *minority* = minority class number.
Set $S_{minority}$ = subset of D contains of instances belong to the minority class;
For i:=1,…,(d-1) **and** $i \neq minority$ **do**
Run the SCA procedure and map instances from D_i into clusters;
Run the PLA on the instances from obtained clusters within D_i and return the reduced subsets of instances S_i contains references instances;
End for
Set *majority* = number of class for which the cardinality of the reduced subset is maximum;
If d=2 **then**
While $\frac{|S_{minority}|}{|S_{majority}|}! \cong 1$ **do**
For each cluster from $S_{majority}$ select its closest cluster and return their reference instances x_1, x_2;
$S_{minority} = S_{minority} \cup \{\mathrm{GAI}(x_1, x_2, k, S_{minority})\}$;
End while
Else
For i:=1,…,(d-1) **and** $i \neq majority$ **do**
While $\frac{|S_i|}{|S_{majority}|}! \cong 1$ **do**
For each cluster from $\bigcup_{j:j\in\{1,..,d\}\setminus\{i\}} S_j$ select their closest clusters and return their reference instances x_1, x_2;
$S_i = S_i \cup \{\mathrm{GAI}(x_1, x_2, k, S_i)\}$;
End while
Return $S_1, \ldots, S_d$;
End

3 Computational Experiment

The proposed approach has been validated experimentally. The main research question was whether the proposed approach performs better than the traditional approach where machine learning algorithms are used for learning from the original imbalanced data.

Classification accuracy of the classifier obtained using the proposed approach, denoted as AOUSID - **A**gent-based **O**ver and **U**nder-**S**ampling for the **I**mbalanced **D**ata, has been compared with the accuracy of:

- *AISAID* – the algorithm originally introduced in [12] assuring balance between minority and majority classes by applying instance selection and a special merging procedure to reduce the cardinality of the majority class instances to the level comparable to the cardinality of the minority class.
- *ALP* - the procedure originally proposed in [25] for data reduction carried-out only in the majority class. The procedure produces clusters of instances in the majority classes using k-means. Next, these clusters are merged to obtain the reduced number of clusters equal to the cardinality of the minority class.
- k-means - in this case, the *k*-means clustering has been implemented using data from the majority class, and next, from thus obtained clusters, prototypes are selected using the agent-based population learning algorithm as in [8].
- C4.5, CART, CNN, 10NN – traditional ML algorithms.

Datasets used in the reported experiment have been obtained from the KEEL dataset repository [25]. Details of these datasets are shown in Table 1. It has been decided to use the 10-cross-validation scheme, and each benchmarking problem has been solved 30 times. The reported values of the quality measure have been averaged over all runs. Classification accuracy has been used as the performance criterion. In the 10-cross-validation scheme, for each fold, the training dataset was reduced using the proposed approach. The learning tool used was the C4.5 algorithm [26]. Details of the parameter settings are shown in Table 2. Values of these parameters included in Table 2 have been set arbitrarily.

Based on the results shown in Table 3, it can be observed that the AOUSID approach assures competitive results in comparison to other algorithms. In several cases, the algorithm performs best including the multi-class imbalanced data sets (wine and balance) and abalone19 and glass2 datasets. One can also observe that the AOUSID outperforms traditional machine learning tools (C4.5, kNN, CART, and CNN - Convolutional Neural Network) when the algorithms have been used on imbalanced datasets.

Table 1. Datasets used in the reported experiment (column IR informs about the ratio of the number of instances of the majority class per instance of the minority class).

Dataset	Number of instances	Number of attributes	Number of classes	IR – the imbalance radio
abalone19	4174	8	2	129.44
shuttle-c0-vs-c4	1829	9	2	13.87
vowel0	988	13	2	9.98

(*continued*)

Table 1. (*continued*)

Dataset	Number of instances	Number of attributes	Number of classes	IR – the imbalance radio
yeast5	1484	8	2	32.73
glass2	214	9	2	11.59
ecoli-0-1-4-6_vs_5	280	6	2	13
glass0	214	9	2	2.06
yeast2	514	8	2	9.08
vehicle2	846	18	2	2.88
wine	178	13	3	1.5
balance	625	4	3	5.88

Table 2. Parameter settings in the reported experiment

Parameter	Value
Number of neighbours k	2
Population size	40
Number of iterations without improvement until the search is stopped	100

Table 3. Results obtained for the AOUSID algorithm and other algorithms on imbalanced datasets and their comparison based on the accuracy (in %)

Dataset	Reduced datasets				Non-reduced datasets			
	AOUSID	AISAID	*k*-means	*ALP*	C4.5	CART	CNN	10NN
abalone19	**82,04**	81,42	74,26	72,45	82,02	–	58,1	48,05
shuttle-c0-vs-c4	97,62	**98,01**	84,25	87,08	97,17	–	84,12	90
vowel0	93,72	91,05	89,21	92,45	94,94	84,67	48,83	**100**
yeast5	88,4	**89,12**	84,45	86,2	87,50	71,45	41,32	79,42
glass2	**71,69**	71,2	54,21	65,45	60,08	43,84	58,24	33,4
ecoli-0-1-4-6_vs_5	80,21	77,13	62,41	77,34	81,36	79,28	82,16	**83,9**
glass0	77,45	**79,24**	72,61	72,14	78,13	74,59	71,61	70,57
yeast2	79,81	68,49	57,82	55,54	62,82	53,96	60,14	**81,63**
vehicle2	94,06	93,67	84,25	82,61	**94,85**	93,51	49,64	88,31
wine	**94,87**	93,04	91,8	92,14	91,42	90,54	89,34	90,71
balance	**73,48**	71,05	68,45	69,33	70,5	67,21	68,06	69,78

4 Conclusions

In the paper a hybrid approach for the imbalanced data learning based on over-sampling and instance selection, is proposed. Both discussed techniques have been integrated and implemented with a view to deal with classifying the imbalanced data by reducing the imbalance ration between minority and majority classes. Over sampling

has been used as a tool for instance duplication in the minority class. Instance selection has been used as a procedure for reducing the number of instances in the majority class. Selection of instances starts with data clustering using the similarity coefficient technique. In the next step, instances are selected from clusters by the team of agents. The proposed approach has been validated experimentally on two and multi-class imbalanced data sets. Based on the results of the computational experiment, one may conclude that the proposed approach can be considered as a promising one with respect to solving the machine learning tasks in case of the imbalanced data.

Future research will focus on studying the influence of different parameters on the performance of the proposed approach as, for example, the number of neighbours in GAI. It is also planned to extend the experiments using additional datasets, as well as to carry out a deeper statistical analysis of the obtained results.

References

1. Chawla, N.V., Japkowicz, N., Drive, P.: Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explor. Newsl. **6**(1), 1–6 (2004)
2. Sun, B., Chen, H., Wang, J., Xie, H.: Evolutionary under-sampling based bagging ensemble method for imbalanced data classification. Front. Comput. Sci. **12**(2), 331–350 (2018)
3. Fernandez, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. Int. J. Approximate Reasoning **50**, 561–577 (2009). https://doi.org/10.1016/j.ijar.2008.11.004
4. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **42**(4), 463–484 (2012)
5. Lin, W.-C., Chih-Fong, T., Hu, Y.-H., Jhang, J.-S.: Clustering-based undersampling in class-imbalanced data. Inf. Sci. **409** (2017). http://doi.org/10.1016/j.ins.2017.05.008
6. Kim, S.-W., Oommen, B.J.: A brief taxonomy and ranking of creative prototype reduction schemes. Pattern Anal. Appl. **6**, 232–244 (2003)
7. Bhanu, B., Peng, J.: Adaptive integration image segmentation and object recognition. IEEE Trans. Syst. Man Cybern. **30**(4), 427–441 (2000)
8. Czarnowski, I., Jędrzejowicz, P.: A new cluster-based instance selection algorithm. In: O'Shea, J., Nguyen, N.T., Crockett, K., Howlett, Robert J., Jain, Lakhmi C. (eds.) KES-AMSTA 2011. LNCS (LNAI), vol. 6682, pp. 436–445. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22000-5_45
9. Tsai, C.-F., Lin, W.-C., Hu, Y.-H., Ya, G.-T.: Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. Inf. Sci. **477**, 47–54 (2019). https://doi.org/10.1016/j.ins.2018.10.029
10. Last, F., Douzas, G., Bacao, F., Oversampling for Imbalanced Learning Based on K-means and SMOTE, p. 19. CoRR abs/1711.00837 (2017)
11. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**(16), 321–357 (2002)
12. Czarnowski, I., Jędrzejowicz, P.: Cluster-based instance selection for the imbalanced data classification. In: Nguyen, N.T., Pimenidis, E., Khan, Z., Trawiński, B. (eds.) ICCCI 2018. LNCS (LNAI), vol. 11056, pp. 191–200. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98446-9_18

13. Czarnowski, I.: Cluster-based instance selection for machine classification. Knowl. Inf. Syst. **30**(1), 113–133 (2012)
14. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. Adv. Intell. Comput. **17**(12), 878–887 (2005)
15. Ma, L., Fan, S.: Cure-smote algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC Bioinf. **18**(1), 169 (2017)
16. Cieslak, D.A., Chawla, N.V., Striegel, A.: Combating imbalance in network intrusion datasets. In: Proceedings of the 2006 IEEE International Conference on Granular Computing, 2006, pp. 732–737. IEEE (2006)
17. Skryjomski, P., Krawczyk, B.: Influence of minority class instance types on SMOTE imbalanced data oversampling. In: Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR, vol. 74, pp. 7–21 (2017)
18. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-Level-SMOTE: safe-level-synthetic minority over-sampling TEchnique for handling the class imbalanced problem. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS (LNAI), vol. 5476, pp. 475–482. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01307-2_43
19. Nejatian, S., Parvin, H., Faraji, E.: Using sub-sampling and ensemble clustering techniques to improve performance of imbalanced classification. Neurocomputing **276**(7), 55–66 (2018)
20. Sowah, R.A., Agebure, M.A., Mills, G.A., Koumadi, K.M., Fiawoo, S.Y.: New cluster undersampling technique for class imbalance learning. Int. J. Mach. Learn. Comput. **6**(3), 205–214 (2016). https://doi.org/10.18178/ijmlc.2016.6.3.599
21. Jędrzejowicz, P.: Social learning algorithm as a tool for solving some difficult scheduling problems. Found. Comput. Decis. Sci. **24**, 51–66 (1999)
22. Talukdar, S., Baerentzen, L., Gove, A., de Souza, P.: Asynchronous teams: co-operation schemes for autonomous, computer-based agents. Technical report EDRC 18-59-96, Carnegie Mellon University, Pittsburgh (1996)
23. Czarnowski, I., Jędrzejowicz, P.: An approach to data reduction and integrated machine classification. New Gener. Comput. **28**(1), 21–40 (2010)
24. Czarnowski, I., Jędrzejowicz, P.: Cluster integration for the cluster-based instance selection. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS (LNAI), vol. 6421, pp. 353–362. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16693-8_37
25. Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., Herrera, F.: KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. J. Multiple Valued Logic Soft Comput. **17**(2–3), 255–287 (2011). Last accessed to the repository 2018/04/10
26. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, SanMateo (1993)