Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 175 (2020) 108–113

Procedia Computer Science

www.elsevier.com/locate/procedia

The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC)
August 9-12, 2020, Leuven, Belgium

Unbalanced data processing using oversampling: Machine Learning

Amelec Viloria [a]*, Omar Bonerge Pineda Lezama[b], Nohora Mercado-Caruzo[c]

[a,b] *Universidad de la Costa, Barranquilla, Colombia.*
[c] *Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras*

Abstract

Nowadays, the DL algorithms show good results when used in the solution of different problems which present similar characteristics as the great amount of data and high dimensionality. However, one of the main challenges that currently arises is the classification of high dimensionality databases, with very few samples and high-class imbalance. Biomedical databases of gene expression microarrays present the characteristics mentioned above, presenting problems of class imbalance, with few samples and high dimensionality. The problem of class imbalance arises when the set of samples belonging to one class is much larger than the set of samples of the other class or classes. This problem has been identified as one of the main challenges of the algorithms applied in the context of Big Data. The objective of this research is the study of genetic expression databases, using conventional methods of sub and oversampling for the balance of classes such as RUS, ROS and SMOTE. The databases were modified by applying an increase in their imbalance and in another case generating artificial noise.

Keywords: Imbalance of classes; Microarray databases, Genetic expression; Deep Learning techniques.

1. Introduction

Recently, the use of Artificial Neural Networks (ANN) has become popular to carry out classification tasks focused on real problems. One of the most used networks is the Multi-Layer Perceptron (MLP) trained with the Back-Propagation method [1]. It is one of the most popular networks due to the advantages it presents, such as:

* Corresponding author. Tel.: +57-3046238313.
E-mail address: aviloria7@cuc.edu.co

1877-0509
10.1016/j.procs.2020.07.018

speed, inherent parallelism, it does not require a priori knowledge of the statistical distribution of the data and tolerance to failures [2].

ANNs are traditionally made up of three layers (input layer, hidden layer and output layer), however, nowadays when an ANN is made up of more than three layers, it is known as an ANN Deep Learning (DL) [3]. The most representative ANN-DL architecture is the MLP with several hidden layers [4]. The main advantages of this type of architecture are three: high performance, robustness to overtraining and high processing capacity.

Conventional classifiers such as the MLP were designed to work with balanced databases, i.e. the number of samples is the same for each class. For this reason, when working with an imbalanced database, optimal results are not achieved due to the high percentage of misclassified samples in the underrepresented or minority classes [5].

In addition, the Back-Propagation method (used to train the MLP) is also affected by the imbalance, since it slows down the convergence of the network [6], which is one of the disadvantages of this training method.

Class imbalance has been extensively studied in two-class problems [7], however, the problems of multiple classes [8] and DL have been little addressed. Work focused on multi-class imbalance commonly uses costs associated with the different classes at the training stage, but this approach is only adequate in training with Back-Propagation when training is in batch mode [9]. In real problem solving, batch training is less used than stochastic training, since the latter is usually faster, achieves better solutions, and can be used to identify changes by passing samples through the MLP.

Traditionally, the methods used to attack class imbalance are based on duplicating or eliminating samples until an equilibrium is reached in the number of samples per class, for example, Random Over-Sampling (ROS) and Random Under-Sampling (RUS) [10]. One of the commonly used methods is the Synthetic Minority Over-sampling Technique (SMOTE), which was proposed by [11], and generates new synthetic samples interpolated into the minority class samples. This method has served as a basis for other sampling methods such as Borderline - SMOTE, Adaptive Synthetic Sampling (ADASYN), SMOTE editing nearest neighbor, among others [12].

On the other hand, in under-sampling techniques, RUS has been reported as one of the most effective techniques [13]. In these techniques, some methods are characterized by including a heuristic mechanism in their operation, which aims to eliminate or change the labels of the samples, whether they are noisy, atypical or redundant [14]. As an example, Neighborhood Clearing Rules and One-sided Selection methods can be mentioned.

Currently, interest has emerged in developing dynamic sampling methods in the context of MLPs, where the objective is to use the appropriate number of samples or increase the size of the minority class when training the perception. An example is the SNOBALL method [15], where the majority class is gradually increased. Another example is D&S [25], which attenuates the imbalance by means of an over-sampling mechanism and identification of the hard-to-learn samples, i.e. it pays more attention to the more difficult-to-classify samples.

In general, class imbalance negatively affects the performance of machine learning (ML) algorithms. The imbalance is also present in the context of Big Data where the use of Boltzman machines, belief networks, convolutional networks and, in general, networks with a DL approach have shown good results [16], while the machine learning algorithms have shown remarkable deficiencies in their performance.

In the context of Big Data there are few proposals to attack the problem of imbalance, currently the automatic learning algorithms are being adapted to perform in this approach. The aim of this work is to identify the behavior of the classifier using databases of gene expression microarrays which have few samples, high dimensionality and, in some cases, class imbalance.

2. Gene expression microarrays

Gene expression microarrays are real-world expression profile data sets used in various types of cancer research. The databases for this study will be Prostate, Ovarian and Breast, which have very few samples, are highly dimensional and are grouped into two classes. It should be mentioned that the number of elements per class is not balanced. These can be obtained from the Kent Ridge Biomedical Data Set Repository (http://leo.ugr.es/elvira/DBCRepository/) [15].

Gene expression microarray databases generally have a limited number of samples, have a high dimensionality and, in some cases, present class imbalances.

3. Multilayer Perception Deep Learning (MLP DL)

An MLP consists of a network containing an input layer, an output layer and one or more hidden layers [17]. In the input layer, the data to be analyzed are entered, passed to the first hidden layer. The results of this layer are passed to the second hidden layer (if any), and finally passed to the output layer. Synaptic weights are assigned to each of the hidden layers. The number of nodes in each hidden layer may vary [18].

The main difference between machine learning and deep learning architectures is the number of hidden layers, conventionally in machine learning architectures are composed of one input layer, one hidden layer and one output layer, unlike deep learning where they are composed of more than 3 layers, because they have more than one hidden layer. When the network has more than 3 layers in its architecture it is classified as deep learning [19].

Many of the advances in deep learning depend largely on the technology used to implement them. Some of the most used libraries for this approach are Theano, PyLearn2, Caffe, Tensorflow and the Apache Spark working platform.

In this study, a neural network with DL focus is applied with a very similar configuration for all the databases, where the only difference is the input layer because the databases do not have the same number of attributes. To run the network, an environment that could perform parallel processing was required, so the Spark platform was used, since it offers good performance when processing large and highly dimensional data, optimizing the execution time and generating reliable results.

4. Experimental design

The purpose of this research is to study the behavior of the MLP DL classifier using high dimensional, low pattern, and high-class imbalance databases, such as gene expression microarray databases. For this purpose, public databases of cancer data available in the Kent Ridge Biomedical Data Set repository were used. Details of the databases can be found in Table 1.

Table 1 shows that in addition to the above-mentioned databases, two more databases are described for each one, which can be identified because the name of the database includes the terms "-Noise" and "-Decreasing" [20].

Table 1. Description of the databases.

Database	Features	No. of Examples	Class 1	Class 0
Ovarian	15214	260	170	92
Ovarian-Noise	15214	260	178	78
Ovarian- Decreased	15214	260	170	86
Prostate	12611	148	78	60
Prostate -Noise	12611	148	89	50
Prostate - Decreased	12611	138	80	50
Breast	25410	100	50	52
Breast -Noise	25410	100	40	63
Breast - Decreased	25410	90	42	54

These databases were generated to obtain an even more significant class imbalance. For the databases identified as "Decreased", 10 samples were randomly subtracted, achieving that the minority class decreased its size and thus had a more relevant imbalance. In the bases identified as "Noise", a way to generate Artificial Noise required, which was achieved by randomly selecting ten samples from the minority class to change its class, thus decreasing the minority class and increasing the majority class.

The configuration used in the neural network is the one established by default in the Spark MLP. However, some parameters were adjusted, such as the input layer with 13520 nodes for the Prostate database, 15547 for Ovarian and

25841 for Breast. Two hidden layers (the first one with 100 nodes and the second one with 90) and finally, an output layer of two nodes. The configuration used in the hidden layers and the output layer was exactly the same for all the databases.

The Hold-Out method [21] was applied for the segmentation of the training and test sets, leaving 60 and 40 percent respectively. The division process was repeated 10 times at random, where each set contained different samples, that is, the samples contained in the training set were not in the test set and vice versa.

The area under the curve (AUC), which is a widely used measure in class imbalance research, was used to evaluate the effectiveness of the model. For the execution, each of the databases were processed ten times, finally obtaining the average of the metrics.

5. Results

Table 2 shows the results obtained by the DLM using AUC to measure the effectiveness of the classifier. Results are shown for each of the sampling techniques used in all of the databases, with values in bold indicating the sampling technique (under or over sampling) with the best performance in each of the databases.

Table 2. Results obtained by the DLM using the AUC metric

Database	ORIG	ROS	RUS	SMOTE
Ovarian	0.8184	0.9541	0.9414	0.9282
Ovarian-Noise	0.7409	0.9167	0.872	0.8861
Ovarian- Decreased	0.8369	0.9651	0.9278	0.9396
Prostate	0.6897	0.8363	0.817	0.8456
Prostate -Noise	0.5671	0.7975	0.7533	0.7764
Prostate - Decreased	0.7083	0.872	0.7772	0.8283
Breast	0.6014	0.6358	0.5402	0.619
Breast -Noise	0.4401	0.644	0.6494	0.648
Breast - Decreased	0.4341	0.6873	0.5591	0.6496

To ensure reliable results, the average score of 10 runs in each of the experiments is presented, figures with four decimal places were used to give further information on the results.

The Ovarian, Prostate and Breast databases are the original databases, and in general, they have a good classification, however, "Breast" presents an accuracy of 0.6014 which is a very low level for efficiency. It should be noted that it is the base with the lowest number of samples and the highest number of attributes, in addition to presenting the smallest imbalance with a difference of five samples between its classes [22].

By applying sampling techniques to the above-mentioned databases, similar results were obtained to those of the original databases when RUS was used. On the other hand, applying ROS and SMOTE oversampling techniques showed a remarkable improvement in the efficiency of the classifier, being ROS the one that obtained the highest efficiency in Ovarian, while SMOTE was the best in Prostate and Breast. Despite this, there was no significant difference in the efficiency of the classifier when using ROS and SMOTE.

In the Ovarian-Decreased, Prostate-Decreased and Breast-Decreased databases, where the class imbalance was increased by subtracting ten samples from the whole minority class, it was observed that they increased their level of effectiveness compared to the original databases, RUS again had the lowest level, even below the original databases. It should be noted that the Breast-Decreased database had a considerably higher level of efficacy than the Breast database.

For Ovarian-Noise, Prostate-Noise and Breast-Noise, which are the databases to which man-made noise was applied, it was found that all sampling methods reached a better level than the original database. The results of RUS are very close to those of the original database. On the other hand, SMOTE obtained the highest efficiency in

Prostate-Noise and Breast-Noise while ROS in Ovarian-Noise; despite this, there was no significant difference in the efficiency of the classifier when using ROS and SMOTE.

It can be noted that using over-sampling techniques, better results are obtained compared to the original databases. In the case of the random sub-sampling (RUS) it is observed that the performance of the classifier is very similar to that of the original database even, in some cases slightly less, therefore, using traditional oversampling techniques, such as ROS and SMOTE, the classifier can obtain better results when working with databases of high dimensionality, few samples and high class imbalance.

6. Conclusions

Classroom imbalance has been recognized as one of the main challenges when training supervised classifiers, because most of them were designed to work with relatively balanced databases. Currently, many of the databases that are being generated present problems of class imbalance, for example, the microarray databases of genetic expression, coupled with this, characteristics such as high dimensionality and scarcity of samples or training patterns characterize these databases.

Deep learning has been an excellent alternative for dealing with large and dimensional databases, however, it has shown notable shortcomings when working with imbalanced databases. This research studied the effectiveness of traditional methods to deal with class balance, in microarray databases of genetic expression which are characterized by having few samples or training patterns and an excessive number of attributes or characteristics.

Results presented in this study show the effectiveness of ROS and SMOTE sampling techniques to treat class imbalance in the classification of gene expression microarray databases; however, a tendency in SMOTE to produce better results is observed. On the other hand, it is shown that it is prohibitive to eliminate samples in this type of database in order to treat class imbalance.

There is no doubt that the topic needs to be studied in depth not only because of its importance but also because of its relationship with other areas of knowledge such as biomedicine and Big Data. For future research, it is suggested to study other classical algorithms for the treatment of class imbalance and in due course to propose a new method that will help to overcome the deficiencies of existing methods in the state of the art.

References

[1] Bolón-Canedo, V., Alonso-Betanzos, A., López-de-Ullibarri, I., & Cao, R. (2019). Challenges and Future Trends for Microarray Analysis. In Microarray Bioinformatics (pp. 283-293). Humana, New York, NY.

[2] Sayed, S., Nassef, M., Badr, A., & Farag, I. (2019). A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets. Expert Systems with Applications, 121, 233-243.

[3] Pal, M.: Extreme learning machine for land cover classification. International Journal of Remote Sensing, 30(14), pp. 3835–3841 (2008)

[4] Guillen, P., & Ebalunode, J. (2016, December). Cancer classification based on microarray gene expression data using deep learning. In 2016 International Conference on Computational Science and Computational Intelligence (CSCI) (pp. 1403-1405). IEEE.

[5] Nene, S.: Deep learning for natural languaje processing. International Research Journal of Engineering Technology, 4, pp. 930–933 (2017)

[6] Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P. E., Verrill, C., ... & Lundin, J. (2018). Deep learning-based tissue analysis predicts outcome in colorectal cancer. Scientific reports, 8(1), 1-11.

[7] Reyes-Nava, A., Sánchez, J. S., Alejo, R., Flores-Fuentes, A. A., & Rendón-Lara, E. (2018, June). Performance analysis of deep neural networks for classification of gene-expression microarrays. In Mexican Conference on Pattern Recognition (pp. 105-115). Springer, Cham.

[8] Viloria, A., & Lezama, O. B. P. (2019). Improvements for determining the number of clusters in k-means for innovation databases in SMEs. In Procedia Computer Science (Vol. 151, pp. 1201–1206). Elsevier B.V. https://doi.org/10.1016/j.procs.2019.04.172.

[9] Flores-Fuentes, A. A., & Granda-Gutiérrez, E. E. (2019, March). Using Deep Learning to Classify Class Imbalanced Gene-Expression Microarrays Datasets. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 23rd Iberoamerican Congress, CIARP 2018, Madrid, Spain, November 19-22, 2018, Proceedings (Vol. 11401, p. 46). Springer.

[10] Ding, L., & McDonald, D. J. (2017). Predicting phenotypes from microarrays using amplified, initially marginal, eigenvector regression. Bioinformatics, 33(14), i350-i358.

[11] Zeebaree, D. Q., Haron, H., & Abdulazeez, A. M. (2018, October). Gene selection and classification of microarray data using convolutional neural network. In 2018 International Conference on Advanced Science and Engineering (ICOASE) (pp. 145-150). IEEE.

[12] Panda, M. (2017). Elephant search optimization combined with deep neural network for microarray data analysis. Journal of King Saud University-Computer and Information Sciences.

[13] Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., ... & Claassen, M. (2018). Automated Gleason grading of prostate cancer tissue microarrays via deep learning. Scientific reports, 8(1), 1-11.

[14] Liu, S., Mocanu, D. C., Matavalam, A. R. R., Pei, Y., & Pechenizkiy, M. (2019). Sparse evolutionary Deep Learning with over one million artificial neurons on commodity hardware. arXiv preprint arXiv:1901.09181.

[15] Shahane, R., Ismail, M., & Prabhu, C. S. R. (2019). A Survey on Deep Learning Techniques for Prognosis and Diagnosis of Cancer from Microarray Gene Expression Data. Journal of Computational and Theoretical Nanoscience, 16(12), 5078-5088.

[16] Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., ... & Litjens, G. (2020). Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology.

[17] Salman, H. K.: Cost-Sensitive Learning of Deep Feature Representations from Imbalanced Data. IEEE Transactions on Neural Networks and Learning Systems (2017)

[18] Nguyen, A.B., Phung, S.L.: A supervised learning approach for imbalanced data sets. In: Proc. of the 19th International Conference on Pattern Recognition, pp. 1–4 (2008)

[19] Shekar, B. H., & Dagnew, G. (2020). L1-Regulated Feature Selection and Classification of Microarray Cancer Data Using Deep Learning. In Proceedings of 3rd International Conference on Computer Vision and Image Processing (pp. 227-242). Springer, Singapore.

[20] Basavegowda, H. S., & Dagnew, G. (2020). Deep learning approach for microarray cancer data classification. CAAI Transactions on Intelligence Technology, 5(1), 22-33.

[21] Khaire, U. M., & Dhanalakshmi, R. (2020). High-dimensional microarray dataset classification using an improved adam optimizer (iAdam). Journal of Ambient Intelligence and Humanized Computing, 1-18.

[22] Viloria, A., Varela, N., Lezama, O. B. P., Llinás, N. O., Flores, Y., Palma, H. H., … Marín-González, F. (2020). Classification of Digitized Documents Applying Neural Networks. In Lecture Notes in Electrical Engineering (Vol. 637, pp. 213–220). Springer. https://doi.org/10.1007/978-981-15-2612-1_20