Using Deep Learning to Classify Class Imbalanced Gene-Expression Microarrays Datasets

A. Reyes-Nava[1(✉)], H. Cruz-Reyes[2], R. Alejo[2], E. Rendón-Lara[2], A. A. Flores-Fuentes[1], and E. E. Granda-Gutiérrez[1]

[1] UAEM University Center at Atlacomulco, Universidad Autónoma del Estado de México, Carretera Toluca-Atlacomulco km. 60, 50450 Atlacomulco, Mexico
adriananava0@gmail.com

[2] Division of Postgraduate Studies and Research, National Institute of Technology of Mexico (TecNM), Campus Toluca, Av. Tecnológico s./n., 52149 Metepec, Mexico

Abstract. Performance of deep learning neural networks to classify class imbalanced gene-expression microarrays datasets is studied in this work. The low number of samples and high dimensionality of this type of datasets represent a challenging situation. Three sampling methods which have shown favorable results to deal with the class imbalance problem were used, namely: Random Over-Sampling (ROS), Random Under-Sampling (RUS) and Synthetic Minority Oversampling Technique (SMOTE). Moreover, artificial noise and greater class imbalance were included in the datasets in order to analyze these situations in the context of classification of gene-expression microarrays datasets. Results show that the noise or separability of the dataset is more determinant than its dimensionality in the classifier performance.

Keywords: Gene-expression microarrays · Deep neural networks · Class-imbalance

1 Introduction

Recently, the use of deep learning to solve a variety of real-life problems has attracted the interest of many researchers because these algorithms usually allow to obtain better results than traditional machine learning methods [14,21]. Multi-Layer Perceptron (MLP), the most common neural network, has been also translated to deep learning context [18]: Deep Learning MLP (DL-MLP) incorporates two or more hidden layers [13], thus increasing the computational cost of processing large size and high dimension datasets [12]. Nevertheless, when efficient frameworks are available, such as Apache-Spark [24] or Tensor-Flow, the advantage of the high performance, robustness to overfitting, and high processing capability of this deep neural networks could be taken.

R. Vera-Rodriguez et al. (Eds.): CIARP 2018, LNCS 11401, pp. 46–54, 2019.
https://doi.org/10.1007/978-3-030-13469-3_6

In the field of biomedical databases classification, the use of deep learning is gaining attention [18]; for instance, in the classification of gene-expression microarrays [6,12,17]. Typical applications with deep neural networks refer to problems in which both: dimensionality and number of samples are high [5,20]. However, in gene-expression microarrays databases, the number of samples is low, and the dimensionality is high, which represent a challenging situation. In some cases, the classes are imbalanced, where one class is highly underrepresented compared to the other class [3].

Class imbalance problem has been a hot topic in machine learning and data-mining and more recently, in deep learning [9]. Usual techniques employed to handle the class imbalance problem have been the random sampling methods (under-sampling - RUS or over-sampling - ROS), mainly due to the independence of the underlying classifier [19]. ROS randomly replicates samples in the minority-class, while RUS randomly eliminates samples from the majority-class, biasing the discrimination process to compensate the imbalance of classes. Synthetic Minority Oversampling Technique (SMOTE) is also a helpful sampling technique which generates artificial samples from the minority class by interpolating existing instances that lie close together [11]; it has motivated the development of other over-samplings methods [16,19]. RUS is one of the most successful under-sampling methods, however, this method loses effectiveness when removes significant samples [19]. Other important under-sampling methods include a heuristic mechanism [16].

Usually in the state-of-the-art, machine learning methods have been used to classify gene-expression microarrays databases [6,10,17], but recent works have been focused in the application of deep learning [1,5,7,15,20,23]. In this scenario, common methods used to face the class imbalance have been ROS, RUS and SMOTE; however, results are not conclusive in different works. For example, in [4] SMOTE is better than others methods, while in [8] RUS presents better results. In this paper, performance of DL-MLP in the classification of gene-expression microarrays databases in presence of class imbalance problem is evaluated. In order to focus this study in the situation of class imbalance and noisy data, artificial scenarios were also included.

2 Deep Learning MLP

MLP is the most conventional neural network architecture, which is commonly based on three layers: input, output, and one hidden layer [18]. MLP can be translated into deep neural networks by incorporating more than two hidden layers, becoming a Deep Learning-MLP (DL-MLP); this allows to reduce the number of nodes per layer and uses fewer parameters, but in return this leads to a more complex optimization problem [12,13]. If an efficient framework, such as Apache-Spark [24] or Tensor-Flow, the advantage of high performance, robustness to overfitting, and high processing capability of DL-MLP could be taken.

Traditionally, a MLP has been trained with the back-propagation algorithm (based in the stochastic gradient descent) and its weights are randomly initialized. However, in the last versions of the DL-MLP, hidden layers are pre-trained

by an unsupervised algorithm and weights are optimized by the back-propagation algorithm [18]. Alternatively, MLP uses sigmoid activation functions, such as the hyperbolic tangent or logistic function, but DL-MLP includes (commonly) the rectified linear unit (ReLU) $f(z) = \max(0, z)$, because typically it learns much faster in networks with many layers, allowing training a DL-MLP without supervised pre-training.

There are three variants of the descending gradient that differ in how many data are used to process the gradient of the objective function: (a) Batch Gradient Descendent, calculates the gradient of the cost function to the parameters for the entire training data set, (b) Stochastic Gradient Descendent, performs an update of parameters for each training example and (c) Mini-batch Gradient Descendent, takes the best of the two previous types and performs an update for each mini-batch of n training examples [22].

The most common algorithms of descending gradient optimization are: (a) Adagrad, it adapts the learning reason for the parameters, making bigger updates for less frequent parameters and smaller for the most frequent, (b) Adadelta, is an extension of Adagrad that seeks to reduce aggressiveness, monotonously decreasing the learning rate, instead of accumulating all the previous descending gradients, restricting accumulation to a fixed size and (c) Adam, calculates adaptations of the learning rate for each parameter and stores an exponentially decreasing average of past gradients. Other relevant algorithms are AdaMax, Nadam and RMSprop [22].

3 Related Works

Nowadays, the treatment of bioinformatic databases is increasingly common, including tasks for disease prediction, treatments and sick classification. Most efforts have focused on feature selection: for example, [6] presents a technique to classify microarrays of genetic expression by selection of characteristics, comparing 7 classifiers with 4 databases, obtaining the best results with HAM (proposed method) and Support Vector Machines (SVM). A similar case is presented by [10], where a cancer database is evaluated by using 6 classifiers, of which the MLP shows the best results with 98% accuracy. Likewise, in [17], several methods of selection and extraction of characteristics are presented to reduce the dimensionality of the microarray databases; this selection is made through filters, wrappers, and integrated techniques.

In literature, machine learning methods have been applied to treat biomedical information; however, the presented databases have high dimensionality. For example, in [20], a series of images of breast cancer is classified into 3 different groups; the method uses Restricted Boltzmann Machines (RBM) in a Deep Neural Network (DNN) that allows classifying faster and more accurately. In [5], multiple Recurrent Neural Networks (RNN) are used to make the classification of people with benign and malignant breast cancer; the proposed model consists of four RNN to extract the characteristics and one RNN to make the final classification. Reference [1] presented an automatic diagnosis system to detect

breast cancer based on a deep belief network for the training phase followed by a back-propagation neuronal network. Reference [15] implemented deep bidirectional recurrent neuronal networks of short-term memory for the reduction of intrinsic protein disorder.

In works such as [7,23], a comparison is made to classify and identify relevant genes and perform cancer detection; the data used are the most important characteristics extracted through different methods, including the use of autoencoders. When the results of both works are analyzed, better classification results are obtained when deep learning methods are used.

Currently, there are several works dealing with the problem of imbalance of classes in bioinformatic databases, which contain biological information such as nucleotide sequence data, protein structures, genomes, genetic expression, metabolism and other similar data. Re-sampling techniques are applied to balance the classes in the set of samples; for example, in the work presented by [4], SMOTE is used to deal with the imbalance of databases of high dimensionality, presented in three databases with real and simulated data, where it is shown that in the case of the low dimensionality database, the accuracy of the classifier increases when applying SMOTE, however, in cases of high dimensionality, it is not advisable.

In problems of imbalance of classes in databases of gene expression microarrays, studies have been carried out by applying under and over sampling techniques. In [8], RUS, ROS and SMOTE were applied to treat the imbalance in bioinformatic databases through classifiers 5-NN and SVM; results show that RUS obtains the best results, in comparison with the work of [4] where they mention that SMOTE is significantly better to RUS when the k-NN classifier is applied.

4 Experimental Set-Up

In order to study the class imbalance problem on the classification of gene-expression microarrays datasets using deep neural networks, four microarray cancer data sets were used (see Table 1). Datasets were obtained from the Kent Ridge Biomedical DataSet Repository (http://datam.i2r.a-star.edu.sg/datasets/krbd). Original datasets were also modified to highlight the class imbalance (+HI) and to include artificial noise (+Noise). The modification consisted in random elimination (+HI) or change the label (+Noise) at ten samples from the minorities classes. Table 1 presents the main features of the new produced benchmarking datasets. For the experimental design, the hold out method was adopted (10 times), with 60% of the samples for training and 40% for testing.

In this work it is used a DL-MLP with two hidden layer, sigmoid (logistic) function in their nodes, and softmax function on the output layer nodes. The framework Apache-Spark [24] was used. The configuration of each hidden layer was 10 and 20 nodes respectively. Back-propagation was used for learning the model, and the logistic loss function for optimization. Then, Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) was employed as an optimization routine.

Table 1. Description of the benchmarking datasets. The imbalance ratio (IR), which corresponds to the ratio of the majority class size to the minority class size is reported in the last column

Database	Features	Samples	Class 1	Class 2	IR
Ovarian	15154	253	162 Cancer	Normal 91	1.78
Ovarian+HI+Noise	15154	253	172 Cancer	Normal 81	2.12
Ovarian+HI	15154	243	162 Cancer	Normal 81	2
Colon	2000	62	22 Positive	Negative 40	1.82
Colon+HI+Noise	2000	62	12 Positive	Negative 50	4.16
Colon+HI	2000	52	12 Positive	Negative 40	3.33
Prostate	12600	136	77 Tumor	Normal 59	1.31
Prostate+HI+Noise	12600	136	87 Tumor	Normal 49	1.78
Prostate+HI	12600	126	77 Tumor	Normal 49	1.57
CNS	7129	60	21 Class1	Class0 39	1.86
CNS+HI+Noise	7129	60	11 Class1	Class0 49	4.45
CNS+HI	7129	50	11 Class1	Class0 39	3.54

In order to deal with the class imbalance problem, ROS, RUS and SMOTE were used to sample the training dataset to reach a relative class distribution balance on class imbalanced gene-expression microarrays datasets.

Area Under the Receiver Operating Characteristic Curve (AUROC) was used as measure criteria for the classifiers performance; it is one of most widely-used and accepted technique for the evaluation of binary classifiers in imbalanced domain [3].

Finally, a non-parametric statistical tests [19] of Friedman and Iman–Davenport was applied in order to detect whether differences in the results exist. If the null-hypothesis was rejected, the Holm-Shaffer post-hoc test was used to find the particular pairwise comparisons that produce statistical significant differences. These test were applied with a level of confidence $\gamma = 0.05$, by using the KEEL software [19].

5 Results and Discussion

In this section, main experimental results of this research are presented. Table 2 exhibits, in term of AUROC and rank values, the experimental results obtained to classify class imbalanced gene-expression microarrays datasets with DL-MLP. It is noted that when the class imbalance is increased and noise is included in the dataset (HI+Noise), the classification performance is harmed; however, increasing the class imbalance does not necessarily affect the performance of DL-MLP (see Prostate and Ovarian datasets). The worst performance of DL-MLP was obtained for original CNS and Colon datasets, in accordance to AUROC values (0.539 and 0.721 respectively), and the performance of the classifier is reduced

when it includes noise (HI+Noise) or more imbalance (HI). It is assumed that these dataset are less separable than Prostate and Ovarian; thus, noise and imbalance do affect the classifier performance. These results agree with the reported in others works [2,3], which show that the class imbalance becomes a problem when the dataset is overlapped or less separable. Also, when SMOTE is applied to CNS, the performance of DL-MLP trained with this dataset exhibits the worst performance compared to the other methods; it occurs because SMOTE generates artificial samples, which could increase the overlapping or noise in the dataset. Table 2 also shows that Ovarian datasets produces good results for all methods: even RUS does not harm the performance of the classifier (remember RUS deletes about 50% of the size dataset), possibly because it is a highly separable dataset.

Table 2 does not show a direct relationship between the dimensionality dataset and its percentage of success. For example, Ovarian is the dataset with highest dimensionality (compared with other datasets used in this work) and the classification performance of the DL-MLP was close to 1. Thus, AUROC values obtained for CNS (which has about 50% less features than Ovarian) are close to 0.5, i.e., the prediction is not much better than just flipping a coin.

In general terms, it is noticed in Table 2 that RUS is not a good option to deal with the class imbalance problem: although it does not considerably deteriorate the classification performance, the tendency is that RUS presents results worse than other methods, even by using the original dataset. In contrast, it is noticeable than ROS and SMOTE improve the DL-MLP performance when class imbalanced gene-expression microarrays datasets are used.

Table 2. Classification performance of the DL-MLP on class imbalanced gene-expression microarrays datasets using the AUROC and average ranks (AR).

	ROS	SMOTE	ORIGINAL	RUS
Colon	0.865(1)	0.814(2)	0.721(3)	0.694(4)
Colon+HI+Noise	0.873(1)	0.866(2)	0.547(4)	0.563(3)
Colon+HI	0.883(2)	0.898(1)	0.667(3)	0.657(4)
CNS	0.642(2)	0.684(1)	0.539(3)	0.506(4)
CNS+HI+Noise	0.870(1)	0.754(2)	0.512(3)	0.462(4)
CNS+HI	0.781(1)	0.751(2)	0.510(4)	0.548(3)
Prostate	0.861(2)	0.896(1)	0.835(4)	0.839(3)
Prostate+HI+Noise	0.823(2)	0.826(1)	0.712(4)	0.774(3)
Prostate+HI	0.897(1)	0.878(2)	0.853(3)	0.798(4)
Ovarian	0.979(1)	0.978(2)	0.963(3)	0.962(4)
Ovarian+HI+Noise	0.942(1)	0.936(2)	0.886(4)	0.893(3)
Ovarian+HI	0.990(1.5)	0.990(1.5)	0.982(4)	0.949(3)
Average rank	1.375	1.625	3.416	3.583

Nevertheless, according to the results shown in this section, it is highly recommended to make a statistical analysis (see Sect. 4). Friedman and Iman−Davenport non-parametrical statistical tests report that considering a performance reduction distributed according to chi-square with 3 degrees of freedom, the Friedman statistic is set to 29.125, and p value computed by Friedman test is 2.108×10^{-6}. Considering a performance reduction distributed according to F-distribution with 3 and 33 degrees of freedom, Iman−Davenport statistic is 46.6, and p value computed by this test is 5.804×10^{-6}. Thus, the null hypothesis is rejected because both: Friedman's and Iman-Davenport's tests, indicate that significant differences exist. Upon these results, it is concluded that a post-hoc statistical analysis is required. Holm and Shaffer statistics values were obtained, as well as p-values (Table 3). Holm and Shaffer procedures rejects those hypotheses with an unadjusted p-value $\leq \{Holm$ and $Shaffer\}$ values, respectively.

In the Table 3, the Holm-Shaffer test demonstrates non-significant statistical differences between ROS and SMOTE; however, there is statistical differences between both methods and Original (i.e., the dataset without any preprocessing) and RUS. In addition, it shows that does not exist significant statistical differences between RUS and Original. These results confirm that the worst method to deal with the class imbalance problem is RUS, while ROS is very competitive with SMOTE; in other words, does not exist evidence of the predominance of any of these two methods.

Table 3. The accepted null hypothesis are typed in bold (p-values for $\alpha = 0.05$).

i	Algorithms	$z = (R_0 - R_i)/SE$	p	Holm	Shaffer
6	ROS vs. RUS	4.190018	0.000028	0.008333	0.008333
5	Original vs. ROS	3.873790	0.000107	0.010000	0.016667
4	RUS vs. SMOTE	3.715676	0.000203	0.012500	0.016667
3	Original vs. SMOTE	3.399448	0.000675	0.016667	0.016667
2	**ROS vs. SMOTE**	**0.474342**	**0.635256**	**0.025000**	**0.025000**
1	**Original vs. RUS**	**0.316228**	**0.75183**	**0.050000**	**0.050000**

6 Conclusion

The classification performance of the deep learning MLP to classify class imbalanced gene-expression microarrays datasets was analyzed. In accordance to the results, does not exist evidence of the predominance of ROS over SMOTE or vise-versa, but the tendency is that RUS presents worse results than other methods, even when the original dataset is used. On the other hand, does not exists evidence (based on the datasets used in this work), of the relationship between its dimensionality and the classifier effectiveness on class imbalance scenarios. However, results show that the noise or separability of the dataset is more determinant than its dimensionality in the classifier performance.

References

1. Abdel-Zaher, A.M., Eldeib, A.M.: Breast cancer classification using deep belief networks. Expert Syst. Appl. **46**, 139–144 (2016)
2. Alejo, R., Monroy-de Jesús, J., Ambriz-Polo, J.C., Pacheco-Sánchez, J.H.: An improved dynamic sampling back-propagation algorithm based on mean square error to face the multi-class imbalance problem. Neural Comput. Appl. **28**(10), 2843–2857 (2017). https://doi.org/10.1007/s00521-017-2938-3
3. Alejo, R., Monroy-de Jesús, J., Pacheco-Sánchez, J., López-González, E., Antonio-Velázquez, J.: A selective dynamic sampling back-propagation approach for handling the two-class imbalance problem. Appl. Sci. **6**(7), 200 (2016). https://doi.org/10.3390/app6070200
4. Blagus, R., Lusa, L.: Smote for high-dimensional class-imbalanced data. BMC Bioinform. **14**(1), 106 (2013). https://doi.org/10.1186/1471-2105-14-106
5. Chen, D., Qian, G., Shi, C., Pan, Q.: Breast cancer malignancy prediction using incremental combination of multiple recurrent neural network. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.S. (eds.) ICONIP 2017. LNCS, vol. 10635, pp. 43–52. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70096-0_5
6. Cleofas-Sánchez, L., Sánchez, J.S., García, V.: Gene selection and disease prediction from gene expression data using a two-stage hetero-associative memory. Progress Artif. Intell. (2018). https://doi.org/10.1007/s13748-018-0148-6
7. Danaee, P., Reza, G., Hendrix, D.A.: A deep learning approach for cancer detection and relevant gene identification. In: Pacific Symposium on Biocomputing, Honolulu, pp. 219–229 (2016)
8. Dittman, D., Khoshgoftaar, T., Wald, R., Napolitano, A.: Comparison of data sampling approaches for imbalanced bioinformatics data. In: Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014, pp. 268–271 (2014)
9. Dong, Q., Gong, S., Zhu, X.: Imbalanced deep learning by minority class incremental rectification. CoRR abs/1804.10851 (2018)
10. Dwivedi, A.K.: Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Comput. Appl. **29**(12), 1545–1554 (2018). https://doi.org/10.1007/s00521-016-2701-1
11. Fernandez, A., Garcia, S., Herrera, F., Chawla, N.V.: SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. Res. **61**, 863–905 (2018)
12. Geman, O., Chiuchisan, I., Covasa, M., Doloc, C., Milici, M.-R., Milici, L.-D.: Deep learning tools for human microbiome Big Data. In: Balas, V.E., Jain, L.C., Balas, M.M. (eds.) SOFA 2016. AISC, vol. 633, pp. 265–275. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-62521-8_21
13. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
14. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S.: Deep learning for visual understanding: a review. Neurocomputing **187**, 27–48 (2016)
15. Hanson, J., Yang, Y., Paliwal, K., Zhou, Y.: Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. Bioinformatics **33**, 685–692 (2016)
16. He, H., Garcia, E.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009). https://doi.org/10.1109/TKDE.2008.239

17. Hira, Z., Gillies, D.F.: A review of feature selection and feature extraction methods applied on microarray data. Adv. Bioinform. **2015**, 1–13 (2015)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
19. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf. Sci. **250**, 113–141 (2013). https://doi.org/10.1016/j.ins.2013.07.007
20. Maqlin, P., Thamburaj, R., Mammen, J.J., Manipadam, M.T.: Automated nuclear pleomorphism scoring in breast cancer histopathology images using deep neural networks. In: Prasath, R., Vuppala, A.K., Kathirvalavakumar, T. (eds.) MIKE 2015. LNCS (LNAI), vol. 9468, pp. 269–276. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26832-3_26
21. Reyes-Nava, A., Sánchez, J.S., Alejo, R., Flores-Fuentes, A.A., Rendón-Lara, E.: Performance analysis of deep neural networks for classification of gene-expression microarrays. In: Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Olvera-López, J.A., Sarkar, S. (eds.) MCPR 2018. LNCS, vol. 10880, pp. 105–115. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92198-3_11
22. Ruder, S.: An overview of gradient descent optimization algorithms. CoRR abs/1609.04747 (2016)
23. Salaken, S.M., Khosravi, A., Khatami, A., Nahavandi, S., Hosen, M.A.: Lung cancer classification using deep learned features on low population dataset. In: IEEE 30th Canadian Conference on Electrical and Computer Engineering, Windsor, pp. 1–5 (2017)
24. Zaharia, M., et al.: Apache spark: a unified engine for Big Data processing. Commun. ACM **59**(11), 56–65 (2016). https://doi.org/10.1145/2934664