Noname manuscript No.
(will be inserted by the editor)

Cost-sensitive Algorithm with Local and Global Consistency

Weifeng Sun · Jianli Sun

Received: date / Accepted: date

Abstract Assuming that misclassification costs among different categories are equal, traditional graph based semi-supervised classification algorithms pursuit high classification accuracy. However, in many practical problems, misclassifying one category as another always leads to higher (or lower) cost than that in turn, such that, higher classification accuracy generally not means lower cost, which is more important in these problems. Cost-sensitive classification methods enable classifiers to pay more attention to data samples with higher cost, and then attempt to get lower cost by ensuring higher classification accuracy of the category with higher cost. In this paper, we bring cost sensitivity to local and global consistency (LGC) classifiers, and propose the CS-LGC (cost-sensitive LGC) methods, which can make better use of semi-supervised classification algorithms, and ensure high classification accuracy on the basis of reducing overall cost. At the same time, since the improved algorithm may bring some problems due to unbalanced data account, we introduce a SMOTE algorithm for further optimization. Experimental results of bank loan and diagnosis problems verify the effectiveness of CS-LGC.

Keywords Graph based semi-supervised classification · Cost-sensitive · Rescale · SMOTE

1 Introduction

Recently, machine learning classification algorithms have achieved sound development in multimedia identification, health care, finance, and many other applications. Among them, Graph-based Semi-Supervised Classification based on perfect

W. Sun
School of Software, Dalian University of Technology
DaLian, 116620 - China
E-mail:

J. Sun
School of Software, Dalian University of Technology
DaLian, 116620 - China
E-mail: sjl_dlut@163.com

graph theory, which is relatively straightforward and easy to understand, has become one of the hotest research focuses. This method can make better use of the information from label data and a great deal of data distribution information mined from unlabeled data, then solves the problem of the less labeled data and the huge labeling costs.

High classification accuracy is the target of GSSC algorithm, which assumes that the cost of misclassification among different categories is the same. However, in many practical problems, the costs of category classifications is different, especially in the fields of medical service, finance and network security and so on.

For example, in the bank loan problem, banks decide whether to approve the loan based on the loan applicants' personal information, which can be viewed as a binary classification problem. In the process of approving a bank loan, the cost of approving an applicant who is unable to repay the loan is much greater than that of approving an applicant who can repay the loan. Similarly, in the disease diagnosis, the cost of misdiagnose unhealthy patients as healthy is much greater than that of misdiagnose healthy patients as unhealthy. When the cost differences are inconsistent among different categories, a classifier paying more attention to the accuracy of higher cost classification problem is more practical than those treat all categories equally.

Cost-sensitive learning method [7–10, 12] takes the inconsistencies of cost among categories into account to lower the overall cost for the target. This method defines the static cost matrix to make classifiers pay more concern to the sample data with higher cost, and improves classification accuracy of higher-cost category. In the semi-supervised classification problems, labels account for a small portion of the data and most data are unlabeled. Actually, the traditional semi-supervised classification treats all categories equally, which cannot guarantee a lower overall cost. Meanwhile, cost-sensitive learning may arise less fit due to limited data set of tags, which could lead to lower classifier accuracy and weaker generalization ability. Besides, in the semi-supervised learning, we often encounter the situation called uneven data problem, in which different types of samples have different numbers of label data.

In summary, researches on cost-sensitive semi-supervised classification algorithms for unbalanced datasets are of great significance to the development of finance, medical fields, network security and many other areas.

2 Related Work

The effectiveness of semi-supervised learning algorithm depends on the three assumptions: manifold hypothesis[1, 2], mlustering hypothesis and smoothness hypothesis. GSSC based on manifold hypothesis builds a figure to describe the data, as well as the relationship between the data. In this figure, nodes represent the data samples while edges with weight represent the relationships between samples. At the same time, the larger the weight is, the higher the similarity of the samples have. The process that GSSC classifiers assign labels to unlabeled data is the process of label propagation in the figure. Label Propagation Algorithm Budyytis et al. proposed in [4] can calculate the probability of transfer among tag data by the topology and the similarity between the samples in figure, and make a label

transfer by combining node out-degree. A method of Gaussian fields and harmonic functions proposed by Zhu et al. in [5] that make discrete prediction function slack to become continuous prediction function considers the transfer probability samples fully and have a label transfer in k-connection diagram. Local and global consistency, LGC proposed by Zhou et al. in the [6] introduced clustering hypothesis and had a label transfer by using local and global consistency. LGC algorithm gives a rigorous mathematical logic derivation and proves the convergence. However, these related works never consider the problem of inconsistent classification cost.

Cost-sensitive learning method is an effective way to solve the problem of the inconsistency of costs. Cost-sensitive learning method introduces cost matrix to describe the inconsistency of costs among categories and get global minimum cost. The cost-sensitive classification methods can be divided into two categories: Rescale and Reweight, depending on the representations of cost. In the method of Rescale, differences in cost are described as differences among the numbers of samples, such as Cost-sensitive sampling[7], Rebalance[8] and Rescale new[9] etc. This method constructs different sample data sets according to the difference in costs which make classifiers' decision face prefer samples with more costly category. Reweight method describes differences in costs by differences in weight among samples of different types. In the method of Reweight, samples with costly category have a higher weight, which make a greater impact to classifier. MetaCost[10] is a typical representative of such Reweight methods. MetaCost based on Bayesian risk theory adds the cost of the sensitive nature for Non-cost-sensitive classification algorithm by using bagging[11].

AdaBoost algorithm was proposed in[13] which offers the possibility for multiple weak classifiers aggregating into a global strong classifier. AdaCost method[12] was proposed as a cost-sensitive classification algorithm based on AdaBoost. AdaCost is based on the Reweight at the same time, and introduces cost performance function for classifiers by heuristic strategy. AdaCost forces classifier to pay more attention to costly samples, hence it shows some advantages in cost-sensitive classification problems. However, cost performance function introduced in the theoretical analysis have not been verified and damages the most important characteristics of Boosting, which makes the algorithm not converge to the Bayesian decision. Qin etc. in [14] did try to combine semi-supervised classification algorithms with cost-sensitive learning methods, and improved classical EM algorithm by introducing misclassification cost in the process of probability assessment.

In this paper, advantages of many unlabeled data are fully taken. We use the method of Rescale to describe cost inconsistency and introduce cost-sensitive nature for LGC algorithm classic semi-supervised classification algorithm. We propose a cost-sensitive LGC methodCS-LGC algorithm. At the same time, we take the impact caused by unbalanced data into account for CS-LGC algorithm, improve the CS-LGC algorithm by proposing the average similarity concept and introduce SMOTE algorithm. The main contribution of this paper is as follows:

(1) Introduce cost-sensitive nature for LGC algorithm, and propose cost-sensitive LGC algorithm;
(2) Propose CSS-LGC algorithm. We take the impact caused by unbalanced data into account for CS-LGC algorithm, and we propose optimized CS-LGC algorithmCSS-LGC algorithm.

(3) Analyze the experimental threshold, and verify the rationality of the threshold.
(4) Demonstrate the effectiveness of the algorithm by experiment on German Credit Data Set and Breast Cancer Data Set.

3 CSS-LGC Algorithm

3.1 Problem Definition

Many problems can be described as binary classification problems (multiple class classification problems can be split into multiple binary classification problem), so this paper simply discusses binary classification problem. The two classes can be defined as positive class whose sample number is denoted as N_+, and negative class with the sample number denoted as N_-. We further assume C_{-+}, the cost that a sample in negative class is classified into positive class is much bigger than that in turn.X $= \{x_1, x_2, \ldots, x_l, x_{l+1}, \ldots, x_{l+u}\}$ is a set of data samples, where l is the number of labeled data, and $\{x_1, x_2, \ldots, x_l\}$ forms the initial labeled data sets. u represents the number of unlabeled data, thus $U = l + u$ denotes the total number of samples. x_i is a multidimensional vector of the i-th features in data, which can be viewed as a point in higher dimensional space. Those data points and their relationships can be described by a fully connected graph $G = (V, E)$, where $V \in R^{N \times P}$ represents the vertex set of G, P is the feature value of each data sample. we have $V = X$ here, and E is the edge set of G. Similarity among data points is important in classification, clustering and other machine learning fields. In LGC algorithm, similarity among data points is described by weight matrix $W \in R^{N \times N}$. Labeled data information is described by labeled information matrix $Y \in L^{N \times 2}$, where $L = \{+1, -1\}$ If data i belongs to category j, then $Y_{ij} = 1$, otherwise $Y_{ij} = 0$. The initial labeled information matrix only has a little labeled information.

Liu and Zhou mentioned that the cost of misclassification can be standardized in [15], which will not affect the best decision when calculation is simplified. In this paper, we let $c_{+-} = 1$, so c_{-+} is a real number larger than 1. In order to make semi-supervised classifier cost sensitive, We introduce sample cost information for the initial label matrix, denote cost difference using different initial labeled information volumes, and process the initial label matrix as Formula 1 shows:

$$YN_{\cdot i} = T_0 \circ Y_{\cdot i} \tag{1}$$

$N_{\cdot i}$ indicates the i-th column of the processed initial label matrix. $\circ$ indicates Adama product. T_0 indicates the cost of sample data at the initial time. In addition

$$T_0(x_i) = \begin{cases} 1/N_+ & \text{if } x_i \text{ is positive class} \\ C_{-+}/N_- & \text{if } x_i \text{ is negative class} \end{cases} \tag{2}$$

CS-LGC method is designed to classify all data samples using the information of graph G with unknown Y and T_0. Our goal is to ensure unlabeled data to obtain class label as much as possible without changing the information of labeled data, and at the same time, minimize the global classification cost without damaging classification accuracy.

3.2 Algorithm Process

CS-LGC method is a boosting process that trains semi-supervised classifiers in each iteration and update labeled data set based on the performance of the classifiers. At the initial time, semi-supervised classifier h_0 is trained using traditional LGC algorithm according to preprocessed initial label matrix, and we further calculate error rate by Formula 3.

$$\varepsilon_0 = \frac{\sum_{i=1}^{N} I\left(y_{i,h_0(x_i)} = 0\right)}{N} \tag{3}$$

Here I is a binary function, whose value is 1 if conditions are satisfied, or is 0 else. $h_0\left(x_i\right)$ shows classification results for x_i at the initial time. Then according to the AdaBoost algorithm, cost information of misclassified samples increases while cost information of correctly classified samples decreases following Formula 4,

$$T_{t+1}\left(x_i\right) = \frac{T_t\left(x_i\right)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } y_{i,h_t(x_i)} = 1 \\ e^{\alpha_t} & \text{if } y_{i,h_t(x_i)} = 0 \end{cases} \tag{4}$$

where T_{t+1} represents the updated cost matrix, and

$$Z_t = \sum_{i=1}^{N} T_t\left(x_i\right) \times \begin{cases} e^{-\alpha_t} & \text{if } y_{i,h_t(x_i)} = 1 \\ e^{\alpha_t} & \text{if } y_{i,h_t(x_i)} = 0 \end{cases} \tag{5}$$

where α_t represents the weight of classifier h_t in the final global classification at time t, and can be calculated by

$$\alpha_t = \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \tag{6}$$

The main objective of Cost-sensitive classifier is to minimize global classification. To avoid the impact of data set scale on the results, the calculation of global cost in our model use the mean cost method brought by Seiffert et al[16]. Since we have to standardize classification cost of our model, the corresponding mean cost needs to be updated according to Formula 7:

$$PEC = \frac{\#f_{pos} + \#f_{neg} \cdot C_{-+}}{\#t_{pos} + \#t_{neg} + \#f_{pos} + \#f_{neg}} \tag{7}$$

where PEC denotes the average cost of classifiers, $\#t_{pos}$ and $\#f_{pos}$ denote the number of positive class samples classified correctly and classified incorrectly respectively. $\#t_{neg}$ and $\#f_{neg}$ represent the corresponding information of negative class. The global cost we refer to here that is indeed the average cost. From Formula 7 we can see that the correct classified samples does not increase the global cost.

Next, we update the label matrix YN according to Rescale Method. Specifically, when the cost of x_i increases, means there is a wrong classification. Then a certain number of data that have the highest similarity with x_i are chosen from unlabeled data set and added into labeled data set. The class and cost information of the newly added data are the same with x_i. On the contrary, when sample data x_i's cost decreases, and the decreasing range is larger than the threshold ts, then we will remove x_i and its label information from the labeled data set.

After then, we train the classifier of the next time according to updated label matrix and repeat the process until the global cost difference between two training time step is less than a certain threshold. Finally, a global classifier can be obtained by the way of weighted voting.

$$Y_{final}\left(x_i\right) = sign\left(\sum_{j=1}^{M} \alpha_j h_j\left(x_i\right)\right) \tag{8}$$

Here M denote the number of classifier when the iteration ends.

In summary, the process of CS-LGC method is shown in Algorithm 1:

Algorithm 1 CS-LGC

Input: data set X, the initial label matrix Y

1: add cost information for the initial label matrix according to Formula 1
2: **while** $\Delta PEC \geq threshold$ **do**
3: train LGC semi-supervised classifier
4: update the cost matrix by Formal 4 according to the result
5: update label data set applying Rescale Method
6: update the initial label matrix according to Formula 1
7: **end while**

Output: $\mathrm{Y}_{final}(x_i) = sign(\sum_{j=1}^{M} \alpha_j h_j(x_i))$, $PEC = \frac{\#f_{pos}+\#f_{neg}\cdot C_{-+}}{\#t_{pos}+\#t_{neg}+\#f_{pos}+\#f_{neg}}$

3.3 The convergence of the algorithm

Proof The convergence of the CS-LGC algorithm depends on the convergence of the LGC, and whether the cost of the sample data by wrong classification has an upper limit. For ease of reading, we use $c_i(\frac{1}{N_+}$ or $\frac{C_{-+}}{N_-}$ or 0) to denote the cost of sample i, and y_i(-1 or +1) to denote the real label of sample i.

Zhou gives the evaluation function based on FIG of the semi-supervised classification algorithm in [6], which is usual composed of smooth function and loss function. According to extremum condition acquired by evaluation function as well as recursion formula of LGC algorithm, Zhou deduces the convergence formula(Formula 9) of LGC algorithm, then proves the convergence of LGC algorithm.

$$F = (I - \beta S)^{-1} Y \tag{9}$$

Here F is the final classification function, $\beta \in (0, 1)$, S is a symmetric matrix, and

$$\mathrm{S} = D^{-1/2} W D^{-1/2} \tag{10}$$

where D is a diagonal matrix, whose diagonal elements is the sums of elements of corresponding row vectors in the weight matrix W.

For the cost matrix of CS-LGC algorithm, Formula 4 can be rewrited as

$$\begin{aligned} T_{t+1}(i) &= T_t(i)\frac{e^{-\alpha_t y_i h_t(x_i)}}{Z_t} \\ &= T_1(i)\frac{e^{-y_i \sum_{j=1}^{M} \alpha_j h_j(x_i)}}{\prod_{j=1}^{S} Z_j} \end{aligned} \tag{11}$$

Because of $\sum_{i=1}^{N} T_{t+1}(i) = 1$, we have

$$\begin{aligned} \prod_{j=1}^{M} Z_j =& \frac{1}{N_+}\sum_{i=1}^{N_+} e^{-y_i \sum_{j=1}^{M} \alpha_j h_j(x_i)} \\ &+ \frac{C_{-+}}{N_-}\sum_{k=1}^{N_-} e^{-y_k \sum_{j=1}^{M} \alpha_j h_j(x_k)} \end{aligned} \tag{12}$$

where x_i denotes the positive class sample, and x_k denotes the negative class sample.

In the binary classification problem, the label of data is either -1 or +1, and when $y_i \neq Y_{final}(x_i)$, $y_i Y_{final}(x_i) \leq 0$ holds, then we have $e^{-y_i Y_{final}(x_i)} \geq 1$. The global cost of the classification can be calculated by Formal 13,

$$\begin{aligned} E &= \sum_{j=1}^{N} c_j \begin{cases} 1 & \text{if } y_j \neq Y_{final}(x_j) \\ 0 & \text{else} \end{cases} \\ &\leq \sum_{j=1}^{N} c_j e^{-y_j Y_{final}(x_j)} \\ &= \sum_{j=1}^{N} c_j \prod_{i=1}^{M} Z_i \frac{T_M(x_j)}{T_0(x_j)} \\ &= d\prod_{i=1}^{M} Z_i \end{aligned} \tag{13}$$

there d denotes the sum of all the data samples' cost initially.

In conclusion, when the cost matrix is decided, LGC algorithm will converge to Formal 9, and the upper bound of cost matrix does exist (Formal 13), thus CS-LGC algorithm can converge. □

References

1. R. Urner, S. Shalev-Shwartz, and S. Ben-David, "Access to unlabeled data can speed up prediction time," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 641–648, 2011.
2. Z.-H. Zhou and M. Li, "Semi-supervised learning by disagreement," *Knowledge and Information Systems*, vol. 24, no. 3, pp. 415–439, 2010.

3. X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.
4. I. Budvytis, V. Badrinarayanan, and R. Cipolla, "Label propagation in complex video sequences using semi-supervised learning.," in *BMVC*, vol. 2257, pp. 2258–2259, Citeseer, 2010.
5. X. Zhu, Z. Ghahramani, J. Lafferty, *et al.*, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, vol. 3, pp. 912–919, 2003.
6. D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, no. 16, pp. 321–328, 2004.
7. Y. Gao and J. Wang, "Active learning method of bayesian networks classifier based on cost-sensitive sampling," in *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on*, vol. 3, pp. 233–236, IEEE, 2011.
8. C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, pp. 973–978, Citeseer, 2001.
9. Z.-H. Zhou and X.-Y. Liu, "On multi-class cost-sensitive learning," *Computational Intelligence*, vol. 26, no. 3, pp. 232–257, 2010.
10. P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 155–164, ACM, 1999.
11. T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Comparing boosting and bagging techniques with noisy and imbalanced data," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 41, no. 3, pp. 552–568, 2011.
12. W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "Adacost: misclassification cost-sensitive boosting," in *ICML*, pp. 97–105, 1999.
13. X. Jin, X. Hou, and C.-L. Liu, "Multi-class adaboost with hypothesis margin," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 65–68, IEEE, 2010.
14. Z. Qin, S. Zhang, L. Liu, and T. Wang, "Cost-sensitive semi-supervised classification using cs-em," in *Computer and Information Technology, 2008. CIT 2008. 8th IEEE International Conference on*, pp. 131–136, IEEE, 2008.
15. X.-Y. Liu and Z.-H. Zhou, "The influence of class imbalance on cost-sensitive learning: An empirical study," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pp. 970–974, IEEE, 2006.
16. C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "A comparative study of data sampling and cost sensitive learning," in *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*, pp. 46–52, IEEE, 2008.
17. K. Bache and M. Lichman, "Uci machine learning repository," 2013.
18. H. Hofmann, "German credit data," 2000.