IRAQI Academic Scientific Journals

University of Kirkuk Journal For Administrative and Economic Science (2026) 16 (1): 222-228

UKJAES

University of Kirkuk Journal For Administrative and Economic Science

ISSN:2222-2995 E-ISSN:3079-3521
University of Kirkuk Journal For Administrative and Economic Science

Mulla Guhdar Abdulaziz Ahmed, Abdullah Ahmed Hassan & Demir Yıldırım. Classification of High Dimensional Imbalanced Data Using PCA, SMOTE Methods and Classification Algorithms. ***University of Kirkuk Journal For Administrative and Economic Science*** **(2026) 16 (1):222-228.**

Classification of High Dimensional Imbalanced Data Using PCA, SMOTE Methods and Classification Algorithms

Guhdar Abdulaziz Ahmed Mulla [1], Ahmed Hassan Abdullah [2], Yıldırım Demir [3]

[1,2] *University of Duhok-Faculty of administration and economic/Department of Statistics and informatic, Duhok, Iraq*
[3]*Van Yuzuncu Yil University-Faculty of Economics and Administrative Sciences/Department of Statistics, Van, Turkey*

guhdar.abdulaziz@gmail.com [1]
Ahmed.h.abdullah@uod.ac [2]
Ydemir.yyu@gmail.com [3]

Abstract: Data mining is a technique for extracting possible trends and correlations from data obtained from various sources in order to uncover secret information. For data visualization, high-dimensionality statistics and dimensionality reduction methods are often used. If the classification groups are not roughly evenly represented, the dataset is imbalanced. In this study, we used SMOTE method to rebalanced data set with Principal Component Analysis (PCA) is proposed to solve the high dimensional data. Four classification algorithms are Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Stochastic Gradient Descent (SGD) and Random Forest Analysis (RFA). In this study we used Cancer dataset were used to check the efficiency of the proposed method and select the result of the classifiers. Respectively, raw datasets, converted datasets by PCA, SMOTE methods, were analyzed with the given algorithms. Analyzes were made using WEKA.

Keywords: Classification, Imbalance, Principal component analysis.

تصنيف البيانات عالية الأبعاد غير المتوازنة باستخدام طرائق PCA وSMOTE وخوارزميات التصنيف

م.م. كهدار عبد العزيز احمد[١]، م.م. احمد حسن عبدالله[٢]، أ.م.د. يلدرم دمير[٣]

[١،٢] *جامعة دهوك-كلية الإدارة والاقتصاد/قسم الإحصاء والمعلوماتية، دهوك، العراق*
[٣] *جامعة فان يوزنجيل يل/كلية العلوم-قسم الإحصاء، فان، تركيا*

guhdar.abdulaziz@gmail.com [1]
Ahmed.h.abdullah@uod.ac [2]
Ydemir.yyu@gmail.com [3]

المستخلص: استخراج البيانات هو تقنية تهدف إلى استخراج الاتجاهات والارتباطات المحتملة من البيانات التي يتم الحصول عليها من مصادر مختلفة لكشف المعلومات المخفية. ولأجل عرض البيانات بصريًا، غالبًا ما تُستخدم إحصاءات عالية الأبعاد وتقنيات خفض الأبعاد. إذا لم تكن مجموعات التصنيف ممثلة بشكل متقارب، يُعدّ ذلك اختلالًا في توازن مجموعة البيانات. في هذه الدراسة، استخدمنا طريقة SMOTE لإعادة موازنة مجموعة البيانات، مع استخدام تحليل المكوّنات الرئيسية (PCA) لحل مشكلة البيانات عالية الأبعاد. تم تطبيق أربع خوارزميات تصنيف وهي: التحليل التمييزي الخطي (LDA)، الشبكة العصبية الاصطناعية (ANN)، الانحدار العشوائي المتدرّج (SGD)، وغابة القرارات العشوائية (RFA). في هذه الدراسة، تم استخدام مجموعة بيانات السرطان للتحقق من كفاءة الطريقة المقترحة واختيار نتائج المصنّفات. تم تحليل مجموعات البيانات الخام، والمحوّلة بواسطة PCA، والمعدّلة بطريقة SMOTE باستخدام الخوارزميات المذكورة. وقد أُجريت التحليلات باستخدام WEKA.

الكلمات المفتاحية: التصنيف، عدم التوازن، تحليل المكوّنات الرئيسية.

Corresponding Author: E-mail: guhdar.abdulaziz@gmail.com

Introduction

Data mining is the process of looking through vast amounts of data or archives for patterns and then using those patterns to forecast future events. One of the data mining techniques for categorizing a specific category of objects into targeted categories is classification. The main aim of classification is to predict the disposition of an object or data based on the groups of objects that are available (Adebayo and Chaubey, 2019). There are several different algorithms for classification in the literature. Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Stochastic Gradient Descent (SGD) and Random Forest Analysis (RFA) are the most important and commonly machine learning algorithms for classification process. The classification of imbalanced data is one of the most complex problems encountered in classification. Because problems arise in binary grouping when there are many instances of one class (majority) and a small number of instances of the other (minority). However, if there is no disparity between samples from the positive and negative classes, this problem may not be so serious (Al-Rousan et al., 2012). In this paper, it is aimed to solve the classification problem for high dimensional imbalanced data. In the study, the classification problem was investigated with simultaneous rebalancing and dimension reduction methods (using PCA and SMOTE methods). In accordance with this purpose, four well-known classification algorithms (LDA, ANN, SGD and RFA) were used for different imbalanced datasets where the number of samples in one class (majority) is substantially higher than the number of samples in the other class (minority).

1st: Materials And Methods

Machine learning deals with different types of learning, and classification algorithms in detail. In this paper discusses LDA, ANN, SGD, and RFA data mining and machine learning algorithms to deal with the class imbalance problem in high dimensional data. In addition, PCA and SMOTE methods were used.

In order to investigate the effectiveness and reliability of the proposed methods, the methods used were applied to the Cancer disease dataset. The dataset was selected based on the imbalanced percentage values between negative (37.25%) and positive (62.75%) class; the imbalance between negative and positive classes is 25.5%. The Wisconsin breast cancer diagnostic dataset consists of 357 benign and 212 malignant cases. Also, the number of variables in the dataset is 32. This data This dataset was taken from https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.on 12.06.2020.

The imbalanced data set is balanced with the SMOTE method and then the dimension is reduced by the PCA method. Of the rebalanced and the dimension reduction dataset, 30% was used as test data and 70% as training data. By the end of this process, the data set was classified using four classification algorithms (LDA, ANN, SGD, RFA). Accuracy, Precision and ROC area measurements were used as evaluation measures. Using tenfold cross-validation for evaluation, the

information was automatically divided into ten equal parts, testing and training procedure was carried out ten times. After the data are prepared for classification and evaluation, the method appoints each test record to the most likely class. In the last stage, the classification performances before and after the PCA and SMOTE processes were compared separately to determine whether there was an improvement in the performance of the classification models and the efficiency of the methods.

2nd: Classification

The classification and prediction tasks deal with predicting the value of one field (the target) based on the values of other fields (attributes or features). The assigned assignment is called classification whether the objective is discrete (e.g. nominal or ordinal). Classification is typically a supervised process in which the model learns to correctly identify new unseen instances based on a previously correctly identified collection of training instances. Predicting whether not to award a credit to a customer is an example of a classification challenge. So, a set of yes or no reflecting a positive and negative judgment respectively, could shape the values of the class C in in this problem. Details about a consumer will be the input to the classification system (that is, for a classifier). If the theory space is made up of rules, the output could be made up of a series of learned rules. (Ławrynowicz and Tresp, 2014).

1- The Random Forest Algorithm (RFA)

Random Forest, as the name suggested, is a tree-based ensemble in which each tree is dependent on a set of random variables. Assuming that there is an undefined co-distribution $P_{XY}(X,Y)$, for a $p-$ dimensional random vector $X = (X_1,...,X_p)T$ representing the real-valued input or predictor variables and a random variable Y representing the real-valued response. The aim is to find $f(X)$, which is a prediction function to predict Y. The prediction function is determined by the loss function $L(Y,f(X))$ and is defined to minimize the expected value of equation 1.

$$E_{XY}\left(L(Y,f(X))\right) \qquad (1)$$

where the subscripts denote expectation with respect to the co-distribution of X and Y (Cutler et al., 2012).

2- Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) The stochastic gradient descent (SGD) algorithm is a significant reduction in complexity. Rather than computing the gradient of $E_n(fw)$ fully, each iteration estimates this gradient to the basis of a single randomly chosen example z_t:

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_t, w_t) \qquad (2)$$

The stochastic process $\{w_t, t = 1,\dots\}$ depends on the samples randomly selected at each iteration (Pham et al., 2016). The stochastic algorithm can process examples on the fly in a deployed system because it does not need to recall which examples were visited during previous iterations. In this case, as the examples are randomly attracted from the basis truth distribution, stochastic gradient descent directly optimizes the inevitable risk (Bottou, 2012).

3- Artificial Neural Network (ANN)

The first artificial neural network was created using a very basic concept of neural connections. The neurobiologist Frank Rosenblatt (1957), inventing the "Mark I Perceptron" machine, suggested that the mismatch between the real and predicted performance of the artificial connections between neurons could be reduced by a supervised learning mechanism. The mismatch between the real and

predicted responses of the network contains crucial information for optimizing learning outcomes. Predicted performance is obtained using training data (Dell'Aversana, 2019).

4- Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis finds an ideal set of discriminant projection vectors W, to map the original data space onto a lower dimensional feature space, by maximizing the fisher criterion $J(W)$ which ensure that the overlap between the classes in lower dimensional feature space is minimum. (Qin et al., 2005). For example, consider two classes in a two-dimensional feature space. W, W', which reflects the distribution of classes and shows two spaces, is shown in figure 1. Here W' indicates significantly class overlap in the projected space, while the W projection shows greatly improved class separation. Thus, LDA is described for dimensionality reduction. Assume that $X = \{x_1, x_2, \ldots, x_p\}$ is a data set of p dimensional vectors. Each data point is associated with one of the C object classes $\{x_1, x_2, \ldots x_i, \ldots, x_c\}$ (Shashoa et al., 2016). LDA, which is a function of W, is given by equation (3).

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \qquad (3)$$

Here, W must be chosen such that it maximizes $J(W)$. SB shows the between-class distribution matrix, while SW shows the within-class distribution matrix (Poston and Marchette, 1998).

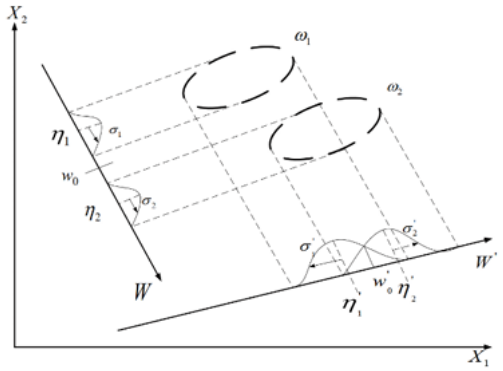

Figure (1): An example linear mapping (Shashoa et al., 2016).

3rd: Confusion matrix

The representation of the following parameters in the form of a simple matrix is expressed as the Confusion matrix. Confusion matrix, in the classical binary classification problem, the classifier units are labeled as positive and negative, and the matrix has four outcomes (Al-Rousan, 2012).

Table (1): Confusion matrix

	Predicted	
Actual	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100 \qquad (4)$$

This measure is proposed as a fitness measure in evaluating each subset generated by data mining classification algorithms.

$$Precision = \frac{TP}{TP + FP} \qquad (5)$$

Another one measure used in classification is precision, and the denominator in precision consists only of the total predicted positive.

$$TPR = \frac{TP}{TP + FN} \quad (6)$$

$$FPR = \frac{FP}{FP + TN} \quad (7)$$

4th: RESULT

In Table 2 according to Accuracy (ACC), Precision (PRE), and ROC area (ROC) measure, the analysis results of the raw data set and the data sets obtained by the methods (PCA and SMOTE) are given in determining the effectiveness of methods and performance of the four classification algorithms.

Table (2): Measurement rates for classification algorithms and methods

Algorithms	Raw Data			PCA			SMOTE		
	ACC	PRE	ROC	ACC	PRE	ROC	ACC	PRE	ROC
LDA	96.4	96.6	*99.6*	97.0	97.1	99.1	97.4	97.5	*99.4*
ANN	97.6	97.7	97.8	*97.7*	97.7	*99.2*	*99.1*	*99.2*	99.3
SGD	*98.2*	*98.3*	97.7	*97.7*	*97.8*	97.0	97.8	97.9	97.9
RFA	94.7	94.8	*99.6*	95.4	95.4	99.0	97.8	97.9	99.0

Table 2. show that, the accuracy, precision and ROC area measures were calculated for checking the performance of the four classification algorithms with and without using dimensionality reduction and rebalancing data methods. For the raw data, the highest Accuracy, Precision rate of 98.2%, 98.3% were obtained in SGD respectively; but the highest Roc area rate of 99.6%, was obtained in LDA and RFA; According to the Accuracy and Precision measurements calculated, the lowest rates as 94.7%, 94.8% were obtained in RFA. On the other hand, when we used PCA the result was improved significant because the number of variables decreased to 12 variables. However, when the PCA dimensionality reduction method was used, we can see that the highest Accuracy rate of 97.7%, were obtained in ANN, SGD respectively; The highest Precision rate of 97.8% was obtained in SGD; on the other hand, the highest Roc area rate of 99.2% was obtained in ANN algorithm. When the SMOTE oversampling method was applied, the highest Accuracy, Precision rate of 99.1% 99.2%, were obtained in ANN, respectively. When Table 1 is examined in more detail, it is seen that the performance of the classification algorithms is very good, although the data is rebalanced and the dimensionality is reduced. Thus, it can be said that the given methods are important to deal with high-dimensional data in the presence of imbalance problem.

Figure 2 shows the performance of each classification algorithms, before and after using PCA and SMOTE, according to Accuracy, Precision and ROC area.
When Figure 2 is examined, when PCA and SMOTE methods are used, a good improvement was observed in the performances of classification algorithms except for SDG, according to almost all three measures. namely, the algorithms gave very good results although 32 variables were reduced to 12 variables with PCA and 25.5% imbalance rate was eliminated with SMOTE. Therefore, this study indicates that both methods are very useful.

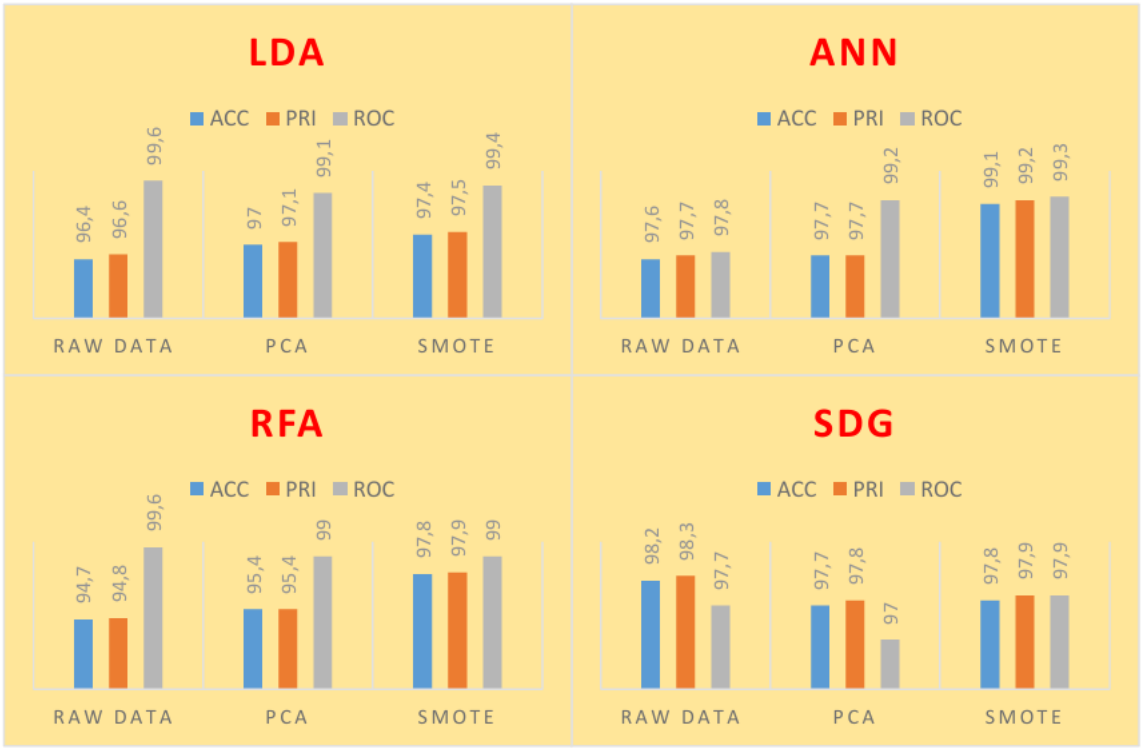

Figure (2): Performance of each classification algorithms, before and after using methods.

According to Figure 2, the LDA figure showed that accuracy and precision increased after using PCA and SMOTE methods, but Roc area was slightly decreased. It was observed that the ANN algorithm gave very good results according to all three measurements in the SMOTE method, and the Roc field in the PCA method gave very good results compared to the first case and the other two measurements.

Considering the ROC area measure, the RFA algorithm performed poorly with both methods compared to the raw data set. However, according to Accuracy and Precision measurements, the algorithm performed very well in both methods, especially in the SMOTE method. According to the three measurement criteria, it was determined that the SDG algorithm gave better results in the raw data set. In general, it can be said that the SDG algorithm does not give bad results, but it does worse than the initial state and other algorithms. This may be due to the fact that the SDG algorithm already gives very good results before the methods are used.

5th: DISCUSSION AND CONCLUSION

The problem of dealing with high dimensional imbalanced data is resolved by removing redundant features (via the PCA method of dimensionality reduction) and rebalancing the results (using the SMOTE method). By defining a faster and more effective model, more ideal solutions and more successful results can be obtained in many different fields, especially in the field of health (Naseriparsa and Kashani, 2014).

Experimental results on the four different classification algorithms for imbalanced high-dimensional data showed that all classification algorithms have enhanced the classification performance of datasets using either PCA, SOMTE methods. For the dataset, the SGD algorithm has provided the best results for raw data before applying any rebalancing or dimensionality reduction methods. On the other hand, when the SMOTE method is used, the ANN classifier gave better results than other algorithms. It was observed that PCA and SMOTE methods did not affect the SGD algorithm results, but these two methods improved the results of LDA, ANN and RFE algorithms.

As a result, it has been observed that classification performances increase in general when PCA and SMOTE methods are used. When these methods are used, it can be said that the best results are obtained with the ANN algorithm and although good results are obtained with the SDG algorithm, no improvement is achieved with the methods.

References

1- Adebayo, A. O., Chaubey, M. S. 2019. Data Mining Classification Techniques on the Analysis of Student's Performance. Global Scientific Journal, 7(4), 79-95.
2- Al-Rousan, N., Haeri, S., Trajković, L. 2012. Feature selection for classification of BGP anomalies using Bayesian models. 2012 International Conference on Machine Learning and Cybernetics, Xi'an, China, 15-17 July 2012, pp. 140-147.
3- Bottou. L. 2012. Stochastic Gradient Descent Tricks. Grégoire Montavon. Geneviève B. Orr and Klaus Robert Müller (Eds.) Neural Networks: Tricks of the Trade (2nd ed.), pp. 421-436. Springer-Verlag, Berlin.
4- Cutler, A., Cutler, D. R., Stevens, J. R. 2012. Random Forests. Cha Zhang and Yunqian Ma (Eds.) Ensemble Machine Learning: Methods and Applications, pp. 157-175. Springer-Verlag, New York. DOI 10.1007/978-1-4419-9326-7
5- Dell'Aversana, P. 2019. A Global Approach to Data Value Maximization: Integration, Machine Learning and Multimodal Analysis. Cambridge Scholars Pub., Newcastle.
6- Ławrynowicz. A., Tresp. V. 2014. Introducing Machine Learning. Jens Lehmann and Johanna Voelker (Eds.) Perspectives on Ontology Learning, pp. 35, 50, IOS Press, Germany.
7- Mulla, G.A.A. 2021. Combination of PCA with SMOTE Oversampling for Classification of High Dimensional Imbalanced Data [M.Sc. Thesis]. Van Yuzuncu Yil University, Institute of Natural and Applied Sciences, Van.
8- Naseriparsa, M., Kashani, M. M. R. 2014. Combination of PCA with SMOTE resampling to boost the prediction rate in lung cancer dataset. International Journal of Computer Applications, 77(3): 33-38.
9- Pham, T. S., Hoang, T. H., Vu, V. C. 2016. Machine learning techniques for web intrusion detection – a comparison. 2016 Eighth International Conference on Knowledge and Systems Engineering (KSE 2016), pp. 291-297, 6-8 October 2016, Hanoi, Vietnam.
10- Poston, W. L., Marchette, D. J. 1998. Recursive Dimensionality Reduction Using Fisher's Linear Discriminant. Pattern Recognition, 31(7), 881-888.
11- Qin, A. K., Shi. S. Y. M., Suganthan. P. N., Loog. M. 2005. Enhanced direct linear discriminant analysis for feature extraction on high dimensional data. Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05) (pp. 851-855).
12- Shashoa. N. A. A., Salem, N. A., Jleta, I. N., Abusaeeda. O. 2016. Classification Depend on Linear Discriminant Analysis Using Desired Outputs. 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA'2016) (pp. 328-332), Sousse, Tunisia, December 19-21.