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 cost among different categories equals to each other, 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.

2 Section title

Text with citations and

2.1 Subsection title

as required. Don't forget to give each section and subsection a unique label (see Sect. 2).

Paragraph headings Use paragraph headings as needed.

$$a^2 + b^2 = c^2 \tag{1}$$

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.