Expert Systems With Applications 60 (2016) 249–257

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Distance-based online classifiers

Joanna Jędrzejowicz [a,*], Piotr Jędrzejowicz [b]

[a] Institute of Informatics, Gdansk University, Wita Stwosza 57, 80-952 Gdansk, Poland
[b] Department of Information Systems, Gdynia Maritime University, Morska 83, 81-225 Gdynia, Poland

ARTICLE INFO

Article history:
Received 27 July 2015
Revised 8 May 2016
Accepted 9 May 2016
Available online 9 May 2016

Keywords:
Online learning
Fuzzy C-means clustering
Kernelized clustering
Rotation forest

ABSTRACT

Main impact of the paper is proposing a family of algorithms for the online learning and classification. These algorithms work in rounds, where at each round a new instance is given and the algorithm makes a prediction. After the true class of the instance is revealed, the learning algorithm updates its internal hypothesis. The proposed algorithms are based on fuzzy C-means clustering and kernel-based fuzzy C-means clustering, followed by a calculation of distances between cluster centroids and the incoming instance for which the class label is to be predicted. In one of the proposed variants, simple distance-based classifiers thus obtained serve as basic classifiers for the implemented Rotation Forest ensemble classifier, which increases the accuracy of classification. In the paper we also propose using kernelized fuzzy C-means clustering method as an alternative approach to constructing distance based online classifiers. The approach allows to construct online classifiers of the polynomial computational complexity which is a significant feature considering potential application to the big data analysis. Using the kernelized clustering is advantageous since it allows for automatic estimation of the number of clusters maintaining the number of the user-defined parameters. The proposed classification algorithms are validated experimentally. Experiment results show that the approach assures good quality of classification, extending the range of the available online approaches.

1. Introduction

One of the basic data mining tasks is the classification of unknown objects progressively identified using a membership phenomenon of objects or class phenomena. In machine learning and statistics, classification is usually understood as the problem of identifying which set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing instances (observations) whose category membership is known. The idea in machine learning is to produce a so called classifier, which can be viewed as the function induced by a classification algorithm that maps input data into one or more categories. In machine learning there are two basic approaches to inducing classifiers: static and dynamic one. Static approach is based on two assumptions. First one assumes that a training set of the adequate size required to induce classifier and consisting of instances with known class labels, is available in advance, that is before a classifier is constructed. Second assumption requires that data instances arriving in the future have a stationary distribution identical with the distribution of data in the training set.

In the case of the dynamic approach, assumptions are weaker. Usually, some limited number of training examples with known class labels is required at the outset, however future incoming instances, after their true class label is revealed, can be added to the available training set extending it, or perhaps replacing some of its earlier arriving members. The second assumption does not have to hold, hence data distribution does not have to be a stationary one. The dynamic approach is a natural way of dealing with important online learning problems. Online learning is considered to be of increasing importance to deal with never ending and usually massive streams of received data, such as: sensor data, traffic information, economic indexes and video streams (Wang, Ji, & Jin, 2013). As a rule, the online approach is required when the amount of data collected over time is increasing rapidly. This is especially true for data stream model, where the data arrive at high speed. The algorithms used for mining the streams must process data using very strict constraints of space and time (Pramod & Vyas, 2012). A data stream can roughly be thought of as an ordered sequence of data items, where the input arrives more or less continuously as time progresses, see for example (Gama & Gaber, 2007). Reviews of algorithms and approaches using data streams mining can be found in Gaber, Zaslavsky, and Krishnaswamy (2005), Gaber, Zaslavsky, and Krishnaswamy (2010), and Pramod and Vyas (2012). Online classifiers are induced from the initially available dataset as

* Corresponding author.
E-mail addresses: jj@inf.ug.edu.pl (J. Jędrzejowicz), pj@am.gdynia.pl (P. Jędrzejowicz).

http://dx.doi.org/10.1016/j.eswa.2016.05.015
0957-4174/

is the case for the static approach. However, in addition, there is also some adaptation mechanism providing for a classifier evolution after the classification task has been initiated and started. In each round a class label of the incoming instance is predicted and afterwards information as to whether the prediction was correct or not, becomes available. Based on this information the adaptation mechanism may decide to leave a classifier unchanged, modify it, or induce a new one.

The usual approach to deal with the online classification problems is to design and implement an online classifier incorporating some incremental learning algorithm (Last, 2002; Sung & Kim, 2009). According to Murata, Kawanabe, Ziehe, Muller, and Amari (2002) an algorithm is incremental if it results in a sequence of classifiers with different hypothesis for a sequence of training requirements. Among requirements an incremental learning algorithm should meet are ability to detect concept drifts, ability to recover its accuracy, ability to adjust itself to the current concept and use past experience when needed (Widmer & Kubat, 1996). Examples of some state of the art classifiers belonging to the discussed class include approaches proposed in Wang et al. (2013) and Bertini, Zhao, and Lopes (2013).

One of possible approaches to deal with the online classification problems is to use the idea of the distance-based classification. To do so one needs the dissimilarity measure or distance which assures that if two objects are similar their representations in the metric or pseudo-metric representation space are close to each other. Classes are represented by prototypes or sets of prototypes with known class labels. To classify an object its distances to the available prototypes are calculated and the prototype to which the distance is smallest determines the class. Examples of the distance based classifiers include the k-nearest neighbor (k-NN) classifier and the nearest class mean classifier (NCM). Both have been successfully applied to large scale classification problems - see, for example, Webb (2002), Boiman, Shechtman, and Irani (2008), Weinberger and Saul (2009), and Mensink, Verbeek, Perronnin, and Csurka (2013).

Distance-based classifiers, often referred to as similarity-based ones, belong to a broad class of the instance-based approaches, see for example Shaker and Hullermeier (2013), Gora and Wojna (2002), Yang, Rundensteiner, and Ward (2013), and Skowron and Wojna (2004). One of the recent approaches to similarity-based data stream classification is Similarity-based Data Stream Classifier (Sim C) proposed by Mena-Torres and Aguilar-Ruiz (2014). There have been also reported interesting dedicated algorithms and applications. Fan, Ye, and Chen (2016) proposed All-Nearest-Neighbor (ANN) classifier for sequence mining for automatic malware detection. An approach to novelty detection through constructing an ensemble of classifiers providing a kind of metric to characterize different classes proximity was suggested by Zhou, Zhou, and Ning (2015). The approach for prediction of recurring concepts on data-streams by means of a meta-model and a fuzzy similarity function was proposed in Abad, Gomes, and Menasalvas (2016). An adaptive ensemble classifier for mining concept drifting data streams was proposed in Farid, Zhang, and Hossain (2013). The model detects novel classes following the idea that data points belonging to the same class should be closer to each other and should be apart from the data points belonging to other classes.

Basic classification of clustering methods assigns them into two groups: crisp and fuzzy. One of the most used fuzzy methods is fuzzy C-means clustering (FCM) (Bezdek, 1981). FCM considers each cluster as a fuzzy set with membership function measuring the possibility for each data to belong to a given cluster. The method proved useful overcoming some drawbacks of crisp C-means clustering. One of the proposed distance-based online classifiers presented in the paper uses fuzzy C-means clustering method. The smallest distance between the reference instance represented by cluster centroids and incoming instance determines prediction of the class label. Simple distance-based classifiers thus obtained serve as basic classifiers for the implemented Rotation Forest ensemble classifier, introduced in (Rodrigeuz, Kuncheva, & Alonso, 2006). Recently, to deal with over-lapping and noisy data kernel methods have been applied to fuzzy clustering (Chiang & Hao, 2003; Li, Tang, Xue, & Jiang, 2001; Zhang & Chen, 2002) and the kernelized version of FCM is referred to as kernel-based fuzzy C-means clustering. Graves and Pedrycz (2010) give a broad report on kernel methods applied to clustering. In the paper we propose using kernelized fuzzy C-means clustering method as an alternative approach to constructing distance based online classifiers.

The paper includes and extends some results presented in earlier papers of the authors. In Jędrzejowicz and Jędrzejowicz (2013), we proposed the online classifier based on fuzzy C-means clustering with two strategies for maintaining the training dataset. Under the first strategy the training dataset is extended in each round by adding a new example after its class label has been revealed. Under the second strategy the training dataset size was kept constant. In Jędrzejowicz and Jędrzejowicz (2014), we have proposed diversified strategies for the training set management and also have come up with the idea of using Rotation Forest to improve accuracy of the classification. Replacing the fuzzy C-means clustering by kernel-based fuzzy C-means clustering technique resulted in proposing the online classifier where the number of clusters can be automatically fixed (Jędrzejowicz & Jędrzejowicz, 2015). In the current paper, we have integrated and unified approaches and carried out the extensive computational experiment which results have been analyzed using the Friedman non-parametric ranking test with Bonferroni–Dunn test.

The paper is organized as follows. Section 1 contains introduction. Sections 2 and 3 contain algorithms descriptions and Section 4 analysis of their complexity. Section 5 presents validation experiment settings and experiment results. Section 6 contains conclusions and suggestions for future research.

2. Online classification using fuzzy C-means clustering and Rotation Forest ensemble

The general data classification problem is formulated as follows. Let C be the set of categorical classes which are denoted $1, \ldots, |C|$. The learning algorithm is provided with the learning instances $LD = \{< d, c > \mid d \in D, c \in C\} \subset D \times C$, where D is the space of attribute vectors $d = (w_1^d, \ldots, w_N^d)$ with w_i^d being a numeric value and N is the number of attributes. The algorithm is used to find the best possible approximation $\hat{f}$ of the unknown function f such that $f(d) = c$. Then $\hat{f}$ can be used to find the class $c = \hat{f}(d)$ for any d such that $(d, c) \notin LD$, that is the algorithm will allow to classify instances not seen in the process of learning.

2.1. Fuzzy C-means clustering

The considered algorithms make use of fuzzy C-means clustering, see Dunn (1973), that is an iterative method which allows one feature vector to belong to two or more clusters. The method is based on minimization of the objective function

$$J_m = \sum_{i=1}^{M} \sum_{j=1}^{noCl} u_{ij}^m \cdot dist(x_i, c_j), \tag{1}$$

where m is a fixed number greater than 1 (in the experiments the value was fixed and equal 2), M is the number of feature vectors, $noCl$ is the number of clusters, c_j is the center of the j-th cluster, u_{ij} is the degree of membership of the i-th feature vector x_i in cluster j and $dist$ is a fixed metric to calculate the distance from the