A CLASS SPECIFIC DIMENSIONALITY REDUCTION FRAMEWORK FOR CLASS IMBALANCE PROBLEM: CPC_SMOTE

T. Maruthi Padmaja[1,2], Bapi S. Raju[2]
[1]*IDRBT Masab Tank, Hyderabad 500 057, (AP), India*
[2]*DCIS, University of Hyderabad, Hyderabad 500 057, (AP), India*

P. Radha Krishna
SET Labs, Infosys Technologies Ltd, Hyderabad 500 057, (AP), India

Keywords: Class imbalance problem, Principle component analysis, SMOTE, Decision tree.

Abstract: The performance of the conventional classification algorithms deteriorates due to the class imbalance problem, which occurs when one class of data severely outnumbers the other class. On the other hand the data dimensionality also plays a crucial role in performance deterioration of classification algorithms. Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction. Due to unsupervised nature of PCA, it is not adequate enough to hold class discriminative information for classification problems. In case of unbalanced datasets the occurrence of minority class samples are rare or obtaining them are costly. Moreover, the misclassification cost associated with minority class samples is higher than non-minority class samples. Capturing and validating labeled samples, particularly minority class samples, in PCA subspace is an important issue. We propose a class specific dimensionality reduction and oversampling framework named CPC_SMOTE to address this issue. The framework is based on combining class specific PCA subspaces to hold informative features from minority as well as majority class and oversample the combined class specific PCA subspace to compensate lack of data problem. We evaluated the proposed approach using 1 simulated and 5 UCI repository datasets. The evaluation show that the framework is effective when compared to PCA and SMOTE preprocessing methods.

1 INTRODUCTION

Classification usually utilizes a supervised learning method and learns a model from already labeled historical data and uses this model to predict the class of unseen test data. Class imbalance problem betides in those datasets, where one class (majority class) of data severely outnumbers the other class of data (minority class). For two-class classification problem, many real world applications such as fraud detection (Phua and Lee, 2004), and rare disease prediction in medical diagnosis (Yoon and Kwek, 2007) that reflect the nature of class imbalance problem. In the literature, effect of class imbalance problem is investigated over many classification algorithms such as decision tree, neural networks, support vector machines, *k* nearest neighbor classifier and proved that the classification algorithms bias towards predicting the majority class in class imbalance scenario (Japkowicz and Stephen, 2002). This bias nature of the classifier performance towards the majority class leaves huge error rates from the minority class.

For high-dimensional data, classification process may also include dimensionality reduction to increase the class discrimination, better data representation and for attaining good computational efficiency. The misclassification rate for classification process drastically increases due to spurious dimensions in the original high dimensional data space, known as the curse of dimensionality problem (Duda and Strok, 2001). Hence there is a need for dimensionality reduction. Principal Component Analysis (PCA) is a quite popularly used technique for dimensionality reduction. PCA linearly transforms high dimensional data into lower dimensional space by maximizing the global variance of the data as well as minimizing least square error for that transformation. However, PCA is an unsupervised dimensionality reduction technique. Therefore, it is not adequate enough to hold the discriminative information for classifica-

Maruthi Padmaja T., S. Raju B. and Radha Krishna P..
A CLASS SPECIFIC DIMENSIONALITY REDUCTION FRAMEWORK FOR CLASS IMBALANCE PROBLEM: CPC SMOTE.
DOI: 10.5220/0003092502370242
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 237-242
ISBN: 978-989-8425-28-7

tion problems when the maximum variance direction of one class is different from another class i.e., unequal covariance matrices (Vaswani and Chellappa, 2006; Xudong, 2009). Finding principal axes directions i.e principle components (PCs) is one of the key steps for PCA and it depends on spread of the data. In case of the unbalanced datasets, the spread is dominated by majority class as its prior probabilities are much higher than minority class samples. Moreover, in real world domains such as intrusion detection systems, the occurrence of intrusion transactions are rare and generating them is a costly process. Miss predicting these rarely occurred intrusion transactions is risky and could lead to financial loss for organizations. Therefore, Capturing and validating labeled samples, particularly non-majority class samples in PCA subspace for classification task is a challenging issue.

In this paper, we propose a class specific dimensionality reduction and oversampling framework named CPC_SMOTE to address class imbalance issue in Principle Component Analysis subspace where there is directional difference in Principal Components (PCs) of two classes. The proposed framework based on capturing class specific features in order to hold major variance directions from individual class and oversampling is to compensate lack of data in under-represented class. Proposed approach is evaluated over decision tree classifier using accuracy and F-measure as evaluation metrics. Experimental evidence show that proposed approach yields superior performance on simulated and real world unbalance datasets compared with classifier learned on reduced dimensions of whole unbalanced datasets as well as on oversampled datasets.

The rest of the paper is organized as follows. Section 2 discusses work related to class imbalance problem. Section 3 provides proposed CPC_SMOTE framework. Section 4 presents the experimental evaluation based on a comparative study, which is done with applying PCA on whole unbalanced dataset as well as applying SMOTE on unbalanced datasets. Finally, conclusions are given in section 5.

2 RELATED WORK

There are several ways to handle class imbalance problem. Among them cost sensitive learning, one class classification and resampling the class distribution are frequently used. However most of the research that addresses the class imbalance problem centered on balancing the class distributions. Resampling the data either by random oversampling or by random undersampling in order to make approximately balance class distributions are quite popularly adopted solutions. But for discriminative learners such as decision tree classifier oversampling causes overfitting, where as the undersampling leads to performance degradation due to loss of informative instances from majority class (Drummond and Holte, 2003). (Weiss and Provost, 2003) concluded that the natural distribution is not usually the best distribution for learning. Study of "whether oversampling is more effective than under-sampling" and "which over-sampling or under-sampling rate should be used" was done by (Estabrooks and Japkowicz, 2004), who concluded that combining different expressions of the resampling approach is an effective solution. (Kubat and Matwin, 1997) did selective under-sampling of majority class by keeping minority classes fixed. They categorized the minority samples into some noise overlapping, the positive class decision region, borderline samples, redundant samples and safe samples. By using Tomek links concept, which is a type of data cleaning procedure they deleted the borderline majority samples.

(Chawla and Kegelmeyer, 2002), proposed Synthetic Minority Over-sampling Technique (SMOTE). It is an oversampling approach in which the minority sample is over-sampled by creating synthetic (or artificial) samples rather than by oversampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing synthetic samples along the line segments joining any/all of the k minority class' nearest neighbors. Depending upon the amount of oversampling required, neighbors from the k nearest neighbors are randomly chosen. This approach effectively forces the decision region of the minority class to become more general.

(Villalba and Cunningham, 2008) evaluate unsupervised dimensionality reduction techniques over one-class classification methods and concluded that Principle Component Analysis (PCA) damages the performance on most of the datasets. (Xudong, 2009) analyzed the role of PCA over unbalanced training sets and concluded that the PCA subspace is biased by the majority class eigen vectors. Further the authors proposed Asymmetric Principle component Analysis (APCA), a weighted PCA to combat the bias issue in PCA subspace. In this paper we propose a class specific dimensionality reduction and oversampling framework in the context of two class classification to combat this problem. Proposed approach yielded superior performance on those datasets where there is directional difference between two classes' principle components.

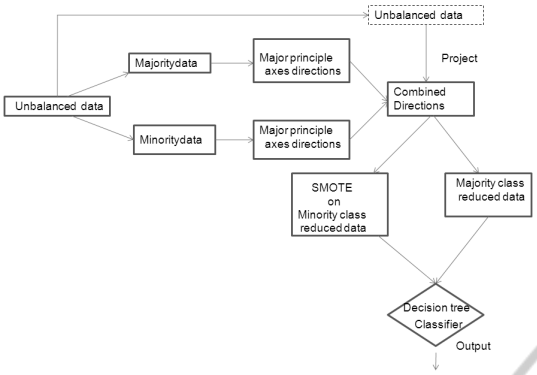

Figure 1: Flow diagram for CPC_SMOTE framework.

3 THE CPC_SMOTE FRAMEWORK

The main goal behind this framework is to combat class imbalance problem in PCA subspace, where the principle component analysis predominantly represents majority class maximum variance directions only. In order to accomplish this goal a class specific principle component analysis and oversampling framework is proposed. Figure 1 depicts the flow diagram for proposed framework.

The class specific PCA is for extracting better informative features from both classes, where there is directional difference between two class PCs. This is done by concatenating class specific principle components that are extracted from each class by applying PCA. Latter SMOTE is applied to this reduced feature space of minority class to alleviate the lack of data problem.

The steps for proposed CPC_SMOTE framework are discussed below.

- Step 1: Extract minority class patterns $(Mi)_{p*d}$ and majority class patterns $(Mj)_{q*d}$ from an unbalanced data X_{n*d}. where $n = p+q, q > p$ and d=number of features in unbalanced data X_{n*d}.
- Step 2: Apply classical PCA on each class, for each class select $r(< d)$ eigen vectors corresponding to first r large eigen values. Let $(E_c)_{d*r}$ be the corresponding eigen vector matrices selected in this step. Where $c = 1$ and 2.
- Step 3: Concatenate eigen vectors $(E_c)_{d*r}$ obtained in step 2 in order to facilitate class specific informative directions.
  $(E_{total})_{d*2r} = [(E_1)_{d*r}, (E_2)_{d*r}]$.
- Step 4: Project unbalanced data X_{n*d} into $(E_{total})_{d*2r}$ to enable the informative feature space of majority and minority classes.
  $(NEW\_X)_{n*2r} = X_{n*d} * (E_{total})_{d*2r}$.
- Step 5: Extract reduced minority class $(NEW\_X\_Mi)_{p*2r}$ and majority class $(NEW\_X\_Mj)_{q*2r}$ patterns from the informative feature space $(NEW\_X)_{n*2r}$ and do SMOTE on $(NEW\_X\_Mi)_{p*2r}$.
  $(T\_NEW\_X\_Mi)_{l*2r} = SMOTE(NEW\_X\_Mi)_{p*2r}$,where $l = (q/p) * p$.
- Step 6: Combine minority class and majority class patterns
  $(T\_NEW\_X)_{t*2r} = [(T\_NEW\_X\_Mi)_{q*2r}, (NEW\_X\_Mj)_{q*2r}]$, where $t = 2q$.

Let X_{n*d} be an unbalanced dataset with n number of records and d number of features. In order to get better separated features we apply PCA independently on each class distribution. Let us suppose that $(Mi)_{p*d}$, $(Mj)_{q*d}$ be the corresponding minority class and majority class distributions, where $q > p$. From each class distribution of $(Mi)_{p*d}$ and $(Mj)_{q*d}$ equal number of eigen vectors are extracted. Let $r(< d)$ are the selected number of eigen vectors and $(E_c)_{d*r}$ be the corresponding eigen vector matrix, and here $c = 1, 2$. The extracted eigen vectors are combined horizontally to get class specific features directions $(E_{total})_{d*2r}$. Now the selected number of features becomes $2r$, where r number of features from each class. Then the whole unbalanced data X_{n*d} is projected into these combined feature space $(E_{total})_{d*2r}$ in order to get combined reduced feature space $(NEW\_X)_{n*2r}$. Since the combined class specific direction matrix $(E_{total})_{d*2r}$ covers the maximum variance directions from all classes, the data can be better discriminate in combined reduced feature space $(NEW\_X)_{n*2r}$. But still there is lack of data problem in reduced feature space of $(NEW\_X)_{n*2r}$ due to the unbalanced nature of the original data. This lack of data problem is alleviated by oversampling the minority class reduced feature space with synthetic samples using synthetic minority oversampling technique (SMOTE). In order to do so, extract minority class feature space $(NEW\_X\_Mi)_{p*2r}$ and majority class feature space $(NEW\_X\_Mj)_{q*2r}$ from $(NEW\_X)_{n*2r}$. Generate l number of synthetic samples where $l = (q/p) * p$ in minority class feature space $(NEW\_X\_Mi)_{p*2r}$ using SMOTE. Finally combine both class reduced feature spaces $(T\_NEW\_X\_Mi)_{q*2r}$, $(NEW\_X\_Mj)_{q*2r}$ to attain balanced class reduced feature space distribution $(T\_NEW\_X)_{t*2r}$ where $t = 2q$. The final matrix $(T\_NEW\_X)_{t*2r}$ obtained is the combined matrix with combination of class specific features and oversampled reduced feature space. This matrix can be directly used for classification tasks. Proposed approach is evaluated using decision tree classifier.

4 EXPERIMENTAL EVALUATION

This section describes the datasets, presents evaluation metric for estimating classifier performance and elaborates the comparative study with other methods for approximating the performance of the proposed framework, in terms of experimental results and discussion.

4.1 Evaluation Metric

Generally the outcomes of any classification model are:

- True positive(TP) Rate is the percentage of correctly classified positive samples.
- True negative(TN) rate is the percentage of correctly classified negative samples.
- False negative(FN) rate is the percentage of incorrectly classified positive samples.
- False positive(FP) rate is the percentage of negative examples predicted as positives.

The goal of any ideal classifier is to maximize TP and TN rates. The following are normally applied measures for evaluating classification performance:

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}. \quad (1)$$

$$TPrate = Recall = \frac{TP}{(TP+FN)}. \quad (2)$$

$$Precision = \frac{TP}{(TP+FP)}. \quad (3)$$

$$F-measure = \frac{2*Precision*Recall}{Precision+Recall}. \quad (4)$$

Proposed approach is evaluated using classification accuracy (eq.1) and F-measure (eq 4). Classification accuracy is designed to reduce the overall classification error rate. Moreover, for classification process, *n* number of dimensions selected as reduced features, based on improvement in overall accuracy over original input space. Similarly *n* number of features selected from PCA. But class imbalance point of view accuracy is not an appropriate measure for evaluating the classifier performance. Consider that there are only 6% of samples from minority class and 94% of the samples are from majority class. If a classifier miss predicts all minority class samples as majority class samples then the accuracy becomes 94% with the contribution of majority class samples only.

F-measure depicts the performance of the target class in terms of tradeoff between precision and recall, where as recall is simply the TP rate of the target class and precision gives the trade-off between TP and FP rates. If both Precision and Recall are high, then F-measure is also high. For unbalanced datasets the precision and recall goals are conflicting, increasing Recall rates without disturbing the precision of the minority class (target class) is a challenging issue.

4.2 Datasets

In order to show how the directional difference between principles axes of PCA affects the performance of unbalanced datasets we have generated a synthetic dataset. The synthetic dataset generated using multivariate normal distributions separately for each class:

Where $p(x)$ is probability density function, x is a M-dimensional vector of features, μ is feature vector mean and Σ is a covariance matrix.

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)] \quad (5)$$

The two class distribution are generated with equal mean $\mu = 0$, M =10 uncorrelated features and with unequal covariance matrices. The two classes covariance matrices are generated in such a way that one class variance dominates the other class variance and with different maximum variance directions, so $(\Sigma_1 > \Sigma_2)$. Moreover, generated distribution contains $\omega_1 = 2000$ samples from majority class and $\omega_2 = 100$ samples from minority class with imbalance ratio of 20%. Figure 2 depicts the structure of gaussians generated for the evaluation purpose, where ω_1 is the majority class distribution, ω_2 is minority class distribution and probability density of $p(\omega_1) > p(\omega_2)$). The diagonal elements for two classes' covariance matrices for which the half diagonal elements are all zeros are depicted as
$\Sigma_1 d_{ii} = [7,5,1,3,4,0.1,0,0.5,0.01,0.01]$,
$\Sigma_2 d_{ii} = [0.5,1,2,0.9,0.7,0,0.5,0.1,0.01,0.01]$ where $i = 1,2,...,10$.

The rest of the datasets Waveform, Mfeat-pixel, Satimage and Musk are taken from UCI repository (http://archive.ics.uci.edu/ml/.,) where as Bronchiolitis dataset is taken from UCD repository (http://mlg.ucd.ie/datasets,). Out of these considered data sets Musk and Bronchiolitis datasets are meant for two-class classification problem. The rest of the datasets have more than two classes, and so they are converted into binary class datasets by considering the class with fewer samples as minority (positive) class and rest of the samples as majority (negative) class, as suggested in (Villalba and Cunningham, 2008). Table 2 shows the description of the datasets used for the experiments.

Table 1: Datasets description.

Dataset	#Min	#Maj	Imb.ratio	#Attri
Simulated	100	2000	20	10
Waveform-1	1647	3353	2	21
Bronchiolitis	37	81	2.18	22
Mfeat-pixel-8	200	1800	9	240
Satimage-4	626	5809	9.27	36
Musk	1017	5581	5.48	166

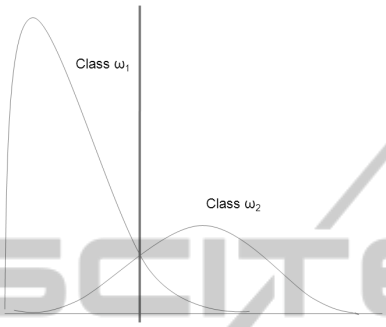

Figure 2: Data from classes ω_1 and ω_2 where the probability of $p(\omega_1) > p(\omega_2)$.
The vertical line indicate class separation.

4.3 Experimental Results and Discussion

Science decision tree classifier prone to be sensitive for class imbalance problem, in this work it is used as a baseline for evaluating the proposed CPC_SMOTE and other methods considered for comparative study. Decision tree classifier learned on proposed CPC_SMOTE is compared with same classifier learned on original data, PCA subspace and on the oversampled data obtained using SMOTE.

Table 3 describes the obtained results on considered methods in terms of accuracy and F-measure. Here θ enables the directional difference between two class's first principal components which contains maximum variance, in terms of cosine angle. If θ is near to one means both classes principle components are in same direction else they are in different directions. The best performance in terms of F-measure that is obtained for that dataset is represented in bold. As mentioned in section 4.1 for evaluating PCA we considered that n number of PC's which can give better accuracy than the decision tree learned on original dataset. The results on simulated dataset reported 95% of accuracy for all methods, but minority class F-measure varied accordingly.PCA on simulated dataset performed badly compared with all methods. Even though the overall accuracy is 95%, but the minority class prediction in terms of F-measure is left with 0.This might be because of two reasons. (1) Due to maximum variance coverage from majority class by ignoring minority class maximum variance direction which contributes less variance to the whole distribution. Here θ=0.137 which shows larger directional difference between two class PC's. (2) The lack of data causes unnecessary overlap in the PCA subspace. This result clearly substantiates the effect of unbalance data on PCA. Further, classifier learned on original data attained 49% of F-measure. As expected SMOTE showed large improvement of 46% over original class F-measure. Proposed CPC_SMOTE yielded superior performance among all methods with 95.2% F-measure.

Comparing the results of real world datasets, for Waveform CPC_SMOTE yielded superior performance of 92% in terms of both accuracy and F-mesure than rest of the methods. The value $\theta = 0.781$ clearly indicates the directional difference between two clases's PC directions. For this dataset PCA showed an improvement of 9.6% F-measure on original data. But compared to the accuracy, minority class F-measure on PCA is less, showing the bias towards majority class. In this dataset SMOTE did not perform well in improving the performance of original dataset both in terms of F-measure and accuracy than rest of the two methods. For Bronchiolitis and Mfeat-pixel datasets PCA subspace did not improve the classification accuracy of original data set. Moreover the minority class prediction in terms of F-measure considerably less than the original datasets as like simulated dataset. Corresponding directional differences $\theta = 0.297, 0.671$ shows larger variation in principle axis directions which substantiate the evidence for lose performance in minority class data due to bias of PCA subspace towards majority class.However, proposed solution CPC_SMOTE yields superior performance on Bronchiolitis dataset with 80.2% minority class prediction in terms of F-measure. But on Mfeat-pixel data set SMOTE yielded superior performance than CPC_SMOTE. For this dataset compared with PCA subspace on original data CPC_SMOTE reduces the bias caused by selecting the majority class maximum variance directions. Even though the two class's PC are in same direction with $\theta = 0.923$ for Satimage dataset CPC_SMOTE is superior to rest of the two methods with 94.7% minority class F-measure and with 94.6% of overall accuracy. For this dataset PCA on decision tree classifier showed consistency in improving the performance on original data in terms of F-measure with 5.7% as well as in total accuracy with 1.4% respectively. For this dataset, SMOTE exhibited 38.3% improvement in minority class prediction in terms of F-measure.

For Musk dataset where the directional differ-

Table 2: Comparison of CPC_SMOTE(C_S) performance with Original Data(OD), PCA, SMOTE(S) in terms of accuracy and minority class F-measure over decision tree classifier.

Dataset	θ	Accuracy%				F-measure%				Selected Features	
		OD	PCA	S	C_S	OD	PCA	S	C_S	PCA	C_S
Simulated	0.137	95	95	95	95	49.1	*0*	95	**95.2**	4	8
Waveform-1	0.781	85	91	88	92	77.4	87	88	**92.2**	5	10
Bronchiolitis	0.297	72	65	79.3	78.2	54.3	*42.3*	79.2	**80.2**	7	14
Mfeat-pixel-8	0.671	93.6	88.3	96.7	93.5	68.8	*32.8*	**96.9**	94	40	80
Satimage-4	0.923	91.8	93.2	94.4	94.6	56.1	61.8	94	**94.4**	7	14
Musk	0.901	96.8	97	97.5	96.5	89.8	90.9	**97.5**	96.2	35	70

ence $\theta = 0.901$ is SMOTE achieved superior performance than rest of the methods with 97.5% minority class prediction and with over all accuracy 97.5. CPC_SMOTE stood in second position with 96.2% F-measure and with 96.5% overall accuracy. PCA showed 1% and 2% consistent improvement on minority class F-meaure and on overall accuracy. From our experiments we observed that the overall accuracy of the classifier learned on original unbalanced data as well as on PCA subspace is biased towards majority class when there is directional difference between principle components of the two classes. Though CPC_SMOTE is computationally costlier than PCA, it is effective in alleviating the bias caused by majority class in PCA subspace and enables better minority class prediction. From the considered 6 datasets CPC_SMOTE outperformed on 4 datasets for which the directional difference is high.

5 CONCLUSIONS

This paper proposed a class specific dimensionality reduction and oversampling framework to combat the class imbalance issue occurred in Principle Component Analysis while selecting maximum variance directions. Proposed approach is compared with classical PCA for dimensionality reduction and SMOTE for oversampling in terms of accuracy and F-measure. Experimental evidence showed that proposed approach yields superior performance in terms of dimensionality reduction and classification of unbalanced data where the maximum variance predominantly represents majority class.

REFERENCES

Chawla, N. V. Bowyer, K. H. L. and Kegelmeyer, W. (2002). Smote: Synthetic minority over-sampling technique. In *Journal of Artificial Intelligence Research.16*.

Drummond, C. and Holte, R. (2003). C4.5, class imbalance,and cost sensitivity: Why under-sampling beats over-sampling. In *In Workshop on Learning from Imbalanced Data Sets II*.

Duda, R.O. Hart, P. and Strok, D. (2001). *Pattern classification and scene analysis*. Wiley.

Estabrooks, A. Jo, T. and Japkowicz, N. (2004). A multiple resampling method for learning from imbalances data sets. In *Computational Intelligence.20*.

http://archive.ics.uci.edu/ml/.

http://mlg.ucd.ie/datasets.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: Systematic study. In *Intelligent Data Analysis Journal 6(5)*.

Kubat, M. and Matwin, S. (1997). Addressing the curse of imbalanced data sets: One-sided sampling. In *ICML'04,Proceedings of the Fourteenth International Conference on Machine Learning*.

Phua, C. Damminda, A. and Lee, V. (2004). Minority report in fraud detection: Classification of skewed data. In *Special Issue on Imbalanced Data Sets 6(1)*. ACM Sigkdd Explorations.

Vaswani, N. and Chellappa, R. (2006). Principal component null space analysis for image and video classification. In *IEEE Trans.Image Processing*.

Villalba, S. and Cunningham, P. (2008). An evaluation of dimension reduction techniques for one-class classification. In *Artificial Intelligence Review, 27(4)*.

Weiss, G. and Provost, F. (2003). Learning when training data are costly: The effect of class distribution on tree induction. In *Journal of Artificial Intelligence Research.19*.

Xudong, J. (2009). Asymmetric principal component and discriminant analyses for pattern classification. In *IEEE Trans. Pattern Analysis and Machine Intelligence 31(5)*.

Yoon, K. and Kwek, S. (2007). A data reduction approach for resolving the imbalanced data issue in functional genomics. In *Neural Computing and Applications 16(3)*.