Neural Networks 80 (2016) 53–66

Contents lists available at ScienceDirect

Neural Networks

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

Binary classification SVM-based algorithms with interval-valued training data using triangular and Epanechnikov kernels

Lev V. Utkin[a,*], Anatoly I. Chekh[b], Yulia A. Zhuk[c]

[a] *Peter the Great Saint-Petersburg Polytechnic University, Russia*
[b] *Saint Petersburg State Electrotechnical University, Russia*
[c] *ITMO University, Russia*

ARTICLE INFO

Article history:
Received 2 November 2015
Received in revised form 6 April 2016
Accepted 11 April 2016
Available online 27 April 2016

Keywords:
Classification
Support vector machine
Interval-valued data
Minimax strategy
Linear programming
Quadratic programming

ABSTRACT

Classification algorithms based on different forms of support vector machines (SVMs) for dealing with interval-valued training data are proposed in the paper. L_2-norm and L_∞-norm SVMs are used for constructing the algorithms. The main idea allowing us to represent the complex optimization problems as a set of simple linear or quadratic programming problems is to approximate the Gaussian kernel by the well-known triangular and Epanechnikov kernels. The minimax strategy is used to choose an optimal probability distribution from the set and to construct optimal separating functions. Numerical experiments illustrate the algorithms.

1. Introduction

The binary classification problem can be formally written as follows. Given n training data (examples, patterns) $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, in which $\mathbf{x}_i \in \mathbb{R}^m$ represents a feature vector involving m features and $y_i \in \{-1, 1\}$ indices the class of the associated examples, the task of classification is to construct an accurate classifier $c : \mathbb{R}^m \to \{-1, 1\}$ that maximizes the probability that $c(\mathbf{x}) = y_i$ for $i = 1, \dots, n$. Generally $\mathbf{x}_i$ may belong to an arbitrary set $\mathcal{X}$, but we consider the special case $\mathcal{X} = \mathbb{R}^m$ for simplicity. One of the ways for classification is to find a real valued separating function $f(\mathbf{x}, \mathbf{w}, b)$ having parameters $\mathbf{w}$ and b such that $\mathbf{w} = (w_1, \dots, w_m) \in \mathbb{R}^m$ and $b \in \mathbb{R}$, for example, $f(\mathbf{x}, \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b$. Here $\langle \mathbf{w}, \mathbf{x} \rangle$ denotes the dot product of two vectors $\mathbf{w}$ and $\mathbf{x}$. The sign of the function determines the class label prediction or $c(\mathbf{x})$. We also introduce the notation $x_i^{(k)}$ for the kth element of the vector $\mathbf{x}_i$.

Most available classification algorithms assume that training data are precise or point-valued. However, training examples in many real applications can be obtained only in the interval form. Interval-valued data stem from imperfection of measurement tools or imprecision of expert information, from missing data. Interval data arise in situations such as recording monthly interval temperatures in meteorological stations, daily interval stock prices, etc. Another source of interval data is the aggregation of huge data-bases into a reduced number of groups (Lima Neto & de Carvalho, 2008). For example, suppose that ages of patients in Hospital A are 56, 27, 36, 45; ages of patients in Hospital B are 51, 76, 82, 84. If we consider an aggregated training set about hospitals in a city, then ages of patients may be represented in interval form, for instance, Hospital A: [27,56]; Hospital B: [51,84]. It is assumed here that we have only the aggregated information about hospitals, but we do not have the initial ages in order to construct a probability distribution over values of intervals. Therefore, the main assumption related to the interval-valued training data considered in the present paper is that a true probability distribution over every interval is unknown, i.e., we have a case of the total ignorance concerning the probability distribution in the interval. The total ignorance means that there is absolutely no information about probability distributions over intervals. Simultaneously, the total ignorance concerning the probability distribution means that arbitrary probability distributions can be constructed over every interval.

The importance of the conditions stimulates to development the corresponding models and algorithms. As a result, many methods in machine learning have been presented for dealing with interval-valued data (Ishibuchi, Tanaka, & Fukuoka, 1990; Nivlet,

[*] Corresponding author.
E-mail addresses: lev.utkin@gmail.com (L.V. Utkin), anatoly.chekh@gmail.com (A.I. Chekh), zhuk_yua@mail.ru (Y.A. Zhuk).

http://dx.doi.org/10.1016/j.neunet.2016.04.005
0893-6080/© 2016 Elsevier Ltd. All rights reserved.

Fournier, & Royer, 2001; Silva & Brito, 2006). When we say about the interval-valued data, we formally assume that there are again two classes, i.e., $y_i \in \{-1, 1\}$, but feature vectors $\mathbf{x}_i$ are interval-valued, i.e., $\mathbf{x}_i \in \mathbf{A}_i$, $i = 1, \dots, n$. Here $\mathbf{A}_i = [\underline{a}_i^{(1)}, \overline{a}_i^{(1)}] \times \cdots \times [\underline{a}_i^{(m)}, \overline{a}_i^{(m)}]$, i.e., $\underline{a}_i^{(k)} \leq x_i^{(k)} \leq \overline{a}_i^{(k)}$, $k = 1, \dots, m$; $\underline{a}_i^{(k)}, \overline{a}_i^{(k)}$ are bounds for values of the kth feature in the ith training example.

All available algorithms and models aimed for learning from interval-valued data can be divided in five groups.

The first group consists of algorithms which are based on the standard interval analysis for constructing the classification and regression models (Angulo, Anguita, Gonzalez-Abril, & Ortega, 2008; Hao, 2009). In fact, these algorithms come to constructing and computing some functions of intervals to get a set of models in the form of interval-valued separating or regression functions. Interesting similar models for dealing with interval-valued and fuzzy observations in classification and regression are proposed in papers Carrizosa, Gordillo, and Plastria (2007a, 2007b) and Forghani and Yazdi (2014). However, these models as well as the standard interval analysis are restricted by considering mainly the linear case, i.e., a case when separating or regression functions are linear. Moreover, many interval-valued parameters may lead to too large intervals of resulting functions.

The second group consists of algorithms which are based on replacement of interval-valued observations by precise values using some additional assumptions, for example, by taking middle points of intervals (Lima Neto & de Carvalho, 2008). It should be noted that many algorithms dealing with missing data impute some precise values instead of missing ones (Garcia-Laencina, Sancho-Gomez, & Figueiras-Vidal, 2010), i.e. they also can be referred to as the second group. The replacement of interval-valued observations by precise values can be successfully used when intervals are not large and the area produced by the interval intersections is rather small. However, if intervals are very large and overlapping, then the replacement of intervals by point-valued data may lead to large classification errors. Another difficulty of using algorithms from the second group is interpretation of classification or regression results. It is often difficult to justify the obtained separating or regression functions in the decision-theoretic framework.

The third group of algorithms uses a very interesting and original idea to consider the distance measure between two interval-valued data points because many classification and regression methods, for example, the support vector machine (SVM) with non-linear kernel functions, k-nearest neighbors, deal only with distances between training data, but not with the training examples themselves. By taking into account this peculiarity of many methods, Do and Poulet (2005) proposed a very simple method based on the replacement of the Euclidean distance between two data points in the Gaussian kernel function used in the SVM by the Hausdorff distance between two hyper-rectangles produced by intervals from sample data. The method can be used in classification as well as in regression analyses. The main condition of its use is the assumption of the Gaussian kernel (or the kernels based on the distance between points) in the corresponding SVM. The Hausdorff distance also was used in clustering with imprecise data, for example, Chavent (2004); Chavent, de Carvalho, Lechevallier, and Verde (2006) proposed a partitional dynamic clustering method for interval data based on adaptive Hausdorff distances. A city-block distance function as the distance of a special form for solving clustering problems under interval-valued data was studied by de Souza and de Carvalho (2004). Pedrycz, Park, and Oh (2008) exploited a concept of the Hausdorff distance that determines a distance between some information granule and a numeric pattern (a point in the highly dimensional feature space) for constructing classifiers by interval and fuzzy data. Schollmeyer and Augustin (2013) illustrated that other distance measures have been successfully applied to machine learning problems instead of the Hausdorff distance, in particular, the authors (Schollmeyer & Augustin, 2013) proposed another distance measure for solving regression problems under interval data. Schollmeyer and Augustin (2013) argued that their measure might be better in some problems because the Hausdorff distance does not match points of two sets, but compares all points of the two sets to each other.

It should be noted that the algorithms from the third group are rather simple because they replace the interval-valued data by precise distances. However, they have important obstacles for their application. First of all, it is difficult to interpret the classification or regression results. Second, by dealing with interval-valued data, we usually implicitly or explicitly select a point in every interval in accordance with some decision strategy, which can be regarded as a "typical" point of the interval under the accepted decision strategy. The method using the Hausdorff distance allows having many different data points in intervals simultaneously, namely, pairwise distances between three intervals may correspond to different points in every interval. This implies that the same interval is represented by its different precise values. Though, this property may be useful sometimes. Another difficulty of using the Hausdorff distance is again the justification of the obtained algorithms in the decision-theoretic framework because we select some points of intervals in fact without taking into account a general aim of algorithms to minimize classification or regression errors.

The fourth group consists of robust algorithms using probabilistic constraints. These algorithm differ from the algorithms using the point-valued representation of intervals. A binary linear classification algorithm which can be referred to the fourth group was proposed by Ghaoui, Lanckriet, and Natsoulis (2003). The authors develop a robust classifier by minimizing the worst-case value of a given loss function over all possible choices of the data in the multi-dimensional intervals. We have to mark out very interesting algorithms dealing with interval-valued data whose key idea is to derive convex constraints in the SVM involving the partial information in the form of intervals (Ben-Tal, Bhadra, Bhattacharyya, & Nath, 2011; Bhadra, Nath, Ben-Tal, & Bhattacharyya, 2009). These algorithms use Bernstein approximation schemes for constructing classifiers which are robust to interval-valued uncertainty in examples. We have to point out here that the Bernstein approximation utilizes both the support (bounds of intervals) and moment information (mean and variance) of random variables. This is an additional information which may be unknown in many applications. The problem of constructing robust classifiers is posed as a chance-constrained program which ensures that the uncertain data points are classified correctly with high probability. Ben-Tal et al. (2011) applied the idea of using Bernstein approximation schemes to SVMs. It should be noted that a very clear and comprehensive survey of SVMs dealing with uncertain data is provided by Wang and Pardalos (2014).

The fifth group consists of robust algorithms which are based on using various forms of SVMs and consider robust strategies in the decision-theoretic framework. The main distinctive feature of the algorithms is to consider an interval of expected risk measures produced by interval-valued learning data. One of the algorithms was proposed by Utkin and Coolen (2011). However, this algorithm uses a weak assumption which restricts its usage. According to this assumption, the separating function f is monotone, for example, linear, because its lower and upper bounds in this case are determined only by the bounds of pattern intervals. However, in spite of the restricted application of the algorithm, it looks for "optimal" points to some extent of the expected classification risk, but not for points of intervals of training data. This is an important peculiarity of the algorithm. Similar approaches have