Journal of
Computer Science & Systems Biology

Rollins and Pankayatselvan, J Comput Sci Syst Biol 2015, 8:1
DOI: 10.4172/jcsb.1000165

Research Article Open Access

A One-Dimensional PCA Approach for Classifying Imbalanced Data

Derrick K Rollins Sr[1,2]* and Varayini Pankayatselvan[3]

[1]Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa 50011, USA
[2]Department of Statistics, Iowa State University, Ames, Iowa 50011, USA
[3]Department of Chemical Engineering, New Mexico State University, Las Cruces, NM 88003, New Mexico

Abstract

Background: Highly complex and computational intensive methods based on Synthetic Minority Over-sampling Technique (SMOTE) and more recently Learning Vector Quantization SMOTE (LVQ-SMOTE) have been proposed for classification problems of imbalanced biomedical data. This works presents a much simpler approach that is not computationally intensive and competes well with existing approaches. It uses principal component analysis (PCA) to generate a pseudo-variable as a linear combination of the features. From this one pseudo-variable, several classification methods are developed that classify directly based on very simple statistics. One method, the Mean Method (MM), classifies cases based on closeness to the means for the two classes from training data sets. When the number of features is very large, a feature reduction (FR) procedure is proposed to reduce misclassifications. In cases where the means of both classes are similar but their spread about their means are different, the Spread Method (SM) is proposed. A unique feature of this method is that one is able to vary the accuracy of classification between the two classes by changing the width of the window for allocation of cases. These proposed methods are found to perform well without the use of over-sampling techniques and multiple-fold cross validation.

Results: The MM or the MM with FR was compared directly to recently published results for LVQ-SMOTE on six (6) data sets and gave better or much better results in every case as measured by adding the percent of true positives to the percent of true negatives. The SM was compared with LVQ-SMOTE on two (2) data sets and operating windows widths were obtained that gave much better results for the SM over LVQ-SMOTE.

Conclusion: Given the simplicity, strengths, and performance of the proposed approach in comparison to current methods, these methods and procedures are recommended for use in classification of imbalanced biomedical data applications.

Keywords: Biomedical data; Over-sampling, Learning vector quantization; Synthetic minority over-sampling technique; Principal component analysis

Abbreviations: MM: Mean Method; SM: Spread Method; STP: Sum of True Percent; TN: True Negative; TP: True Positive, FP: False Positive; FN: False Negative; SE: Sensitivity; SP: specificity; TR: Training; Ts: testing; CS: Class size; DL: Diving line; FR: Feature reduction; PCA: Principle Component Analysis

Background

Sets of data in which the distribution of the two classes are not equal are referred to as imbalanced [1,2]. These data sets can range from the fields of bioinformatics [3] to telecommunications management [4]. A standard classification method, in which a balanced distribution is assumed, has been difficult to apply to these imbalanced data sets. Furthermore, many classification methods link classifiers mainly to the majority class, thus producing poor classification results in the minority class and decreasing the overall performance of the classifier.

Common techniques to bypass this problem are via under or over sampling [5]. Under sampling is when the majority class in the data set is reduced to equal that of the minority class. Oversampling increases the minority class to equal that of the majority class. Each method, however, produces a bias when training the data.

Complex classification algorithms such as SMOTE, random forest and tree-based learning, are then applied to these under or over sampled data sets. A clustering algorithm is often implemented beforehand to produce distinct groups. K-means is a common clustering algorithm [6]. In this method, fixed centroids are initially randomly placed in the data set as far away from one another as possible. Each data point is then grouped into these centroids based on relative proximity. Problems with this method include whether or not the squared error function-to determine random proximity-will converge to a global minimum, rather than a local minimum [7]. Also, because the initial points at which the centroids are placed are at random, each titration will produce different results. To get the best results, the k-means algorithm can be run over and over, but this approach may take some time.

Others have modified these techniques, one of which includes the LVQ-SMOTE method of Nakamura et al. [8]. This article compares results obtained from the work of Nakamura et al. to our proposed one-dimensional PCA-based approach. We have developed two classification techniques in which the original data does not need to be over or under sampled.

Thus, this article presents a new approach to classification problems of imbalanced biomedical data. This approach is based on principal component analysis (PCA) and partially follows PCA methodologies in [9,10]. In the proposed approach, PCA is used to create p pseudo-variables that are linear combinations of the p features for data sets of

***Corresponding author:** Derrick K Rollins Sr, Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa 50011, USA, Tel: 515-708-2557; E-mail: drollins@iastate.edu

Received October 10, 2014; **Accepted** November 04, 2014; **Published** January 03, 2015

Citation: Derrick K Rollins, Pankayatselvan V (2015) A One-Dimensional PCA Approach for Classifying Imbalanced Data. J Comput Sci Syst Biol 8: 005-011. doi:10.4172/jcsb.1000165

the two classes [11-13]. One, and only one, of the p pseudo-variables are selected for use in the methodologies. This is the one that gives the greatest difference between the mean values for the two classes when using the proposed mean method (MM) or the greatest difference between the spread for the two classes when using the proposed spread method (SM). The MM is not used when the two means are very close together and the SM gives sufficient differences in spread that can provide accurate and acceptable classification accuracy.

The results of our proposed methods are compared directly to the results in Nakamura et al. (i.e., their Learning Vector Quantization-Synthetic Minority Over-sampling Technique, or "LVQ-SMOTE"). Nakamura et al. reported results on eight bench-mark data sets. This work is compared to LVQ-SMOTE on seven of these data sets. The data set that is not evaluated in this work was not done because the amount of missing data was too large to obtain any meaningful results.

The classification statistics are determined from training data with m and n samples for each class [14]. Since m and n are usually quite large, these statistics will have small standard errors and will be close to the population (true) values. Thus, in these cases k-fold cross-validation is unnecessary. We demonstrate this strength of the proposed approach by showing excellent agreement for two studies that compare 10-fold cross-validation results and results that use only the means from the two classes without any cross validation [15].

Strength of the proposed approach is its ability to treat imbalanced data sets directly. Since the proposed methods are based on the exploitation of differences in centrality and spread for a selected pseudo variable, as long as m and n are sufficiently large, their effectiveness do not depend on m and n being the same as long as both are sufficiently large. Thus, preconditioning for treating imbalance data is not needed for this approach. Moreover, we compare the results of this work that use the original imbalanced data sets with LVQ-SMOTE that used over-sampling to obtain balanced data as indicated in Nakamura et al.

When the number of features is very large, for example in gene expression data, the standard error of the selected pseudo variable can be inflated by the inclusion of many features with negligible impact. Removing these features can reduce the standard error of the pseudo variable and thus, improve the classification accuracy. This works presents a feature reduction approach that follows the work of Rollins et al. and Rollins and Teh [9,10] and is based on ranking the contribution of each feature and eliminating features that give less than a certain amount of contribution. This procedure is applied in two case studies involving the MM.

Methods

This section describes in detail the MM and the SM. In addition, a methodology will also be presented for ranking the features and removing ones with lower rank. This procedure is called the feature reduction (FR) technique and can be applied prior to using the MM or the SM.

The approach in this work is based on use of PCA to find one pseudo-variable that is a linear combination of the features that gives large separation between the two classes or gives very different variances about the means of the two classes. Thus, both proposed method use basic statistical concepts and are one dimensional in application.

Mean Method (MM)

The MM will be described first. Let $\mathbf{X}_{tr}$ be a given p_{tr} by q matrix with $p_{tr} - n$ rows of Class 1 data and n rows of Class 2 data that are stacked one on top of the other one, and with q columns of features. Let $\mathbf{X}_{ts}$ be a given p_{ts} by q matrix with $p_{ts} - m$ rows of Class 1 data and m rows of Class 2 data that are stacked one on top of the other one, and with the same order of the q columns of features. The steps for the applying the MM to $\mathbf{X}_{tr}$ for one fold or multiple folds in a cross-validation study are given as:

1. Standardize the columns of $\mathbf{X}_{tr}$ and obtain the matrix $\mathbf{Z}_{tr}$. With the number of rows in the training data as p_{tr} and the number of features as q, then $\mathbf{Z}_{tr}$ is a p_{tr} by q matrix.
2. Standardize the columns of $\mathbf{X}_{ts}$ by using the means and standard deviations from the same columns of the training data set and obtain the matrix $\mathbf{Z}_{ts}$. With the number of rows in the training data as p_{ts} and the number of features as q, then $\mathbf{Z}_{ts}$ is a p_{ts} by q matrix.
3. Note that all steps below are done on standardized data sets.
4. Do PCA (Johnson and Wichern [1]) on the Class 1 Training (Tr) data in the $\mathbf{Z}_{tr}$ matrix. Obtain loading scatter plots for the 1st k principal components (PCs).
5. Do PCA on the Class 2 Tr data in the $\mathbf{Z}_{tr}$ matrix. Obtain scatter plots for the 1st k PCs.
6. For the PCs from Step 4, use them to obtain scores for the Class 2 Tr data in the $\mathbf{Z}_{tr}$ matrix and obtain k score plots from these scores.
7. For the PCs from Step 5, use them to obtain scores for the Class 1 Tr data in the $\mathbf{Z}_{tr}$ matrix and obtain k score plots from these scores.
8. Examine all the Score plots. Select the PC that gives the greatest separation of Classes 1 and 2.
9. Use the selected PC and plot the Tr Scores for both classes on one plot to confirm the separation.
10. Obtain $\bar{x}_1$ and $\bar{x}_2$ from the Scores in the plot from Step 9.
11. Using $\bar{x}_1$ and $\bar{x}_2$, classify the test data using the MM procedure given by Equation 3 below.
12. After classifying every sample in the Ts set, determine the summary statistics (i.e., TN, FP, TP, FN) for this fold using the results from Step 11.
13. Repeat Steps 1-12 for each fold.
14. After getting the individual results for each fold, obtain summary statistics for all the folds combined.

The classification rule needed for Step 11 is determined as follows. Johnson and Wichern [1] give the following classification rule under the assumption that both classes come from normal distributions with the same variance σ^2

Allocate x_0 to Class 1 if

$$(\mu_1 - \mu_2)x_0 - \frac{1}{2}(\mu_1^2 - \mu_2^2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_1}{p_2}\right)\right] \quad (1)$$

Allocate x_0 to Class 2 otherwise,

where x_0 is the pseudo observation to be classified, μ_1 and μ_2 are the true means for Class 1 and Class 2, respectively, $c(i|j)$ is the cost when observation from Class j is incorrectly classified as Class i, p_i

is the prior probability of Class i, $i = 1, 2$, with $p_1 + p_2 = 1$, and in this context of normalized variables, $\sigma = 1$. With $c(1|2) = c(2|1)$, $p_1 = p_2$ and substitution of the sample statistics, Equation 1 becomes

$$(\bar{x}_1 - \bar{x}_2)x_0 - \frac{1}{2}(\bar{x}_1^2 - \bar{x}_2^2) \geq 0 \quad (2)$$

such that, upon rearranging, gives the allocation rule in this work as

$$\text{Allocate } x_0 \text{ to } \begin{cases} \text{Class 1 if } x_0 \geq \frac{1}{2}(\bar{x}_1 + \bar{x}_2) \\ \text{Class 2, otherwise} \end{cases} \quad (3)$$

For PC j, let $l_{i,j}$ be the value of its i^{th} loading, $i = 1, \ldots, q$; then the m^{th} value for x_0 computed from this PC is obtained by

$$x_0 = l_{1,j}z_{m,1} + \cdots + l_{q,j}z_{m,q} \quad (4)$$

Figure 1 is a graphical example of the MM for the Breast [16] data set. In this figure, x_0 for the test data are plotted in sequence with the negative class first followed by the values of the positive class. In addition, lines are plotted representing $\bar{x}_1$ ("Mean Negative Class"), $\bar{x}_2$ ("Mean Positive Class"), and $(\bar{x}_1 + \bar{x}_2)/2$ "Mid-Point/Dividing Line"). The difference in the mean levels is quite apparent. Values below the "Dividing Line" (DL) are allocated to the negative class and values above the DL are allocated to the positive class. As shown in this plot, the percent of correct allocations for both classes is quite high. These results will be given later in the Results Section.

Spread Method (SM)

The steps for the applying the SM to $\mathbf{X}_{tr}$ for one fold or multiple folds in a cross-validation study are given as:

Follow Steps 1-7 in the MM procedure.

1. Examine all the Score plots. Select the PC that gives the tightest spread for one class while giving the greatest spread for the other class around the same mean for both classes.
2. Use the selected PC and plot the Tr Scores for both classes on one plot to confirm the very tight and very wide spread relationship between the two classes.
3. Set the upper and lower limits for the tightest spreading class.
4. Classify each sample vector in the Ts set. Allocate x_0 to the tightest spreading class if it is within the limits; otherwise allocate it to the other class.
5. After classifying every sample in the Ts set, determine the summary statistics (i.e., TN, FP, TP, FN) for this fold using the results from Step 5.
6. Repeat Steps 1-6 for each fold.
7. After getting the individual results for each fold, obtain summary statistics for all the folds combined.

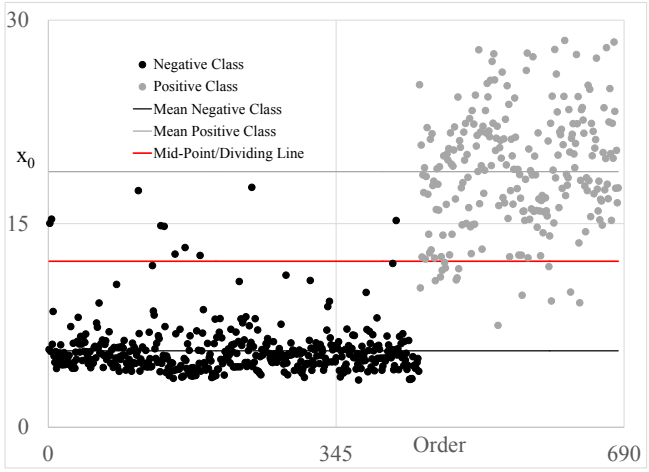

Figure 1: An example of the MM. This test data set is for Breast. The values of x_0 are plotted for negative class (black filled circles) first and then for the positive class (gray filled circles). Lines for the means of both classes are shown as well as their average value (the red line). Values above the red line are classified as being in the positive class and values below the line are classified as belonging to the negative class.

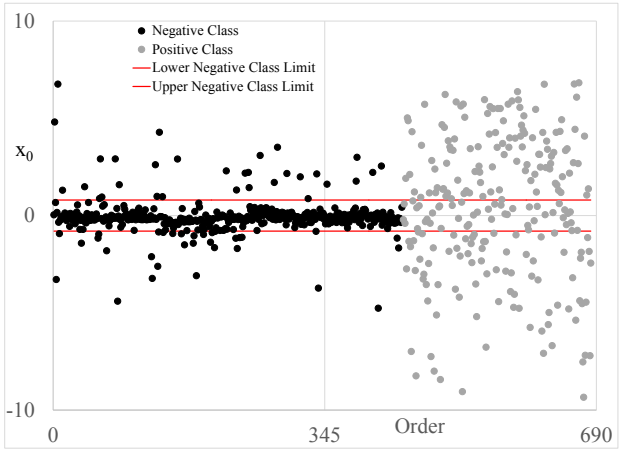

Figure 2: An example of the SM. This test data set is for Breast. The values of x_0 are plotted for negative class (black filled circles) first and then for the positive class (gray filled circles). The red lines are the upper and lower limits for being classified as belonging to the negative class. Thus, values inside the red lines are classified as belonging to the negative class and values outside the red lines are classified as belonging to the positive class. The percent of total variation for this PC is 8.6%.

Figure 2 is a graphical example of the SM for the Breast data set. As in Figure 1, in this figure, x_0 for the test data are plotted in sequence with the negative class first followed by the values of the positive class. In addition, lines are plotted representing the lower and upper limits for allocation to the negative class. The differences in the spread for the two classes are quite apparent. Values of x_0 that fall within the limits are allocated to the negative class; otherwise they are allocated to the positive class. While the accuracy of the allocations shown is not as high as the MM in Figure 1, it is still quite high and on par with LVQ-SMOTE as discussed in the Results Section later.

Results and Discussion

The Mean Method (MM)

The MM results for six data sets are given in Table 1 along with LVQ-SMOTE results from Nakamura et al. The results are given in terms of the number of true negatives (TN) and number of true positives (TP) as well as the percent of TN and TP. This is not how Nakamura et al. reported their results and we had to calculate TN and TP from the following two statistics that they used to report their results:

$$SE = \frac{TP}{TP + FP} \quad (5)$$

Data Set (PC % of Total Variation for MM)	# Features	# of Folds[1]	Size of Data Set		Testing Size		Current Method		Mean Method (MM)	
			Positive	Negative	Positive	Negative	TN (%)	TP (%)	TN (%)	TP (%)
Satimage (39.1%)	36	One	3397	3397	1530	1530			878 (57.4)	1495 (97.7)
	36		3397	3397	1530	1530	1159 (75.8)	1159 (75.8)		
Breast (65%)	(9)[2]	10	239	444					436 (98.2)	229 (95.8)
	(9)[2]	None	239	444					434 (97.7)	225 (94.1)
	10	10	444	444			301 (67.8)	409 (92.1)		
Blood (36.1%)	(3)[3]	None	178	570					343 (60.2)	136 (76.4)
	4	10	570	570			103 (18.1)	565 (99.1)		
Yeast (26.2%)	8	10	1433	1433					1150 (80.3)	1164 (81.2)
	8	None	1433	1433					1200 (83.7)	1125 (78.5)
	8	10	1433	1433			1301 (90.8)	329 (23.0)		
Colon Cancer (5.2 %)	(1788)[4]	None	40	22					20 (90.9)	34 (85.0)
	2000		40	22					18 (81.8)	30 (75.0)
	2000	10	40	40			29 (72.5)	Unknown[4]		
Leukemia (4.5%)	(2000)[4]	One	11	27	14	20			16 (80.0)	5 (35.7)
	7129		11	27	14	20			14 (70.0)	7 (50.0)
	7129		11	27	14	20	20 (100.0)	Unknown[4]		

Table 1: Results for the LVQ-SMOTE and MM for six (6) data sets

[1]The results are based on 10-fold, 1-fold or no (represented by "None") cross validation. "None" is applicable to MM only.

[2]For this data set there were only three true features, i.e., $\mathbf{X}_{tr}$ has a rank of 3.

[3]The number of features was reduced using the FR technique.

[4]It was not possible to determine TP for LVQ-SMOTE from the value of the statistics reported in Nakamura et al.

[5]These numbers represent the sizes of the classes for the training data sets.

[6]The balanced data sets for LVQ-SMOTE are determined by over-sampling except in the case of Colon Cancer.

[7]Oversampling is not used to determine the LVQ-SMOTE because the results are not possible.

Data Set	TN % + TP %	
	LVQ-SMOTE	MM
Satimage	151.7	155.1
Breast	160.0	[1]191.8
Blood	117.3	136.6
Yeast	114.1	[1]162.2
Colon Cancer	157.5	[2]175.9
Leukemia	Not Determinable	[2]115.7

Table 2: The sum of TP % and TN % reported in Table 1.

$$SP = \frac{TN}{TN + FP} \tag{6}$$

where SE and SP are called the "Sensitivity" and "Specificity," respectively, and FP is the number of false positives. Note that the number of false negatives is defined as "FN." When the class size (CS) of both classes are the same (i.e., balanced),

$$CS = TP + FN = TN + FP \tag{7}$$

Then from substituting Equation 7 into Equation 6 and solving for TN gives

$$TN = SP \cdot CS \tag{8}$$

From Equation 7

$$FP = CS - TN \tag{9}$$

Solving Equation 5 for TP gives

$$TP = \frac{SE}{1 - SE} FP \tag{10}$$

Therefore, again from Equation 7

$$FN = CS - TP \tag{11}$$

When CS_p=TP + FN, CS_n=TN + FP, where $CS_p > CS_n$, and CS_p and CS_n are the positive and negative class sizes, respectively, then

$$TN = SP \cdot CS_n \tag{12}$$

$$FP = CS_n - TN \tag{13}$$

$$FN = CS_p - TP \tag{14}$$

Equations 8 to 14 are used to obtain the LVQ-SMOTE results in this article. We prefer to report our results in terms of TP and TN because given these statistics and the class sizes one can obtain any summary statistic based on these results such as SE and SP. However, the reverse is not true. Depending on the summary statistics presented, it is not necessarily possible to obtain TN and TP. We will illustrate this limitation from the results reported by Nakamura et al. later in this section.

To simplify the comparisons with LVQ-SMOTE, we created a summary statistic that sums the % of TN and the % of TP. We call this statistic the sum of the true % or STP. STP values are reported in Table 2 and these results will be primarily used to compare the two approaches. Unless otherwise stated, all the results given are Ts results. Also note that STP is used in this work only as a way to compare the methods because it is based strictly on TP % and TN %. We are not advocating it as the sole measure of accuracy in all applications or situations. It is one way to measure accuracy and nothing more. If one prefers some other measure of performance such as G-means (i.e., the SE + SP divided by 2), the information is provided to obtain the statistic since the TP and TN values are given. We do not report G-means here, for example, because of the dependence on statistics with limitations which makes it limited also.

The first data set in Table 1 is called "Satimage" [16]. It consists of

fixed Tr and Ts sets. Thus, only one-fold cross validation was done and under balanced data as given. For LVQ-SMOTE, the TN % and TP % are both 75.8% and for the MM they are 57.4% and 97.7%, respectively. STP for LVQ-SMOTE and MM are 151.6 and 155.1, respectively. Hence, although the TN and TP percentage are very different, the STP values are very close.

The next data set in Table 1 is called "Breast" [16]. For the MM, results were obtained two ways: using 10-fold cross validation and directly from using the sample means based on all the pseudo values for each class. A *k*-fold cross validation study can provide information on not only the mean level of classification accuracy but also its variability in accuracy for the size of data in the Tr set. For confidence in the estimate of the variability *k* must be sufficiently large. However, when classification allocation is based solely on means, as in the MM, accuracy is maximized by using all the available data to estimate the means for the classes since the standard errors for a sample mean decreases with increasing sample size. Nonetheless, to show that it is unnecessary to do 10-fold validation for the MM, this is one of two data sets that give results with 10-fold cross validation and without cross validation. For the MM without cross validation, the TN% and TP% for 10-fold cross validation and the means based on all the data without cross validation are 98.2% and 95.8%, and 97.7% and 94.1%, respectively. The closeness of these results confirms that cross validation is not necessary. Actually, since the results without cross validation are based on larger sample sizes to estimate the means for each class, they are more reliable.

For the Breast data set, the TN % and TP % for LVQ-SMOTE are 67.8% and 92.1%, respectively. Both values are lower than the MM and TN is considerably lower. STP for LVQ-SMOTE and the MM based on averages for all the data in the classes are 160.0 and 191.8, respectively, as shown in Table 2.

We found the Blood [16] data set to have only three (3) independent features and eliminated one from our analysis. For this case the means were determined for the MM using all the data for each class only (i.e., we did not do 10-fold cross validation). The TN % and TP % are 60.2% and 76.4%, respectively. For LVQ-SMOTE they are 18.1% and 99.1%, respectively. Thus, LVQ-SMOTE results are more at the extremes while the MM is more balanced in these values. Thus, for LVQ-SMOTE, it seems to have such a high TP level, but the TN level is sacrificed. STP for LVQ-SMOTE and the MM are 117.2 and 136.6, respectively, as shown in Table 2.

The next data set is called "Yeast" [16] For the MM, 10-fold cross validation and no cross validation results for TN % and TP % are reported in Table 1 as 80.3% and 81.2%, and 83.7% and 78.5%, respectively. Thus, these results are in excellent agreement and support our conclusions given for the Breast data set regarding the sufficiency of determining the results for the MM without the need for cross validation.

For the Yeast data set, LVQ-SMOTE TN % and TP % are 90.8% and 23.0%, respectively. Thus, like the previous data set, although in a reverse manner, LVQ-SMOTE results are more at the extremes and the MM is more balanced in its classifications. Thus, for LVQ-SMOTE, it seems to have such a high TN level, the TP level is sacrificed. STP for LVQ-SMOTE and MM based on averages for all the data in the classes are 114.1 and 162.2, respectively, as shown in Table 2.

The Colon Cancer [17] data set is the first of two with a very large number of features and much smaller number of samples in both classes. The features in this and the Leukemia [18] data set are genes in microarray data sets. For both these data sets we apply the MM to all the features and to a reduced set of features. Our feature reduction (FR) technique is based on the work of Rollins et al. and Rollins and Teh [9,10]. This technique obtains loadings by applying PCA to the standardized matrix, $\mathbf{Z}_{tr}$. For any PC, the magnitude of a loading is its relative contribution to the pseudo-variable $x_{0.}$ For the selected PC, the loadings are ranked and plotted against its rank as shown in Figure 3 in this case of Colon Cancer data set. As shown, this plot is highly nonlinear with decreasing rank from left to right (i.e., the higher the rank number the lower the contribution). As shown, as the rank decreases, the highest contributing loadings drop off rapidly, then fairly linearly for a large number of loadings, and then drops off quite rapidly again. From this plot we chose to eliminate all loadings above rank numbers 1788, the point where the drop off is very rapidly after the linear period.

The Colon Cancer case in Table 1, for the MM, contains results with and without the elimination of features (i.e., FR). The FR results are significantly better than the results with all the features used.

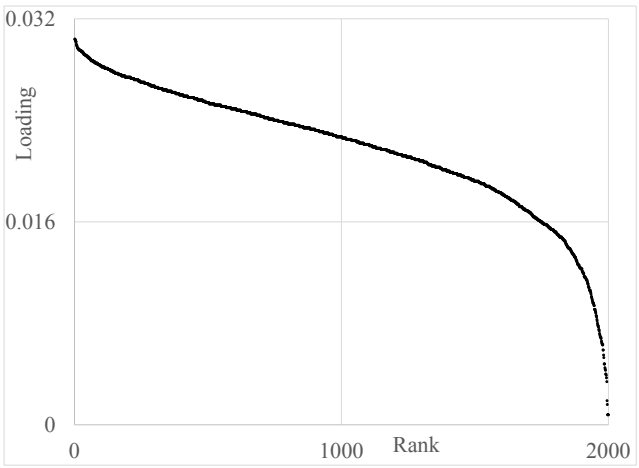

Figure 3: An example of loadings for a PCA plotted against their magnitude which is proportional to their contribution. The loadings are for the selected PC for the MM in the Colon Cancer case. The highest ranks are on the left and the rank decreases as the value on the x-axis increase.

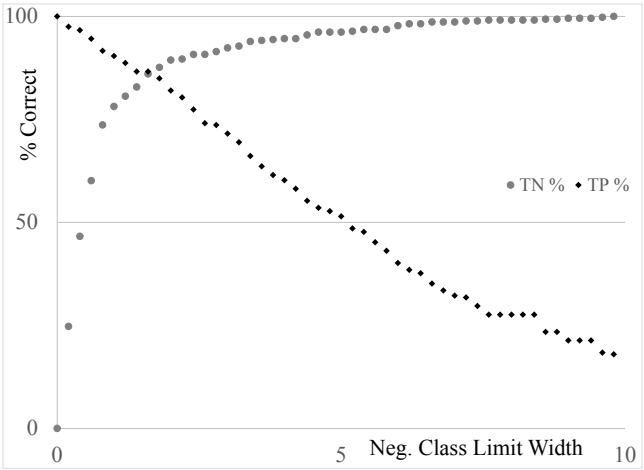

Figure 4: SM results for the Breast case showing how TN % and TP % vary with increasing range for the negative class limits.

The MM FR TN % and TP % are 90.9% and 85%, respectively. The corresponding LVQ-SMOTE results are 72.5% and 85%, respectively. However, we are uncertain of these values based on the information contained in Nakamura et al. Upon assuming balanced data (i.e., 40 samples in both classes), it is not possible to get realistic numbers for TP % and TN % using the SE and SP values reported by Nakamura et al. Therefore, to get realistic numbers we used the original sample sizes for both classes. STP for LVQ-SMOTE and the MM with FR are 157.5 and 175.9, respectively, as shown in Table 2.

The last data set in Table 1 is Leukemia (18). This data set has the most features (7129) and the fewest samples in each class; 11 and 27 for the negative and positive classes, respectively. Thus, the information for classification is very poor. In an effort to improve the quality of the selected pseudo-variable, the FR technique was applied and only the top 2000 features with the greatest contribution were kept for use in the MM. This modification resulted in an increase in the TN % from 70% to 80%. However, the TP percentage decreased from 50% to 35%. Nonetheless, this still could be a more accurate number since there are only 11 samples in the positive class and hence, not much information to achieve very high accuracy. Thus, the FR technique appears to also be improving the accuracy of classification here. For LVQ-SMOTE it is not possible to determine TP for this case because their SE value was reported to be 1.0 and Eq. 10 is not solvable. STP for the MM is 115.7 as given in Table 2. This is the lowest value for all the cases for the MM and it is the only one with a value below 50% for any statistic. The poor TP accuracy is consistent with this case having a very small number of samples, 11, for the positive class for training. In addition, the information quality is also weakened by the very large set of features.

The Spread Method (SM)

When a PCA pseudo-variable shows very different spread about class means, the SM may provide accurate classification. For the data sets evaluated in Nakamura et al. using their LVQ-SMOTE method, we found two of them to be good choices for the SM. The first one is the Breast data set. As discussed above, it has already been evaluated by the MM. The second one, Ionosphere [16], was not found to be suited for the MM. This data set consists of 34 features, 225 samples in the negative class and 126 samples in the positive class. SM results for the Breast and Ionosphere data sets are given in Figure 4 and 5, respectively, where the TN % and TP % are plotted against the width of the limits for the negative class, the one with the tightest spread. The crossover point where TN % = TP % is about 86% and 76% for Breast and Ionosphere, respectively, as shown in these figures. Thus, the SM gave better classification results for the Breast data set.

Figures 6 and 7 are plots of TP % versus TN% for the results in Figures 5 and 6, respectively. It is desirable for this type of plot to be in the upper right corner where both TP % and TN % are high. From comparing these plots, the higher accuracy of the Breast data set is seen by its curve being higher in the upper right corner. In addition, these figures have the LVQ-SMOTE result plotted as a point on these graph. For the Breast case, this result is TN % = 67.8% and TP % = 92.1%. This point is right on the SM curve in Figure 6. The two results for the MM (TN % = 98.2% and TP % = 95.8%; and TN % = 97.7% and TP % = 94.1%) are also plotted on Figure 6. These two points are high in the upper right hand corner indicating very high accuracy. For the Ionosphere case, the LVQ-SMOTE results are TN % = 92.4% and TP % = 47.1%. This point is significantly below the SM curve in Figure 7, and indicates lower accuracy of the LVQ-SMOTE method for this data set.

Conclusion

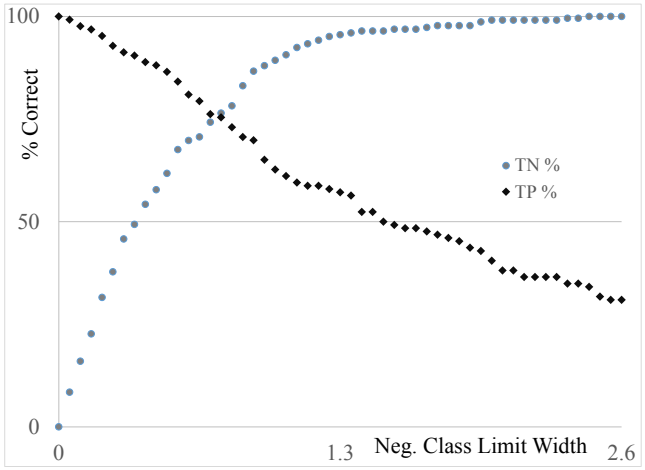

Figure 5: TP % versus TN% for the results in Figure 4. The MM and LVQ-SMOTE results are also given on this plot.

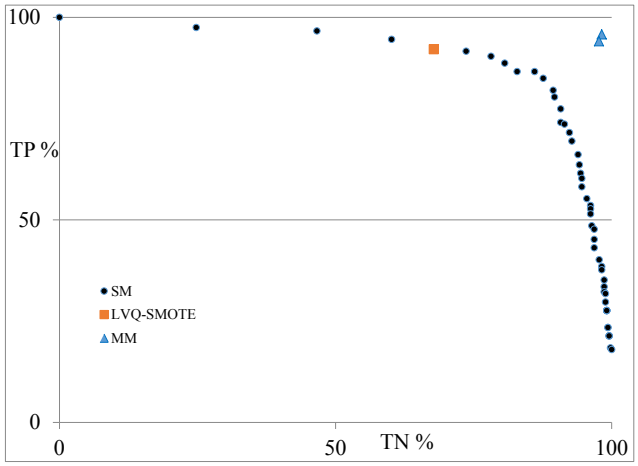

Figure 6: SM results for the Ionosphere case showing how TN % and TP % vary with increasing range for the negative class limits. The percent of total variation for this PC is 1.1%.

This article presented a new one-dimensional PCA approach for classifying subjects or objects in binary population studies in imbalanced biomedical data sets in particular, but also for any classification problem of this type in general. This approach performs PCA on the training data set, and then selects a pseudo-variable (x_0) that is linear combination of the features. For the MM, this pseudo-variable is the one that gives the greatest mean difference between the pseudo-data in the two classes. For the SM, this variable is the one that gives the greatest difference in spread between the two classes. This variable and method is selected when it can give high accuracy for the training data. A noteworthy and unique strength of the SM is the ability to change the accuracy levels for the classes by changing the range of the limits for the class with the tightest spread. Thus, depending on the cost of misclassification, the SM allows maximization of the class with the greatest misclassification cost over the other class. In every case the proposed approach was shown to be as accurate or more accurate as the LVQ-SMOTE reported results in Nakamura et al. Other strengths of the proposed approach includes a procedure to reduce the number of

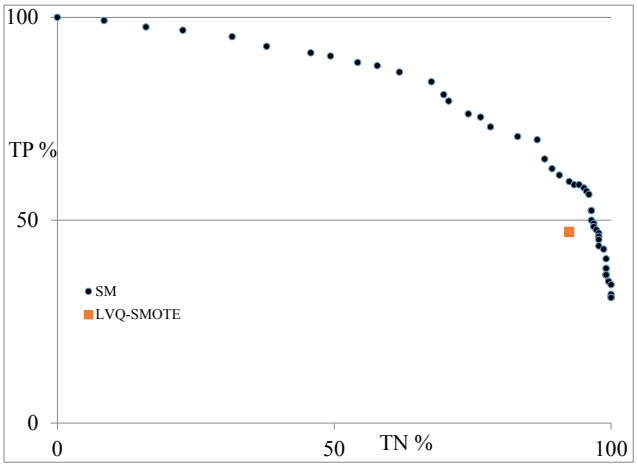

Figure 7: TP % versus TN% for the results in Figure 6. The LVQ-SMOTE result is also given on this plot.

features called the feature reduction (FR) technique based on the work in [9,10]. This approach uses PCA to rank the features and keeps only the most highly ranked ones. Future work will involve how to better optimally determine the reduced set of features. FR is a technique that can help to improve accuracy by eliminating features that contribute more noise than signal to value of the pseudo-variable which helps to reduce its standard error and improves classification accuracy. For the proposed approach, cross validation and techniques to balance data for the two classes is not needed. When the sizes of training data for both classes are sufficiently large to estimate the means under a small standard error, the methods should give accurate classification as evidence by the cases studied in this work. Thus, our overall recommendation is for the use of these techniques in these types of classification studies. Their simplicity alone is a critical advantage that can likely justify their use.

When the mean method (MM) fails this is seen in its inability to separate the classes in the training data. We were fortunate, in that, when this occurred in the data sets in this study, the spread method (SM) came through. In practice, this may not always be the case. Future work could consist of consideration of more than one PCA when the MM and the SM fails. Future work could also consist of Monte Carlo Simulation Studies that could provide guidance on the best approach to use in a particular situation. The situations would have to be defined by experts in this area and this type of study would likely be quite extensive and evolving as more knowledge and understanding of the different situations grow. Given the performance of the proposed methods in this work, it appears that their inclusion into such studies will be of merit.

References

1. Johnson AR, Wichern WD (1998) Applied Multivariate Statistical Analysis. 6th Edition, New Jersey: Prentice-Hall, Inc.
2. Cherkassky V, Ma Y (2009) Another look at statistical learning theory and regularization. Neural Netw 22: 958-969.
3. Davis JC (1996) Statistics and Data Analysis in Geology. New York: John Wiley and sons.
4. Weiss MG (2005) Data Mining in Telecommunications. Data Mining and Knowledge Discovery Handbook.
5. Dey TK, Giesen J, Goswami S, Hudson J, Wenger R, et al. (2001) Undersampling and Oversampling in Sample Based Shape Modeling. In IEEE Visualization. IEEE Computer Society 83-90.
6. Kanungo T, Mount D, Netanyahu N, Piatko C, Silverman R, et al. (2002) An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24: 881-892.
7. A tutorial on clustering algorithms.
8. Nakamura M, Kajiwara Y, Otsuka A, Kimura H (2013) LVQ-SMOTE - Learning Vector Quantization based Synthetic Minority Over-sampling Technique for biomedical data. BioData Min 6: 16.
9. Rollins DK, Teh A (2010) An extended data mining method for identifying differentially expressed assay-specific signatures in functional genomic studies. BioData Min 3: 11.
10. Rollins DK, Zhai D, Joe AL, Guidarelli JW, Murarka A, et al. (2006) A novel data mining method to identify assay-specific signatures in functional genomic studies. BMC Bioinformatics 7: 377.
11. Multiscale Principal Component Analysis.
12. Smith IL (2006) A tutorial on principal components analysis. Cornell University, USA 51: 52.
13. Annotated SPSS Output Principal Components Analysis.
14. Akobeng AK (2007) Understanding diagnostic tests 1: sensitivity, specificity and predictive values. See comment in PubMed Commons below Acta Paediatr 96: 338-341.
15. An Introduction to Cross Validation.
16. Frank A (2010) Asuncion A: UCI Machine Learning Repository. Irvine, CA.
17. Alon U, Barkai N, Notterman D, Gish K, Barra S, et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96: 6745-6750.
18. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537.