Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Cluster-based Weighted Oversampling for Ordinal Regression (CWOS-Ord)

Iman Nekooeimehr, Susana K. Lai-Yuen *

Industrial and Management Systems Engineering, University of South Florida, 4202 East Fowler Avenue, ENB 118, Tampa, FL 33620, USA

ARTICLE INFO

Article history:
Received 1 February 2016
Received in revised form
26 June 2016
Accepted 11 August 2016
Communicated by: Sato-Ilic Mika

Keywords:
Imbalanced dataset
Ordinal regression
Clustering
Oversampling

ABSTRACT

A new oversampling method called Cluster-based Weighted Oversampling for Ordinal Regression (CWOS-Ord) is proposed for addressing ordinal regression with imbalanced datasets. Ordinal regression is a supervised approach for learning the ordinal relationship between classes. In many applications, the dataset is highly imbalanced where the instances of some classes (majority classes) occur much more frequently than instances of other classes (minority classes). This significantly degrades the classification performance as classifiers tend to strongly favor the majority classes. Standard oversampling methods can be used to improve the dataset class distribution; however, they do not consider the ordinal relationship between the classes. The proposed CWOS-Ord method aims to address this problem by first clustering minority classes and then oversampling them based on their distances and ordering relationship to other classes' instances. The final size to oversample the clusters depends on their complexity and their initial size so that more synthetic instances are generated for more complex and smaller clusters while fewer instances are generated for less complex and larger clusters. As a secondary contribution, existing oversampling methods for two-class classification have been extended for ordinal regression. Results demonstrate that the proposed CWOS-Ord method provides significantly better results compared to other methods based on the performance measures.

1. Introduction

Ordinal regression is a supervised approach for learning ordering or ranking patterns, and has the properties of both multi-class classification and metric regression [1]. It has properties of multi-class classification because the outcome is a finite set but it considers the ordinal relationship between classes. Ordinal regression also has properties of metric regression as it assumes the outcome variable is a latent continuous variable where the number of ranks is finite and the difference between ranks is not defined. Ordinal regression has applications in many areas such as information retrieval [2], credit rating [3,4], visual recognition [5,6], and preference learning [7] because very often, people represent their preferences via ranks and ordered classes. As an example, consider a clinical diagnosis where patients can be categorized to stages ranging from 0 to 4. Higher stages indicate higher severity of the condition so the misclassification error between different stages should be penalized differently. For instance, the misclassification error between stages 0 and 4 should be much higher than the error between stages 0 and 1. On the other hand, the stages are not continuous and the difference between adjacent stages is not equal making this problem different from standard regression.

The class imbalance problem is a big challenge for classification algorithms. In real applications, there are many datasets that are highly imbalanced where the majority of the data represents one or few classes while less data represents minority classes. The imbalanced dataset problem can weaken the performance of classifiers significantly as classifiers normally favor larger classes while ignoring smaller ones [8]. Various approaches have been presented to address the imbalance problem such as data preprocessing, algorithmic modification, cost-sensitive learning, and ensemble of classifier sampling methods [9,10]. Data preprocessing techniques modify the data distribution to address the problem of the skewed class distribution in the learning phase [11–13]. Algorithmic modification approaches modify the existing algorithms to give significance to minority instances [14–17]. Cost-sensitive methods combine both algorithm and data modification approaches to give different misclassification costs for each class in the learning process [18,19]. Finally, ensemble of classifier sampling methods modify the ensemble learning algorithm to address the imbalance problem without normally changing the base classifier [20–22].

* Corresponding author.
E-mail addresses: nekooeimehr@mail.usf.edu (I. Nekooeimehr), laiyuen@usf.edu (S.K. Lai-Yuen).

http://dx.doi.org/10.1016/j.neucom.2016.08.071
0925-2312/

Sampling methods have shown great potential as they attempt to improve the dataset itself rather than the classifier [13,23]. They change the distribution of each class observation by either oversampling the minority samples (generating new minority instances) [11,13] or under-sampling the majority samples (removing some majority instances) [24–26]. Under-sampling methods may eliminate important information from the dataset by deleting some majority instances from the dataset, especially in small datasets [27,28]. Most of the research conducted to address the imbalanced dataset problem focus mainly on two-class and multi-class imbalanced problems [29–31]. However, very limited works have addressed the imbalanced dataset problem for regression or ordinal regression [32,33].

In this paper, we propose a new oversampling method called Cluster-based Weighted Oversampling for Ordinal Regression (CWOS-Ord) to address the imbalanced dataset problem in ordinal regression. CWOS-Ord identifies clusters of data by first clustering all classes except the largest class using hierarchical clustering to ensure that all clusters are considered for oversampling. The set of all classes except the largest class will be referred to as the *smaller* classes. The largest class is not considered for oversampling. A modification of the traditional hierarchical clustering is presented that clusters the instances of smaller classes by considering other class instances to reduce overlapping between the generated instances and instances of other classes. Then, the final size to oversample the clusters depends on their complexity and initial size so that more synthetic instances are generated for more complex and smaller clusters while fewer instances are generated for less complex and larger clusters. Consequently, the clusters will not necessary have the same size after oversampling but in general, all the classes will be of equal size. This is particularly practical for ordinal regression as it contains multiple classes and oversampling the clusters of each class to the size of the largest majority cluster can result in a very large dataset. CWOS-Ord avoids over-generalization and mislabeling errors in terms of the ordinal label scale by oversampling instances of smaller classes based on their average Euclidean distance and rank differences to other class instances. Finally, well-known oversampling methods designed for two-class classification have been extended to the ordinal regression problem for performance comparison.

The contribution of this paper is three-fold. First, a modified agglomerative hierarchical clustering is introduced to reduce the generation of overlapping synthetic instances during oversampling. This is achieved by iteratively merging clusters of the same class while considering clusters of instances of other classes. Second, a new measure is proposed that quantifies the trade-off between cluster complexity and the initial size of the cluster. The new measure is used to determine the number of oversampled instances for each cluster. Finally, a new probability distribution is proposed that incorporates the distance as well as rank distance to other-class instances so that instances closer to the non-adjacent classes are oversampled more. As an additional contribution, existing oversampling methods for binary classification have been extended to ordinal regression.

The remainder of this paper is organized as follows. In the next section, previous work related to the imbalanced dataset classification problem is explained. The notation used in the paper is defined in Section 3. In Section 4, the proposed CWOS-Ord methodology is described. A description of our extension of well-known oversampling methods to ordinal regression is presented in Section 5 for subsequent method comparison. Section 6 presents the results and discussion while Section 7 provides the conclusions.

2. Previous work

Given that most of the research on oversampling methods for imbalanced dataset classification has focused mainly on two-class imbalance problems, this section first provides an overview of oversampling methods for two-class classification followed by techniques for ordinal regression. The simplest oversampling method is random oversampling where the instances generated are identical to the original instances. This method may cause severe overfitting as a result of no diversity among instances. To address this problem, SMOTE [11] was proposed where new instances are generated between randomly selected minority instances and their k-nearest neighbors. However, this may cause over-generalization as the new instances are generated without considering the majority instances thus increasing the overlap between classes. Methods like ADASYN [34] and MWMOTE [13] were proposed to overcome the over-generalization problem by considering the instances of the majority class while oversampling the minority class. ADASYN [34] assigns weights to minority instances so that those that have more majority instances in their neighborhood have higher chance to be oversampled. MWMOTE [13] presents a two-step procedure to find candidate majority border instances and then candidate minority border instances. Then, weights are assigned to candidate minority instances with respect to the candidate majority border instances so that those with higher weights have a higher chance to be oversampled. However, small clusters of minority instances that are far from the majority class may not be detected even if they may contain important information. In general, it is necessary to find hard-to-learn instances to be used for oversampling because they may contain important information for the classifier. These instances are usually near the decision boundary or belong to small clusters [27,35]. In [36], a method was presented that identifies small clusters for all classes in the dataset and then oversamples each cluster so that its size is equal to the size of the largest cluster of the majority class. Given that the majority class is also oversampled and the largest cluster of the majority class may be very large, the resultant dataset after oversampling can become very large thus increasing the computational time for training the model. Also, instances close to the borderline, which may contain important information for the model, are not identified. Finally, all clusters are oversampled to the same size even if different clusters may need to be oversampled differently.

In our previous work [12], a new oversampling method called Adaptive Semi-Unsupervised Weighted Oversampling (A-SUWO) was presented for imbalanced binary dataset classification. A-SUWO addresses the aforementioned challenges through the identification of hard-to-learn instances by clustering the minority instances and identifying those instances that are closer to the borderline. Moreover, it avoids generating minority instances that overlap with the majority class during clustering, and adaptively determines the size to oversample each minority cluster using its classification complexity.

To the best of our knowledge, very few research works have addressed the imbalanced dataset problem in ordinal regression. In [33], an Ordinal Graph-based Oversampling (OGO) framework is proposed to generate synthetic instances by considering the ordering relationship between the classes. The framework consists of three versions: OGO-NI, OGO-SP, and OGO-ISP. OGO-NI first finds the instances on the border of the adjacent classes and then, it creates synthetic instances for the minority class between the minority class instances and the instances in the border of the adjacent classes. In OGO-ISP and OGO-SP, minority instances that are along the shortest path of their adjacent classes are identified and those that are not on the shortest path are removed from the dataset to avoid oversampling outliers. The difference is that in

Please cite this article as: I. Nekooeimehr, S.K. Lai-Yuen, Cluster-based Weighted Oversampling for Ordinal Regression (CWOS-Ord), Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.08.071