Multimed Tools Appl (2018) 77:10835–10851
https://doi.org/10.1007/s11042-017-5359-5

Cost-sensitive collaborative representation based classification via probability estimation with addressing the class imbalance

Zhenbing Liu[1] · **Chao Ma**[1] · **Chunyang Gao**[1] · **Huihua Yang**[1] · **Rushi Lan**[1] · **Xiaonan Luo**[1]

Received: 10 August 2017 / Revised: 22 October 2017 / Accepted: 25 October 2017 /
Published online: 3 November 2017

Abstract Collaborative representation has been successfully used in pattern recognition and machine learning. However, most existing collaborative representation classification methods are to achieve the highest classification accuracy, assuming the same losses for different misclassifications. This assumption, however, may not hold in many real-word applications as different types of misclassification could lead to different losses. Meanwhile, the class distribution of data is highly imbalanced in real-world applications. To address this problem, a novel Cost-Sensitive Collaborative Representation based Classification (CSCRC) method via Probability Estimation with Addressing the Class Imbalance was proposed. Unlike traditional methods, the class label of test samples is predicted by minimizing the misclassification losses which are obtained via computing the posterior probabilities. In this paper, a Gaussian function was defined as a probability distribution of collaborative representation coefficient vector and the probability distribution was transformed into collaborative representation framework via logarithmic operator. The experiments show that our proposed method performs competitively compared with existing methods.

Keywords Collaborative representation · Cost-sensitive learning · Probability estimate · Loss function

✉ Rushi Lan
rslan2016@163.com

Chao Ma
machao199271@sina.cn

[1] Guangxi Colleges and Universities Key Laboratory of Intelligent Processing of Computer Images and Graphics, Guilin University of Electronic Technology, No. 1 Jinji Road, Qixing Strict, Guilin 541000, China

1 Introduction

In recent years, sparse coding and representation have been widely studied to solve the various computer vision and machine learning problems [21]. In [20], Wright J et al. proposed a sparse representation based classification (SRC) method to solve the face recognition under varying illumination. In this method, an input test image is represented as a sparse linear combination of training images, and then the classification is performed by checking which class yields the least coding error. Such a SRC scheme has achieved a great success in face recognition and has been widely studied in the community. It is widely believed that the l_1-norm sparsity constraint on coding coefficients plays a key role in the success of SRC [16]. However, Zhang et al. [25] argued that the success of SRC should be largely attributed to the collaborative representation of a test sample by the training samples across all classes. To solve the shortage of training samples, they further proposed an effective collaborative representation based classifier (CRC) by utilizing l_2-norm regularization. Moreover, Lei Z et al. devoted to analyze the working mechanism of SRC in [19], and proposed a very simple yet much more efficient face classification scheme, namely collaborative representation based classification with regularized least square (CRC_RLS) [3]. The probability based classifiers is a popular type of classifier widely used in various visual recognition tasks, e.g., Probabilistic Support Vector Machine (PSVM) [14], Probabilistic Principal Component Analysis (PPCA) [8, 17] and Probabilistic Linear Discriminant Analysis (PLDA) [15]. Motivated by the work of probabilistic subspace methods [9, 12, 13], S Cai et al. analyzed the classification mechanism of CRC from a probabilistic viewpoint and proposed a Probabilistic Collaborative Representation based approach for pattern Classification (ProCRC) in [2], which jointly maximized the likelihood that a test sample belongs to each of the multiple classes. The final classification is performed by checking which class has the maximum likelihood.

Traditional classification algorithms [4, 10], including mentioned above, are designed to achieve the lowest recognition errors and assume the same losses for different types of misclassifications. However, this assumption may not be suitable for many real-world applications. For example, it may cause inconvenience to a gallery who is misclassified as an impostor and not allowed to enter the room controlled by a face recognition system, but may result in a serious loss if an impostor is misrecognized as a user and allowed to enter the room.

Cost-sensitive learning always co-exists with class imbalance in most applications with the goal of minimizing the total misclassification cost [22]. Class-imbalance has been considered as one of the most challenging problems in machine learning and data mining. The ratio of imbalance (the size of majority class to minority class) can be as huge as 100, even up to 10,000. Much work has been done in addressing the class imbalance problem. Because the cost that the positive class is misclassified as negative is higher than opposite, cost-sensitive learning is an effective method to deal with the imbalance data classification problem. Published solutions to the class imbalance problem can be categorized as data level and algorithm level approaches. At the algorithm level, solutions try to adapt existing classifier learning algorithms to bias towards the small class [7]. H Lu et al. constructed the filtering deep convolutional network and got a better result on marine organism classification than other methods [11].

In recent year, cost-sensitive learning has been studied widely and become one of the most important topics for solving the class imbalance problem. In [26], Zhou et al. studied empirically the effect of sampling and threshold-moving in training cost-sensitive neural networks, and revealed threshold-moving and soft-ensemble are relatively good choices in

training cost-sensitive neural networks. In [18], Sun et al. proposed cost-sensitive boosting algorithms which are developed by introducing cost items into the learning framework of AdaBoost. In [6], Jiang et al. proposed a novel Minority Cloning Technique (MCT) for class-imbalanced cost-sensitive learning. MCT alters the class distribution of training data by cloning each minority class instance according to the similarity between it and the mode of the minority class. In [5], a new cost-sensitive metric was proposed by George to find the optimal tradeoff between the two most critical performance measures of a classification task-accuracy and cost. Generally, users focus more on the minority class and consider the cost of misclassifying a minority class to be more expensive. In our study, we adopted the same strategy to address this problem.

Motivated by the probabilistic collaborative representation based approach for pattern classification [2], we proposed a new method to handle misclassification cost and class-imbalance problem called Cost-Sensitive Collaborative Representation based Classification (CSCRC) via Probability Estimation. In Zhang's cost-sensitive learning framework, posterior probabilities of a testing sample are estimated by KLR or KNN method. In [2], ProCRC is designed to achieve the lowest recognition errors and assume the same losses for different types of misclassifications, it is difficult to resolve the class imbalance problem. For this case, we introduce cost-sensitive learning framework into ProCRC, which not only derive the relationship between Gaussian function and collaborative representation but also resolve the cost-sensitive problem. Firstly, we used the probabilistic collaborative representation framework to estimate the posterior probabilities. The posterior probabilities were generated directly from the coding coefficients by using a Gaussian function and applying the logarithmic operator to the probabilistic collaborative representation framework, this explained clearly the l_2-norm regularized representation scheme used in CRC. Secondly, calculate all the misclassification losses using Zhang's cost-sensitive learning framework. At last, the test sample is assigned to the class whose loss is minimal. Experimental results on UCI databases validate the effectiveness and efficiency of our methods.

The rest of this paper is organized as follows. Section 2 outlines the details of the relevant method. Section 3 presents the details of the proposed algorithm. Section 4 reports the experiments. Finally, section 5 concludes the paper and offers suggestions for future research.

2 Related work

2.1 Cost-sensitive learning

In multiclass cost-sensitive learning, considering c gallery subjects with their class labels $G = \{G_i\}, i = 1, 2, \cdots, c$, the labels of impostor are i. In [23], Zhou et al. categorized the costs into three types: the cost of accepting the one which should be rejected is C_{IG}; the cost of rejecting the one which should be accepted is C_{GI}; the cost of misidentifying the one as another is C_{GG}. Cost-sensitive learning usually sets the misclassification cost as objective function and identify the label by minimizing loss function. Given a test sample y and its predicted class label $\phi(y)$, respectively. The label is obtained by minimizing the objective function:

$$L(y) = \underset{\phi(y)\in\{G_1,\cdots,G_c,I\}}{\arg\min} loss(y, \phi(y)) \tag{1}$$

where

$$loss(y,\phi(y)) = \begin{cases} \sum_{i=1}^{c} P\ (G_i|y)C_{GI} & if\ \phi(y) = I \\ \sum_{\substack{i=1 \\ i\neq\tau}}^{c} P(G_i|y)C_{GG} + P(I|y)C_{IG}, & if\ \phi(y) = G_\tau \end{cases} \tag{2}$$

$\hat{\phi}(y)$ is the optimal prediction of y, c represents the gallery subjects in classification problem.

2.2 Collaborative representation based classification (CRC_RLS)

Suppose that we have K classes of subjects $X = [X_1, X_2, \cdots, X_K]$, and each class has enough training samples. For a query sample y, we code it collaboratively over the dictionary of all samples X under the l_1-norm sparsity constraint. We can write it as $y = x + e$, where $x = X\alpha$ is the component we want to recover from y for classification use and e is the residual (e.g., noise, occlusion and corruption) we want to remove from y. Then, $y = X\alpha + e$. To recover a stable coding coefficient vector $\hat{\alpha}$ from y and X, the regularization method is the best choice. If we assume that the model error e follows a Gaussian distribution, then the optimization problem can be written as follows:

$$\hat{\alpha} = \text{argmin}_\alpha \left\{ \|y - X\alpha\|_2 + \lambda\|\alpha\|_2 \right\} \tag{3}$$

where λ is the regularization parameter. The solution of collaborative representation with regularized least square in Eq. (3) can be easily and analytically derived as:

$$\hat{\alpha} = \left(X^T X + \lambda \cdot I\right)^{-1} X^T y \tag{4}$$

Let $P = (X^T X + \lambda \cdot I)^{-1} X^T$, then $\hat{\alpha} = Py$. Clearly, P is independent of y so that it can be pre-calculated as a projection matrix. For a query sample y, we can simply project it onto P via P_y. In addition to using the class-specified representation residual $\|y - X_i\hat{\alpha}_i\|_2$ for classification, where $\hat{\alpha}_i$ is the coding vector associated with class i, the l_2-norm "sparsity" $\|\hat{\alpha}_i\|_2$ also brings some discrimination information. We propose to use both of them in the decision making. We then compute the residual $r_i(y)$ as:

$$r_i(y) = \|y - X_i \cdot \hat{\alpha}_i\|_2 / \|\hat{\alpha}_i\|_2 \tag{5}$$

The query sample y is identified by minimizing $r_i(y)$ as follow:

$$L(y) = \arg\min_i \ \{r_i(y)\} \tag{6}$$

3 Proposed approach

Different data points x have different probabilities of $l(x) \in l_X$, where $l_{(x)}$ means the label of x, l_x means the label set of all candidate classes in X, and $P(l(x) \in l_X)$ should be higher if the l_2-norm

of $\alpha$is smaller, vice versa. One intuitive choice is to use a Gaussian function to define such a probability:

$$P(l(x)\in l_X)\propto\exp\left(-c\|\alpha\|_2^2\right) \tag{7}$$

where c is a constant and data points are assigned with different probabilities based on α, where all the data points are inside the subspace spanned by all samples in X. For a sample y outside the subspace, the probability as:

$$P(l(y)\in l_X) = P(l(y) = l(x)|l(x)\in l_X)P(l(x)\in l_X) \tag{8}$$

$P(l(x) \in l_X)$ has been defined in Eq. (7). $P(l(y) = l(x)|l(x) \in l_X)$ can be measured by the similarity between x and y. Here we adopt the Gaussian kernel to define it:

$$P(l(y) = l(x)|l(x)\in l_X)\propto\exp\left(-k\|y-x\|_2^2\right) \tag{9}$$

where k is a constant, with Eq. (7)–(9), we have

$$P(l(y)\in l_X)\propto\exp\left(-\left(k\|y-X\alpha\|_2^2 + c\|\alpha\|_2^2\right)\right) \tag{10}$$

In order to maximize the probability, we can apply the logarithmic operator to Eq. (10). There is:

$$\begin{aligned}\max P(l(y)\in l_X) &= \max\ln(P(l(y)\in l_X))\\ &= \min_\alpha k\|y-X\alpha\|_2^2 + c\|\alpha\|_2^2\\ &= \min_\alpha \|y-X\alpha\|_2^2 + \lambda\|\alpha\|_2^2\end{aligned} \tag{11}$$

where $\lambda = c/k$. Interestingly, Eq. (11) shares the same formulation of the representation formula of CRC [19], but it has a clear probabilistic interpretation.

A sample x inside the subspace can be collaboratively represented as: $x = X\alpha = \sum_{k=1}^{K} X_k\alpha_k$, where $\alpha = [\alpha_1; \alpha_2; \cdots; \alpha_k]$ and α_k is the coding vector associated with X_k. Note that $x_k = X_k\alpha_k$ is a data point falling into the subspace of class k. Then, we have

$$P(l(x) = k|l(x)\in l_X)\propto\exp\left(-\delta\|x-X_k\alpha_k\|_2^2\right) \tag{12}$$

where δ is a constant. For a query sample y, we can compute the probability that $l(y) = k$ as:

$$\begin{aligned}&P(l(y) = k)\\ &= P(l(y) = l(x)|l(x) = k)\cdot P(l(x) = k)\\ &= P(l(y) = l(x)|l(x) = k)\cdot P(l(x) = k|l(x)\in l_X)\cdot P(l(x)\in l_X)\end{aligned} \tag{13}$$

Since the probability definition in Eq. (9) is independent of k as long as $k \in l_X$, we have $P(l(y) = l(x)|l(x) = k) = P(l(y) = l(x)|l(x) \in l_X)$. With Eq. (11)–(12), we have

$$\begin{aligned}P(l(y) = k) &= P(l(y)\in l_X)\cdot P(l(x) = k|l(x)\in l_X)\\ &\propto\exp\left(-\|y-X\alpha\|_2^2 + \lambda\|\alpha\|_2^2 + \gamma\|X\alpha-X_k\alpha_k\|_2^2\right)\end{aligned} \tag{14}$$

where $\gamma = \delta/k$. Applying the logarithmic operator to Eq. (14) and ignoring the constant term, we have:

$$(\hat{\alpha}) = \text{argmin}_{\alpha}\left\{\|y - X\alpha\|_2^2 + c\|\alpha\|_2^2 + \|X\alpha - X_k\alpha_k\|_2^2\right\} \tag{15}$$

Refer to Eq. (15), let $X_k^{'}$ be a matrix which has the same size with X, while only the samples of X_k will be assigned to $X_k^{'}$ at their corresponding locations in X, i.e., $X_k^{'} = [0, \cdots, X_k, \cdots, 0]$. Let $\overline{X}_k^{'} = X - X_k^{'}$. We can then compute the following projection matrix offline:

$$T = \left(X^T X + \left(\overline{X}_k^{'}\right)^T \overline{X}_k^{'} + \lambda I\right)^{-1} X^T \tag{16}$$

where I denotes the identity matrix. Then, $\hat{\alpha} = Ty$.

With the model in Eq. (15), a solution vector $\hat{\alpha}$ is obtained. The probability P($l(y) = k$) can be computed by:

$$P(l(y) = k) \propto \exp\left(-\left(\|y - X\hat{\alpha}\|_2^2 + \lambda\|\hat{\alpha}\|_2^2 + \|X\hat{\alpha} - X_k\hat{\alpha}_k\|_2^2\right)\right) \tag{17}$$

Note that $(\|y - X\hat{\alpha}\|_2^2 + \lambda\|\hat{\alpha}\|_2^2)$ is the same for all classes, and thus we can omit it in computing P($l(y) = k$). Then we have:

$$P_k = \exp\left(-\left(\|X\hat{\alpha} - X_k\hat{\alpha}_k\|_2^2\right)\right) \tag{18}$$

In cost-sensitive learning, the loss function (Eq. (2)) is regarded as an objective function to identify the label of a test sample. In binary classification problem, there are two misclassification costs, and we denote the cost that misclassify positive class as negative class by C_{10}, and the cost by C_{01} conversely. Then a cost matrix can be constructed as shown in Table 1, where G_1, G_0 represents the label of minority class and majority class, respectively.

It is well known that the loss function can be related to the posterior probability$P(\phi(y)|y) \approx P(l(y) = k)$. Then the loss function can be rewritten as follow:

$$loss(y, \phi(y)) = \begin{cases} \sum_{i=G_1} P_i C_{10} \; if \; \phi(y) = G_0 \\ \sum_{j=G_0} P_j C_{01} \; if \; \phi(y) = G_1 \end{cases} \tag{19}$$

The test sample y belongs to the class with higher probability. We can obtain the label of test sample y by minimizing Eq. (19):

$$L(y) = \arg \min_{i \in \{0,1\}} loss(y, \phi(y)) \tag{20}$$

Table 1 The cost matrix

	G_0	G_1
G_0	0	C_{01}
G_1	C_{10}	0

The whole process of CSSRC is described in Algorithm 1.

Algorithm 1: CSCRC algorithm

Input: dictionary X ∈ $R^{m\times n}$, test sample $y \in R^m$

Output: the label $L(y)$ of test sample y

1: Normalize the columns of X to have unit l_2-norm.

2: Code y over X by

$$\hat{\alpha} = Ty$$

where $T = (X^T X + (\overline{X}'_k)^T \overline{X}'_k + \lambda I)^{-1} X^T$ in Eq. (16)

3: Compute the posterior probability P_k

4 Experiments

4.1 Data sets and experimental setting

We tested the proposed method on 10 UCI data sets [1]. Detail information about these data sets is summarized in Table 2.

In cost-sensitive learning, false positive (actual negative but predicted as positive, denoted as FP), false negative (actual positive but predicted as negative, FN), true positive (actual positive and predicted as positive, TP) and true negative (actual

Table 2 Description of data sets

Dataset	Size	Target	Ratio	min/maj
Abalone	4117	Ring = 7	9.7	391/3786
Housing	506	[20, 23]	3.8	106/400
Nursery	12,960	very-recom	38.5	328/12632
Letter	20,000	A	24.3	789/19211
Pima	786	Class1	1.7	268/500
Cmc	1473	Class2	3.4	333/1344
Car	1728	acc	3.5	384/1344
Ionosphere	351	bad	1.8	126/225
Balance	625	balance	11.8	49/576
Haberman	306	Class2	2.8	81/225

Table 3 Confusion matrix

	Positive Class	Negative Class
Positive Class	TP	FN
Negative Class	FP	TN

negative and predicted as negative TN) can be given in a confusion matrix, as shown in Table 3:

To binary classification problems, four kinds of misclassification cost are needed, which were referred as CTP, CFP, CTN, and CFN, respectively. CTP and CTN are the costs of true positive (TP) and true negative (TN). In order to simplify the cost matrix, we set CTP = 0, CTN = 0. CFN and CFP are the costs of false negative (FN) and false positive (FP). We always assume that the cost of misclassifying positive class instances is much higher than the cost of misclassifying negative class instances, so we set CFN> > CFP. In this paper, CFP is set to be a unit cost of 1, CFN is assigned as 10. For class imbalance experiment, the imbalance ratio is set as 1, 2,…, 10, respectively. In our experiments, we repeated an experiment for 50 times and got

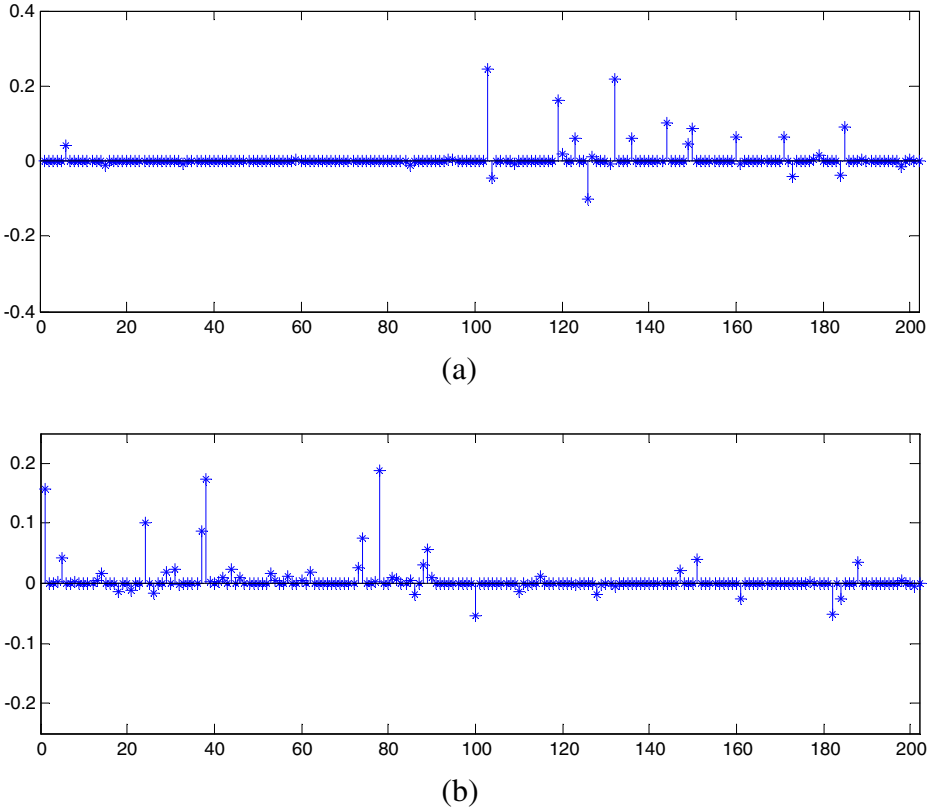


Fig. 1 The coding coefficients of positive query sample and negative query sample

Table 4 The Classification Accuracy for the five methods on 10 data sets (where bold entries are the methods with highest classification accuracy on each data set)

Accuracy	CRC	SRC	SVM	ProCRC	CSCRC
Letter	0.9932	0.9945	0.9690	0.9653	**0.9950**
Balance	0.9829	0.9138	0.9786	0.9357	**0.9919**
Abalone	0.9147	0.9984	0.9982	0.9994	**0.9997**
Car	0.9982	0.9781	0.9995	0.9902	**1.0000**
Pima	0.9395	**0.9779**	0.8665	0.9053	0.9377
Nursery	0.9994	0.9356	0.9718	0.9947	**0.9998**
Cmc	0.9729	0.9900	0.9981	0.9994	**0.9997**
Haberman	0.9088	0.9905	0.9506	0.9478	**0.9997**
Housing	0.9934	**1.0000**	0.6750	0.9888	0.9919
Ionosphere	0.9794	0.6738	**0.9913**	0.9581	0.9716
Average	0.96824	0.94526	0.93986	0.96847	**0.98870**

the average results. In our experiments, four evaluation criteria are adopted to evaluate the classification performance: Average Cost (AC), F-measure, G-mean and classification Accuracy. They are defined as follows [24]:

$$Recall = Acc_+ = \frac{TP}{TP + FN}$$
$$Acc_- = \frac{TN}{TN + FP}$$
$$Accuracy = Acc_- + Acc_+$$
$$Precision = \frac{TP}{TP + FP}$$
$$G\text{–}mean = \sqrt{Acc_+ \times Acc_-}$$
$$F\text{–}measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$
$$AC = \frac{C_{10}FP + C_{01}FN}{N}$$

The experiments were performed on Matlab 2014a, and the computer with a 2.6GHz Intel Xeon CPU.

Table 5 The F-measure for the five methods on 10 data sets (where bold entries are the methods with highest F-measure on each data set)

F-measure	CRC	SRC	SVM	ProCRC	CSCRC
Letter	0.9928	0.9944	0.9689	0.9652	**0.9951**
Balance	0.9827	0.9017	0.9727	0.9459	**0.9923**
Abalone	0.9131	0.9984	0.9982	0.9994	**0.9997**
Car	0.9982	0.9773	0.9995	0.9901	**1.0000**
Pima	0.9430	**0.9760**	0.8688	0.9053	0.9419
Nursery	0.9939	0.9302	0.9725	0.9947	**0.9998**
Cmc	0.9725	0.9897	0.9980	0.9994	**0.9997**
Haberman	0.9075	0.9899	0.9509	0.9473	**0.9997**
Housing	0.9932	**1.0000**	0.6942	0.9891	0.9921
Ionosphere	0.9793	0.5927	**0.9912**	0.9559	0.9720
Average	096816	0.93530	0.94149	0.96923	**0.98923**

Table 6 The Average Cost for the five methods on 10 data sets (where bold entries are the methods with lowest average cost on each data set)

Average Cost	CRC	SRC	SVM	ProCRC	CSCRC
Letter	0.0663	0.0548	0.1616	0.2118	**0.0050**
Balance	0.0943	0.8619	0.0343	0.0643	**0.0081**
Abalone	0.3335	0.0016	0.0018	0.0050	**0.0003**
Car	0.0032	0.2179	0.0034	0.0737	**0.0000**
Pima	0.1694	0.1658	0.8173	0.5084	**0.0623**
Nursery	0.0065	0.6435	0.2416	0.0271	**0.0002**
Cmc	0.2056	0.1000	0.0048	0.0006	**0.0003**
Haberman	0.5075	0.0952	0.2941	0.2772	**0.0003**
Housing	0.0516	**0.0000**	2.0828	0.0113	0.0081
Ionosphere	0.1191	2.5719	**0.0313**	0.3906	0.1044
Average	0.1557	0.4713	0.3673	0.1570	**0.0189**

4.2 Experimental results not considering the imbalance

The main idea of collaborative representation is to represent the query samples using training samples in binary classification problem. Figure 1a and b show the coding coefficients of a positive query sample and a negative query sample. We can easily find that the query samples are much related to the samples from the same class and observe the class label of the query samples.

We compared the performance of these five methods (SRC, CRC, SVM, ProCRC, CSCRC) on 10 UCI data sets, and the results are summarized in Tables 4, 5 and 6. The last row of Tables 4 and 5 is the average Accuracy and F-measure value for the method on ten data sets. We selected 31 positive samples and 31negative samples randomly from data sets Haberman, Housing, Ionosphere and Balance as test samples, 41 positive samples and 41 negative samples as training samples; 61 positive samples and 61 negative samples as test samples, 101 positive samples and 101 negative samples as training samples from the other 6 data sets. The cost ratio (the cost of false acceptance respect to false rejection) was set as 10. We performed the process for 50 times and get the average results.

On Letter, Balance, Abalone, Car, Nursery, Cmc and Haberman, our method achieved very high Accuracy and F-measure value respect to the other four methods. One of the three data sets does not get the highest value of Accuracy and F-measure, but we achieve the highest value of

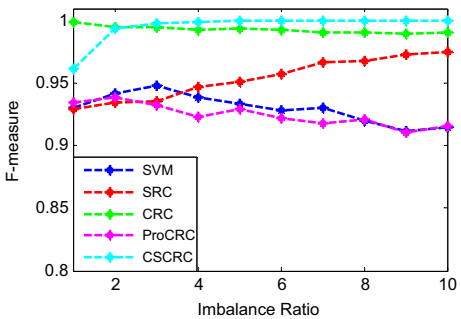


Fig. 2 The result of F-measure on Letter with different imbalance ration

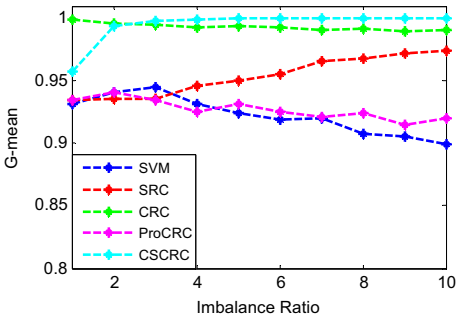


Fig. 3 The result of G-mean on Letter with different imbalance ration

average Accuracy and F-measure. The values of Accuracy and F-measure are higher than 0.93. In other words, our method have better performance than SRC, CRC, SVM andProCRC.

We calculated the misclassification cost of these five methods on 10 UCI data sets and summarized as Table 6. On Letter, Balance, Abalone, Car, Pima, Nursery, Cmc and Haberman, our method achieves very low average misclassification cost. In Tables 4 and 5, SRC has the highest value on Pima, but CSCRC has the highest value of Average Cost on Pima. Obviously, CSCRC classify the positive samples correctly. Furthermore, the value of Accuracy and F-measure is lower than CRC on Housing and Ionosphere, but the value of Average Cost is inverse.

4.3 Experimental results considering the imbalance

Similarly, we compared the performance of these five methods (SRC, CRC, SVM, ProCRC, CSCRC) on Letter, and evaluated the performance via F-measure, G-mean and Average Cost for the class-imbalance problem. In this experiment, we set the imbalance ratio as [1, 2,…, 10], respectively. The size of minority class is 30 and the majority class is 30 multiply the imbalance ratios in train set, accordingly. We selected 61 positive samples and 61 negative samples as test set. The cost was set as mentioned in section 4.1.

Note that there are also situations in which CSCRC is preferred. From the results on Figs. 2, 3 and 4, we can see that CSCRC has higher F-measure and G-mean than the other four methods except when the imbalance ratio is 1. Meanwhile, CSCRC achieves the lowest Average Cost respect to the other methods. This suggests that CSCRC can focus on more useful data. With the increasing of imbalance ratio, we have more training samples, and the proposed method can

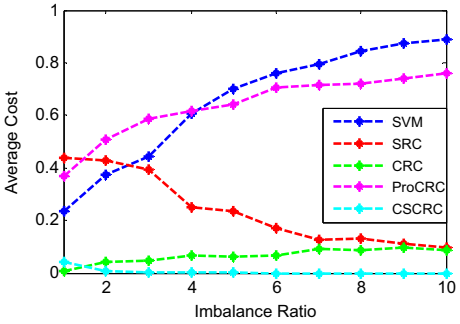


Fig. 4 The result of Average Cost on Letter with different imbalance ration

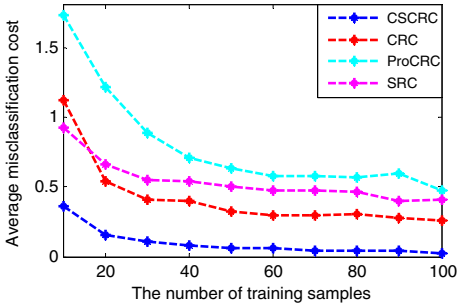


Fig. 5 The average misclassification cost on YaleB

classify the samples correctly when the imbalance ratio is up to 4. Generally speaking, class-imbalance is affect by the proposed method. Concretely, CSCRC is not influenced by the distribution of samples, we can also get a better classify result when the imbalance ratio is high.

4.4 Experimental results on face recognition

This section, we selected two persons from the dataset YaleB for this experiment and performed on Matlab 2014a. We compared the performance of these 4 methods (SRC, CRC, ProCRC, CSCRC), and evaluated the performance via Average misclassification cost and classification accuracy for the cost sensitive problem. In this experiment, the training samples consist of 10, 20, …, 100 images and the rest of this two persons' images are test samples. We set the misclassification cost as 10 for the error that misclassify negative class sample as positive class sample, the opposite is 1, and training 100 times for per subsets and the results as follows:

The results has summarized in Figs. 5 and 6. With the increasing of training set, the average misclassification cost has reduced and classification accuracy has increased for this four methods. Although the classification accuracy of CSCRC is lower than some of the other methods, we can obtain the lowest misclassification cost. In our method, we pursue the lowest misclassification cost and regard it as the objective function. Traditional methods pursue the highest classification accuracy, but it is unsuitable for solving the cost sensitive problem. CSCRC combines the CRC with cost sensitive learning can well deal with the cost sensitive and class imbalance problem.

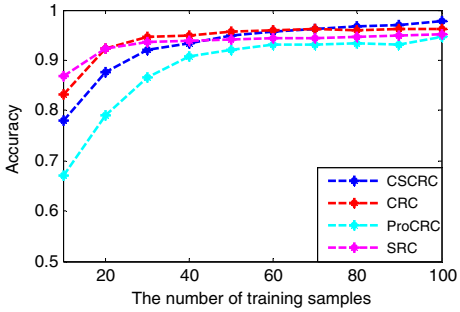


Fig. 6 The classification accuracy on YaleB

5 Conclusions

This paper, we proposed a novel method to handle misclassification cost and class imbalance problem simultaneously called Cost-Sensitive Collaborative Representation Classification based Probability Estimation. The proposed approach adopted probabilistic model and sparse representation coefficient matrix to estimate the posterior probability and then obtained the label of a testing sample by minimizing the misclassification losses. The experimental results show that the proposed CSCRC has a comparable or even lower average cost with higher accuracy compare to the other four classification algorithm.

6 Acknowledgements

The authors want to thank the anonymous reviewers and the associate editor for helpful comments and suggestions. This work is supported by the National Natural Science Foundation of China (Grant Nos. 61562013, 61320106008), Guangxi Colleges and Universities Key Laboratory of Intelligent Processing of Computer Images and Graphics (Grant No. LD16096x), the Center for Collaborative Innovation in the Technology of IOT and the Industrialization (Grant No. WLW20060610), Innovation Project of GUET Graduate Education, the study abroad program for graduate student of Guilin University of Electronic Technology. The authors declare that they have no conflict of interest.

References

1. Blake C, Keogh E, Merz CJ UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html, Department of Information and Computer Science, University of California, Irvine, CA
2. Cai S, Zhang L, Zuo W et al (2016) A Probabilistic Collaborative Representation based Approach for Pattern Classification. Comput Vis Pattern Recogn 2016:2950–2959
3. Cheng Y, Jin Z, Gao T, Chen H, Kasabow N (2016) An improved collaborative representation based classification with regularized least square (CRC–RLS) method for robust face recognition. Neurocomputing 215(C):250–259
4. EthemAlpayd (2011) Wiley Interdisciplinary Reviews Computational Statistics. Mach Learn 3(3):195–203
5. George NI, Lu TP, Chang CW (2016) Cost-sensitive performance metric for comparing multiple ordinal classifiers. Artif Intell Res 5(1):p135
6. Jiang L, Qiu C, Li C (2015) A novel minority cloning technique for cost-sensitive learning. Int J Pattern Recognit Artif Intell 29(4):18. https://doi.org/10.1142/S0218001415510040
7. Lan R, Yang J, Jiang Y, Song Z, Tang YY (2012) An affine invariant discriminate analysis with canonical correlation analysis. Neurocomputing 86:184–192
8. Lawrence N (2005) Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. J Mach Learn Res 6(3):1783–1816
9. Liu J, Wu Z, Li J et al (2015) Probabilistic-Kernel Collaborative Representation for Spatial–Spectral Hyperspectral Image Classification. IEEE Trans Geosci Remote Sens 54:1–14
10. Lu H, Li Y, Chen M, Kim H, Serikawa S (2017) Brain Intelligence: Go Beyond Artificial Intelligence. Mob Netw Appl 7553:1–8
11. Lu H, Li Y, Uemura T, Ge Z, Xu X et al (2017) FDCNet: filtering deep convolutional network for marine organism classification. Multimed Tools Appl 2017:1–14
12. Moghaddam B (2002) Principal manifolds and probabilistic subspaces for visual recognition. IEEE Trans Pattern Anal Mach Intell 24(6):780–788
13. Moghaddam B, Pentland A (2002) Probabilistic Visual Learning for Object Representation. IEEE Trans Pattern Anal Mach Intell 19(7):696–710
14. Polson NG, Scott SL (2011) Data augmentation for support vector machines. Bayesian Anal 6(1):43–47

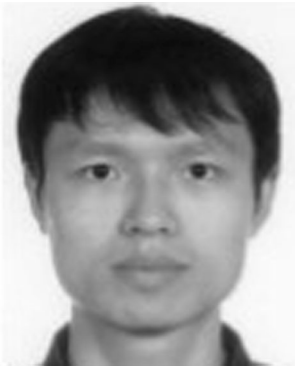
15. Prince SJD, Elder JH (2017) Probabilistic Linear Discriminant Analysis for Inferences about Identity. IEEE Int Conf Comput Vis 2007:1–8
16. Rushi L, Yicong Z (2016) Quaternion-Michelson Descriptor for Color Image Classification. IEEE Trans Image Process 25(11):5281–5292
17. Sadeghian A, Huang B (2016) Robust probabilistic principal component analysis for process modeling subject to scaled mixture Gaussian noise. Comput Chem Eng 90:62–78
18. Sun Y, Kamel MS, Wong AKC et al (2007) Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn 40(12):3358–3378
19. Waqas J, Zhang Y, Lei Z (2013) Collaborative neighbor representation based classification using l 2-minimization approach. Pattern Recogn Lett 34(2):201–208
20. Wright J, Yang AY, Ganesh A et al (2008) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227
21. Xu X, He L, Lu H, Shimada A, Taniguchi RI (2016) Non-linear Matrix Completion for Social Image Tagging. IEEE Access 99:1–1
22. Yen SJ, Lee YS (2009) Cluster-based under-sampling approaches for imbalanced data distributions. Expert Syst Appl 36(3):5718–5727
23. Yin Z, Zhi-Hua Z (2010) Cost-sensitive face recognition. IEEE Trans Pattern Anal Mach Intell 32(10): 1758–1769
24. Zhang X, Hu B (2014) A New Strategy of Cost-Free Learning in the Class Imbalance Problem. IEEE Trans Knowl Data Eng 26(12):2872–2885
25. Zhang L, Yang M, Feng X (2012) Sparse representation or collaborative representation: Which helps face recognition. IEEE Int Conf Comput Vis 2011(5):471–478
26. Zhou ZH, Liu XY (2006) Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng 18(1):63–77

Zhenbing Liu received Ph.D. in Institute of Image Recognition & Artificial Intelligence at Huazhong University of Science and Technology in 2010. Do a visiting scholar in the University of Pennsylvania in 2015. Now he is professor and master supervisor in School of Computer and Information Security, Guilin University of Electronic Technology, China. His main research interests include image processing, machine learning and pattern recognition.

Chao Ma is currently pursuing the M.S. degree in School of Electronic Engineering and Automation, Guilin University of Electronic Technology (GUET), China. He received his Bachelor degree in School of Information Engineering from Henan University of Science and Technology (HUST) in 2016. His researches focus on machine learning and optimization.

Chunyang Gao is a Master graduate student in School of Electronic Engineering and Automation, Guilin University of Electronic Technology (GUET), China. He received his Bachelor degree in School of Basic Science from Harbin University of Commerce in 2014. His researches focus on pattern recognition, machine learning.

Huihua Yang received Ph.D. in East China University of Science and Technology in 2002. He was engaged in postdoctoral research work at Analysis Center of Tsinghua University from 2002 to 2007. He is currently a Professor and Doctoral Supervisor in Guilin University of Electronic Technology and an Adjunct Professor, Doctoral Supervisor in Beijing University of Posts and Telecommunications, China. Now, he is senior member of China computer society, committee member of High Performance Computing and associate chairman of China Instrument and Control Society Near infrared spectroscopy club.



Rushi Lan received the B.S. and M.S. degrees from the Nanjing University of Information Science and Technology, Nanjing, China, in 2008 and 2011, respectively. He received the Ph.D. degree from the Department of Computer and Information Science, University of Macau, Macau, China. Now he is a lecturer of School of Computer and Information Security, Guilin University of Electronic Technology, China. His current research interests include image classification, image denoising, and metric learning.

Xiaonan Luo is a professor of School of Computer and Information Security, Guilin University of Electronic Technology, China. He won the National Science Fund for Distinguished Young Scholars granted by the National Nature Science Foundation of China. His research interests include computer graphics, CAD, image processing, mobile computing.