Engineering Applications of Artificial Intelligence 64 (2017) 440–461

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Distance based resampling of imbalanced classes: With an application example of speech quality assessment

Drasko Furundzic [a,b,*], Srdjan Stankovic [a], Slobodan Jovicic [a,c], Silvana Punisic [c], Misko Subotic [c]

[a] School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73, Belgrade 11000, Serbia
[b] Mihajlo Pupin Institute, Volgina 15, 11000 Belgrade, Serbia
[c] Life Activities Advancement Center, Gospodar Jovanova 35, Belgrade 11000, Serbia

ARTICLE INFO

Keywords:
Distance based balancing
Imbalanced learning
Binomial distribution
Nearest neighbors
Neural network
Speech signal classification

ABSTRACT

This paper presents a new general methodological approach to imbalanced learning as one of the challenging problems in pattern classification. The presented method is founded on maximization of the sample entropy. The method involves detection of distributive properties of ideally balanced regular lattice sample and acceptable transfer of these properties to an arbitrary imbalanced sample increasing its representativeness. The proposed procedure assumes undersampling applied on areas of high probability density in the sample space combined with oversampling in the areas of low density. The main achievement of this method is the increased sample class entropy which reduces the inductive learner's tendency to favor prominent class, or cluster. In addition to class balancing, this method can be useful for function approximation, clustering, and sample dimension reduction. The high degree of generality of the method implies its applicability on data of various complexity and imbalance. The presented theoretical foundation of the method was verified on a set of proper synthetic samples. The method's practical usability is confirmed by a comparative classification of a large set of databases including speech signal samples.

1. Introduction

Here we present a new general methodological approach to imbalanced learning, based on detection and transfer of convenient distributive properties of the regular lattice sample to the arbitrary imbalanced sample. These properties of the lattice are presented in the condensed form of the ratio established between the volume value occupied by the sample instance and its local mean distance from the neighbors. According to its distributive properties, we consider the regular lattice a paradigm of uniformity, representativeness and internal balance. Given the fact that a uniformly distributed sample has good representativeness in relation to the target population, transfer of these characteristics to an arbitrary distributed training sample should show a positive effect on its representativeness. Standard inductive learning classifiers, known as data driven models, are generally designed to operate with data of "well balanced" class distributions and/or equal misclassification costs. In reality inductive learners operate under conditions of an uneven distribution of data, both within the classes and between the classes, and also in terms of different misclassification costs. Within-class imbalance influences performance of the classifier while an imbalance between the classes (CI) represents a small problem when there are acceptable within-class balances (representativeness) (Japkowicz, 2003). In the context of imbalanced learning, there are crucial facts and *concepts* essential for this problem, so we draw attention to some of them. For a detailed analysis, understanding, and recent actual approaches to the problem we recommend to readers the works presented in: Japkowicz (2003), Weiss (2004), Chawla et al. (2004), He et al. (2008), He and Garcia (2009), and Japkowicz and Stephen (2002). The concept of *class* assumes a, more or less, homogeneous group of instances characterized by a vector of discriminatory feature values which distinguish them from other, more or less, similar groups. Each instance (example) represents a single point in the feature space. In practice, a class usually encompasses just a part of the target population (sample). The concepts, *majority* (negative) class and *minority* (positive) class are generally accepted in the field earlier, as the terms which indicate that one class seriously outnumbers the other one (between-class imbalance) (He and Garcia, 2009) that is expressed by Imbalance Ratio (IR). In the modern approach, the

* Corresponding author at: Mihajlo Pupin Institute, Volgina 15, 11000 Belgrade, Serbia
E-mail addresses: drasko.furundzic@pupin.rs (D. Furundzic), stankovic@etf.rs (S. Stankovic), jovicic@etf.rs (S. Jovicic), silvanapunisic@hotmail.com (S. Punisic), ifp2@ikomline.net (M. Subotic).

http://dx.doi.org/10.1016/j.engappai.2017.07.001
Received 18 May 2016; Received in revised form 27 June 2017; Accepted 3 July 2017
0952-1976/

concepts of majority and minority classes are replaced by the terms *underrepresented* and *overrepresented* classes, as more descriptive terms (Japkowicz, 2001). Term *sampling frame* represents the source from which a sample is drawn, representing a list of the available examples of the population. Various practical constraints limit the quantity and quality of the sampling frame causing this frame to, more or less evenly, cover different areas of the target population space. These facts have a determining influence on the representativeness of the sampling frame and, consequently, the training sample. There are two basic approaches to imbalanced learning problem: (a) algorithmic approaches (Pazzani et al., 1994; Japkowicz et al., 1995; Kubat et al., 1998; Estabrooks et al., 2004) that emphasize the importance of the inductive learning algorithms such as: cost-sensitive methods, support vector machines (SVM) (Wu and Chang, 2003), neural networks (NN) (Lawrence et al., 1998), k nearest neighbors (kNN) (Wilson, 1972), genetic algorithms (GA) (Garcia et al., 2009), decision trees and active learning, and (b) data level that changes the original distribution of data in order to reduce the negative effect of their CI (Lewis and Gale, 1994; Kubat and Matwin, 1997; Ling and Li, 1998; Drummond and Holte, 2003; Maloof, 2003; Weiss and Provost, 2003; Chawla et al., 2004). There are several widespread heuristic and non-heuristic sampling techniques of which the *concept complexity* based methods are the group dealing with the essence of imbalanced learning. This term implies the appearance of different structures within the sample space as a consequence of uneven distribution of instances, known as sub concepts, sub classes, clusters, etc. (Japkowicz, 2001, 2003; Nickerson et al., 2001). The pioneering work dealing with this method (Holte et al., 1989) analyzed the relationship between the sub clusters, small disjuncts, and classification error. The concept of *error concentration* (Weiss, 2003) establishes an empirical correspondence between small disjunct size and classification error. Probability distribution of the available source of instances determines the degree of chosen sample representativeness. A greater degree of uniformity distribution, as well as more occupied space, corresponds to greater representativeness of the sample. This important statement is the subject of proving and discussion in further presentation. Great variability of the density of instances corresponds to a high degree of complexity. During selection of the sample, the primary objective is to obtain a high level of its representativeness. This statement means that the selected sample should as evenly as possible and as much as possible cover the available sample space. Upper request is one of the main tasks of imbalanced learning, and for that purpose we have designed a new resampling algorithm for within-class balancing, named Distance Based Balancing (DBB) algorithm. Theoretical basis of the attitude about the importance of the uniformity distribution of the training sample lies in maximization of its entropy (Jaynes, 1957). Uniform distribution of instances within classes strongly reduces the effect of imbalance between the classes, so the primary objective of resampling strategies is elimination of internal imbalance of the classes. This problem requires the use of a combined resampling strategy which, on one hand, reduces the number of instances in the areas of high density and, on the other, increases the number of instances in the areas of low density. Our resampling strategy is based on these just stated proposals. The paper is organized in the following way. Section 2 is devoted to clarification of the concept "well balanced classes" which is commonly found in the literature, but not clearly defined in the sense of the relationship among the within-class imbalance and between-classes imbalance. Here we emphasize the primary importance of internal class imbalance as the generator of general imbalance learning problem. We also present a correspondence among the internal imbalance and representativeness of the involved sample, as the basis for the development of a new balancing algorithm. In this section, the necessary attention is paid to formalization of the concept of representativeness in the light of information theory. Section 3 presents the main part of the paper where we introduce a new balancing algorithm, resulting in quasi-uniform probability density of the processed training sample. Necessary attention is devoted to the mathematical foundations of this model. Section 4 describes a practical evaluation of the DBB algorithm through a comparison with standard balancing methods. This chapter analyzes the theoretical assumptions about the correspondence between the distribution characteristics of databases, as indicators of representativeness, and the underlying performance of the classifier subjected to these databases. Here is presented a critical review of the advantages and disadvantages of the proposed method and its potential impact on further research. The conclusions and suggestions are given in Section 5.

2. The concept of a well-balanced sample

Many authors, talking about the problems of classification, mention the assumption of "ideally balanced classes" but they do not explain in detail this concept, so, further clarification of this issue is one of the important objectives of this work. In many papers, imbalanced data are defined as data with impaired balance in the number of representatives of the actual classes. On the other hand, Japkowicz argues that in the case of very prominent within-class imbalance, insisting on between-class imbalance does not make sense (Japkowicz, 2003). That is one of the reasons why we focus our attention on the within-class imbalance as the basic generator of imbalance in the context of representativeness.

2.1. Representativeness of the sample

A representative sample can be intuitively defined as an unbiased indicator of the target population state. According to Pan et al. (2005) and Ma et al. (2011), a representative sample is a carefully designed subset of the population, including three properties: it is significantly reduced in size compared to the original source, it better covers the main information on the source than other subsets of the same size, and it has low redundancy (R). Due to a fast growth of the quantity of information, the need for an effective detection and reduction of redundancy of databases is of great interest. Representative sample as an informational category has a dual nature: high coverage rate of the sample space and low sample redundancy. The related authors combine entropy measurement based techniques, such as mutual information (Pan et al., 2005; Ma et al., 2011) and Kullback–Leibler (KL) divergence (Pan et al., 2005; Ma et al., 2011; Paek and Hsu, 2011) and dissimilarity indexes (Bertino, 2006), to quantify the representativeness of samples (Liu, 2007). The approaches to determination of a representative sample are based on evaluation of its information coverage rate of the population (Pan et al., 2005; Zhai et al., 2003) and redundancy (Pan et al., 2005; Zhang et al., 2002; Carbonell and Goldstein, 1998; Kubat and Matwin, 1997). A sample that covers the most information should have a large mutual information value with respect to the target population and low redundancy. In the present work, we use the coverage rate and redundancy, as a measure of the representativeness.

2.1.1. Measure of the coverage rate

To explain the concept of coverage rate, let us present the basic notion of the KL divergence, also known as relative entropy. For two probability distributions $P = P(X) = \{p(x_i) = p_i, \forall x_i \in X\}$ and $Q = P(Y) = \{p(y_i) = q_i, \forall y_i \in Y\}$ over variables X and Y, the KL divergence of P from Q is defined as: $D_{KL}(P \parallel Q) = -\sum_i p_i \log[p_i/q_i]$. This divergence has the following properties: $D_{KL}(P \parallel Q) \geq 0; D_{KL}(P \parallel Q) = 0 \; iff \; P(x) = Q(x), \forall x \in X$. Consider discrete random variables X and Y with the joint probability function $p(x, y)$ and marginal probabilities $p(x)$ and $p(y)$. The mutual information $I(X;Y)$ is defined as relative entropy between the $p(x, y)$ and product $p(x)\,p(y)$: $I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x, y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$. The mutually independent variables do not share any common information, so: $p(x, y) = p(x)\,p(y) \Rightarrow \log(1) = 0 \Rightarrow I(X;Y) = 0$. Further we obtain the following relations : $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) = H(X,Y) - H(X|Y) - H(Y|X)$, where $H(X,Y)$ is the joint entropy, $H(X)$ and $H(Y)$ are the marginal entropies, $H(X|Y)$ and $H(Y|X)$ are the conditional entropies. These relations show analogy with the