

A Cost-sensitive Learning Strategy for Feature Extraction from Imbalanced Data

Ali Braytee¹, Wei Liu² and Paul Kennedy¹

¹ Quantum Computation and Intelligent Systems, School of Software, University of Technology Sydney, New South Wales 2007, Australia.

² Advanced Analytics Institute, University of Technology Sydney, New South Wales 2007, Australia.

{Ali.Braytee, Wei.Liu, Paul.Kennedy}@uts.edu.au

Abstract. In this paper, novel cost-sensitive principal component analysis (CSPCA) and cost-sensitive non-negative matrix factorization (CSNMF) methods are proposed for handling the problem of feature extraction from imbalanced data. The presence of highly imbalanced data misleads existing feature extraction techniques to produce biased features, which results in poor classification performance especially for the minor class problem. To solve this problem, we propose a cost-sensitive learning strategy for feature extraction techniques that uses the imbalance ratio of classes to discount the majority samples. This strategy is adapted to the popular feature extraction methods such as PCA and NMF. The main advantage of the proposed methods is that they are able to lessen the inherent bias of the extracted features to the majority class in existing PCA and NMF algorithms. A series of experiments on twelve public datasets with different levels of imbalance ratios show that the proposed methods outperformed the state-of-the-art methods on multiple classifiers.

1 Introduction

The class imbalance issue is caused by unequal distributions of the data between class labels [5]. It occurs due to a paucity of cases, for example, patients with a rare disease, or difficulties in collecting samples due to high cost or privacy. The imbalanced class is considered as a crucial issue in machine learning and data mining due to two reasons: firstly, learning from an imbalanced dataset leads to poor classification because classical data mining algorithms tend to favor classifying examples as belonging to the majority class (negative class). Thus, these base learning algorithms would be incapable of classifying the instances of the minority class (positive class), which are considerably the class of interest. Secondly, it is a common problem in many real world domains including those related to biomedical, financial data and others.

In the literature, three categories of techniques have been proposed to solve the class imbalance issue: the data-level, the cost-sensitive, and the algorithm-level. Firstly, the data-level category or sampling techniques modify the distribution of dataset to make it balanced by using under-sampling or over-sampling methods. The former reduces the majority class samples such as “ABC-Sampling” [3] and others. However, this may exclude important information by removing

useful majority samples [15]. The latter approach increases the number of minority class samples by duplicating or creating new synthetic samples. Oversampling approach does not lose information since all majority samples are retained [6]. However, duplicating samples approach is more prone to over-fitting. Several methods proposed in oversampling approach such as “SMOTE” [6], Adaptive Synthetic Sampling (ADASYN) [9], and others. Secondly, cost-sensitive supervised learning term has been proposed in classification environment to increase the cost on the misclassified instances [8]. Finally, the algorithm-level category modifies existing learning algorithms to tend towards the minority samples. For example, Class Confidence Proportion Decision Tree (CCPDT) [13] has been proposed to modify the classical decision trees algorithm. However, these algorithms solve the class imbalance problem at the classifier level.

On the other hand, feature extraction techniques are considered as a preprocessing method to reduce the number of dimensions in the dataset. In the case of imbalanced data, the extracted features are biased to predict the majority class samples that lead to poor performance on classification [14]. Principal component analysis (PCA) [11] and non-negative matrix factorization (NMF) [12] are very well-known feature extraction methods. The unsupervised PCA algorithm seeks the orthogonal feature extractors that maximize the total variance. Therefore, the extracted features favors majority class because there number is much more than the minority class.

NMF has recently been shown to be a very effective matrix factorization technique in approximating the high dimensional data [12]. It is a vector space method that uses matrix factorization to find two non-negative reduced-dimension matrices W and H [12]. The factorized matrices W and H will be affected by the imbalance problem and the basis matrix W will be biased to represent the majority class samples, because the magnitude of the residual squared error of negative class instances is much more than the positive ones.

Recently, supervised feature extraction methods are proposed to enhance the classification capability of the factorized matrices by utilizing the existence of class labels [10]. However, to our knowledge, there is no cost-sensitive learning approach to solve the imbalance problem in the supervised feature extraction techniques. In fact, we are unaware of any previous studies that address the imbalanced class problem in the most widely used feature extraction approaches namely NMF and PCA.

In this paper, we propose a cost-sensitive approach for classical PCA and NMF feature extractions which is able to solve the imbalanced class problem without modifying the existing base classifiers, or changing the original information of the training datasets. Our approach is validated on highly imbalanced datasets, and the extracted features are classified on popular classifiers such as SVM, k -NN, Naive Bayes and decision trees (CART).

Specifically, **the contributions** of this paper are:

1. We propose an effective cost-sensitive strategy that can improve general feature extraction methods for imbalanced data.

2. We apply the cost-sensitive learning strategy to the popular classical feature extraction methods PCA and NMF to handle the imbalance class problem.
3. We provide comprehensive evaluations of our methods on many real-world imbalanced datasets which show the advantages of our methods.

This rest of this paper is organised as follows. Section 2 describes a motivating example. Section 3 presents the theoretical analysis. Section 4 reports experiments before concluding in Section 5.

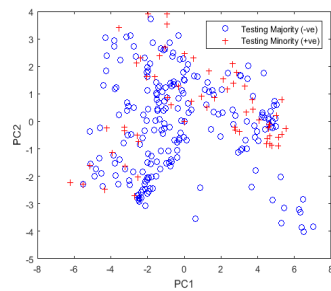
2 A Motivating Example

A training dataset is comprised of instances $\mathbf{X} \in \mathbb{R}^{n \times m}$ with m features. In a binary case, these instances belong to two different classes which are known as the majority (negative) and minority (positive) class respectively. When the data is imbalanced, the number of majority instances greatly outnumbers the instances of minority class. In this section, a vehicle dataset is used and we postpone the description of the characteristics of this dataset to Table 1.

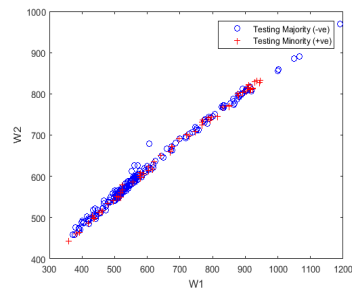
A major motivation for proposing a cost-sensitive feature extraction method is that the classical PCA and NMF ignore the minority samples during the feature construction that negatively affects the performance of classifiers and leads to a overlap class problem that increases with the imbalance ratio.

Figures 1a and 1b show the distribution of the testing samples on original unbalanced data after a classical PCA and NMF is applied on it respectively. It clearly shows that the classifiers will find it very difficult to distinguish between classes due to the overlap class problem.

Thus, integrating the feature extraction methods with the labelling information can often improve the performance of supervised learning algorithms as shown in section 4, because it generates extracted features that consider both classes and corrects the overlap class problem by separating the two classes.



(a) Testing data after projecting the instances using PCA



(b) Testing data after projecting the instances using NMF factors

Fig. 1: Applying PCA and NMF on imbalanced Vehicle data leads to overlapping problem

3 Theoretical Analysis

In this section, we present the theoretical analysis of our proposed cost-sensitive PCA (CSPCA) and cost-sensitive NMF (CSNMF).

3.1 Imbalance cost ratio

Different cost ratios are used in all training examples for the majority and minority classes:

$$C_i = \begin{cases} C_i^- = \frac{(1-\alpha)}{N_-}, & \text{If } y_i = -1, (\text{Negative class}) \\ C_i^+ = \frac{(1+\alpha)}{N_+}, & \text{If } y_i = +1, (\text{Positive class}) \end{cases} \quad (1)$$

for $i = 1, \dots, N$ samples, N_- and N_+ are the total number of negative and positive samples respectively. $0 \leq \alpha < 1$ is a parameter to weight the majority class. If $\alpha = 0$, the majority class is weighted by the ratio of the two class sizes in training data. In the case, where the α value gets close to 1, it gradually represents the learning from the positive examples only.

3.2 Cost-sensitive Principal Component Analysis (CSPCA)

Principal component analysis (PCA) [11] is one of the most popular feature extraction techniques. It is defined as a statistical procedure that uses an orthogonal transformation to construct a low-dimensional representation of the data known as principal components. The linear transformation aims to maximize the global variance of the data as well as to minimize the least square error of the transformation. The first principal component represents the largest variance of the data; the remaining principal components have smaller variance and orthogonal to the preceding ones. Consider a data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, where each of the n rows represent the instances or observations, and m columns are the dimensions.

Mathematically, the first loading \mathbf{w}_1 is computed by:

$$\mathbf{w}_1 = \arg \max \sum_{ij} (\mathbf{X}_{ij} \cdot \mathbf{w}_{1j})^2 \quad (2)$$

where i and j are the index of rows and columns of \mathbf{X} respectively, \mathbf{w}_1 is the first principal component of p dimensions and $p \ll m$. In the case of the class imbalance issue, the spread of data is dominated by the majority samples, because when the directions of principal axes (components) of both classes are different, the reduced space found by PCA represents the majority space and under-represents the minority one.

Geometrically, the first step in PCA is to centre the data by subtracting the mean of the data from all points. However, in the case of imbalance class, the global mean may be shifted to the majority samples space. Moreover, PCA computes the covariance matrix of the data which captures the variance of the dataset. But, in the case of highly skewed data, the covariance matrix mostly represents the variance of majority class samples, and the largest variance direction of the data may be captured mostly from the majority space.

Therefore, we propose a cost-sensitive PCA technique (CSPCA) to improve the computations of the principal components with consideration of the imbalanced class issue. In the binary case, assume that the negative and positive samples are discounted by imbalance cost ratio C^- and C^+ respectively. The weighted first principal component becomes:

$$\mathbf{w}_1 = \arg \max \sum_{i: y_i = -1, j} (C_i^- \mathbf{x}_{ij} \cdot \mathbf{w}_{1j})^2 + \sum_{i: y_i = +1, j} (C_i^+ \mathbf{x}_{ij} \cdot \mathbf{w}_{1j})^2 \quad (3)$$

where C_i^- and C_i^+ from Eq. (1), and j is the dimension index. Using the different cost ratios for the negative and positive class lead to lessening the dominant effect of the negative samples on the extracted features.

3.3 Cost-sensitive Non-negative Matrix Factorization (CSNMF)

Non-negative Matrix Factorization (NMF) [12] is a matrix factorization technique under the constraint that the values of the input matrix are non-negative. NMF can be described by the following factorization form

$$X_{n \times m} \simeq W_{n \times p} H_{m \times p}^T \quad (4)$$

where n is the number of observations, m is the dimension of the data, p is the desired rank such that $p < \min(m, n)$ and $X \in \mathbb{R}^{n \times m}$, $W \in \mathbb{R}^{n \times p}$, $H \in \mathbb{R}^{m \times p}$.

To find the approximate matrix factorization (4), an optimization function is defined by [12] to *minimize* $\|X - WH\|^2$ with respect to W and H .

In the case of imbalanced data, we propose a cost-sensitive NMF (CSNMF), which injecting the classical unsupervised NMF with labelling information to take into consideration the imbalance class problem.

Our CSNMF function modifies the original matrix to alleviate the effectiveness of the negative samples, a new matrix X' is defined by

$$X' = \begin{bmatrix} C_i^- X_i, & \text{If } y_i = -1 \\ C_i^+ X_i, & \text{If } y_i = +1 \end{bmatrix} \quad (5)$$

where C^- and C^+ is defined in (1), $X' \in \mathbb{R}^{n \times m}$. CSNMF aims to find two non-negative matrices $W = [w_{ip}] \in \mathbb{R}^{n \times p}$ and $H = [h_{jp}] \in \mathbb{R}^{m \times p}$ whose products can estimate the balanced matrix X' . The objective function is the Euclidean distance between two matrices, it can be written as:

$$O = \|X' - WH^T\|^2 = (X' - WH^T) (X' - WH^T)^T \quad (6)$$

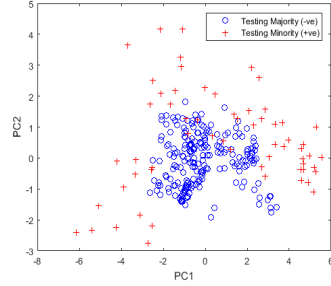
The objective function O is convex with respect to W and H separately. Lee and Seung [12] proposed iterative multiplicative update rules to minimise the error of O in Eq. (6):

$$h_{jp} \leftarrow h_{jp} \frac{(X'W)_{jp}}{(HW^TH)_{jp}} \quad w_{ip} \leftarrow w_{ip} \frac{(X'^TH)_{ip}}{(WH^TH)_{ip}} \quad (7)$$

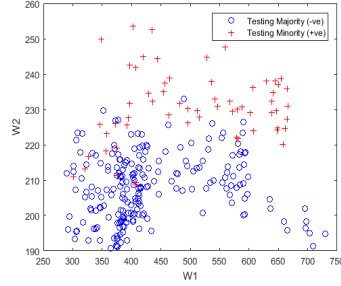
The convergence of the objective function with X' is the same as the objective with X proof in [12].

3.4 Revisiting the Motivating Example

This section visualizes the classification improvements after applying our two proposed methods CSPCA and CSNMF on the same dataset that we used in section 2. Fig. 2 shows how the classes can be better distinguished compared to Fig. 1.



(a) Testing data after projecting the instances using CSPCA



(b) Testing data after projecting the instances using CSNMF factors

Fig. 2: Applying CSPCA and CSNMF on imbalanced Vehicle dataset leads to improve the classification performance.

4 Experiments and Analysis

In this section, we analyse and compare the performance of CSPCA and CSNMF against classical PCA and NMF, existing ADASYN algorithm [9], random under-sampling (RU), maximum likelihood cost-sensitive (ML-CST) [7] and CCPDT [13]. We will make use of four classifiers, namely SVM, Naive Bayes, decision trees (CART), and k -NN as instance based learning approach.

4.1 Experimental Framework

In the first place, we define a set of parameter values for classifiers that have been used in all experiments. Without loss of generality, we select $k = 5$ neighbors in k -NN classifier. We choose the number of principal components which represent the 90% of the original data, and the desired rank for *CSNMF* equal the number of class labels because it is highly related to the cluster structure of the data [4]. Also, we set balance factor $\alpha = 0$ for imbalance cost ratio C in (1) to represent equal class proportions in training set.

We evaluated the methods on 12 datasets each with two classes. These datasets are selected from the UCI repository [2] and KEEL-datasets [1]. Details of the data sets are shown in Table 1. The AUC metric is used to measure

the accuracy in the experiments, which is widely used metric for evaluations on imbalanced data. Estimates of AUC was averaged over 5-fold cross-validation.

Dataset	#Instance	#Attributes	#IR	Dataset	#Instances	#Attributes	#IR
Diabetes	768	20	1.86	Blood	748	5	3.2
Glass	214	9	6.38	Colon	62	2000	1.82
Yeast	1484	8	28.1	Survival	306	3	2.77
Spambase	4601	57	1.53	Adeno	86	76	5.34
Breast	198	34	3.21	Ecoli	336	7	8.6
Vehicle	846	18	2.99	Abalone	4174	8	129.43

Table 1: Characteristics of the data sets. Column #IR is the imbalance ratio (i.e., Neg/Pos)

4.2 Analysis and Results

In this section, we analyse the behaviour of classical PCA and NMF with existing algorithms that proposed to handle the imbalanced class issue against our proposed CSPCA and CSNMF methods. Firstly, the state-of-the-art methods are composed of two methods from data sampling techniques: ADASYN [9] and random undersampling, one method from cost-sensitive level: ML-CST [7], and one method from algorithm level: CCPDT [13].

The obtained results in Tables 1 through 5 show that the performance of the classifiers substantially improves for the cost-sensitive version of PCA and NMF. Moreover, the proposed CSPCA and CSNMF with base classifiers outperformed the classical PCA and NMF with the state-of-the-art methods. Therefore, our proposed method can solve the imbalanced class issue at feature extraction level without the need of changing the data distribution or modifying the existing algorithms.

We conduct t -tests between vector results of our methods (Base) against the compared methods for each classifier, under the null hypothesis that the AUC on vectors of the used methods is not significantly different. As shown on the bottom line of Tables 1 through 5, the p -values reject the null hypothesis, as most values of our methods (base) are lower than 0.01. This indicates that the proposed methods CSPCA and CSNMF have significantly improved the performance of the classifiers.

Effectiveness of CSPCA and CSNMF on classification performance

We analyse the effects of our CSPCA and CSNMF algorithm on the base classifiers, and we compare it with baseline and existing algorithms that proposed for the imbalanced class problem. Firstly, we apply the data sampling algorithms on the training sets to balance the datasets using ADASYN oversampling method and undersampling method. Then, we apply the PCA and NMF on the training sets of imbalanced datasets to construct the extracted features. The projected test data is classified using the base classifiers such as SVM, Naive Bayes, Decision Trees and k -NN. Secondly, we apply our proposed CSPCA and CSNMF

Dataset	PCA	ADASYN - PCA	RU - PCA	CSPCA
Diabetes	.695	.706	.576	.762
Spambase	.869	.868	.872	.904
Breast	.562	.586	.556	.620
Ecoli	.860	.828	.857	.879
Abalone	.669	.627	.666	.683
Survival	.607	.546	.536	.611
Blood	.676	.676	.677	.660
Glass	.945	.933	.900	.951
Yeast	.850	.831	.839	.871
Vehicle	.634	.620	.628	.681
Colon	.512	.490	.517	.531
Adeno	.579	.525	.497	.610
t-test	3.4×10^{-3}	9.7×10^{-5}	5.3×10^{-3}	Base

(a) SVM Performance

Dataset	PCA	ADASYN - PCA	RU - PCA	CSPCA
Diabetes	.641	.694	.633	.71
Spambase	.857	.703	.815	.841
Breast	.503	.556	.557	.611
Ecoli	.722	.857	.864	.881
Abalone	.500	.645	.692	.711
Survival	.515	.532	.510	.590
Blood	.546	.655	.651	.741
Glass	.914	.922	.928	.912
Yeast	.514	.825	.821	.842
Vehicle	.508	.619	.638	.650
Colon	.532	.510	.552	.640
Adeno	.542	.414	.597	.620
t-test	1×10^{-3}	3.1×10^{-3}	2×10^{-3}	Base

(b) Naive Bayes Performance

Dataset	PCA	ADASYN - PCA	RU - PCA	CSPCA
Diabetes	.624	.645	.616	.689
Spambase	.811	.802	.819	.836
Breast	.515	.561	.542	.591
Ecoli	.563	.775	.849	.852
Abalone	.503	.522	.529	.615
Survival	.514	.474	.498	.600
Blood	.577	.575	.606	.641
Glass	.821	.867	.882	.891
Yeast	.543	.683	.808	.812
Vehicle	.555	.565	.595	.625
Colon	.587	.392	.590	.630
Adeno	.578	.518	.538	.565
t-test	4.2×10^{-3}	7.5×10^{-4}	1.4×10^{-3}	Base

(c) Decision Trees Performance

Dataset	PCA	ADASYN - PCA	RU - PCA	CSPCA
Diabetes	.632	.633	.536	.671
Spambase	.855	.828	.848	.862
Breast	.550	.590	.495	.620
Ecoli	.580	.823	.844	.856
Abalone	.500	.573	.621	.641
Survival	.509	.481	.458	.570
Blood	.609	.626	.623	.681
Glass	.930	.850	.925	.941
Yeast	.535	.721	.834	.880
Vehicle	.546	.581	.613	.600
Colon	.600	.472	.615	.631
Adeno	.476	.589	.503	.591
t-test	6.7×10^{-3}	1.1×10^{-3}	4.2×10^{-3}	Base

(d) 5-NN Performance

Table 2: CSPCA classification performance on imbalanced datasets compared to PCA and ADASYN-PCA.

on the training sets of the imbalanced datasets. Then, we use the same base classifiers as in the first case on the projected test data. Tables 2 and 3 show the preference of our CSPCA and CSNMF method over the existing data sampling techniques. The best average values per approach are highlighted in bold.

On the other hand, we compare our CSPCA and CSNMF method with the ML-CST using logistic regression as a base classifier. Then, we conduct another experiment, to compare CSPCA and CSNMF with CCPDT using decision trees as a base classifier, and in both cases our method outperformed the above two methods. Tables 4 and 5 show the quality of using our proposed solution for applying feature extraction on imbalanced datasets, as there is a significant difference between the results of our methods and the standard PCA, NMF and the other compared algorithms. One may also observe the generalization of our new method by improving the performance of classification on a set of the well-known classifiers.

As highlighted earlier, CSPCA and CSNMF methods improve the performance of classification for two reasons. Firstly, it tackles the overlap class problem by shrinking the majority samples. Secondly, the extracted features of CSPCA and CSNMF respectively consider the minority samples, even in highly imbalanced datasets where the imbalance ratio $\gg 20$.

Dataset	NMF	ADASYN - NMF	RU - NMF	CSNMF
Diabetes	.695	.699	.691	.721
Spambase	.627	.665	.636	.825
Breast	.573	.484	.516	.597
Ecoli	.868	.855	.840	.869
Abalone	.674	.577	.670	.693
Survival	.577	.523	.533	.621
Blood	.679	.566	.559	.756
Glass	.911	.832	.790	.925
Yeast	.844	.833	.817	.890
Vehicle	.645	.637	.600	.740
Colon	.752	.772	.752	.798
Adeno	.537	.620	.613	.723
t-test	5.5×10^{-3}	1.1×10^{-4}	1.9×10^{-4}	Base

(a) SVM Performance

Dataset	NMF	ADASYN - NMF	RU - NMF	CSNMF
Diabetes	.670	.693	.686	.743
Spambase	.580	.583	.582	.622
Breast	.547	.484	.516	.583
Ecoli	.850	.853	.833	.882
Abalone	.500	.669	.616	.712
Survival	.604	.489	.592	.621
Blood	.553	.561	.544	.680
Glass	.536	.760	.808	.870
Yeast	.548	.727	.726	.781
Vehicle	.621	.647	.643	.660
Colon	.687	.780	.672	.791
Adeno	.500	.475	.471	.660
t-test	1.8×10^{-3}	6.5×10^{-4}	2.5×10^{-4}	Base

(b) Naive Bayes Performance

Dataset	NMF	ADASYN - NMF	RU - NMF	CSNMF
Diabetes	.612	.613	.630	.750
Spambase	.709	.711	.699	.770
Breast	.526	.507	.523	.630
Ecoli	.684	.748	.882	.890
Abalone	.497	.543	.623	.651
Survival	.566	.558	.550	.680
Blood	.589	.624	.611	.710
Glass	.807	.760	.770	.830
Yeast	.593	.753	.734	.771
Vehicle	.617	.578	.617	.651
Colon	.692	.760	.727	.751
Adeno	.520	.471	.582	.620
t-test	3.8×10^{-5}	6.4×10^{-5}	2.6×10^{-4}	Base

(c) Decision Trees Performance

Dataset	NMF	ADASYN - NMF	RU - NMF	CSNMF
Diabetes	.631	.643	.658	.700
Spambase	.722	.724	.723	.800
Breast	.520	.500	.519	.591
Ecoli	.733	.831	.850	.880
Abalone	.500	.616	.682	.691
Survival	.530	.563	.535	.591
Blood	.598	.628	.615	.674
Glass	.788	.800	.782	.791
Yeast	.497	.789	.795	.821
Vehicle	.591	.581	.647	.691
Colon	.777	.727	.672	.791
Adeno	.460	.570	.586	.630
t-test	1.3×10^{-3}	5.9×10^{-5}	1.9×10^{-4}	Base

(d) 5-NN Performance

Table 3: CSNMF classification performance on imbalanced datasets compared to NMF, NMF-ADASYN and NMF-RU.

Logit	PCA- MLCST	CSPCA	NMF- MLCST	CSNMF
Diabetes	.604	.640	.625	.680
Spambase	.729	.810	.716	.720
Breast	.523	.656	.500	.580
Ecoli	.541	.728	.500	.610
Abalone	.499	.660	.500	.730
Survival	.543	.610	.500	.690
Blood	.510	.610	.500	.581
Glass	.915	.660	.500	.590
Yeast	.496	.574	.500	.570
Vehicle	.511	.634	.630	.691
Colon	.590	.664	.602	.642
Adeno	.510	.610	.500	.650
t-test	4.3×10^{-2}	Base	2.9×10^{-4}	Base

Table 4: CSPCA and CSNMF on logistic regression classifier compared to ML-CST

Decision Trees	PCA- CCPDT	CSPCA- DT	NMF- CCPDT	CSNMF- DT
Diabetes	.676	.689	.704	.750
Spambase	.918	.836	.756	.770
Breast	.487	.591	.487	.630
Ecoli	.498	.852	.498	.890
Abalone	.481	.615	.481	.651
Survival	.492	.600	.492	.680
Blood	.721	.641	.552	.710
Glass	.854	.891	.767	.830
Yeast	.491	.812	.634	.771
Vehicle	.652	.625	.671	.651
Colon	.472	.630	.528	.751
Adeno	.433	.565	.443	.620
t-test	3.3×10^{-2}	Base	9.5×10^{-4}	Base

Table 5: CSPCA and CSNMF CART classification performance compared to CCPDT

5 Conclusions and Future Work

Many existing studies have been proposed to solve the skewed data issue at the learning algorithm level. But, in the case of applying feature extraction techniques before classifying the data, these solutions cannot help. Applying classical feature extraction techniques on imbalanced data will yield biased features that favor the majority class.

This paper proposes a cost-sensitive learning strategy to address the imbalanced class problem that can be applied to well-known feature extraction techniques such as PCA and NMF. We have integrated the cost-sensitive strategy in PCA and NMF and proposed two new methods CSPCA and CSNMF. Our proposed method embeds the labelling information in the classical feature extraction methods to extract the balanced features which improve the accuracy of classification and reduce the overlapping between the classes. Our results show the high-performing quality of our proposed methods on multiple popular classifiers and benchmark datasets. They can deal with different levels of imbalance and sizes of the datasets. In future, we will extend the idea to multi-label classification problems. We also plan to adapt the strategy to other feature extraction techniques.

References

1. Alcalá, J., Fernández, A., et al.: Keel data-mining software tool: Data set repository. *Journal of Multiple-Valued Logic and Soft Computing* 17(2-3), 255–287 (2010)
2. Asuncion, A., Newman, D.: Uci machine learning repository (2007)
3. Braytee, A., Hussain, F.K., et al.: Abc-sampling for balancing imbalanced datasets based on artificial bee colony algorithm. In: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). pp. 594–599. IEEE (2015)
4. Carmona-Saez, P., Pascual-Marqui, R.D., Tirado, F., Carazo, J.M., Pascual-Montano, A.: Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC bioinformatics* 7(1), 1 (2006)
5. Chawla, N., Japkowicz, N., Kolcz, A.: Special issue on learning from imbalanced datasets, sigkdd explorations. In: ACM SIGKDD (2004)
6. Chawla, N.V., Bowyer, K.W., et al.: Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* pp. 321–357 (2002)
7. Dmochowski, J.P., Sajda, P., Parra, L.C.: Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. *The Journal of Machine Learning Research* 11, 3313–3332 (2010)
8. Elkan, C.: The foundations of cost-sensitive learning. In: International joint conference on artificial intelligence. vol. 17, pp. 973–978. Citeseer (2001)
9. He, H., Bai, Y., et al.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: Neural Networks, 2008. IJCNN 2008. IEEE International Joint Conference on. pp. 1322–1328. IEEE (2008)
10. Huh, S.i., Gupta, M.D., Xiao, J.: Supervised nonnegative matrix factorization (Jul 30 2013), uS Patent 8,498,949
11. Kirby, M., Sirovich, L.: Application of the karhunen-loeve procedure for the characterization of human faces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 12(1), 103–108 (1990)
12. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems. pp. 556–562 (2001)
13. Liu, W., Chawla, S., et al.: A robust decision tree algorithm for imbalanced data sets. In: SDM. vol. 10, pp. 766–777. SIAM (2010)
14. Wang, J., You, J., et al.: Extract minimum positive and maximum negative features for imbalanced binary classification. *Pattern Recognition* 45(3), 1136–1145 (2012)
15. Yen, S.J., Lee, Y.S.: Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications* 36(3), 5718–5727 (2009)