Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 25):681
https://doi.org/10.1186/s12859-019-3255-x

BMC Bioinformatics

RESEARCH Open Access

Learning misclassification costs for imbalanced classification on gene expression data

Huijuan Lu[1], Yige Xu[1], Minchao Ye[1*], Ke Yan[1], Zhigang Gao[2] and Qun Jin[3]

From 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference
Wuhan and Shanghai, China. 15-18 August 2018, 3-4 November 2018

Abstract

Background: Cost-sensitive algorithm is an effective strategy to solve imbalanced classification problem. However, the misclassification costs are usually determined empirically based on user expertise, which leads to unstable performance of cost-sensitive classification. Therefore, an efficient and accurate method is needed to calculate the optimal cost weights.

Results: In this paper, two approaches are proposed to search for the optimal cost weights, targeting at the highest weighted classification accuracy (WCA). One is the optimal cost weights grid searching and the other is the function fitting. Comparisons are made between these between the two algorithms above. In experiments, we classify imbalanced gene expression data using extreme learning machine to test the cost weights obtained by the two approaches.

Conclusions: Comprehensive experimental results show that the function fitting method is generally more efficient, which can well find the optimal cost weights with acceptable WCA.

Keywords: Cost-sensitive, Misclassification cost, Weighted classification accuracy, Parameter fitting

Background

Classification of gene expression data reveals tremendous information in various application fields of biomedical research, such as cancer diagnosis, prognosis and predictions [1–3]. However, the gene expression data is composed of high-dimensional, noisy and imbalanced data samples [4]. The characteristic of imbalanced data is serious imbalance in the proportion of positive and negative samples [5, 6]. Gene expression data exacts a series of pre-processing steps to eliminate misleading classification results [7]. Moreover, the classification of gene expression data is a cost-sensitive problem, although both positive and negative classifications of cancer genes provide important evidences for doctors to make the treatment plan.

Traditional machine learning algorithms usually assume that the training set is balanced. For imbalanced datasets, such as the gene expression datasets, the classical classification algorithms with the correct classification rates (CCR) may bias towards the majority classes. However, the misclassifications of minority classes usually contribute the higher influences than those of majority classes. Therefore, The introduction of cost sensitive learning (CSL) is necessary to eliminate the defects of traditional classification algorithms for imbalanced datasets. Traditionally, oversampling the minority class, undersampling the majority class, and synthesizing new minority classes can be used to handle this problem. In this work, we utilize a more sophisticated way to search for the optimal weights, and the proposed methods are more advanced than ever.

* Correspondence: yeminchao@cjlu.edu.cn
[1]Key Laboratory of Electromagnetic Wave Information Technology and Metrology of Zhejiang Province, College of Information Engineering, China Jiliang University, Hangzhou, China
Full list of author information is available at the end of the article

In CSL, misclassification cost is an important factor to evaluate the classification performance of imbalanced datasets. However, solving the misclassification cost matrix is not a trivial task in many situations [8–10]. A direct solution for finding the misclassification costs is to assign them manually according to user expertise or inversely calculate the costs based on class distribution [11–13]. More sophisticated solutions can be found by fitting the importance of features to adaptive equations.

In this paper, we learn the misclassification cost from the evaluation functions of cost-sensitive algorithms, using weighted classification accuracy as the measurement of cost-sensitive classification performance. The cost weights that lead to optimal classification performance are learned by grid searching strategy. It will help the researchers to obtain a reference weight. Then, three fitting functions will be found to represent the optimal cost weights. A series of comprehensive experimental results show that the function fitting approach is an effective way of finding the optimal cost weights, targeting at high weighted classification accuracy (WCA). Fitting functions can accurately locate optimal weights. Appropriate weights will greatly improve the accuracy of the model.

Imbalanced data greatly affects the accuracy of classification. We discuss the cost-sensitive classification algorithms in the imbalance problem. CSL is one of the most hot topics in the field of machine learning. Many works have studied on CSL and embedded the misclassification costs into various classifiers, such as the decision trees (DTs), support vector machines (SVMs) and extreme learning machines (ELMs). Chai et al. [14] considered the testing costs of missing values in naive Bayes (NB) and DT algorithms. Feng [15] defined a customized objective function for misclassification costs and designed a score evaluation based cost-sensitive DT. For multi-class classification problems, Feng's method generally achieves higher classification accuracy or lower misclassification costs. Zhao and Li [16] extended the evaluation function by including weighted information gain ratio and the test cost for the cost-sensitive DT. The proposed cost-sensitive DT algorithm not only reduced the misclassification cost, but also improved the classification efficiency of the original C4.5 algorithm [17, 18]. Lu et al. [19] made use of the cost-sensitive DTs as base classifiers and constructed a cost-sensitive rotational forest. Two kinds of DTs, i.e., EG2 and C4.5, are considered and tested [20]. These experiments show that integrating cost-sensitive to classification algorithms can effectively improve classification efficiency.

Cost sensitivity and classification algorithms combine to form efficient classification methods. Cao et al. [21] proposed to embed evaluation measures into the objective function for to improve the performance of a cost-sensitive support vector machine (CS-SVM). He et al. [22] integrated the Gaussian Mixture Model (GMM) into the CS-SVM to deal with the imbalanced classification problem. Cheng and Wu [23] added weights to features and introduced a weighted features cost-sensitive SVM (WF-CSSVM). The WF-CSSVM algorithm showed significant performance improvement on both aspects of accuracy and cost. Silva et al. [24] combined CS-SVM with semi-supervised learning method to form a hybrid classification algorithm. The effectiveness of the proposed hybrid method is shown in the experimental results on Earth monitoring and landscape mapping. Cao et al. [25] tackled the problem of multi-labeled imbalanced data classification problem. They successfully assigned different misclassification costs to different label sets for reducing the overall misclassification cost.

CS-ELM has been studied by many researchers in various aspects. Zong et al. [26] introduced a weighted extreme learning machine (WELM) for imbalanced data learning. It was claimed that the WELM can be extended to a cost-sensitive ELM (CS-ELM). Zheng et al. [27] formally applied the concept of the cost-sensitivity to extreme learning machine (ELM). Yan et al. [28, 29] extended Zheng et al.'s work and introduced a cost-sensitive dissimilar ELM (CS-D-ELM). Compared to traditional ELM algorithms, the CS-ELM algorithms guarantee the classification accuracy and reduce the misclassification cost. More recently, Zhang and Zhang [30] solved the problem of defining and optimizing the cost matrix for CS-ELM to make it more robust and stable [31, 32]. Zhu and Wang [33] treated CS-ELM as a base classifier to solve a semi-supervised learning problem. Incremental results show that the CS-ELM has better performance in terms of accuracy, cost, efficiency and robustness over other existing classifiers.

Classical definition of cost matrix

Considering the binary classification problem, the confusion matrix shows four types of classification results according to the prediction values, namely, true positive, false positive, false negative and true negative (Table 1) [34, 35].

The CSL seeks the overall minimum cost by introducing sensitive costs, rather than only aiming at high CCR. While there are several types of classification costs, it should be noted that this work only focuses on the misclassification cost.

Misclassification cost can be viewed as penalties for errors in the classification process. In binary classification problems, costs caused by different types of errors may be different. We define the minority class as positive (P), the majority class as negative (N), and construct the cost matrix C as shown in Table 2.

In Table 2, C_{00} and C_{11} show the cost of correct classification. By default, we set the costs of correct classifications as 0. C_{01} and C_{10} show the costs of

Table 1 The confusion matrix for binary classification

	Prediction of Positive	Prediction of Negative
Positive samples	True Positive TP	False Negative FN
Negative samples	False Positive FP	True Negative TN

error classifications, where C_{01} denotes the misclassification costs of samples from P class, and C_{10} denotes the misclassification costs of samples from N class. Therefore, the cost matrix in Table 2 can be simplified as:

$$C = \begin{bmatrix} 0 & C_{01} \\ C_{10} & 0 \end{bmatrix} \tag{1}$$

Correct classification rates versus weighted classification accuracy

For classical machine learning problems, the classification accuracy always refers to the correct classification rate (CCR) [36–38], or called overall accuracy (OA) [39–42], which is the proportion of all correctly classified samples:

$$OA = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \tag{2}$$

However, for imbalanced datasets where the numbers of positive and negative samples differ significantly, the CCR might be misleading [43, 44]. Considering a test set containing 99 negative samples but with only one positive sample [45, 46], a poorly designed classifier that simply puts all samples as negative will achieve an overall accuracy of 99/100 = 0.99, even though the accuracy for positive class is 0. To resolve this issue, we introduce the notion of adaptive classification accuracy (ACA) defined as follows:

$$ACA = \frac{1}{2} \cdot \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \tag{3}$$

By embedding a weight w_{i} into the i-th class, we get the weighted classification accuracy (WCA) as:

Table 2 Cost matrix

Predicted Actual	*P*	*N*
P	C_{00}	C_{01}
N	C_{10}	C_{11}

$$WCA = \frac{w_1}{w_1 + w_2} \cdot \frac{TP}{TP + FN} + \frac{w_2}{w_1 + w_2} \cdot \frac{TN}{TN + FP} \tag{4}$$

By enforcing $w_1 + w_2 = 1$, Formula (9) is reduced to:

$$WCA = w_1 \cdot \frac{TP}{TP + FN} + w_2 \cdot \frac{TN}{TN + FP} \tag{5}$$

Formula (10) can be easily extended to multi-classification problems:

$$WCA_n = \sum_{i=1}^{n} w_i \frac{CM_i}{M_i}, \sum_{i=1}^{n} w_i = 1 \tag{6}$$

where n denotes the number of classes, M_i ($i = 1, 2,..., n$) denotes the number of samples belonging to the i-th class, and CM_{i} ($i = 1, 2,..., n$) denotes the number of correctly classified samples within i-th class. Since the WCA is more accurate describing the classification accuracy, we use the WCA to evaluate the classification performance of cost-sensitive classifiers in the problem of gene expression data classification.

Methods

Optimal cost weights searching

From the University of California Irvine (UCI) standard classification dataset, we choose Leukemia, Colon, Prostate, Lung and Ovarian gene as the datasets for cost weights searching and further test, i.e., the Leukemia cancer dataset, the Colon cancer dataset, the Prostate cancer dataset, the Lung cancer dataset, and the Ovarian cancer in the tumor data respectively. All details of aforementioned datasets are shown in Table 3.

Optimal cost weights searching by grid searching strategy

The optimal weights are searched by an adaptive algorithm using grid searching. There are two crucial factors to consider: the sample importance w and sample categorical distribution p. The sample categorical distribution p is the proportion between the number of positive

Table 3 Specifications of datasets

Dataset	Sample number	Feature dimension	Classification number
Leukemia	34	7130	2
Colon	62	2000	2
Prostate	136	12600	2
Lung	181	12533	2
Ovarian	253	15154	2

Table 4 Grid Searching Strategy

Grid Searching Strategy

```
procedure GRIDSEARCHING($M$, $T$, $P_0$)
    $P = P_0$
    $f$ = WCA($P$)
    if $P < M$ then
        $P = P + T$
        if $f > f_{max}$ then
            $f_{max} = f$
            $P_{max} = P$
        end if
    end if
    return $P_{max}$, $f_{max}$
end procedure
```

class and negative class in test sets. Test set is constructed by random sampling. As such, it is necessary to study the relationship between the three factors, namely, *w*, *p* and WCA, where WCA is the fitness value for the grid searching strategy. In general, the grid searching strategy can be described as follows (the detailed algorithm steps are listed in Table 4):

1) Set the searching region as *M*, grid searching step size as *T*, and the initial position as P_0;

2) Calculate the fitness of the current position, record the position P_{max} that has the best fitness f_{max} (f_{max} = WCA);

3) Update current location, P=$P + T$;

4) if the current fitness value is greater than f_{max}, update f_{max} and P_{max};

5) return f_{max} and $P_{max.}$

Extreme learning machine is an effective single hidden-layer feed-forward neural network (SLFN) learning algorithm. Cost-sensitive extreme learning machine (CS-ELM) is a kind of ELM, which attaches a cost matrix on output layer. In this research, we set the number of hidden neurons at 10. Less neurons will make the result more sensitive to observe the change of weights. And seven different gene expression datasets are used to obtain the classification results with CS-ELM as the classifier. CS-ELM minimizes the conditional risk by embedding misclassification cost in ELM.

$$argmin\ R(i|x) = argmin \sum_{j} P(j|x) \bullet C(i,j) \tag{7}$$

where *R(i*|*x)* is the conditional risk when the sample *x* is assigned to the class *i*, and *P(j*|*x)* is the conditional probability that *x* belongs to *j*, *C(i, j)* is the risk of misclassifying *j* to class *i*, where *i*, *j* ∈ $\{c_1, c_2, \ldots, c_m\}$ and *m* is the number of classification categories.

Results

Optimal cost weights searching by function fitting

In this subsection, we use *w* and *p* as independent variables, and define a function fitting problem as:

$$w_c = f(w, p) \tag{8}$$

where $w_c = C_{01}/C_{10}$, $w = w_1/(w_1 + w_2)$ and *p* represents the proportion of positive and negative classes. We set C_{10} to 1 to reduce the complexity of calculation, i.e., $f_c = C_{01}$.

The sample distribution *p*, the optimal weight $w_c = C_{01}/C_{10}$ and the highest fitness value of each dataset are listed in Table 6.

We use an automatic fitting software named 1STOPT to do the function fitting [47]. In 1STOPT, Levenberg-Marquardt and Universal Global Optimization are used to fit functions. We compared 500 functions with different types, and selected the three functions with the highest correlation coefficient:

Table 5 Optimal weights for different data set

Data set	Sample categorical distribution *p*	Influence factors		Optimal weights w_c	WCA
		$w_1/(w_1+w_2)$	$w_2/(w_1+w_2)$		
Colon	1	0.2	0.8	1.03	0.6167
Leukemia	1.33	0.9	0.1	0.9	0.9179
Ovarian1	1.68	0.9	0.1	1.65	0.9055
Prostate1	2	0.9	0.1	1.06	0.939
Prostate2	2.5	0.9	0.1	1.04	0.9372
Lung1	3	0.9	0.1	0.93	0.92
Ovarian2	4	0.1	0.9	3.45	0.9094
Lung2	5	0.1	0.9	4.26	0.9078
Ovarian3	6.5	0.9	0.1	0.8	0.9075
Lung3	8	0.9	0.1	0.92	0.9009

Table 6 Datasets, cost weights and WCAs with the two approaches proposed

Dataset			Cost weight				WCA				
type	*p*	*w*	optimal	w_{c1}	w_{c2}	w_{c3}	optimal	w_{c1}	w_{c2}	w_{c3}	*ECSELM*
ovarian	1.68	0.1	1.65	1.63	1.53	1.58	0.9055	0.9695	0.1966	0.2084	0.1017
Prostate	2.5	0.9	1.04	1.05	1.05	1	0.9372	0.9815	0.9509	0.9869	0.8985
Lung1	5	0.1	4.26	4.03	4.1	3.94	0.9078	0.9778	0.9786	0.9779	0.875
Lung2	8	0.9	0.92	0.9	0.66	0.61	0.9009	0.9564	0.9762	0.9675	0.9

$$w_{c1} = f_1(w,p) = \frac{a_1 + a_2 \cdot w + a_3 \cdot w^2 + a_4 \cdot w^3 + a_5 \cdot a_{12} \cdot \ \ln p + a_6 \cdot (a_{12} \cdot \ \ln p)^2}{1 + a_7 \cdot w + a_8 \cdot w^2 + a_9 \cdot a_{12} \cdot \ \ln p + a_{10} \cdot (a_{12} \cdot \ \ln p)^2 + a_{11} \cdot (a_{12} \cdot \ \ln p)^3} \tag{9}$$

where $a_1 = 1.323$, $a_2 = -2.278$, $a_3 = 3.047$, $a_4 = -1.286$, $a_5 = -1.746$, $a_6 = 0.998$, $a_7 = -0.400$, $a_8 = 0.369$, $a_9 = -2.606$, $a_{10} = 2.544$, $a_{11} = -0.818$, $a_{12} = 0.482$. The correlation coefficient R_1 of f_1 is 0.96346.

$$w_{c2} = f_2(w,p) = \frac{b_1 + b_3 \cdot w + b_5 \cdot \ \ln p + b_7 \cdot w^2 + b_9 \cdot \ \ln^2 p + b_{11} \cdot w \cdot \ \ln p}{1 + b_2 \cdot w + b_4 \cdot \ \ln p + b_6 \cdot x^2 + b_8 \cdot \ \ln^2 p + b_{10} \cdot w \cdot \ \ln p} \tag{10}$$

where $b_1 = 1.008$, $b_2 = 2.618$, $b_3 = 1.743$, $b_4 = -0.808$, $b_5 = 0.297$, $b_6 = 2.327$, $b_7 = 4.605$, $b_8 = 0.406$, $b_9 = 0.699$, $b_{10} = -2.343$, $b_{11} = -4.984$. The correlation coefficient R_2 of f_2 is 0.95903.

$$w_{c3} = f_3(w,p) = \frac{c_1 + c_3 \cdot \ \ln w + c_5 \cdot p + c_7 \cdot \ \ln^2 w + c_9 \cdot p^2 + c_{11} \cdot p \cdot \ \ln w}{1 + c_2 \cdot \ \ln w + c_4 \cdot p + c_6 \cdot \ \ln^2 w + c_8 \cdot p^2 + c_{10} \cdot p \cdot \ \ln w} \tag{11}$$

where $c_1 = 1.279$, $c_2 = 0.574$, $c_3 = 0.943$, $c_4 = -0.152$, $c_5 = -0.291$, $c_6 = 0.113$, $c_7 = 0.154$, $c_8 = 0.009$, $c_9 = 0.018$, $c_{10} = -0.062$, $c_{11} = -0.250$. The correlation coefficient R_3 of f_3 is 0.95244.

We compare the fitting functions with the optimal weights in Figs. 1, 2 and 3.

Figures 1, 2 and 3 show the comparison results of the three-dimensional interpolation of optimal weights and fitting functions. The red surface represents the optimal weights. The green, yellow, blue planes are fit surfaces of f_1, f_2 and f_3. The correlation coefficient *R* of f_1, f_2 and f_3 identified that the overall fitness of the function f_1 is better than other two. The function f_2 gradually deviates from optimal weights while we increase the value of *w*, and decrease the value of *p*. The function f_3 is slightly coarser than the function f_1 in general.

Discussion

Comparison with grid searching and function fitting

Using different gene expression datasets, we compared the optimal cost weights obtained from the grid searching strategy and fitted functions f_1, f_2 and f_3. In Table 6, we compared the WCAs with four different datasets, namely, Ovarian, Prostate, Lung1 and Lung2. The majority over minority class proportion of the four datasets

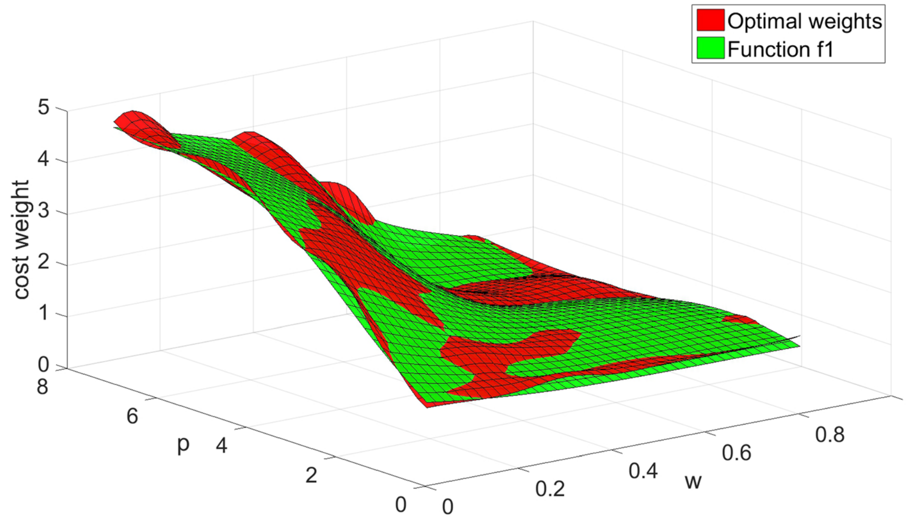

Fig. 1 The values of function w_{c1} compared with the optimal weights

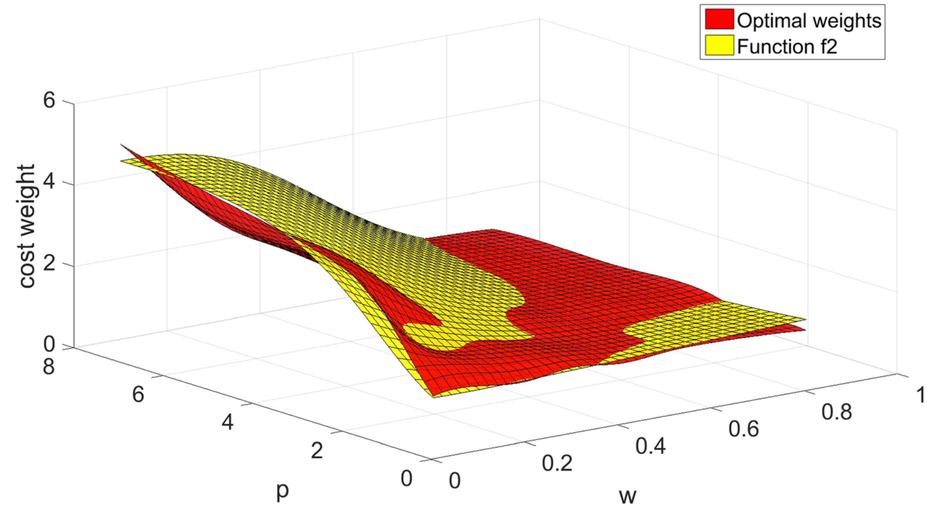

Fig. 2 The values of function w_{c2} compared with the optimal weights

are 1.68, 2.5, 5 and 8 respectively. All WCAs are computed using ELM as the base classifier. We also compare the two approaches with ECSELM. The best fit datasets are listed in Table 6.

For each dataset, we plot the weight variance with different values of *w*. For different dataset, the fittest function (choice from f_1, f_2 and f_3) might be different (Fig. 4).

Figure 4 shows that the more unbalanced the dataset is, the higher degree of fitness we can get; and the cost weights obtained from the fitting functions are closer to the optimal weights. In addition, the cost weights from function f_1 and f_3 are slightly superior to f_2. We put all cost weights obtained by different methods in a three-dimensional picture and show the results in Fig. 5.

For each dataset, we also illustrate the comparison of WCAs against different *w* values (Fig. 6). Besides, we compare WCAs of optimal weights and f_{1-3} with ECSELM [48].

In Fig. 6, we can see that the WCAs of the three fitting functions are lower than the optimal accuracy when *w* is less than 0.5. The reason is that the fitting degree of the cost weights in this range is lower. Moreover, it can be seen from Fig. 6 that the WCAs of the fitting functions approach to the optimal accuracy with the increment of *p*. Furthermore, the WCAs of our approaches is better than ECSELM in most field. Compared with ECSELM,

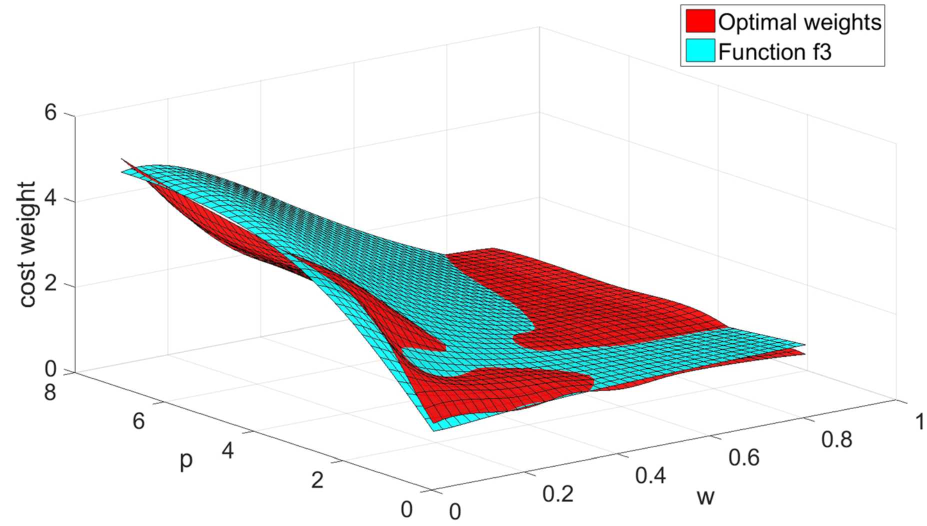

Fig. 3 The values of function w_{c3} compared with the optimal weights

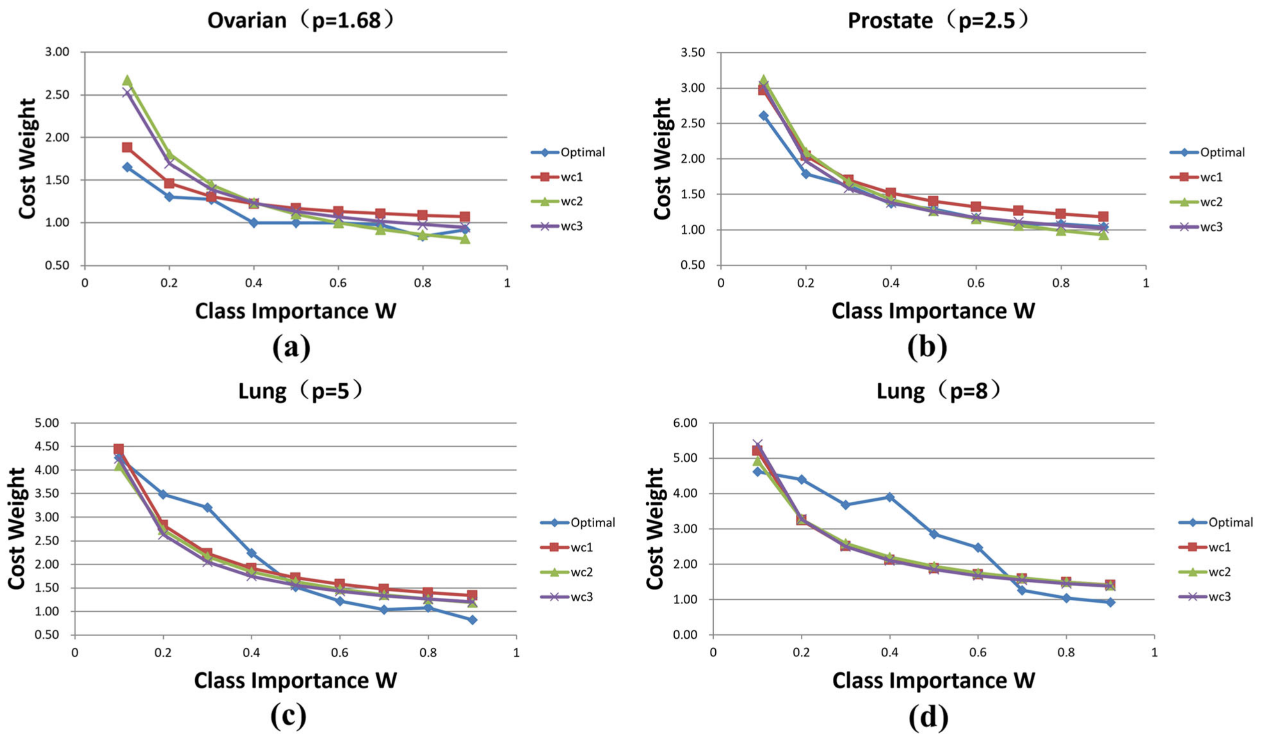

Fig. 4 Cost weight comparison using Ovarian, Prostate, Lung1, Lung2 dataset ($p = 1.68$, 2.5, 5, 8)

our methods are more stable, and meanwhile can guarantee high WCA. This proves the robustness of our strategy. Similar to the case of cost weights, we ensemble all WCAs obtained by different methods in a three-dimensional picture (Fig. 7). In summary, we find that the function f_1 provides better classification performance than the other two functions in general; and the fitting function f_3 and f_2 have better performance while the valuable p is large (when p above 5).

Conclusions

In this paper, we have proposed two approaches to calculate the optimal cost weights for gene expression data. The two approaches include a grid searching strategy and a function fitting method. They enrich the ways of calculating the cost weights for imbalanced data classification problems. In general, the function fitting approach is more efficient than the grid searching strategy. The experimental results also

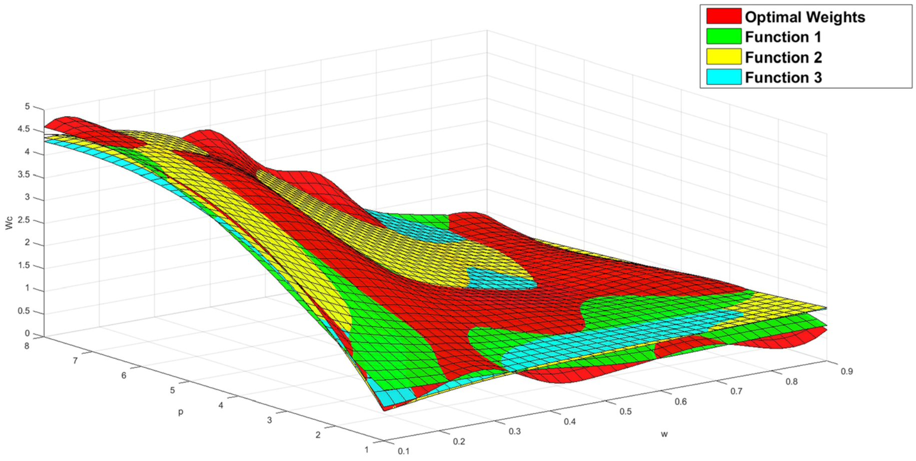

Fig. 5 Cost weight comparison in overall

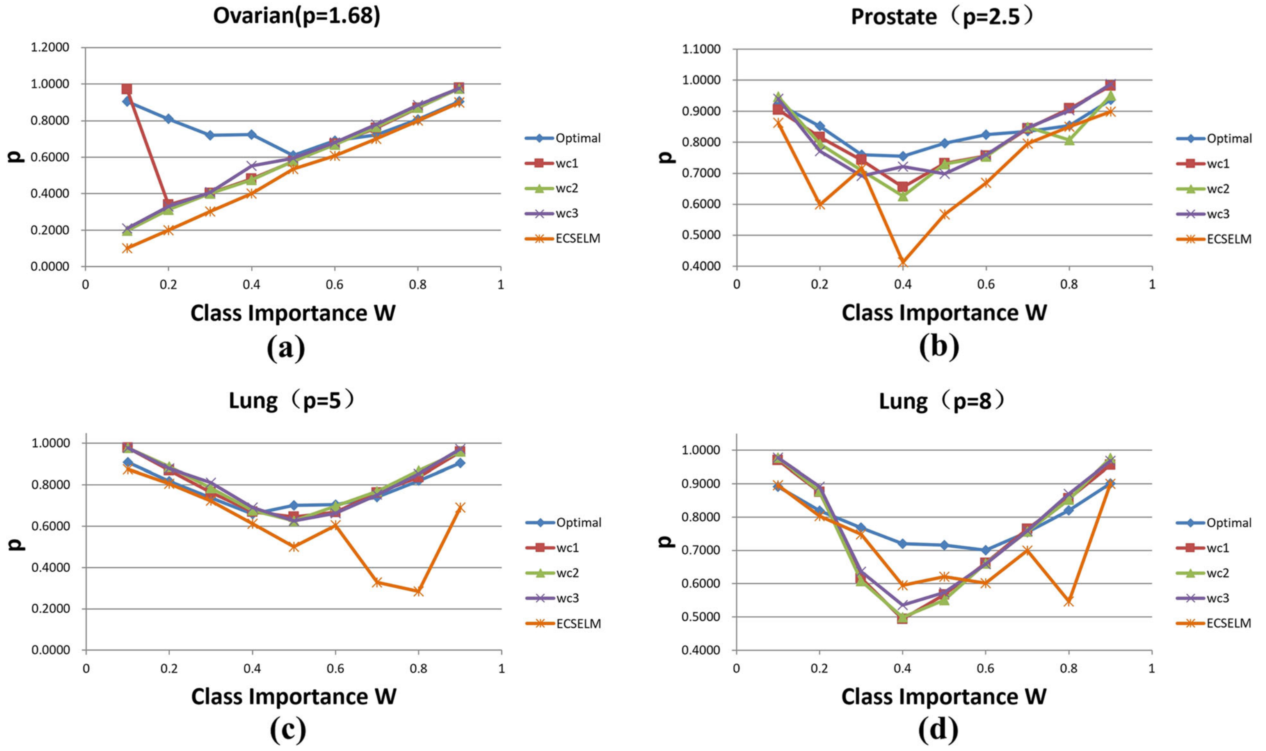

Fig. 6 WCA comparison with Ovarian, Prostate, Lung1, Lung2 dataset ($p = 1.68$, 2.5, 5, 8)

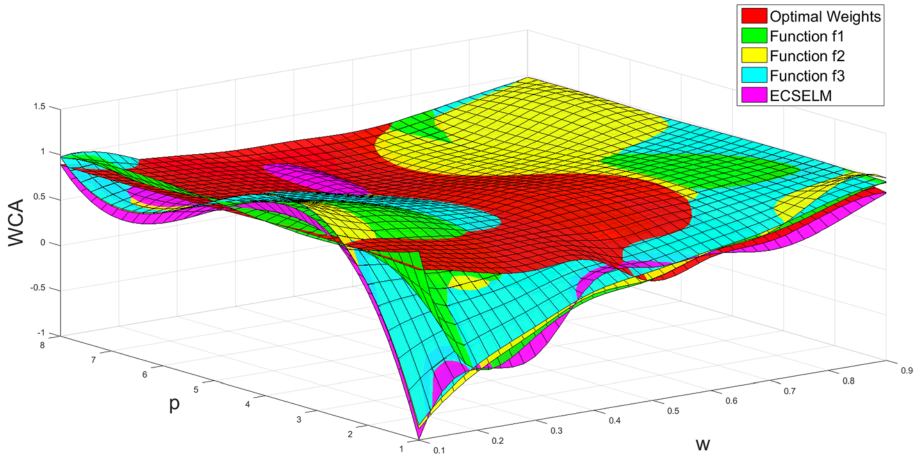

Fig. 7 The WCA comparison in 3-dimension

show that the function fitting approach can accurate find the optimal cost weights for imbalanced gene expression datasets.

The limitation of this work is that, although the ELM classifier is tested, the stability of the function fitting method is not proven, especially for other significantly different datasets. The exploration of the proposed algorithm's stability is left as future work.

Abbreviations

ACA: Adaptive classification accuracy; CCR: Correct classification rates; CSL: Cost sensitive learning; OA: Overall accuracy; WCA: Weighted classification accuracy

Acknowledgments

We thank T. R. Golub, D. K. Slonim et al. providing us the gene expression data.

About this supplement

This article has been published as part of *BMC Bioinformatics Volume 20 Supplement 25, 2019: Proceedings of the 2018 International Conference on Intelligent Computing (ICIC 2018) and Intelligent Computing and Biomedical Informatics (ICBI) 2018 conference: bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-25.

Authors' contributions

HL and MY conceived the project. YX developed the methodology, analyzed the results and wrote the manuscript. KY, ZG and QJ provided administrative and technical support. All authors read and approved the final manuscript.

Funding

This study is supported by National Natural Science Foundation of China (Nos. 61272315, 61402417, 61602431, 61701468, 61850410531, 61572164 and 61877015), Zhejiang Provincial Natural Science Foundation (Nos. Y1110342, LY15F020037 and LY19F020016) and International Cooperation Project of Zhejiang Provincial Science and Technology Department (No. 2017C34003).

Availability of data and materials

The datasets analyzed in this manuscript are available from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Key Laboratory of Electromagnetic Wave Information Technology and Metrology of Zhejiang Province, College of Information Engineering, China Jiliang University, Hangzhou, China. [2]College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. [3]Faculty of Human Sciences, Waseda University, Tokorozawa, Japan.

Published: 24 December 2019

References

1. Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(2):531–7.
2. Shipp M, Ross K, Tamayo P, Weng A, Kutok J, Aguiar R, Gaasenbeek M, Angelo M, Reich M, Pinkus G. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002;8(1):68–74.
3. Veer L, Dai H, Vijver M, He Y, Hart A, Mao M, Peterse H, Kooy K, Marton M, Witteveen A. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–6.
4. Deng S, Zhu L, Huang D. Predicting hub genes associated with cervical cancer through gene co-expression networks. IEEE Comput Soc Press. 2016;13(1):27–35.
5. Deng S, Zhu L, Huang D. Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks. BMC Genomics. 2015;16(Suppl 3):1–10.
6. Wang W, Lou B, Li X, Lou X, Jin N, Yan K. Intelligent maintenance frameworks of large-scale grid using genetic algorithm and k-mediods clustering methods. World Wide Web. 2019;2019(7):1573–413.
7. Deng S, Cao S, Huang D, Wang Y. Identifying stages of kidney renal cell carcinoma by combining gene expression and DNA methylation data. IEEE/ACM Trans Comput Biol Bioinform. 2016;14(5):1147–53.
8. Elkan C. The foundations of cost-sensitive learning. In: Seventeenth International Joint Conference on Artificial Intelligence; 2001. p. 973–8.
9. Zhou Z, Liu X. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng. 2006;18(1):63–77.
10. Yan K, Zhong C, Ji Z, Huang J: Semi-supervised learning for early detection and diagnosis of various air handling unit faults 2018, 181(12):75–83.
11. Liu X, Zhou Z. The influence of class imbalance on cost-sensitive learning: An empirical study. In: International Conference on Data Mining; 2007. p. 970–4.
12. Maheshwari S, Jain R, Jadon R. An insight into rare class problem: analysis and potential solutions. J Comput Sci. 2018;14(8):777–92.
13. Hu M, Li W, Yan K, Ji Z, Hu H: Modern machine learning techniques for univariate tunnel settlement forecasting: A comparative study 2019, 2019(4):1–12.
14. Chai X, Deng L, Yang Q, Ling C. Test-cost sensitive naive bayes classification. In: IEEE International Conference on Data Mining; 2004. p. 51–8.
15. Feng S. A cost-sensitive decision tree under the condition of multiple classes. In: International Conference on Logistics Engineering, Management and Computer Science; 2015. p. 1212–8.
16. Zhao H, Li X. A cost sensitive decision tree algorithm based on weighted class distribution with batch deleting attribute mechanism. Inf Sci. 2016;378(2):303–16.
17. Quinlan J. C4.5: Programs for machine learning: Morgan Kaufmann Publishers Inc; 1992.
18. Liu K, Huang D. Cancer classification using rotation forest. Comput Biol Med. 2008;38(5):601–10.
19. Lu H, Yang L, Yan K, Xue Y, Gao Z. A cost-sensitive rotation forest algorithm for gene expression data classification. Neurocomputing. 2017;228(C):270–6.
20. Turney P. Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm. AI Access Foundation. 1994;2(1):369–409.
21. Cao P, Zhao D, Zaiane O. An optimized cost-sensitive svm for imbalanced data learning. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining; 2013. p. 280–92.
22. Yuan H, Zhang X. Multiscale fragile watermarking based on the gaussian mixture model. IEEE Trans Image Process. 2006;15(10):3189–200.
23. Cheng D, Wu M. A novel classifier - weighted features cost-sensitive svm. In: IEEE International Conference on Internet of Things; 2017. p. 598–603.
24. Silva J, Bacao F, Caetano M. Specific land cover class mapping by semi-supervised weighted support vector machines. Remote Sens. 2017;9(2):1–16.
25. Cao P, Liu X, Zhao D, Zaiane O. Cost sensitive ranking support vector machine for multi-label data learning. In: International Conference on Hybrid Intelligent Systems; 2016. p. 244–55.
26. Zong W, Huang G, Chen Y. Weighted extreme learning machine for imbalance learning. Neurocomputing. 2013;101(3):229–42.
27. Zheng E, Zhang C, Liu X, Lu H, Sun J. Cost-sensitive extreme learning machine. In: International Conference on Advanced Data Mining and Applications; 2013. p. 478–88.
28. Liu Y, Lu H, Yan K, Xia H, An C. Applying cost-sensitive extreme learning machine and dissimilarity integration to gene expression data classification. Comput Intell Neurosci. 2016;2016(8):1–9.
29. Yan K, Ji Z, Lu H, Huang J, Shen W, Xue Y. Fast and accurate classification of time series data using extended elm: application in fault diagnosis of air handling units. IEEE Trans Syst Man Cybern Syst. 2017;49(7):1–8.
30. Zhang L, Zhang D. Evolutionary cost-sensitive extreme learning machine. IEEE Trans Neural Netw Learn Syst. 2017;28(12):3045–60.

31. Wang S, Li X, Zhang S, Gui J, Huang D. Tumor classification by combining pnn classifier ensemble with neighborhood rough set based gene reduction. Comput Biol Med. 2010;40(2):179–89.
32. Wang S, Zhu Y, Jia W, Huang D. Robust classification method of tumor subtype by using correlation filters. IEEE/ACM Trans Comput Biol Bioinform. 2012;9(2):580–91.
33. Zhu H, Wang X. A cost-sensitive semi-supervised learning model based on uncertainty. Neurocomputing. 2017;251(8):106–14.
34. Ailijiang A, Charapko A, Demirbas M. Consensus in the cloud: Paxos systems demystified. In: 25th International Conference on Computer Communication and Networks (ICCCN). Waikoloa: IEEE; 2016. p. 1–10.
35. Zheng C, Huang D, Kong X, Zhao X. Gene expression data classification using consensus independent component analysis. Genomics Proteomics Bioinformatics. 2008;6(2):74–82.
36. Yan K, Ji Z, Shen W. Online fault detection methods for chillers combining extended kalman filter and recursive one-class svm. Neurocomputing. 2017;228(3): 205–12.
37. Wang X, Wang J, Yan K. Gait recognition based on gabor wavelets and (2d) 2 pca. Multimed Tools Appl. 2017;77(10):1–17.
38. Pei S, Huang D. Cooperative competition clustering for gene selection. J Clust Sci. 2006;17(4):637–51.
39. Zheng C, Zhang L, Ng V, Shiu S, Huang D. Molecular pattern discovery based on penalized matrix decomposition. IEEE/ACM Trans Comput Biol Bioinform. 2011;8(6):1592–603.
40. Zheng C, Zhang L, Ng T, Shiu C, Huang D. Metasample-based sparse representation for tumor classification. IEEE/ACM Trans Comput Biol Bioinform. 2011;8(5):1273–82.
41. Zheng C. Tumor clustering using nonnegative matrix factorization with gene selection. IEEE/ACM Trans Comput Biol Bioinform. 2009;4(13):599–607.
42. Huang D, Zheng C. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics. 2006;22(15):1855–62.
43. Huang D, Yu H. Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(2):457–67.
44. Zhao X, Dheung Y, Huang D. Analysis of gene expression data using rpem algorithm in normal mixture model with dynamic adjustment of learning rate. Int J Pattern Recognit Artif Intell. 2010;24(4):651–66.
45. Zheng C, Huang D, Shang L. Feature selection in independent component subspace for microarray data classification. Neurocomputing. 2006;69(16):2407–10.
46. Zheng C, Huang D, Sun Z, Lyu M, Lok T. Nonnegative independent component analysis based on minimizing mutual information technique. Neurocomputing. 2006;69(7):878–83.
47. Cheng X, Chai F, Gao J, Zhang K. 1stopt and global optimization platform-comparison and case study. In: Proceedings of the 4th IEEE International Conference on Computer Science and Information Technology; 2011. p. 18–21.
48. Alejo R, Sotoca JM, García V, Valdovinos RM. Cost-sensitive neural networks and editing techniques for imbalance problems. In: Mexican Conference on Pattern Recognition. Berlin, Heidelberg: Springer; 2010. p. 180–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.