PreCLAS: An Evolutionary Tool for Unsupervised Feature Selection

Jessica A. Carballido, Ignacio Ponzoni(✉), and Rocío L. Cecchini

Institute for Computer Science and Engineering (UNS - CONICET),
Department of Computer Science and Engineering, Universidad Nacional del Sur,
Bahía Blanca, Argentina
{jac,ip,rlc}@cs.uns.edu.ar
https://icic.conicet.gov.ar/

Abstract. Several research areas are being faced with data matrices that are not suitable to be managed with traditional clustering, regression, or classification strategies. For example, biological so-called omic problems present models with thousands or millions of rows and less than a hundred columns. This matrix structure hinders the successful progress of traditional data analysis methods and thus needs some means for reducing the number of rows. This article presents an unsupervised approach called PreCLAS for preprocessing matrices with dimension problems to obtain data that are apt for clustering and classification strategies. The PreCLAS was implemented as an unsupervised strategy that aims at finding a submatrix with a drastically reduced number of rows, preferring those rows that together present some group structure. Experimentation was carried out in two stages. First, to assess its functionality, a benchmark dataset was studied in a clustering context. Then, a microarray dataset with genomic information was analyzed, and the PreCLAS was used to select informative genes in the context of classification strategies. Experimentation showed that the new method performs successfully at drastically reducing the number of rows of a matrix, smartly performing unsupervised feature selection for both classification and clustering problems.

Keywords: Clustering tendency · Classification strategies · Evolutionary algorithm · Unsupervised feature selection · Microarray data analysis

1 Introduccion

A well-known and extensively studied subject in machine learning, statistics and information theory is the dimensionality problem: matrices with a structure "$n \ll N$" where N (number of rows) is much larger than n (number of columns).

This work is supported by CONICET (Grant number 112-2017-0100829) and Secretaría de Ciencia y Tecnología (UNS) (Grant number 24/N042).

E. A. de la Cal et al. (Eds.): HAIS 2020, LNAI 12344, pp. 172–182, 2020.
https://doi.org/10.1007/978-3-030-61705-9_15

In this work, the objective is to present a novel unsupervised manner for selecting a manageable amount of rows suitable for classification and clustering methods.

A vital issue rarely discussed in data mining studies is the fact that clustering techniques always find clusters, whether or not real groups of coherent values genuinely exist in the structure of data under analysis. Then, clustering methods can contradict the proverb "from where there is nothing, you cannot get something". If the k-means algorithm is given a randomly generated matrix to find three clusters, the method yields three clusters, even though no real groups exist. Therefore, it should be required to perform some previous study to the structure of the matrix, to establish whether it is coherent to look for clusters in it. This study is called "clustering tendency". The Hopkins statistic [15,17] constitutes an appropriate measure for evaluating the clustering tendency of a matrix. In this context arose the main inspiration for the design of the method: an unsupervised instance reduction method that diminishes the number of rows using as a measure of choice the idea of keeping those rows that show some submatrix structure suitable for grouping techniques, whenever this submatrix truly exists. Obtaining a submatrix with this structure is also beneficial for classification methods, as it will be shown later. To the best of our knowledge, no strategy performs unsupervised feature selection based on this criterion.

The method, called PreCLAS, is implemented as an evolutionary algorithm that funds the selection of features based on clustering tendency studies. In this context, the Hopkins statistic is used to analyze whether the data are uniformly distributed. The measure is simple and intuitive. It is based on the difference between the distance from a real point to its nearest neighbor (d_r) and the distance from a randomly artificial generated point to its nearest neighbor (d_a). In this work, the implementation of the Hopkins statistic corresponds to the one provided by the `<clustertend> R package`. The idea is the following: when data contains no group structure, the distance among real points and their real neighbor points is approximately the same as the distance from uniformly distributed random artificial points to their nearest real neighbor points. On the other hand, if data contains some cluster structure, that distance increases.

The ultimate goal of this research is to apply the new method to bioinformatics problems of microarray data analysis. Microarray experiments obtain expression data from thousands of genes for a few samples, presenting this data as a matrix that exhibits the "$n \ll N$" form. Microarray data analysis includes statistics, supervised and unsupervised techniques, generally categorized as class discovery, class prediction, or class comparison. In this context, our method arises for selecting some informative genes in a matrix **without using information about the samples**. This is carried out because many times, microarray experiments present information about SOME of the samples, but not about all of them. Then it is important to count with a tool that analyses the matrix in an unsupervised manner. The final aim is to demonstrate whether those genes are representative of the classes that are defined in the dataset to constitute a useful input for classification algorithms. It is well known that a classification

algorithm will perform accordingly to the quality of the examples that are used to train it.

It is also important to remark that a low number of samples makes it very difficult to create predictive models. First, some machine learning algorithms, such as linear discriminant analysis (LDA), cannot be applied if the number of observations is less than the number of predictors. Secondly, even though all the genes can be incorporated in the model (SVM), if many of them do not contribute more than noise to the model, this diminishes the predictive capacity when applied to new observations (overfitting). Moreover, most genes in microarray experiments are not of interest. Less than 10% manifest the hidden phenotypes, and therefore, are immersed in large amounts of noise. The uncertainty about which genes are relevant hinders the process of selecting informative genes.

On the whole, the idea is the following: differentially expressed genes between samples naturally generate a structure of groups in the matrix. Then, if the algorithm can select a reduced group of differentially expressed genes from matrices with dimension problems, it constitutes an unsupervised feature selection method that can be used before classification and clustering techniques. The importance of this unsupervised step lies in the fact that many times the microarray datasets do not contain class information for all the samples that are in general, very few. As far as we know, no widely accepted method performs feature selection for classification and clustering in an unsupervised manner.

2 Related Work

Data preprocessing is one of the most critical stages in data mining and knowledge discovering processes. Nowadays, data is massively being produced and stored to be studied and analyzed. As it is well known, this massive production of data also carries substantial amounts of redundant, noisy, faulty, and irrelevant information. This scenario has yielded the development of different preprocessing steps that can imply tasks such as data filtering, outlier detection, feature extraction, feature selection, and instance reduction. These kinds of functions have extensively been studied in machine learning, statistics, and information theory. In particular, the so-called feature extraction, feature selection, and instance selection problems are aimed at reducing the dataset size to facilitate the inference process of the machine learning method to be applied. The feature selection problem has been approached in different contexts by several authors [10,19]. Most basic techniques assume that the dataset at hand can be represented for a percentage of the total data, and select this portion in a heuristic way or by some systematic way [16], but numerous methods that are more sophisticated have been proposed. However, many of those methods are formulated and evaluated in the data classification scenario [1,22], or the regression scenario [2,11] but not for clustering, and even less in the case of the matrices $n \ll N$. On the other hand, several works are mainly focused on normalization, noise reduction, or faulted data correction [6,14]. Alternatively, different authors have designed other methods concentrating primarily on the selection

of a few features that may be considered as the representative of the whole dataset [18,20]. Unlike these works, the PreCLAS method is directly intended to reduce the number of rows by finding the submatrix with the best group structure, i.e., not necessarily discarding objects that are similarly representative of the dataset structure, but trying to keep all the subsets of features that can be grouped in clearly defined clusters, whenever they exist.

Regarding the use of Evolutionary Algorithms on feature selection problems, there are also several works to mention. In [7], the authors reviewed some of the most relevant methods, and, since then, other evolutionary inspired methods have been proposed [4,25]. As in the case of the works mentioned in the previous section, many of the evolutionary techniques are thought and evaluated in the classification scenario [4,13,25] and aiming at selecting some few representative instances [13,24], or in the regression scenario [2]. As it can be seen from the available literature, the feature selection problem has been approached from diverse angles and with different purposes, but, to the best of our knowledge, they are not directly comparable with the one proposed in the present article, mainly because it works without information about the classes.

3 The Method: PreCLAS

The method presented in this article was implemented as a Genetic Algorithm (GA). GAs are metaheuristic adaptive methods used to solve NP search and optimization problems. They are based on the genetic process of living organisms. Across generations, populations evolve in nature following the principles of natural selection and the survival of the fittest, postulated by Darwin [5]. Simulating that process, GAs can create near-optimal solutions for many real-world problems. The individuals of the population represent solutions to the addressed problem. Basic principles of GAs were posed by Holland [12], and are well described in several texts, such as Goldberg [9]. In this context, PreCLAS is a GA implemented in R that receives a matrix of real numbers and an optional parameter for the number of rows of the resulting submatrix. Details on the algorithm are given below.

Individuals are vectors of 50 integer numbers that vary from 1 to the number or rows of the original matrix; they are values corresponding to feasible row indices. In this manner, each individual is a list of indices of fixed length that indicates which rows should be kept. As individuals represent different reductions of that matrix, the evaluation consists of calculating the Hopkins statistics of each submatrix. An initial population of 50 individuals is created with a substantial restriction: they must overcome a clustering tendency threshold of quality, which means that very inconvenient individuals (according to their fitness) are not allowed in the initial population. The creation of each individual is repeated until an acceptable value of the statistical threshold is obtained. Then, the fittest individuals of the population are kept by a binary tournament procedure, where two individuals are randomly selected, and the one with the best fitness value passes to the next generation.

The combination of genetic information is performed using ad-hoc designed set operations. Parents are selected, and the indices of rows contained in each of them are combined into a "super father". An intersection is also performed to see how many and which indices appear repeated between parents. Afterward, the first son is constructed selecting some indices from the "super father" and completing the number of indices with the ones that were found repeated in the intersection. The second son contains the remaining indices, also filled with the repeated ones. The crossover probability is 0.7. Finally, the mutation operation ends the process of obtaining the new descendant. In this implementation, individuals selected for mutation are randomly replaced. In this way, the operator introduces the right level of randomness for exploration purposes. The mutation probability is 0.3. A maximum number of 200 generations is established. In addition, the algorithm informs whether a submatrix with a reasonable clustering tendency could be obtained. Parameters of population size, probability of mutation and crossing, and the number of generations were empirically established.

4 Results and Discussion

4.1 Simple Performance Assessment

The functioning of this prototype was verified, analyzing a benchmark dataset used for conglomerate studies. The study case Ruspini [21] consists of 75 observations on two variables, x, and y. As can be seen in Fig. 1 (left), the separation into four clusters is visually recognizable. Some noise was added to the matrix, enlarging it to double its size with random values and maintaining the cluster structure (see Fig. 1). This new matrix was called RuspBIG. In this context, the hypothesis was about PreCLAS being able to find the submatrix from RuspBIG that exhibits a good clustering tendency.

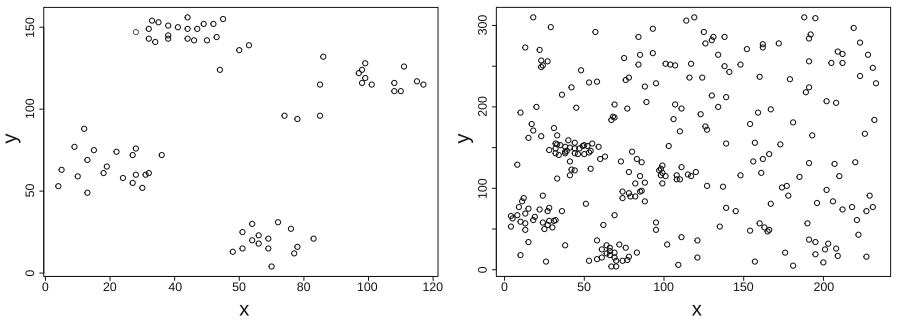

Fig. 1. Benchmark data distribution.

Starting from RuspBIG, two algorithms were used to reduce it: The PreCLAS and a RANDOM search algorithm. Both algorithms reduced the matrix

to 70 rows. One hundred independent runs were performed for each of them. As a first measure, we obtained the mean Hopkins values from the 100 resulting sub-matrices (the best of each run). The results were: mean Hopkins value of the 100 PreCLAS sub-matrices = 0.2912351 and for RANDOM sub-matrices = 0.4611668.

As it was expected, the sub-matrices found by PreCLAS exhibited a better Hopkins value since it is closer to 0, and the confidence intervals showed a statistically significant difference between the mean values, as they do not overlap each other. This result was also expected because this measure precisely guides the evolution of the genetic algorithm. For visual aims, we randomly selected one submatrix yielded by each algorithm to see how the visualization method VAT [3] represents each of them. Figure 2 shows that a clustering tendency is more evident in the matrix obtained from PreCLAS. The next stage consisted of applying the K-means algorithm, with $k = 4$ to each reduced matrix. Final values found by each of the methods from all the trials were: PreCLAS: Hopkins mean = 0.29 (CI [0.28, 0.3]) and Silhouette mean = 0.51 (CI [0.5, 0.51]); RANDOM: Hopkins mean = 0.46 (CI [0.45, 0.47]) and Silhouette mean = 0.4 (CI [0.44, 0.45]).

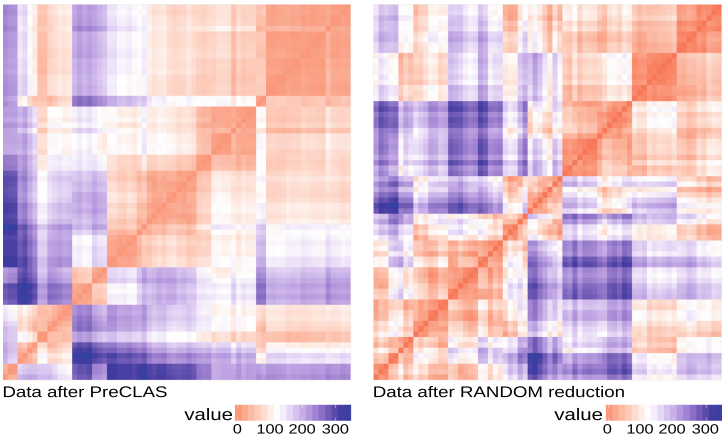

Fig. 2. VAT for data after PreCLAS reduction (left-hand side) and VAT for data after random modification (right-hand side).

A better performance was achieved by the clustering algorithm when the PreCLAS yielded the submatrix. When selecting one result after clustering from each sub-matrix obtained by PreCLAS and RANDOM, it can be graphically shown that the clusters found after PreCLAS (Fig. 3, left-hand side) look more coherent than those found after the RANDOM reduction (Fig. 3, right-hand side).

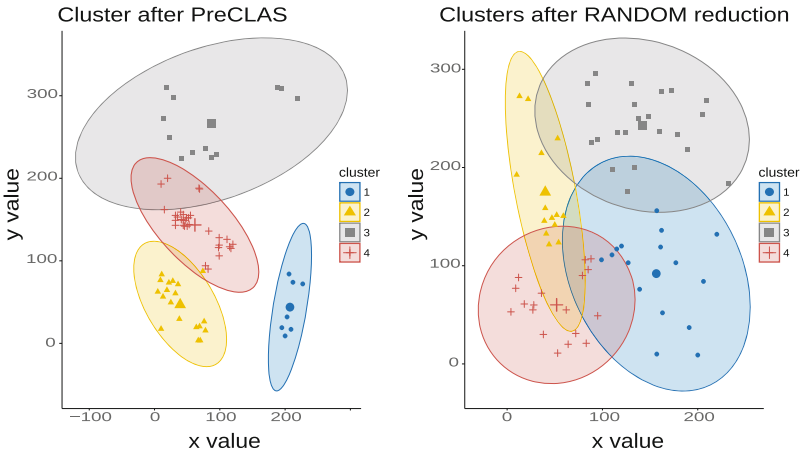

Fig. 3. Clusters after PreCLAS reduction (left-hand side) and after RANDOM reduction (right-hand side)

4.2 A Bioinformatics Example

The dataset used for this stage of the experimentation is the GSE43346 [23]. This dataset was obtained from GEO [8], a public functional genomics data repository. It belongs to a study of human small cell lung cancer (SCLC), with gene expression data represented in a matrix of 54675 probes and 68 samples, of which 23 are clinical SCLC, 42 are normal tissue samples, and 3 are SCLC cell lines. First, this original matrix was reduced to eliminate rows with no variance. With this initial standard reduction method, the number of rows decreased from 54675 to 27284 (called from now Original *Reduced* Matrix). This is one of the most common mathematical (unsupervised) manners of reducing the number of rows. However, note that this ranking is a very rudimentary way to select a subset of genes.

Moreover, this amount of genes is still not practicable for most classifiers. Hypothesis: Is a submatrix yielded by the PreCLAS a good alternative for classification purposes? The analysis was performed as follows:

1. The 50 genes (rows) with the highest variance are selected from the original reduced matrix. Then, nine genes are randomly selected from this matrix and form a new 9 by 68 matrix (from now on, this matrix will be called matrix A).
2. On the other hand, the PreCLAS was executed over the original reduced matrix to select 50 rows. Then, nine genes are also randomly selected from this other matrix and form a new 9 by 68 matrix (from now on, this matrix will be called matrix B).
3. The differential expressions of matrices A and B are calculated and showed in Figs. 4 and 5.

As can be seen in Figs. 4 and 5, the genes selected by PreCLAS (Fig. 5) are **more representative of the classes, even though the method is unsupervised**.

Two classification methods trained and tested with matrix B worked well: a purely statistical method LDA (linear discriminant analysis), and machine learning SVM (support vector machine) method.

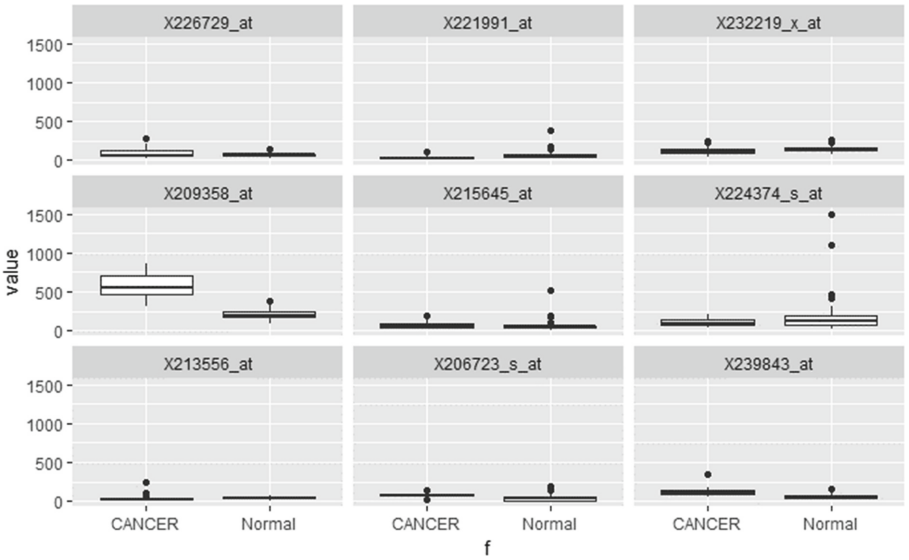

Fig. 4. Differential expression of 9 randomly selected genes from matrix A (50 rows with the highest variance).

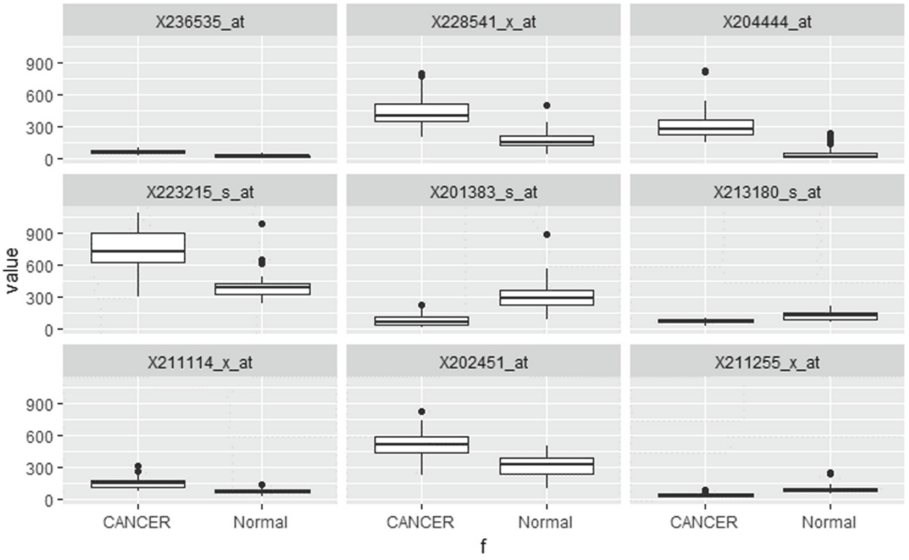

Fig. 5. Differential expression of 9 randomly selected genes from matrix B (PreCLAS).

5 Conclusions

In unsupervised classification, more precisely clustering, interactions between objects are intensely affected by matrices with dimensionality problems, since a vast majority of the features are likely to be uninformative, but will however contribute to the computed similarity metrics. In supervised classification, the effect might be even worse: a program is trained to identify classes based on unauthentic differences found in any combination of the input variables. It is thus essential, for both supervised and unsupervised classification, to perform some feature selection before applying any data mining strategy.

In this paper, we present the PreCLAS, a method that aims at reducing the number of rows of matrices with dimension problems tailing clustering and classification purposes. In other words, a submatrix is searched so that, if there exists a structure in the original matrix that presents a good clustering tendency, the PreCLAS tends to find it. With this aim, the PreCLAS was implemented as a genetic algorithm where the fitness function maximizes the Hopkins statistic, and evolutionary operators, as well as the selection process, were designed in an ad-hoc manner. Given a matrix with a considerable number of rows, PreCLAS returns a submatrix with a reasonable clustering tendency. For testing purposes, a first study case was constructed, enlarging with noise a benchmark clustering dataset. It could be observed that clustering results were better for matrices reduced by the PreCLAS, compared with clustering results of matrices reduced by a random algorithm. To the best of our knowledge, there is no other unsupervised method that reduces a matrix to find a submatrix presenting coherent groups. Finally, it was revealed that the reduction performed by PreCLAS is also useful as a filter method for classification strategies. This feature was assessed with a real-world study case of lung cancer, where the classifiers trained with the resulting reduced submatrix exhibited excellent results. Further studies and discussion remain, but this constitutes a promising achievement that will be approached in-depth to implement a pipeline. The ultimate goal is to use PreCLAS as a filtering step in the context of a complete platform for microarray data analysis.

References

1. Alvar, A.S., Abadeh, M.S.: Efficient instance selection algorithm for classification based on fuzzy frequent patterns. In: 2016 IEEE 17th International Symposium on Computational Intelligence and Informatics (CINTI), pp. 000319–000324 (2016)
2. Antonelli, M., Ducange, P., Marcelloni, F.: Genetic training instance selection in multiobjective evolutionary fuzzy systems: a coevolutionary approach. Trans. Fuzzy Sys. **20**(2), 276–290 (2012)
3. Bezdek, J.C., Hathaway, R.J.: VAT: a tool for visual assessment of (cluster) tendency. In: Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN 2002 (Cat. No. 02CH37290), vol. 3, pp. 2225–2230 (2002)
4. Chen, Z.-Y., Tsai, C.-F., Eberle, W., Lin, W.-C., Ke, S.-W.: Instance selection by genetic-based biological algorithm. Soft. Comput. **19**(5), 1269–1282 (2014). https://doi.org/10.1007/s00500-014-1339-0

5. Darwin, C.: On the Origin of Species by Means of Natural Selection. Murray, London (1859)
6. Delany, S.J., Segata, N., Mac Namee, B.: Profiling instances in noise reduction. Knowl.-Based Syst. **31**, 28–40 (2012)
7. Derrac, J., García, S., Herrera, F.: A survey on evolutionary instance selection and generation. Int. J. Appl. Metaheuristic Comput. **1**(1), 60–92 (2010)
8. Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**(1), 207–210 (2002)
9. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning, 1st edn. Addison-Wesley Longman Publishing Co. Inc., Reading (1989)
10. Grochowski, M., Jankowski, N.: Comparison of instance selection algorithms II. Results and comments. In: Rutkowski, L., Siekmann, J.H., Tadeusiewicz, R., Zadeh, L.A. (eds.) ICAISC 2004. LNCS (LNAI), vol. 3070, pp. 580–585. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24844-6_87
11. Guillen, A., Herrera, L.J., Rubio, G., Pomares, H., Lendasse, A., Rojas, I.: New method for instance or prototype selection using mutual information in time series prediction. Neurocomputing **73**(10–12), 2030–2038 (2010)
12. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1975). 2nd edn, 1992
13. Ishibuchi, H., Nakashima, T., Nii, M.: Learning of neural networks with GA-based instance selection. In: Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569), vol. 4, pp. 2102–2107, August 2001
14. Jamjoom, M., El Hindi, K.: Partial instance reduction for noise elimination. Pattern Recogn. Lett. **74**(C), 30–37 (2016)
15. Kassambara, A.: Practical Guide To Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, Factoextra, vol. 2. STHDA (2017)
16. Kuri-Morales, A., Rodríguez, F.: A search space reduction methodology for large databases: a case study. In: Perner, P. (ed.) ICDM 2007. LNCS (LNAI), vol. 4597, pp. 199–213. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73435-2_16
17. Lawson, R.G., Jurs, P.C.: New index for clustering tendency and its application to chemical problems. J. Chem. Inf. Comput. Sci. **30**(1), 36–41 (1990)
18. Mirisaee, S.H., Douzal, A., Termier, A.: Selecting representative instances from datasets. In: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10 (2015)
19. Olvera-López, J.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Kittler, J.: A review of instance selection methods. Artif. Intell. Rev. **34**(2), 133–143 (2010). https://doi.org/10.1007/s10462-010-9165-y10.1007/s10462-010-9165-y
20. Olvera-López, J.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: Object selection based on clustering and border objects. In: Kurzynski, M., Puchala, E., Wozniak, M., Zolnierek, A. (eds.) Computer Recognition Systems. AINSC, vol. 45, pp. 27–34. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75175-5_4
21. Ruspini, E.H.: Numerical methods for fuzzy clustering. Inf. Sci. **2**(3), 319–350 (1970)
22. Samuels, E.: Fantasies of Identification: Disability, Gender, Race. NYU Press, New York (2014)

23. Sato, T., et al.: PRC2 overexpression and PRC2-target gene repression relating to poorer prognosis in small cell lung cancer. Sci. Rep. **3** (2013). Article number: 1911
24. Triguero, I., García, S., Herrera, F.: Differential evolution for optimizing the positioning of prototypes in nearest neighbor classification. Pattern Recogn. **44**(4), 901–916 (2011)
25. Tsai, C.F., Eberle, W., Chu, C.Y.: Genetic algorithms in feature and instance selection. Know.-Based Syst. **39**, 240–247 (2013)