African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 7, Issue 4, 2024 (pp. 263-271)
www.abjournals.org

A REVIEW ON THE EFFECT OF IMBALANCED DATASET ON LINEAR DISCRIMINANT ANALYSIS

Owoyi M. C.[1*], and Okwonu F. Z.[2]

[1]Department of Mathematics, Dennis Osadebay University, Asaba, Nigeria.

[2]Department of Mathematics, Delta State University, Abraka, Nigeria.

[*]Corresponding Author's Email: mildred.owoyi@dou.edu.ng

Cite this article:

Owoyi, M. C., Okwonu, F. Z. (2024), A Review on the Effect of Imbalanced Dataset on Linear Discriminant Analysis. African Journal of Mathematics and Statistics Studies 7(4), 263-271. DOI: 10.52589/AJMSS-ZOZBNYPR

Manuscript History

Received: 16 Sep 2024

Accepted: 14 Nov 2024

Published: 25 Nov 2024

ABSTRACT: *Imbalanced data are often delegated issues in data sets as it has the power to affect the result and the performance of the classification algorithm. Such problems, if not handled well with good sampling techniques could lead to biased results, overfitting as well as a high rate of misclassification thereby favouring just one class among the two classes. Usually, when assigning sampling techniques, it is necessary to look at the nature of the dataset being studied. It is of a truth that the LDA classifier looking for an efficient performance when presented with imbalanced instances is not suitable to deal with imbalanced learning tasks, since it tends to classify all the data into the majority class, which is usually the less important class. This work explains the different approaches which have been employed by different researchers to resolve the issue of imbalanced data in LDA and the effect of the results obtained both positively and negatively. It should be noted that this single article cannot completely review all the works or research done on the topic, hence we hope that the references which was dually cited will be of help to the major theoretical issues.*

KEYWORD: Imbalanced data; Oversampling; Undersampling; Classification; Metric evaluation.

Article DOI: 10.52589/AJMSS-ZOZBNYPR
DOI URL: https://doi.org/10.52589/AJMSS-ZOZBNYPR

African Journal of Mathematics and Statistics Studies
ISSN: 2689-5323
Volume 7, Issue 4, 2024 (pp. 263-271)

www.abjournals.org

INTRODUCTION

The issue of imbalanced datasets has been a major problem for concern and has gotten more emphasis in recent years. An imbalanced dataset occurs when one group of a data set appears to be larger than the other group. Imbalanced data sets are found in many real-world situations, such as fraudulent classification (Chawla et al., 2002), oil spill detection (Kubat et al., 1998) web mining (Costa et al., 2012; Yeh et al., 2009; Ting, 2008), fraud data (Brockett et al., 2002; Kale et al., 2021), pattern recognition(Szabo et al., 2002; Declerck et al., 2021; Bicciato et al., 2003; Romualdi et al.., 2003), gene expression (Li et al., 2017; Kim et al., 2020), cancer genomic data (Li et al., 2017) and intrusion detection (Garcia-Pedrajas et al., 2012) and information sorting. LDA aims at projecting the features in higher dimensional classification space onto a lower dimensionality, creating rules to differentiate between populations, and making classification based on the rule. Among several tasks in statistics and machine learning, one of the most important tasks is classification. Different solutions to the class-imbalanced problem have been proposed both at the data and algorithmic levels. Classification of such datasets is one of the challenging factors in machine learning. Data set with such an imbalanced sample size problem, the majority group terms to be favoured when the performance of the classification algorithm is tested and such a problem is seen as a class imbalance problem (Das et al., 2018). The equitability of the classification algorithm is tempered by learning relatively more from the majority group than the minority group. To get over such issues, three primary methods have been formed, which include: (i) data pre-processing method, which temper with the imbalanced data before the classification; (ii) algorithm-oriented method, which is made up of several methodologies which are used for improving the already existing classification rules algorithmically to produce simple and easy going rules; and (iii) hybrid method, it involves combining both data pre-processing and algorithm-oriented methods (Kaur et al., 2018). A Resampling method is a cluster of methods within the data pre-processing method which involves the transformation of a data set with an imbalanced group problem (imbalanced data set) into a data set with balanced group sizes (balanced data set). Resampling methods constitute oversampling, undersampling, and a combination of both oversampling and undersampling (which is also termed hybrid resampling). The most common and mostly used oversampling method is random oversampling (ROS) and the most common and mostly used undersampling method is random undersampling. (Xie and Qiu, 2007)Demonstrated the negative effects of imbalanced data sets on the performance of LDA theoretically. This analysis is confirmed theoretically by the experimental results: applying different sampling techniques to balance the imbalanced data sets, the result confirmed that the efficiency of the performances of LDA on balanced data sets is better than those of LDA on imbalanced data sets.

The LDA when faced with the problem of the imbalanced dataset is always a problem since it depends on the sample mean and sample covariance matrix. Limitations like overfitting, misclassification, biased results, etc. are often encountered. It has also been stated that different methods employed in handling the imbalanced data have also influenced the LDA classifier. Methods like cross validation strategy which involve getting some samples of objects employed to train a classification algorithm and validating the efficiency of performance of the classification model alternatively. It is a popular method used as a result of the fact that it can data split heuristics, and the most basic cross-validation strategy is lqocv (Later, in order to reduce the high computational cost of lqocv, kfcv was introduced by Geisser in 1975). Generally, a metric evaluation is needed to verify the performance of a developed classification algorithm. Some performance measures such as sensitivity or specificity, accuracy rate, and

Article DOI: 10.52589/AJMSS-ZOZBNYPR
DOI URL: https://doi.org/10.52589/AJMSS-ZOZBNYPR

misclassification rate are computed using some evaluation metrics like; f1 score, confusion metric, benchmark evaluation, etc. Often, the aim of the research determines the evaluation metric to be applied and the metric used also tells how well the LDA classifier is to perform in a particular dataset.

It should be noted that a single article cannot be a comprehensive review. Instead, our goal has been to provide a sample of existing lines of research in each technique. In each of our listed areas, many other papers have more comprehensive detailed relevant work. In this work, we will be showing some works which have been done on imbalanced data sets which have both negative and positive effects on LDA theoretically. Also, some data level methods employed to balance the data sets and some evaluation metrics used will be considered.

Methods of handling imbalanced data and Result obtain with LDA classifier:

There are different methods of handling an imbalanced data set which include the data-level methods that involve the collection of examples to balance distributions and remove difficult samples, Algorithm-level methods which involve modifying directly existing learning algorithms to reduce the bias towards majority objects and getting used to mining data with skewed distributions and also the Hybrid methods that combine the advantages of the two methods. Here we will be looking at the data-level methods.

Mean-variance cloning technique (MVCT)

According to 0kwonu et al. (2024), the mean-variance cloning technique (MVCT) method works just like the over-sampling method. However, the MVCT procedure applies information from the minority group to generate data sets that behave like the original minority data set. Combining the sample size from the minority data set, and the sample size of the cloned data set which has an equal sample size together with their dimension, they form a new majority group. The data set emerged from the minority group that shares similar characteristics with the given data set. The effect of this is that the possibility of influential observations generated from the original data will be minimized. However, the MVCT procedure may behave like the SMOTE procedure (Chawla, 2002) but with different data extraction and generation procedures. In general, the MVCT method relies on the internal mechanism of the minority data set to compute the mean and standard deviation of the new data set. Also, the metric evaluation applied has an effect on the classification algorithm. In 0kwonu et al. (2024), the results using the benchmark evaluation threshold (0kwonu et al., 2022) are given as equation 1.

$$\delta = \left[\frac{1-\alpha}{2\times\alpha}\right] \alpha \qquad (1)$$

where α is the probability of correct classification from the confusion matrix and

δ is the probability of misclassification. Therefore, the BETH value is

$\cap = \mathsf{C} - \delta, (\mathsf{C} = 1)$

The performance of the classification model is calculated as

$$\rho = \frac{\alpha}{\cap}$$

Article DOI: 10.52589/AJMSS-ZOZBNYPR
DOI URL: https://doi.org/10.52589/AJMSS-ZOZBNYPR

This reveals that the Fisher linear classification method (FLCM) performed comparably for imbalanced and balanced data and outperformed the nearest mean classifier (NMC) and the independent classification rule (ICR). The study demonstrated that the MVCT effects on the classifiers are data-dependent. Therefore, the research showed that sample size balancing irrespective of the data dimension does not have a strong impact on the classifier's performance. This analysis concluded that for the k>m classification problem, the FLCM classifier has comparable performance on the imbalanced and balanced data.

Resampling techniques:

The sampling method involves changing the size of either of the classes (majority or minority class) so as to obtain a dataset that is balanced. There are three different types of sampling methods: over-sampling, under-sampling, and hybrid techniques. The oversampling method, it entails the addition of training data which is synthesised by randomly generating a minority data set of the attributes to form samples in the minority class. It focuses on the minority group. The synthetic minority over-sampling technique (SMOTE) algorithm is a common over-sampling method (Chawla et al., 2002). This method involves reconstructing the minority class by randomly interpolating between two neighbouring minority points. The SMOTE can be enhanced by regenerating the minority class only in the border region of minority clusters(Han et al., 2005) or in the minority class point with the highest safe area(Bunkhumpornpat et al., 2009). Some over-sampling methods combine several clustering techniques to create minority clusters and then add a new minority class such as DBSCAN, clustering using representatives (CURE)-SMOTE (Ma and Fan, 2017), k-means SMOTE (Douzas et al., 2018), radius-SMOTE (Paradipta et al., 2021), and Gaussian Kernels with diagonal smoothing matrices (Menardi and Torelli, 2014).

In Xie and Qiu (2007) they compared the performance of LDA on imbalanced data sets with that of LDA on balanced data sets. To make this comparison objective, they employed the method area under the ROC curve (AUC). Four sampling techniques were considered. These include Random undersampling, random oversampling, Tomek link and synthetic minority oversampling techniques. The theoretical analysis confirmed by the experimental results: using the four sampling methods to rebalance the imbalanced data sets found that the performances of LDA on balanced data sets are better than those of LDA on imbalanced data sets. The experimental results also showed that the two over-sampling methods are more effective than the two under-sampling(Tomek link and random undersampling) methods in improving the performance of LDA.

In Xue and Titterington (2008) they used two metrics evaluation to examine the performance of the model, which is the AUC and ER (error rate) through a hold-out validation strategy. Their investigation claimed that random oversampling or random undersampling methods which are both resampling methods yielded not much improvement in AUC, but alleviated the ER significantly. In recent times, research by Jamaluddin and Mahat (2019) revealed positive findings that laid more emphasis on the findings from both previous pieces of research which were contradicting.100 bivariate was used in their empirical study which was distributed normally and simulated with four different real data sets. Through a 10-fold cross-validation strategy based on TPR and TNR, they revealed that the performance increment effect in the classification of the positives was more significant than the performance decrement effect in the classification of the negatives. Also, they discovered that the LDA was significantly biased towards the majority group, as such, they concluded that class imbalance has a negative effect

Article DOI: 10.52589/AJMSS-ZOZBNYPR
DOI URL: https://doi.org/10.52589/AJMSS-ZOZBNYPR

on the performance of the LDA model. Applying the TPR and TNR helped in relating the findings from the earliest works as both measures allowed the performance estimation of LDA in learning from the minority group objects and the majority group objects individually. The research has revealed completely the effect of the resampling method on the performance of the LDA model. The researcher suggested that the findings could be worked on and verified in the future using different techniques as of validation in order to encourage variableness analysis of the methods through other techniques like the loocv, kfcv, rkfcv, B, and B632. The findings could be enhanced by researching the consistency of the estimates between the strategies, which would eventually help in the effectiveness of the analysis of a resampling method towards LDA, especially when considering unbiasedness in discrimination learning irrespective of the class sizes.

In (Drummond and Holte, 2003; Barandela et al., 2003; Roy et al., 2018), in various domains, they applied Classification tasks. The binary classification was considered and the Gaussian-based Bayes rule was assigned to samples with variables, and then employed the resampling method before constructing the linear discriminant rule. For the case of samples of a dataset in the majority and minority groups, ROS randomly was used to select a sample. Generally, a 2X2 confusion matrix is used to represent the decision-making of an algorithm for binary classification. The positives and negatives in the matrix are named. The positives refer to the sample with the event of interest and the negatives represent the objects without the event of interest. The main cells of the confusion matrix are divided into four having four different elements, which include the true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*). *TP*, *TN*, *FP*, and *FN* are made up of correctly classified positives, correctly classified negatives, incorrectly classified negatives, and incorrectly classified positives, respectively. It is given as;

	P	N
P	TP	FP
N	FN	TN

Precision= $\frac{TP+TN}{TP+FP}$

Accuracy= $\frac{TP+TN}{TP+FP+FN}$

Recall= $\frac{TP}{TP+FN}$

Misclassification Rate: $\frac{FP+FN}{TP+TN+FP+FN}$

Note that TP+TN implies correct classification while FN+FP is incorrect classification.

Both accuracy and error rate tested measured negativity in class imbalance examination due to the unfairness achieved as a result of favour towards the majority group (Branco et al., 2016). Such disfavour led to the initiation of balanced accuracy.

In (Jamaluddin and Mahat, 2019) they succeeded in proving that irrespective of the sample size, the resampling method can enhance the overall unbiasedness of the LDA model in the classification of a dataset. Their work encouraged further research to be carried out so as to explore how much effect can a resampling method have on LDA with varying class imbalance ratios. This is necessary as it will help be of help in quantifying a solution irrespective of the

Article DOI: 10.52589/AJMSS-ZOZBNYPR
DOI URL: https://doi.org/10.52589/AJMSS-ZOZBNYPR

reason for the class imbalance to avoid computations with unnecessary costs and also ensure that the original features of the imbalanced data are kept. Apart from that, separating the groups of samples could also have a negative effect on the behaviour of class imbalance. Even though the groups are imbalanced, making the groups far from each other, can ease the separation. They also suggested in their investigation that the methods that can successively solve several issues at once in a dataset such as the curses of class imbalance and high dimensionality (either in a large number of instances or a high-dimensionality-small-sample-size problem). In summary, They studied the effect of a resampling method (ROS or RUS) on the performance of LDA based on true positive rate and true negative rate through five validation strategies, i.e. leave-one-out cross-validation, k-fold cross-validation, repeated k-fold cross-validation, naive bootstrap, and 632+ bootstrap and the method was tested on four real data sets. The result of the location and dispersion statistics of the metric measures for performance were further enlightened on:(i) how the resampling method affects LDA performance and (ii) the learning unbiasedness of LDA on the samples irrespective of the data size, hence alleviating the effect of the reason of class imbalance.

Some Problems with an imbalanced dataset

Class imbalances are not the only problem to struggle with when dealing with classification tasks. The method employed in distributing the data set within each class (between-class versus within-class imbalance is also important). Japkwicz (2001), Zadrozny and Elkan (2001), and Prati et al. (2004) researched a systematic study to verify whether class imbalances have limitations on classifier performance or whether these limitations might be explained in other ways. In the study, the artificial dataset was used in order to have absolute control over all the variables they wanted to work with. The results of their investigation applying the discrimination-based inductive scheme, showed that the problem is not entirely caused by class imbalance, rather it is also related to the level of overlapping dataset among the classes. Different studies have discussed the interaction between class imbalances. In (Kasemtweechoki and Suwannik, 2023), three approaches to handling imbalanced data were studied: over-sampling, under-sampling, and hybrid approach. The over-sampling method involves duplicating data in the minority class, the under-sampling involves methods that eliminate majority class data and Hybrid methods combine the noise-removing benefits of under-sampling the majority class with the synthetic minority class-creating process of over-sampling. In their research, they employed a dimensionality-reducing model which is the principal component (PC) analysis, which is often applied to datasets in order to reduce dimensionality, and also to alleviate the majority class data amount. They also compared the proposed method which had eight state-of-the-art under-sampling methods across three different classification models: support vector machine, random forest, and AdaBoost. In their research conducted on 35 datasets, the proposed method used had values with higher averages for sensitivity, G-mean, the Matthews correlation coefficient(MCC), and receiver operating characteristic curve (ROC curve) compared to the other methods of under-sampling. Generally, it is important to know that when faced with imbalanced dataset limitations like overlapping, overfitting, biased result, high rate of misclassification rate and many others irrespective of the classification algorithm used, these problems can be handled by choosing a good method of handling imbalanced data that suit your dataset as it has been shown by many studies that large data set works better with oversampling.

Article DOI: 10.52589/AJMSS-ZOZBNYPR
DOI URL: https://doi.org/10.52589/AJMSS-ZOZBNYPR

CONCLUSION

Practically, it is often seen that the oversampling method performs better in imbalanced data when working with LDA than the undersampling as seen in (Xie and Qiu, 2007; Xue and Titterington, 2008). Resampling methods tend to perform well in balancing datasets in LDA but due to some limitations encountered in trying to balance the dataset like in the oversampling technique which can favour only the minority and lead to a biased result and the undersampling which could remove vital information from the dataset, the performance of the LDA model is affected. Also, the MVCT which is similar to the SMOTE as seen in [19] shows that the MVCT does not have an effect on the LDA classifier and also shows that the oversampling method seems to be better in balancing the dataset.

REFERENCES

Barandela, R., Sánchez, J.S., García, V., and Rangel, E. (2003): Strategies for learning in class imbalance problems, Pattern Recognition 36(3) 849-851

Bicciato, S., Pandin, M., Didonè, G., di Bello, C. (2003): Pattern identification and classification in gene expression data using an autoassociative neural network model. Biotechnol. Bioeng.. Bioeng. **81**(5), 594–606 https://doi.org/10.1002/bit.10505

Branco, P., Torgo, L., & Ribeiro, P. R. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, *49*(2), 31:1–31:50. https://doi.org/10.1145/2907070

Brockett, P.L., Derrig, R.A., Golden, L.L., Levine, A., Alpert, M. (2002).: Fraud classification using principal component analysis of RIDITs. J. Risk Insur.Insur. **69**(3), 341–371 .https:// doi.org/10.1111/1539-6975.00027

Bunkhumpornpat C., Sinapiromsaran K., and Lursinsap C., (2009). "Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in Lecture Notes in Computer Science (including subseries Lecture Notes in *Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5476 LNAI, Springer Berlin Heidelb*erg, pp. 475–482.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002): SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. Artif. Intell. Res. **16**, 321–367. https:// doi.org/10.1613/jair.953

Costa, E., (2012).: A framework for building web mining applications in the world of blogs: a case study in product sentiment analysis. Expert Syst. Appl. **39**(5), 4813–4834 https:// doi.org/10.1016/j.eswa.2011.09.135

Das, S., Datta, S., & Chaudhuri, B. B. (2018). Handling data irregularities in classification: Foundations, trends, and future challenges. PatternRecognition, 81, 674–693.

Declerck, K., Novo, C.P., Grielens, L., vanCamp, G., Suter, A., and Vandenberghe, W. (2021): Echinacea purpurea (L.) Moench treatment of monocytes promotes tonic interferon signaling, increased innate immunity gene expression and DNA repeat hypermethylated silencing of endogenous retroviral sequences. BMC Complement. Med. Therap. **21**(1), 14). https://doi.org/10. 1186/s12906-021-03310-5

Douzas, G., Bacao, F. , and Last, F., (2018) ."Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," Information Sciences, vol. 465, pp. 1–20, Oct. doi: 10.1016/j.ins.2018.06.056.

Article DOI: 10.52589/AJMSS-ZOZBNYPR
DOI URL: https://doi.org/10.52589/AJMSS-ZOZBNYPR

Drummond, C., and Holte, R. C. (2003): C4.5, Class Imbalance, and Cost Sensitivity: Why Un-der-sampling beats Over-sampling. In Workshop on Learning from Imbalanced Data Sets II.

García-Pedrajas, N., Pérez-Rodríguez, J., Ortiz-Boyer, D., Fyfe, C. (2012). Class imbalance methods for translation initiation site recognition in DNA sequences. Knowl. Based Syst. -Based, **25**(1),22–34 https://doi.org/10.1016/j.knosys.2011.05.002

Han, H., Wang, W. Y. and Mao, B. H. : (2005) "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Lecture *Notes in Computer Science*, vol. 3644, no. PART I, Springer Berlin Heidelberg, pp. 878–887

Jamaluddin, A.H & Mahat, N. I. (2019). The effects of resampling methods on linear discriminant analysis for data set with two imbalanced groups: An empirical evidence. *Advances and Applications in Statistics*, *59*(1), 17 –42. https://doi.org/ 10.17654/AS059010017

Jamaluddin, A. H., & Mahat, N. I. (2021). Validation assessments on resampling method in imbalanced binary classification for linear discriminant analysis. *Journal of Information and Communication Technology*, 20(1), 83-102.https://doi.org/10.32890/jict.20.1.2021.6358

Japkowicz. N. (2001).Concept-learning in the presence of between-class and within-class imbalances. *In Proceedings of the Fourteenth Conference of the Canadian Society for Computational Studies of Intelligence*, 67-77.

Kale, N., Kochrekar, S., Mote, R., Dholay, S. (2021).: Classification of fraud calls by intent analysis of call transcripts. In: 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6. IEEE https://doi.org/10.1109/ICCCNT 51525.2021.9579632

Kasemtaweechok1, C. and Suwannik W. (2023): Under-sampling technique for imbalanced data using minimum sum of Euclidean distance in principal component subset, *IAES International Journal of Artificial Intelligence (IJ-AI)*,13(1), 305~318 ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i1.pp305-318

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Comput. Surv., 52(4). https://doi.org/10.1145/334344

Kim, B.H., Yu, K., Lee, P.C. (2020).: Cancer classification of single-cell gene expression data by neural network. Bioinformatics **36**(5), 1360–1366 https://doi.org/10.1093/bioinformatics/btz772

Kubat, M., Holte, R.C., Matwin, S. (1998).: Machine learning for the detection of oil spills in satellite radar images. Mach. Learn. **30**(2–3), 195–215 https://doi.org/10.1023/a:100745222 3027

Li, Y., Umbach, D. M., Li, L. (2017): A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. BMC Genom. **18**(1), 1–13 https://doi.org/10.1186/s12 864-017-3906-0

Ma L. and Fan S., (2017)."CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," BMC Bioinformatics, 18, (1) doi: 10.1186/s12859-017-1578-z.

Menardi G. and Torelli N., (2014) ."Training and assessing classification rules with imbalanced data," Data Mining and Knowledge Discovery, 28(1), pp. 92–122, doi:10.1007/s10618-012-0295-5

Article DOI: 10.52589/AJMSS-ZOZBNYPR
DOI URL: https://doi.org/10.52589/AJMSS-ZOZBNYPR

Okwonu, F.Z., Ahad N. A., Okoloko I.F. Apanapudor J.S. Kamaruddin S.A. and Arunaye .F.T (2022). Robust Hybrid classification methods and applications. *Journal of Science and Technology*, https://doi.org/10.47836/pjst.30.4.29

Okwonu F.Z, AhadN. A., Apanapudor J. S., Arunaye F. I and Sharipov O.S. (2024). Application of mean-variance cloning technique to investigate the comparative performance analysis of classical classifiers on imbalance and balanced data, I*ntelliSys 2023, LNNS 824, pp. 284–300*, https://doi.org/10.1007/978-3-031-47715-7_19

Pradipta, G. A., Wardoyo, R, Musdholifah, A. and Sanjaya, H.N.I, (2021). Radius-SMOTE: A new oversampling technique of minority samples based on radius distance for learning from imbalanced data," IEEE Access, vol. 9, pp. 74763–74777, doi:10.1109/ACCESS.2021.3080316.

Prati, R. C., Batista, G. E. A. P. A., and Monard, M. C. 2004). Class Imbalances versus ClassOverlapping: *an Analysis of a Learning System Behavior*. In MICAI pp. 312–321.LNAI 2972

Romualdi, C., Campanaro, S., Campagna, D., Celegato, B., Cannata, N., Toppo, S., Lanfranchi, G. (2003).: Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. Human Molecul. Genet. **12**(8), 823–836 https://doi.org/10.1093/hmg/ddg093

Roy, S., Ahmed, M., & Akhand, M. A. H. (2018). Noisy image classification using hybrid deep learning methods. *Journal of Information and Communication Technology*, *17*, 233–269

Szabo, A., Boucher, K., Carroll, W.L., Klebanov, L.B., Tsodikov, A.D., Yakovlev, A.Y. (2002).: Variable selection and pattern recognition with gene expression data generated by the microarray technology. Math. Biosci.Biosci. **176**(1), 71–98 https://doi.org/10.1016/S0025-556 4(01)00103-1

Ting, i. (2008).: Web-mining applications in e-commerce and e-services. Online Inf. Rev. **32**(2),129–132. https://doi.org/10.1108/14684520810879773

Xie J. & Qiu Z. (2007).The effect of imbalanced data sets on LDA: A theoretical and empirical analysis, Journal of Pattern Recognition (40) 557 – 562

Xue, J.-H., & Titterington, D. M. (2008). Do unbalanced data have a negative effect on LDA? Pattern, 41(5), 1575–1588. https://doi. org/10.1016/j.patcog.2007.11.008

Yeh, I.C., Lien, C., Ting, T.M., Liu, C.H. (2009).: Applications of web mining for marketing of online bookstores. Expert Syst. Appl. **36**(8), 11249–11256 https://doi.org/10.1016/j.eswa. 2009.02.068

Zadrozny B. and Elkan. C. (2001). Learning and making decisions when costs and probabilities are both unknown. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 204-213.

Article DOI: 10.52589/AJMSS-ZOZBNYPR
DOI URL: https://doi.org/10.52589/AJMSS-ZOZBNYPR