# Cluster-Based Instance Selection for the Imbalanced Data Classification

Ireneusz Czarnowski(✉) and Piotr Jędrzejowicz

Department of Information Systems, Gdynia Maritime University,
Morska 83, 81-225 Gdynia, Poland
{irek, pj}@am.gdynia.pl

**Abstract.** Instance selection, often referred to as data reduction, aims at deciding which instances from the training set should be retained for further use during the learning process. Instance selection is the important preprocessing step for many machine leaning tools, especially when the huge data sets are considered. Class imbalance arises, when the number of examples belonging to one class is much greater than the number of examples belonging to another. The paper proposes a cluster-based instance selection approach for the imbalanced data classification. The proposed approach bases on the similarity coefficient between training data instances, calculated for each considered data class independently. Similar instances are grouped into clusters. Next, the instance selection is carried out. The process of instance selection is controlled and carried-out by the team of agents. The proposed approach is validated experimentally. Advantages and main features of the approach are discussed considering results of the computational experiment.

**Keywords:** Instance selection · Clustering · Imbalanced data · Team of agents

## 1 Introduction

In the current era quantity of data produced by various information systems can be, by some estimates, measured in zetabytes. Since traditional techniques of the analytical processing are not fit to effectively deal with such a massive datasets, their owners more and more often require applying data mining or machine learning techniques enabling discovery and extraction of yet undiscovered knowledge and useful patterns. Responding to such a requirement becomes more and more important for taking effective decisions. Thus, mining information and knowledge from a big data sources is regarded as an important research area.

Main tools for such mining remain machine learning algorithms. Research work in the field of machine learning has resulted in development of numerous approaches and algorithms for classification problems [1]. One of the recent research directions focuses on methods for data reduction. Data reduction performed without losing the extractable information is considered as an important step increasing effectiveness of the learning process when the available data sets are large [2]. Moreover, removing some instances from the training set reduces time and memory complexity of the learning process [3].

N. T. Nguyen et al. (Eds.): ICCCI 2018, LNAI 11056, pp. 191–200, 2018.
https://doi.org/10.1007/978-3-319-98446-9_18

The main goal of the data reduction techniques is to decrease the complexity of computation needed to learn a high quality classifiers. Data reduction can be carried-out by selection of instances, selection of attributes or by simultaneous reduction in both these dimensions [4]. The paper focuses on the data reduction problem through instance selection.

Instance selection aims at identifying and eliminating irrelevant and redundant information, and finding patterns or regularities within the training dataset structure, allowing to induce the so-called prototypes or reference vectors, which can be effectively used in the further steps of the machine learning process. Instance selection is considered especially useful as a mean to increasing effectiveness of the machine learning process when the available datasets are large, since overcoming storage and complexity constraints might become computationally very expensive [5, 6].

Another important research direction focuses on methods for data analysis, when the data collected for analysis are class-imbalanced. Class imbalance arises when the number of data instances in one class is much greater than the number of such instances belonging to another class. This situation is particularly troublesome, when the problem of classification is considered and the classifier training has to be carried-out. In such case some traditional machine learning algorithms focusing on optimizing the overall classification accuracy tend to achieve poor classification performance, especially for the minority class, which might be of a special interest to the user [7, 18].

The problem of the imbalanced data may cause difficulty for many classification algorithms. Classifiers are designed to maximize the classification accuracy and hence, tend to correctly predict the majority class labels. The minority class in such process might be ignored in favor of the majority one. However, in many real problems the correct predictions are crucial and they should focus on the minority class, like for example in case of fraud detection, medical diagnostics, or software fault prediction.

To deal with the problem, many different approaches have been, so far, proposed. Among them one can mention the bagging ensemble methods, cost-sensitive methods or approaches based on the sampling techniques [7] and methods based on the, so-called, data level [8]. The data level methods transform the data into a more balanced datasets, so that the data mining tools can be effectively used. The above approaches aim at reducing the imbalance ration between the majority and minority classes [9].

In this paper a cluster-based instance selection approach for imbalanced data classification is proposed. Main contribution of the paper is proposing and evaluating through computational experiment, the algorithm generating a reduced dataset consisting of relatively equal number of the training dataset instances in each class. The proposed algorithm selects instances from the generated clusters. It is assumed that from each cluster a single prototype is obtained. Clusters are produced using the similarity coefficient as the criterion for grouping instances. The process of clustering is carried-out independently for instances from different classes. Next, the prototypes are selected from the induced clusters. When the number of instances selected in domain of one class is substantially greater than the number of instances belonging to other class, then clusters in a majority class are merged. The cluster merging leads to reduction of the number of cluster and, in the consequence, the number of prototypes and finally to ensure a balanced class distribution in the reduced training dataset. The process of instance selection and cluster merging is integrated with the learning phase executed by the team of agents. Performance of the proposed approach has been evaluated using several benchmark datasets from the KEEL repository [10].

The paper is organized as follows. The instance selection problem for the imbalanced data classification is formulated and shortly discussed in Sect. 2. Section 3 contains a detailed description of the proposed method and algorithms involved. Section 4 provides details on the computational experiment setup and discusses experiment results. Finally, the last section contains conclusions and suggestions for future research.

## 2 Problem Formulation

In general, an instance selection task can be seen as a problem of removing a number of instances from the original data set $D$ and thus producing the reduced training set $S$, with a view to optimize some criterion or criteria. In case of the unsupervised machine learning from the reduced data the task of the learner $L$ is to output the hypothesis $h \in H$ optimizing performance criterion $F$ using dataset $S$ which is a subset of the set $D$, such that $|S| < |D|$ (ideally $S = S_{opt}$), and where $H$ is the hypothesis space., i.e. a set of all possible hypotheses, that the learner can draw.

In case of the imbalanced data set, the instance selection task can be defined as follows: Assume $D$ is the multiclass data set $D = D_1 \cup D_2 \ldots \cup D_d$, where $d$ is the number of different classes, $D_{minority}$ is the subset of $D$, which contains the minority class dataset, and $S_{minority}$ is the subset of $D_{minority}$, which represents the reduced set of instances from the minority class. The task of instance selection is to remove a number of instances from each subset $D_1, \ldots, D_d$ in such way that (ideally):

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

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

$$\bigcup\nolimits_{i=1}^{d} |S_i| < |D| \tag{3}$$

To cope with the assumption that no extractable information is lost it is obvious that data reduction process should be carried-out starting with identification and, eventually reduction of the minority class dataset. Thus, the first step of the data reduction process concerns the minority class dataset. After having performed its reduction or after deciding not to reduce it, datasets representing the remaining classes are considered. We propose to apply the following heuristic rule: cardinality of datasets representing classes other than the minority class dataset, should be reduced to the level not exceeding cardinality of the reduced dataset representing the minority class. So, in case of learning from the imbalanced data, when the data reduction process is 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 $D$, such that $\forall_{1=1,\ldots,d} |S_i|$ is not greater than the number of instances belonging to the minority class.

## 3 An Approach to the Cluster-Based Instance Selection for the Imbalanced Data

This section contains an overview of the proposed approach, including its main features, and gives a detailed description of the proposed cluster-based instance selection algorithm for the imbalanced data. Details of the dedicated agent-based architecture used for the cluster-based instance selection are also presented.

### 3.1 Instance Selection and Data Clustering

The instance selection method proposed in this paper is an extension of the approach introduced in [11]. In this paper the process of data reduction is also carried-out based on clustering. However, in the present paper the process of data reduction is not carried-out with the aim of generating a representative dataset of the required size, but with the goal of ensuring the balanced classes in the reduced training dataset, in case the original dataset contains the imbalanced data.

In the proposed approach the clustering is used to identify groups of a similar instances and the process of clustering is carried-out independently for each considered class. After that the representative instances are selected from the induced clusters. It is also assumed that a single prototype is obtained 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 proposed in [5].

The clusters are generated using the similarity coefficient as the criterion for the instance grouping. The similarity-based clustering algorithm (SCA) produces clusters, where the similarity coefficients are calculated in accordance with the scheme proposed in [5]. Clusters contain instances with identical similarity coefficient and the number of clusters is determined by the value of this coefficient across all instances belonging to the considered class. Thus the clusters are initialized automatically. Under this procedure all the required operations including data transformation, calculation of the similarity coefficient values and vector mapping are carried-out without the user intervention. Next, from thus obtained clusters the prototypes are selected forming, finally, the reduced dataset.

Additional feature of the proposed approach is cluster merging. Since the number of clusters produced from the original training dataset has a direct influence on the number of the reference instances selected for each class, we need some balancing mechanism. In the proposed approach when the number of the reference instances belonging to a class is much greater than the number of instances representing another class than clusters in a majority class are merged under the learning process. The cluster merging has the purpose to reduce the respective number of clusters with a view to ensure the balanced class distribution for the classifier training phase. The process of cluster merging is integrated with the learning process and is executed by the team of agents, as it is in case of the prototype selection.

### 3.2 The Agent-Based Framework for Instance Selection

In general, the instance selection problem is an example of the combinatorial optimization problem belonging to the class of the computationally difficult problems [13]. Additionally, the cluster merging problem belongs to the same class of problems. So far the most effective approach to solve computationally difficult combinatorial optimization problems is to apply some metaheuristics or hybrid algorithms with the metaheuristics components. In the paper the population-based metaheuristics known as the population-learning algorithm and originally proposed in [12] has been implemented. In this implementation the so called solution improvement procedures are executed by the set of agents cooperating and exchanging information within an asynchronous team of agents (A-Team).

Applying the population-based approach with optimization procedures implemented as agents within the asynchronous team of agents (A-Team) for solving data reduction has been proposed in several earlier papers of the authors [5, 14], where the population-based approach has been successfully used for the prototype selection. Agents working in the A-Team achieve an implicit cooperation by sharing the population of solutions, also called individuals, 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 [15]. All agents can work asynchronously and in parallel. Agents cooperate to construct, find and improve solutions which are read from the shared common memory.

In the proposed approach the agent-based population learning is used to perform both task, that is selecting prototypes from clusters and merging them, if needed. Task execution is carried-out independently for each class. Evolutionary steps performed by the optimization agents are executed in parallel.

Basic assumptions behind the proposed Agent-based Instance Selection Approach for the Imbalanced Data classification (*AISAID*) can be summarized as follows:

- During the first phase instances are selected from clusters of instances through the population-based search carried-out by the optimizing agents. Clusters are produced using the procedure based on the similarity coefficient.
- The second phase involves merging of clusters obtained at the initial stage. Clusters are merged using the population-based search and the merging algorithm is carried-out in case the number of clusters obtained at the first phase exceeds the required bound.
- A feasible solution is represented by the two data structures: a string and a binary 3-dimensional matrix (cube) of bits. A string contains numeric labels of instances selected as prototypes from all clusters and classes. The total length of the string is equal to the number of clusters from which reference instances are drawn. This number is bounded by the number of clusters induced from the training dataset composed of instances from the minority class (denoted by $t_{minority}$) multiplied by the number of classes. The length of the discussed string equals $d \cdot t_{minority}$. The binary square matrix denotes whether or not the respective clusters from the respective classes, induced at the cluster initialization stage, need to be merged with a view to comply with the constraint on the number of reference vector allowed.

The binary matrix does not include bits referring to the minority class, since for this particular class merging of clusters is, by assumption, not needed. The submatrix of bits $M_{k:k=1,\ldots,d} = \left[m_{ij}^{k}\right]_{t_k \times t_k}$, where $t_k$ is the initial number of clusters induced for the $k$-th class, denotes whether or not clusters, induced at the cluster initialization stage, have been merged. The element $m_{ij} = 1$, which lies in the $i$-th row and the $j$-th column of the matrix $M_k$, denotes that clusters $i$ and $j$ are merged. In addition each matrix has the following properties:

$$\forall_k \sum\nolimits_{ij} m_{ij}^{k} = t_{minority}, \tag{4}$$

$$\forall_k \forall_i \sum\nolimits_{j} m_{ij}^{k} = 1, \tag{5}$$

$$\forall_k \text{ the element } m_{i,j:i=j}^{k} \text{ of a matrix } M_k \text{ is an artificial “missing value”.} \tag{6}$$

- Initially, for each individual in the population of solutions, the corresponding submatrix $M_k$ conforming with properties (4), (5) and (6) is generated randomly. Likewise, each individual, which contains numbers of instances selected as prototypes is generated through randomly selecting exactly one single instance from each of the considered clusters.
- Each solution from the population is evaluated and the value of its fitness is calculated. The evaluation is carried-out by estimating classification accuracy of the classifier, which is constructed using instances (prototypes) indicated by the solution as the training dataset.

To solve the instance selection problem for imbalanced data classification, the following two groups of optimizing agents dedicated to carrying out different improvement procedures have been implemented:

- The first group includes agents executing procedures responsible for instance selection. These procedures modify a solution by replacing, adding or removing a selected reference instance from the improved solution. Among these procedures we use a local search with the tabu list for instance selection and a simple local search. Both procedures were proposed and described in [16].
- The second group includes agents responsible for merging clusters. In this case the optimizing agents execute a simple local search procedure and modify the current solution changing the composition of clusters, which undergo the merging procedure. The detailed description of the merging procedure can be found in [5].

The cluster-based instance selection algorithm for imbalanced data classification is shown as Algorithm 1.

**Algorithm 1** Agent-Based Instance Selection for the Imbalanced Data

```
Input:
The training set $D$.
Output:
$S=S_1 \cup S_2 ....\cup S_d$ - sets of prototypes - the reduced training set.

Begin
    Set $minority$ = minority class number.
    Run the SCA procedure and map input vectors from $D_{miniority}$ into clusters $D_1^{miniority}, \dots, D_{t_{miniority}}^{miniority}$ and return the outcome.
    Set $t_{miniority}$ = number of cluster obtained from $D_{miniority}$.
    For $k$:=1 to $d$ do
        If (k != $minority$) then
            Run the SCA procedure and map input vectors from $D_k$ into clusters $D_1^k, \dots, D_{t_k}^k$ and return the outcome.
        End If
    End For
    Generate initial population $P$ of individuals randomly, in accordance with properties (4) – (6).
    Activate optimizing agents.
    While (stopping criterion is not met) do {in parallel}
        Read individual from common memory.
        Execute improvement procedures by optimizing agents i.e. execute the merging procedures on clusters and the instance selection procedures.
        Store individual back in the common memory.
        Evaluate the fitness of the newly arriving individual in the common memory.
    End while
    Take the best solution in terms of the fitness from the population of individuals as the final result.
    Return $S_1,...,S_d$.
End
```

## 4 Computational Experiment

To validate the proposed approach it has been decided to carry-out the computational experiment. Experiments aimed at answering the question whether the proposed *AISAID* algorithm performs better than the traditional machine learning algorithms used for learning from the imbalanced data and some selected approaches for imbalanced learning based on instance selection.

Classification accuracy of the classifier obtained using the proposed approach has been compared with the accuracy of:

- *ALP* - the procedure originally proposed in [11] for data reduction. In this paper the data reduction is carried-out only in majority classes. The procedure bases on clustering of instances in majority classes using k-means. Next, the clusters from the majority classes are merged to obtain a reduced number of clusters equal to the cardinality of the minority class. The procedure calculates the distance between two clusters as the average of the Euclidean distances between all the pairs of instances from two clusters. The clusters, where the distance is minimal, are merged, then the prototypes have been selected using the agent-based population learning algorithm as in [11].
- k-means - Results obtained using the set of prototypes produced through selection based on the *k*-means clustering. 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 [5].
- C4.5, CART, CNN, 10NN – traditional algorithms.

To validate the proposed approach several benchmark classification problems have been solved. Datasets for each problem have been obtained from the KEEL dataset repository [10]. Characteristics 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 [17].

**Table 1.** Datasets used in the reported experiment (column IR informs about 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 |
| 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 |

Population size for each implementation of the A-Team, following earlier experiment results presented in [16], has been set to 40. The process of searching for the best solution of each A-Team has been stopped either after 100 iterations or after there has been no improvement of the current best solution for one minute of computation. Values of these parameters have been set arbitrarily.

Based on the above results (Table 2), it can be observed that the proposed approach assures competitive results in comparison to other algorithms. On the average, the proposed AISAID outperforms majority of traditional machine learning tools when dealing with the imbalanced datasets including C4.5, kNN, CART and CNN - Convolutional neural network.

**Table 2.** Results obtained for the AISAID algorithms on imbalanced datasets and their comparison with performance of several different competitive approaches.

| Dataset | Reduced datasets | | | Non-reduced datasets | | | |
|---|---|---|---|---|---|---|---|
| | AISAID | *k*-means | *ALP* | C4.5 | CART | CNN | 10NN |
| abalone19 | 81.42 | 74.26 | 72.45 | **82.02** | – | 58.1 | 48.05 |
| shuttle-c0-vs-c4 | **98.01** | 84.25 | 87.08 | 97.17 | – | 84.12 | 90 |
| vowel0 | 91.05 | 89.21 | 92.45 | 94.94 | 84.67 | 48.83 | **100** |
| yeast5 | **89.12** | 84.45 | 86.2 | 87.50 | 71.45 | 41.32 | 79.42 |
| glass2 | **71.2** | 54.21 | 65.45 | 60.08 | 43.84 | 58.24 | 33.4 |
| ecoli-0-1-4-6_vs_5 | 77.13 | 62.41 | 77.34 | 81.36 | 79.28 | 82.16 | **83.9** |
| glass0 | **79.24** | 72.61 | 72.14 | 78.13 | 74.59 | 71.61 | 70.57 |
| yeast2 | 68.49 | 57.82 | 55.54 | 62.82 | 53.96 | 60.14 | **81.63** |
| vehicle2 | 93.67 | 84.25 | 82.61 | **94.85** | 93.51 | 49.64 | 88.31 |

## 5 Conclusions

In this paper a new cluster-based instance selection method designed for dealing with a classification of the imbalanced data is proposed. The approach is based on using the similarity coefficient calculated for instances from each considered data class independently. Next, based on the value of the coefficient the instances are grouped into clusters from which the prototypes are selected forming finally the reduced dataset. The process of selection of the prototypes is carried-out by the team of agents. The proposed approach is validated experimentally.

Results of the computational experiment allow to conclude that the proposed approach can be considered as a worthy and promising method for solving problems of the machine learning from the imbalanced data. Future research will focus on finding some more refined rules for cluster induction in presence of the imbalanced data. It is also planned to extend the experiments using additional datasets and to carry-out a deeper statistical analysis of the results to obtained a better insight into properties of the proposed approach.

## References

1. Wolper, D.H.: The supervised learning no free lunch theorems. Technical report, NASA Ames Research Center, Moffett Field, California, USA (2001)
2. Kim, S.-W., Oommen, B.J.: A brief taxonomy and ranking of creative prototype reduction schemes. Pattern Analy. Appl. **6**, 232–244 (2003)
3. Wilson, D.R., Martinez, T.R.: Reduction techniques for instance-based learning algorithm. Mach. Learn. **33**(3), 257–286 (2000)
4. Bhanu, B., Peng, J.: Adaptive integration image segmentation and object recognition. IEEE Trans. Syst. Man Cybern. **30**(4), 427–441 (2000)
5. 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
6. Uno, T.: Multi-sorting algorithm for finding pairs of similar short substrings from large-scale string data. Knowl. Inf. Syst. **25**, 229–251 (2009). https://doi.org/10.1007/s10115-009-0271-6
7. 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)
8. 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)
9. Lin, W.-C., Chih-Fong, T., Hu, Y.-H., Jhang, J.-S.: Clustering-based undersampling in class-imbalanced data. Inf. Sci. **409**, 17–26 (2017). https://doi.org/10.1016/j.ins.2017.05.008
10. Alcalá-Fdez, J., et al.: 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). Accessed 10 Apr 2018
11. 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
12. Jędrzejowicz, P.: Social learning algorithm as a tool for solving some difficult scheduling problems. Found. Comput. Decis. Sci. **24**, 51–66 (1999)
13. Hamo, Y., Markovitch, S.: The COMPSET algorithm for subset selection. In: Proceedings of the Nineteenth International Joint Conference for Artificial Intelligence, Edinburgh, Scotland, pp. 728–733 (2005)
14. Czarnowski, I.: Distributed learning with data reduction. In: Nguyen, N.T. (ed.) Transactions on Computational Collective Intelligence IV. LNCS (LNAI), vol. 6660, pp. 3–121. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21884-2_1
15. 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)
16. Czarnowski, I., Jędrzejowicz, P.: An approach to data reduction and integrated machine classification. New Gener. Comput. **28**(1), 21–40 (2010)
17. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, SanMateo (1993)
18. 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