Neural Networks 69 (2015) 99–110

Contents lists available at ScienceDirect

Neural Networks

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

A new robust model of one-class classification by interval-valued training data using the triangular kernel

Lev V. Utkin [a,*], Anatoly I. Chekh [b]

[a] Department of Control, Automation and System Analysis, Saint Petersburg State Forest Technical University, Russia

[b] Department of Computer Science, Saint Petersburg State Electrotechnical University, Russia

ARTICLE INFO

Article history:
Received 17 October 2014
Received in revised form 7 February 2015
Accepted 27 May 2015
Available online 9 June 2015

Keywords:
One-class classification
Novelty detection
Support vector machine
Kernel
Interval-valued data
Minimax strategy
Linear programming
Extreme points

ABSTRACT

A robust one-class classification model as an extension of Campbell and Bennett's (C–B) novelty detection model on the case of interval-valued training data is proposed in the paper. It is shown that the dual optimization problem to a linear program in the C–B model has a nice property allowing to represent it as a set of simple linear programs. It is proposed also to replace the Gaussian kernel in the obtained linear support vector machines by the well-known triangular kernel which can be regarded as an approximation of the Gaussian kernel. This replacement allows us to get a finite set of simple linear optimization problems for dealing with interval-valued data. Numerical experiments with synthetic and real data illustrate performance of the proposed model.

1. Introduction

One of the problems of the statistical machine learning is to classify some objects into classes in accordance with their properties or features. At the same time, we need often to detect abnormal examples or to solve a one-class classification (OCC) or novelty detection problem. A lot of papers are devoted to this important problem (Campbell, 2002; Campbell & Bennett, 2001; Cherkassky & Mulier, 2007; Manevitz & Yousef, 2001; Scholkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001; Scholkopf, Williamson, Smola, Shawe-Taylor, & Platt, 2000; Zhang & Zhou, 2013). Various reviews of the OCC can be found in the machine learning literature, for example, reviews provided by Markou and Singh (2003), Bartkowiak (2011), Khan and Madden (2010), and Hodge and Austin (2004). The OCC aims to detect anomalous or abnormal observations and separate them from the so-called normal examples (Chandola, Banerjee, & Kumar, 2007, 2009; Steinwart, Hush, & Scovel, 2005).

A common way for solving the OCC problem is to model the support of the unknown data distribution directly from data, that is, to estimate a binary-valued function f that is positive in a region where the density is high, and negative elsewhere. Sample points outside this region can be regarded as anomalous observations.

Some models of the OCC are based on using the framework of the support vector machine (SVM). These models are called OCC SVMs. We mark out three main approaches for constructing the OCC SVMs. The first approach is proposed by Tax and Duin (1999, 2004). This is one of the well-known OCC models, which can be regarded as an unsupervised learning problem. According to this approach, the training of the one-class SVM consists in determining the smallest hyper-sphere containing training data. An alternative way to geometrically enclose a fraction of the training data is via a hyperplane and its relationship to the origin proposed by Scholkopf et al. (2001, 2000). Under this approach, a hyperplane is used to separate the training data from the origin with the maximal margin, i.e., the objective is to separate off the region containing the data points from the surface region containing no data. It should be noted that both the approaches provide the same results when a symmetric kernel is used. The third approach which will be considered in detail in this paper is the linear programming approach to the OCC proposed by Campbell and Bennett (2001). The model proposed by Campbell and Bennett uses linear programming techniques.

It should be noted that there are other interesting novelty detection or OCC models (see for instance, Bicego & Figueiredo, 2009; Hodge & Austin, 2004; Kwok, Tsang, & Zurada, 2007; Li, 2011). Every model can be applied in various applications.

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

http://dx.doi.org/10.1016/j.neunet.2015.05.004
0893-6080/

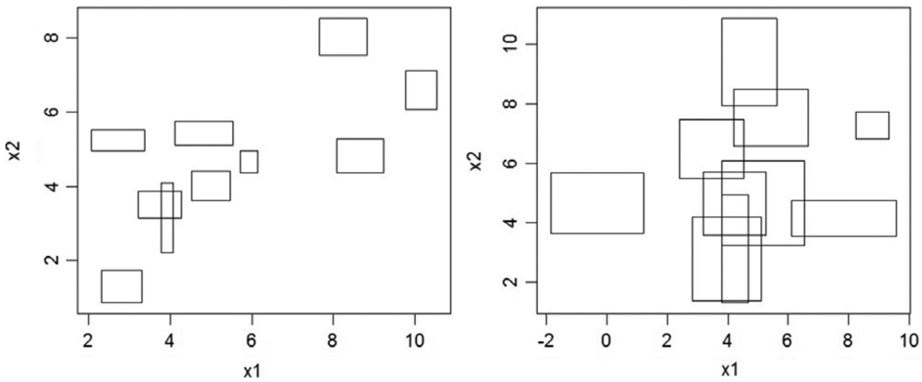

Fig. 1. Examples of interval-valued data with small and large intervals.

All these OCC models are based on using a training set consisting of precise or point-valued data. However, training examples in many real applications can be obtained only in the interval form. Interval-valued data may result from imperfection of measurement tools or imprecision of expert information. There may also be some missing data when some features of an example are not observed (Pelckmans, De Brabanter, Suykens, & De Moor, 2005).

Many methods in machine learning have been presented for dealing with interval-valued data (Ishibuchi, Tanaka, & Fukuoka, 1990; Nivlet, Fournier, & Royer, 2001; Silva & Brito, 2006) due to the importance of this condition. In some methods, interval-valued observations are replaced by precise values based on some additional assumptions, for example, by taking middle points of intervals (Lima Neto & de Carvalho, 2008). This approach can be successfully used when intervals are not large and the area produced by the interval intersections is rather small (see, for example, the left picture in Fig. 1). However, if the intervals are very large (see, for example, the right picture in Fig. 1), then the replacement of intervals by point-valued data may lead to large classification errors.

Another part of methods use the standard interval analysis for constructing the classification and regression models (Angulo, Anguita, Gonzalez-Abril, & Ortega, 2008; Hao, 2009). A series of interesting models for dealing with interval-valued and fuzzy observations in classification and regression can be found in works by 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 only the linear case, i.e., when separating or regression functions are linear.

Do and Poulet (2005) proposed an interesting and very simple method based on the change of the Euclidean distance between two data points in the Gaussian kernel function by the Hausdorff distance between two hyper-rectangles produced by intervals from sample data. The method can be used in classification and regression analyses, in OCC problems. 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. In spite of its simplicity, the method has an important obstacle for its application. It is not known how to interpret the classification results. Moreover, 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. Another disadvantage of the approach based on using the Hausdorff distance is lack of some justified strategy of decision making by dealing with imprecise data. In other words, it is not obvious in using the Hausdorff distance how to interpret the points which determine the distance between intervals from the classification point of view. 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. It should be noted that other distance measures have been successfully applied to machine learning problems. For example, Schollmeyer and Augustin (2013) proposed another distance measure for solving regression problems under interval data. The authors (Schollmeyer & 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.

Another interesting approach to constructing a classifier under interval-valued data was proposed by Bhadra, Nath, Ben-Tal, and Bhattacharyya (2009). The authors presented a novel methodology using Bernstein bounding schemes for constructing classifiers which are robust to interval-valued uncertainty in examples. According to the methodology, the uncertain examples are classified correctly with high probability. A binary linear classification model under interval data different from the models using the point-valued representation of intervals 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.

Following the idea of the robust model provided by Ghaoui et al. (2003), we propose a robust model which is based on three main ideas implemented in order to construct a new OCC model dealing with interval-valued training data.

1. Interval-valued observations produce a set of expected classification risk measures such that the lower and upper risk measures can be determined by minimizing and by maximizing the risk measure over values of intervals.
2. There are many variants of OCC SVMs. It is proposed to use linear programming OCC SVM by Campbell and Bennett (2001) for which constraints in its dual form do not depend on vectors of observations. This allows us to represent the dual optimization problem as a set of simple optimization problems.
3. It is proposed to replace the Gaussian kernel by the well-known triangular kernel which can be regarded as an approximation of the Gaussian kernel. This replacement allows us to get a set of linear optimization problems with variables $\mathbf{x}_i$ restricted by intervals $\mathbf{A}_i$, $i = 1, \ldots, n$.