Information Fusion 39 (2018) 1–24

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Adaptive multi-objective swarm fusion for imbalanced data classification

Jinyan Li[a,*], Simon Fong[a], Raymond K. Wong[b], Victor W. Chu[b]

[a] Department of Computer Information Science, University of Macau, Macau SAR
[b] School of Computer Science and Engineering, University of New South Wales, Australia

ARTICLE INFO

Article history:
Received 21 September 2016
Revised 4 March 2017
Accepted 26 March 2017
Available online 28 March 2017

Keywords:
Swarm fusion
Swarm intelligence algorithm
Multi-objective
Crossover rebalancing
Imbalanced data classification

ABSTRACT

Learning a classifier from an imbalanced dataset is an important problem in data mining and machine learning. Since there is more information from the majority classes than the minorities in an imbalanced dataset, the classifier would become over-fitted to the former and under-fitted to the latter classes. Previous attempts to address the problem have been focusing on increasing the learning sensitivity to the minorities and/or rebalancing sample sizes among classes before learning. However, how to efficiently identify their optimal mix in rebalancing is still an unresolved problem. Due to non-linear relationships between attributes and class labels, merely to rebalance sample sizes rarely comes up with optimal results. Moreover, brute-force search for the perfect combination is known to be NP-hard and hence a smarter heuristic is required. In this paper, we propose a notion of swarm fusion to address the problem – using stochastic swarm heuristics to cooperatively optimize the mixtures. Comparing with conventional rebalancing methods, e.g., linear search, our novel fusion approach is able to find a close to optimal mix with improved accuracy and reliability. Most importantly, it has found to be with higher computational speed than other coupled swarm optimization techniques and iteration methods. In our experiments, we first compared our proposed solution with traditional methods on thirty publicly available imbalanced datasets. Using neural network as base learner, our proposed method is found to outperform other traditional methods by up to 69% in terms of the credibility of the learned classifiers. Secondly, we wrapped our proposed swarm fusion method with decision tree. Notably, it defeated six state-of-the-art methods on ten imbalanced datasets in all evolution metrics that we considered.

1. Introduction

Imbalance dataset is referred to the phenomenon where there are far more samples in majority classes than minorities. Data mining applications suffer from the imbalanced data problem arisen from this phenomenon ubiquitously, such as in big data analytics and text mining [1], forecasting natural disasters [2], fraud detection in transactions [3], target identification from satellite radar images [4], classifying biological anomalies [5] as well as computer-assisted medical diagnosis and treatment [6], etc. In particular, classifier learning from imbalanced data has long been an important and challenging problem in data mining and machine learning [7]. Conventional supervised learning algorithms by greedy search are usually designed to embrace the imbalanced dataset without regards to the class balance ratio. Moreover, most of classification learning models are designed without the consideration of imbalanced date problem. The models obtained from these training methods are commonly suffered from over-fitting owing to the sheer volume of majority training data. Besides, the recognition power for identifying rare test samples is limited. It is because the models are also suffered from under-fitting due to the lack of sufficient training from minority samples.

Preprocessing-styled rebalancing schemes have been proposed in the past to address these problems. They mainly focus on the aspects of artificially inflating the minority class data, resampling down the volume of the majority class data, or a combination of the two. However, it has been shown that merely matching the quantities of the majority and minority data does not yield the highest possible classification performance [8]. In parallel, performance metrics have been developed for the evaluation of whether a classification model is inadequate owing to imbalance training. They go beyond just for accuracy measurements. Some useful metrics can be found in recent literature, which are based on the counts of true-positive and/or false-positive, e.g., G-mean [8], F1 measure [9,10], Kappa statistics [11], AUC/ROC [12], Matthews correlation coefficient(MCC) [13] and Balance error rate(BER) [14].

* Corresponding author.
E-mail addresses: yb47432@umac.mo (J. Li), ccfong@umac.mo (S. Fong), wong@cse.unsw.edu.au (R.K. Wong), wchu@cse.unsw.edu.au (V.W. Chu).

http://dx.doi.org/10.1016/j.inffus.2017.03.007
1566-2535/

This paper proposes a notion of swarm fusion (an adaptive rebalancing model) to address imbalanced classification problem. It effectively avoids the drawbacks of imbalance datasets and absorbs the advantages of current rebalancing techniques. Most importantly, the classifiers learned from our rebalanced model are found to have higher performance as shown by reliability measures. In the process of rebalancing, the parameters are automatically and adaptively controlled. It joints the rebalancing techniques of increasing minority class samples and decreasing majority class samples, and simultaneously, it completes a swarm fusion operation till achieving the current best performance within a reasonable time. Given that the volume of data can be enormous in a dataset, finding the best ratio between majority and minority classes of data is a challenging combinational optimization problem. Without resorting to brute-force search, swarm optimization is applied on two aspects of rebalancing. One on searching for the appropriate amount of majority instances, and the other one on estimating the best combo of control parameters — the intensity and how far that the neighbors of the minority samples are to be synthesized — with respect to enlarging the minority population size.

Our proposed unified rebalancing approach is called Adaptive Multi-objective Swarm Crossover Optimization (AMSCO) [15]. In AMSCO, the optimization of majority instances is called Swarm Instance Selection (SIS(O_maj)) and the optimization of minority instances is called Optimized Synthetic Minority Oversampling Technique (OSMOTE). Our proposed rebalancing method couples these two optimizations together as a unified iterator. It progressively enhances the mixtures of the optimized data from the two swarm optimizations by crossing over their optimized results, generation-after-generation, until a good quality dataset is produced. In this new approach, we used Particle Swarm Optimization (PSO) as the core optimizer whose searching particles represent the solution candidates. A summary of the optimization processes is as follows. The original dataset will become the current dataset after it is loaded for the first time. The current dataset will be checked with respect to its quality by inferring a candidate classifier from it. The current dataset will be subject to two parallel swarm optimizations for optimally increasing the minority samples and decreasing the majority samples until the performance of the candidate classifier meets our threshold. The two swarms operate independently because their candidate solutions are different in nature. However, their outputs are crossed over by selectively merging instances from the most competent optimized datasets into one, which is formed by the size of the original dataset. The selected dataset in turn becomes the current dataset when the optimization cycle iterates. The dataset is checked by a goodness measure and passed to two swarm operations again if it does not meet requirements. Classification algorithm which works like a wrapper approach is used to test the goodness of the current dataset.

The remainder of this paper is structured as follows. Section 2 reviews popular approaches to address imbalanced dataset classification problem. In addition, it contains a briefly introduction of swarm intelligence algorithms. In Section 3, we elaborate the method and the design of our proposed Adaptive Multi-objective Swarm Crossover Optimization (AMSCO) method. Our datasets, and our extensive experiments and their results are presented in Section 4. Section 5 concludes this paper.

2. Related works

Imbalanced classification problem is an enthusiastic topic in the fields of data mining, machine learning and pattern recognition. There are many leading conferences held special workshops for the discussion and study of this problem, like in ACM SIGKDD 2004 [16], AAAI 2000 [17], ICML 2003 [18] [19], etc. At present, the researches for solving imbalanced classification can be roughly into two categories: i) data level and ii) algorithm level. Previous researcher proposed that there are four main factors for tackling imbalanced classification problem. They are i) training set size, ii) class priors, iii) cost of errors in different classes, and iv) placement of decision boundaries [20]. The data level aim at reducing the imbalanced ratio of imbalanced classification model by adjusting the distribution of samples in dataset. Another level of the algorithm makes the classifier more inclined to the minority class through modifying conventional classifier.

Since the design of most conventional classifiers, previous researchers found that the performance of balanced dataset is better than imbalanced classification performance [18]. Therefore, people proposed many methods for sampling the imbalanced dataset, in order to change the distribution of samples and rebalancing the imbalanced dataset. Over-sampling and down-sample respectively increase the number of minority class samples and decrease the number of majority class samples. Random over-sampling through randomly repeat minority class samples to increase the number of minority class samples, but this method will easily cause overfitting. Chawla proposed synthetic minority over-sampling technique (SMOTE) [21] is the most widely and effectively used oversampling method. It synthetizes new minority samples through learning several neighbors in the same class of each minority class sample in order to generate minority samples and rebalancing the imbalanced dataset. SMOTE synthesizes N times new minority class samples and each minority class sample $x_i \in S_{minority}$. K neighbors of x_i in minority class samples are examined, then to randomly select x_t from the K neighbors using Eq. (1) to generate the synthetic data $x_{new,N}$, i.e.,

$$x_{new,N} = x_i + rand\,[0,1] \times (x_t - x_i) \tag{1}$$

In Eq. (1) *rand* [0,1] generates a random number between 0 and 1. N and K influence SMOTE to generate a suitable number of characteristic minority class samples. Following is the pseudo code of SMOTE.

Algorithm SMOTE

Input the number of minority class data T; Define the oversampling rate N and K nearest neighbors
1. **if** $N < 100\%$
2. **do** randomly select minority class samples as the rate N to do the oversampling
3. **end**
4. $N = \text{int}(N)$
5. **for** j = 1:N
6. **for** i = 1:T
7. Compute the K nearest neighbors of minority class sample x_i
8. Randomly select a neighbors x_t from the K nearest neighbors
9. **for** m = 1: (number of attributes = A)
10. Compute: $x_{\text{j, i\_new(attribute\_m)}} = x_{\text{i(attribute\_m)}} + \text{rand}(0,1)^* (x_{t\,(attribute\_m)} - x_{i(attribute\_m)})$
11. **end**
12. synthetic $x_{j,\ i\_new\ (\text{attribute\_1},\ \text{attribute\_2},\ \ldots,\ attribute\_A)}$
13. **end**
14. $x_{j\_new} = [x_{j,1\_new};\ x_{j,2\_new};\ \ldots; x_{j,T\_new}]$
15. **end**
16. synthesized minority class samples = $[x_{1\_new};\ x_{2\_new};\ \ldots; x_{N\_new}]$

Although over-sampling is able to reduce the imbalanced ratio, but the original minority class samples may be diluted by a large amount of synthetic samples. Down-sampling discards a part of majority class samples to rebalancing the dataset. Random down-sampling could be losing some valuable and characteristic samples. Balance Cascade [22] is a classical under-sampling method. Through iteration strategy, it guidance removes the useless majority class samples step by step.

The algorithm level contains two main approaches to improve imbalanced classification, cost-sensitive learning and ensemble learning. In the classification process, they make the base clas-