Engineering Applications of Artificial Intelligence 57 (2017) 38–49

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Feature selection for high dimensional imbalanced class data using harmony search

Alireza Moayedikia[a,b,*], Kok-Leong Ong[b], Yee Ling Boo[c], William GS Yeoh[a], Richard Jensen[d]

[a] Department of Information Systems and Business Analytics, Deakin University, Victoria 3125, Australia
[b] SAS Analytics Innovation Lab, La Trobe University, Victoria 3086, Australia
[c] School of Business IT and Logistics, RMIT University, Victoria 3000, Australia
[d] Department of Computer Science, IMPACS, Aberystwyth University, Wales, UK

ARTICLE INFO

Keywords:
Feature selection
Harmony search
High-dimensionality
Imbalanced class
Symmetrical uncertainty

ABSTRACT

Misclassification costs of minority class data in real-world applications can be very high. This is a challenging problem especially when the data is also high in dimensionality because of the increase in overfitting and lower model interpretability. Feature selection is recently a popular way to address this problem by identifying features that best predict a minority class. This paper introduces a novel feature selection method call SYMON which uses symmetrical uncertainty and harmony search. Unlike existing methods, SYMON uses symmetrical uncertainty to weigh features with respect to their dependency to class labels. This helps to identify powerful features in retrieving the least frequent class labels. SYMON also uses harmony search to formulate the feature selection phase as an optimisation problem to select the best possible combination of features. The proposed algorithm is able to deal with situations where a set of features have the same weight, by incorporating two vector tuning operations embedded in the harmony search process. In this paper, SYMON is compared against various benchmark feature selection algorithms that were developed to address the same issue. Our empirical evaluation on different micro-array data sets using G-Mean and AUC measures confirm that SYMON is a comparable or a better solution to current benchmarks.

1. Introduction

The presence of imbalanced data is a problem for classification algorithms (López et al., 2013; Cerf et al., 2013). An imbalanced data set is one where at least one class is under-represented compared to the others. Such data creates many challenges to the process of knowledge discovery and has many implications in real-world applications (Van Hulse and Khoshgoftaar, 2009). Addressing these issues brings about many good solutions, such as MIROS[1] that is used to detect the possibility of oil spilling (Topouzelis et al., 2007), or to detect malicious activities of users in the context of network intrusion as seen in the AIDE (Subba et al., 2015) environment.[2] In this paper, we investigate the imbalanced class problem further by considering cases where the data set is also high in dimensionality (e.g. large-scale data sets Silva et al., 2016), thus making the problem more pronounced as the efficacy of learning algorithms is further reduced (Van Hulse et al., 2009; Yin et al., 2013; Maldonado et al., 2014).

Different approaches have been proposed to address the imbalanced learning problem, including resampling (Chawla et al., 2004; Deepa and Punithavalli, 2011), one-class learning (Chawla et al., 2004), cost-sensitive learning (He and Garcia, 2009; Sun et al., 2007) and feature selection (Yin et al., 2013; Maldonado et al., 2014; Alibeigi et al., 2012). In resampling, the two most common techniques used for the imbalanced data problem are (i) random oversampling and (ii) random undersampling (Van Hulse et al., 2009; Maldonado et al., 2014). In the former, random duplicates of instances from the minority class are added to the original data set, leading to longer classifier training time. With the latter, the instances from the majority classes are randomly discarded, thus the information loss (Van Hulse et al., 2009) usually leads to sub-optimal learning outcomes.

One example of random oversampling is **S**ynthetic **M**inority **O**versampling **TE**chnique (SMOTE) proposed by Chawla et al. (2004) and Deepa and Punithavalli (2011). The algorithm generates artificial examples for the minority class by interpolating the current minority instances and has been shown Anil Kumar and Ravi (2008) to improve classification performance over imbalanced data. Unfortunately, creat-

* Corresponding author.
E-mail address: amoayedi@deakin.edu.au (A. Moayedikia).
[1] http://goo.gl/aR5elt
[2] http://aide.sourceforge.net/

http://dx.doi.org/10.1016/j.engappai.2016.10.008
Received 10 May 2016; Received in revised form 10 September 2016; Accepted 15 October 2016

ing artificial examples may not always be possible, such as in critical applications like medical diagnosis tools that rely on real data for diagnosis (Tho Thong et al., 2015). In this case, artificial data might affect the accuracy of diagnosis adversely. Hence, solutions that do not attempt to alter the original data in the learning process remain desirable.

One-class learning is an exemplar that does not alter the original data during the learning process. It operates by classifying each instance based on a similarity threshold (Chawla et al., 2004). This approach minimises overfitting seen in other classifiers when one class is significantly overrepresented than the others in the data. Consequently, one-class learners may lead to better predictive performance but that accuracy is dependent on the similarity threshold, which needs to be empirically tuned (Chawla et al., 2004) to achieve the desired performance.

Another way to address the challenge of classification from imbalanced data is cost-sensitive learners. As the name suggests, these learners consider the cost of misclassification and therefore seek to minimise the likelihood of misclassifying a minority class through a cost matrix. They are also quite effective for large data sets as they concurrently minimise learning time (Maldonado et al., 2014) and the misclassification of minority classes.

Recently, researchers are gaining interest in using feature selection (Chen and Wasikowski, 2008; Yin et al., 2013; Wang et al., 2015; Koprinska et al., 2015) as a way to address the imbalanced class problem. Previous approaches (i.e. resampling, one-class learning and cost-sensitive learning) have focused on the samples of the training data. Feature selection on the other hand, takes a different view by shifting the focus to the features (i.e. dimensions) rather than the training examples. The key idea is to find a subset of features that optimise the contrast between classes in the data.

Within feature selection, there are three further sub-approaches: filter, wrapper and hybrid (also known as embedded). Generally, filter approaches will find a subset that is good but may not be optimal for a specific classifier (Chen and Wasikowski, 2008). Hence, wrapper (Yang et al., 2013; Yin et al., 2013) and embedded approaches (Maldonado et al., 2014) were proposed to produce a more targeted feature subset. These approaches can be based on the ranking of features, where the criteria is often a loss function, e.g. the contribution of a feature to the classification rate (Van Hulse et al., 2009; Maldonado et al., 2014), or the discriminative power of features (Yin et al., 2013).

We argue that selecting features based on a loss function does not always yield the best learning outcome for the classifier. Rather, ranking features with respect to their dependency towards a class label and using that information to select the feature subset would give better performance, especially in predicting the minority class. As a result, the algorithm that we propose, called SYMON, is unique in the following ways:

- SYMON is a wrapper approach. Hence the chosen subset will be more relevant to the induced classifier (Chen and Wasikowski, 2008).
- SYMON uses Symmetrical Uncertainty to rank features, giving insight to how relevant a feature is to a class label. This differs from other feature selection algorithms (addressing the imbalanced class problem) which select features based on a loss function. As seen later in the experimental results, this gives SYMON better overall performance.
- SYMON handles high dimensionality well. This is important especially when there is a large number of features and finding the best feature combination should be computationally efficient. SYMON uses Harmony Search to reduce the complexity of the search process (Diao and Shen, 2012), thus ensuring its relevance in practice.
- With high dimensionality, the likelihood of multiple features sharing the same rank is high. This presents a challenge as most feature selection algorithms lack a mechanism to pick the best subset from identically-ranked features. SYMON does not suffer from this issue as it incorporates vector tuning operations.

In the next section, we review recent feature selection algorithms that deal with learning from high dimensional imbalanced data. Once this context is established, we will introduce SYMON in Section 3 with discussion of the experimental results in Section 4. We then conclude with future work for SYMON in Section 5.

2. Related works

Given the specific focus of this paper, we shall only discuss the relevant literature that use feature selection in the context of tackling the imbalanced class problem in high dimensional settings.

One solution by Yin et al. (2013) is a variation of using Bayesian learning as a solution to the imbalanced class problem. The approach proposed by Yin et al. works on the assumption that samples in the majority classes have a dominant influence on general feature selection techniques. The first step is thus to decompose classes with large examples into smaller pseudo-subclasses. Feature selection is then performed on the decomposed data, where the pseudo-subclasses balance the skew across classes, thus neutralising the influence the majority class examples have on feature selection algorithms. Their evaluation over synthetic data showed better feature selection performance once the imbalanced data has been decomposed.

Alibeigi et al. (2012) proposed a different approach with an algorithm called Density-based Feature Selection (DBFS). As the name suggests, features are ranked by their estimated probability density. This is done by exploring the contribution of each feature, taking into account the features' corresponding distributions over all classes and their correlations. This method has been evaluated in the context of high dimensional but low sample data sets, and is shown to be effective over well-known filter-based feature selection algorithms (e.g., Pearson Correlation Coefficient, Signal to Noise Correlation Coefficient, Chi-square and Information Gain).

Chen and Wasikowski (2008) also studied the small sample imbalanced data problem. Their approach is encapsulated in an algorithm called FAST (Feature Assessment by Sliding Thresholds), which is based on the area under the receiver operating characteristic (ROC) curves generated from setting different decision boundaries for a single feature. The algorithm is inspired by the observation that most single feature classifiers set the decision boundary at the mid-point between the mean of the two classes. By moving this decision boundary, different numbers of true/false positives are obtained. In doing so, the algorithm will be able to measure which decision boundary will provide the best area under the ROC curve and then select the one that would yield the best predictive results. By computing this over all features, the algorithm will be able to select the best mix of features.

Maldonado et al. (2014) also considered the problem of high-dimensional and imbalanced data learning but in the context of binary classification. In this case, a family of algorithms inspired by the backward feature selection strategy in Support Vector Machines (SVM) was proposed. Different strategies of backward elimination of features were developed and used with SVM and SMOTE fitted with different loss functions: (a) standard 0–1; (b) balanced loss; and (c) predefined loss. The various algorithms were tested over six imbalanced micro-array data with their algorithms showing better predictive performance over well-known feature selection algorithms (e.g. l_0, l_1 norm SVM, SVM Recursive Feature Elimination (SVM-RFE), Fisher + SVM, etc.) – while also using fewer features.

Not all feature selection algorithms for high-dimensional data work on the basis of a single ranking function or an inductive algorithm. Yang et al. (2013), for example, proposed to create multiple balanced data sets using random under/over-sampling from the original imbalanced data. Feature subsets are then evaluated over an ensemble of