Handling leukaemia imbalanced data using synthetic minority oversampling technique (SMOTE)

F Y Chin¹, C A Lim² and K H Lem³
¹,²,³Department of Physical and Mathematical Science, Universiti Tunku Abdul Rahman, Jalan Universiti, 31900, Kampar, Perak, Malaysia

Email: ¹chinfy@utar.edu.my, ²limchienai@1utar.my, ³lemkh@utar.edu.my

Abstract. High dimensional data always lead to overfitting in the prediction model. There are many feature selection methods used to reduce dimensionality. However, previous studies in this area of research have reported that an imbalanced class raises another issue in the prediction model. The existence of the imbalanced class can lead to low accuracy in the minority class. Therefore, high dimensional data with imbalanced class not only increase the computational cost but also reduce the accuracy of the prediction model. Handling imbalanced class in high dimensional data is still not widely reported in the literature. The objective of the study is to increase the performance of the prediction model. We increased the sample size using the Synthetic Minority Oversampling Technique (SMOTE) and performing the dimension reduction using minimum redundancy and maximum relevance criteria. The support vector machine (SVM) classifier was used to build the prediction model. The leukaemia dataset was used in this study due to its high dimensionality and imbalanced class. Consistent with the literature, the result shows that the performance of the shortlisted features is better than those without undergoing the SMOTE. In conclusion, a better classification result can be achieved when high dimensional feature selection coupled with the oversampling method. However, there are certain drawbacks associated with the use of a constant amount of synthesis of SMOTE, further study on different amounts of synthesis might provide different performances.

1. Introduction
In classification, one of the crucial challenges is the high dimensionality of the dataset [1]. The excessive features in high dimensional data are not always beneficial when building the predictive model because many of the features are likely to be irrelevant and redundant [2]. The high dimensional data can be tackled by using feature selection which is capable to select the best subset of features for the construction of the classification model [3]. But feature selection still has its issue dealing with an imbalanced dataset [4,5]. An imbalanced dataset is a dataset that presents skewed distribution in the label class. The skewness will result in biased prediction and misleading accuracy [6]. Consequently, the predictive model may only able to predict the majority class as the model does not learn enough information about the minority class. Therefore, there is a high chance of misclassification due to the huge difference between the label class [7].

Due to an imbalanced dataset, most of the machine learning algorithms will perform inefficiently and even only focus to predict the majority class. Some performance evaluation methods such as the receiver operating characteristic curve (ROC) and area under the curve will lose their ability to...
evaluate the performance of algorithms [8]. Besides, the recall from the predictive model that gets from imbalanced data might be low. Therefore, there is a need to balance the dataset before building any classification model or doing any further analysis. Previous studies have shown that features selection play an important role in high dimensional dataset and several features selection techniques have been developed [7,9,10]. However, the past studies only focus on the dimension reduction on the high dimensional data and not paying too much attention to the imbalanced high dimensional data.

This study focuses on improving the performance of the prediction model by increasing the sample size of the minority class in imbalanced dataset before dimension reduction. An oversampling technique known as Synthetic Minority Oversampling Technique (SMOTE) was implemented to increase the sample size of the minority class [10,11,12]. Then a mutual information-based feature selection method was applied to the ‘balanced’ data to select the optimal number of significant features [13,14,15]. By hybridizing these two methods in the imbalanced high dimensional dataset, better performance has been achieved compared to past researches that used only the features selection method [16,17].

2. An Overview on SMOTE and Feature Selection

2.1. SMOTE

The SMOTE method is a widely used oversampling technique to solve the imbalanced problem. It creates synthetic data points from the minority class [18,19]. Consider there are \( n \) samples in the training dataset and \( m \) features in the feature space in the dataset. Firstly, a sample \( x \) from the minority class is randomly selected. Next, consider its \( k \) nearest neighbours in the feature space, which have the smallest Euclidean distance. Then, one of the nearest neighbours was randomly chosen and denoted as \( X_{NN} \). The new synthetic data point is calculated by using the following formula:

\[
x_{SMOTE} = x + u \times (X_{NN} - x)
\]

where \( u \) represents the random number from Uniform distribution, which has the value from 0 to 1. The process repeated until it reaches the desired amount of the synthetic samples.

2.2. Feature Selection

The feature selection called mRMR used to rank the importance of the features and also used to measure the relevance between features and the redundancy among features and class vector for a given set of features [20,21]. They used mutual information to estimate how one vector is related to another. One of the mRMR conditions is maximum relevance [22]. The idea of this is to select the features that are relevant to the class label. The maximum relevance conditions:

\[
V_i = \frac{1}{|S|} \sum_{i \in S} I(h,i)
\]

The \( S \) denotes the subset of features we are looking for. \( I(h,i) \) represents the mutual information between the target classes, \( h \) and the features, \( i \). \(|S|\) is the number of features in \( S \).

There is a possibility that features which satisfy the Max-Relevance condition could be redundant [23]. Therefore, the next step is to select the features which are mutually exclusive to each other. The minimum redundancy conditions:

\[
W_i = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j)
\]

where \( I(i,j) \) represents the mutual information between two features, \( i \) and \( j \). The features with the high relevance to class label and low redundancy among the features will be recorded as higher scores. The feature with the highest importance score will be added to the selected feature set \( S \) in this mRMR process.
3. Method and data

3.1. The dataset
The leukaemia dataset was used in this study because of its high dimensional and imbalanced class. The column of the dataset represents gene expression levels and the rows of the dataset represent the sample. This dataset contains 72 samples (rows) and 7129 features (columns). There are 47 samples of Acute Lymphoblastic Leukaemia (ALL), which is labelled as “1” and 25 samples of Acute Myeloid Leukaemia (AML), which is labelled as “0”. Each sample has 7129 features (gene expression levels) obtained from the microarray.

3.2. The proposed method
The general flow of the proposed method is indicated in figure 1. The (leukaemia) data was first divided into two subsets: the training set and the test set according to the ratio of 7:3. The training set underwent the SMOTE to increase the samples in the minority class so that the classes in the data became balanced. Later, the mRMR algorithm was applied to this high dimensional data to select the important features. A support vector machine classifier (SVM) was used to build the predictive model.

![Flowchart of the proposed method](image)

**Figure 1.** Flowchart of the proposed method.

3.3. The performance indexes
The performance of the predictive model of the SMOTE-mRMR hybrid method was compared with the following three methods which were taken to be a control group. They were (i) random selection, (ii) SMOTE with random selection, and (iii) mRMR (without SMOTE). The cross-validation accuracy of the training set and the accuracy of the test set were plotted to compare the performance of the proposed SMOTE-mRMR model with the control groups. The predictive model with better accuracy in the cross-validation and accuracy of the test set would be considered as a better predictive model. The performance of the predictive model was also evaluated by other commonly used indexes such as the ROC curve and confusion matrix (CM). The accuracy of the predictive model is calculated using the following formula:

\[
\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total Number of Predictions}} = \frac{TN + TP}{TN + TP + FP + FN}
\]

where \(TN\) denotes true negative, \(TP\) is true positive, \(FP\) is false positive, and \(FN\) is false negative. The performance of the predictive model was also assessed using the recall, false-positive rate (FPR), false-negative rate (FNR) and the area under curve (AUC). The recall which lies between 0% and 100% with a higher percentage indicating better predictive performance. The FPR and FNR also fall within the same range but on the contrary, the smaller the percentage, the better the predictive performance. The value of the AUC extends from 0 to 1 with its value closer to 1 implies higher accuracy.
4. Experimental Result and Discussion

The comparison of performance among the SMOTE-mRMR method and the controlled methods reveals that SMOTE-mRMR outperforms the others both in terms of cross validation accuracy of the training set (figure 2) and test accuracy (figure 3). The performance of the random selection method ranks the lowest because the randomly selected features may be irrelevant. Since the data is of high dimension, SMOTE will not perform well in the prediction model independently. Although mRMR is meant to handle high dimension data, it carries weakness for unbalanced data, hence the hybrid method using the SMOTE-mRMR method can improve the performance of the predictive model.

![Figure 2. Comparison between the Cross-Validation accuracy of selected features by using mRMR and randomly selected features.](image)

![Figure 3. Comparison between the test accuracy of selected features by using mRMR and randomly selected features.](image)
The proposed method gives the highest performance in figures 2 and 3 compare to the control group when three features selected in the predictive model. The number of features, $k$ uses in figures 2 and 3 is the number of selected features from the set $S$ used in equation (2), where $k$ is subset of $S$. Besides, the poor performance of randomly selected features is mainly due to the fact that it does not consider relevancy and redundancy during the selection. This suggests that feature selection ought to be applied before building the predictive model. The mRMR selects the features that have a strong relationship with the label class but when working on imbalanced data, the prediction power is still lower than the proposed hybrid method.

The performance evaluation methods that have been used are the receiver operating characteristics (ROC) curve and the confusion matrix (CM). The average AUC value and other performance evaluation methods are summarised in table 1. Table 1 shows the average performance measurements for each of the four method combinations. The highest average accuracy was obtained in the mRMR+SMOTE (98.193%), whereas the lowest average accuracy was achieved in the Random Features (55.067%). When the dataset is imbalanced, it is no longer reliable to use the accuracy as the performance index. Instead, we should rely more upon the recall to measure performance for an imbalanced dataset. Obviously, the recall value achieved by the hybrid mRMR+SMOTE predictive model was the highest.

Based on the measurements listed in table 1, it is interesting to notice that the accuracy performance of these four methods can be arranged in the ascending order as Random Features $\rightarrow$ Random Features + SMOTE $\rightarrow$ mRMR $\rightarrow$ mRMR+SMOTE. SMOTE can improve the Random Feature by balancing the data, but mRMR still outperforms these two methods. This implies that not only is the data balancing important, but feature selection considering relevancy and redundancy is more crucial. When data appears to be both imbalanced and high dimensional, mRMR+SMOTE will be the best choice.

**Table 1.** Summary from Confusion Matrix and ROC Curve for selected features by mRMR and randomly selected features.

|               | mRMR+SMOTE | mRMR    | Random Features+SMOTE | Random Features |
|---------------|------------|---------|-----------------------|----------------|
| **Accuracy (%)** | 98.193     | 97.200  | 76.967                | 55.067         |
| **Recall (%)**  | 96.913     | 92.533  | 82.687                | 20.367         |
| **FPR (%)**     | 0.860      | 0.767   | 23.300                | 71.953         |
| **FNR (%)**     | 3.087      | 7.467   | 17.313                | 79.633         |
| **AUC**         | 0.999      | 0.993   | 0.821                 | 0.519          |

5. **Conclusions**

The imbalance in data is a very common issue that causes the deviation in the performances of the classifiers. On a set of leukaemia data, we applied the SMOTE oversampling technique followed by the mRMR feature selection method to tackled both the imbalance and the high dimension issues. This SMOTE-mRMR coupling method managed to give a better performance in term of accuracy, recall and area under curve. On this basis, the SMOTE-mRMR method seems to offer a better alternative for classifying imbalanced high dimension data. Nevertheless, more investigation on the influence of the difference between label class in the predictive model is necessary to validate the kinds of conclusions drawn here.

**Acknowledgments**

The authors gratefully acknowledge the financial support of Universiti Tunku Abdul Rahman.
References
[1] Kalina J and Rensova D 2015 How to reduce dimensionality of data: robustness point of view *Serbian Journal of Management* **10**(1) pp 131-140
[2] Ladha L and Deepa T 2011 Feature selection methods and algorithms *International Journal on Computer Science and Engineering* **3**(5) pp 1787-1797
[3] Radha R and Muralidhara S 2016 Removal of redundant and irrelevant data from training datasets using speedy feature selection method *International Journal of Computer Science and Mobile Computing* **5**(7) pp 359-364
[4] Olaolu A M, Abdulsalam O, Mope I R and Kazeem G A 2018 A comparative analysis of feature selection and feature extraction models for classifying microarray dataset *Computing and Information Journal* **22**(2) pp 29-37
[5] Chawla N, Bowyer K, Hall L and Kegelmeyer W 2002 SMOTE: Synthetic minority oversampling technique *Journal of Artificial Intelligence Research* **16** pp 321-357
[6] Garba A S and Zurinahni Z 2020 Feature selection for high-dimensional and imbalanced biomedical data based on robust correlation based redundancy and binary grasshopper optimization algorithm *Genes* **11**(7) p 717
[7] Blagus R and Lusa L 2012 Evaluation of SMOTE for high-dimensional class-imbalanced microarray data *Proceedings-2012 11th International Conference on Machine Learning and Applications* pp 89-94
[8] Branco P, Torgo L and Ribeiro R P 2016 A survey of predictive modeling on imbalanced domains *ACM Computing Surveys* **49**(2) pp 1–50
[9] Veni C V and Rani T 2011 On the classification of imbalanced datasets *International Journal of Computer Science & Technology* **2**(1) 145-148
[10] Bai H, Edwoardo A G, He H and Li S 2008 ADASYN: Adaptive synthetic sampling approach for imbalanced learning *2008 International Joint Conference on Neural Networks* pp 1322-1328
[11] Padmaja T, Raju P and Bapi S 2008 Majority filter-based minority prediction (MFMP): An approach for unbalanced data sets *IEEE Region 10 Conference TENCON 2008* pp 1–6
[12] Xie W, Liang G, Dong Z, Tan B and Zhang B 2019 An improved oversampling algorithm based on the samples’ selection strategy for classifying imbalanced data *Mathematical Problems in Engineering* 2019 1-13
[13] Battiti R 1994 Using mutual information for selecting features in supervised neural net learning *IEEE Transactions on Neural Networks* **5**(4) pp 537-550
[14] Kwak N and Choi C 2002 Input feature selection for classification problems *IEEE Transactions on Neural Networks* **13**(1) 143-159
[15] Chow T and Huang D 2005 Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information *IEEE Transactions on Neural Networks* **16**(1) pp 213-224
[16] Bidgoli A M, Nasiriparsa M and Varace T 2013 A hybrid feature selection method to improve performance of a group of classification algorithms *International Journal of Computer Applications* **69** 28-35
[17] Zhao Z, Radhika A and Wang M 2019 Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* pp 442-452
[18] El-Sayed A A, Mahmood M A M, Meguid N A and Hefny H A 2015 Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE) *2015 Third World Conference on Complex Systems (WCCS)* pp 1-5
[19] Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A and Hussain A 2016 Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study *IEEE Access* **4** 7940-7957
[20] Ding C and Peng H 2005 Minimum redundancy feature selection from microarray gene expression data Journal of Bioinformatics and Computational Biology 03(02) pp 185-205

[21] Jo I, Lee S and Oh S 2019 Improved measures of redundancy and relevance for mRMR feature selection Computers 8(2) p 42

[22] Radovic M, Ghalwash M, Filipovic N and Obradović Z 2017 Minimum redundancy maximum relevance feature selection approach for temporal gene expression data BMC Bioinformatics 18(9)

[23] Estevez P, Tesmer M, Perez C and Zurada J 2009 Normalized mutual information feature selection IEEE Transactions on Neural Networks 20(2) pp 189-201