Neurocomputing 174 (2016) 194–202

Contents lists available at ScienceDirect

Neurocomputing

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

An algorithm for classification over uncertain data based on extreme learning machine ☆

Keyan Cao [a,b], Guoren Wang [b,c], Donghong Han [b], Mei Bai [b], Shuoru Li [b]

[a] Shenyang Jianzhu University, Liaoning, Shenyang 110168, China
[b] Northeastern University, Liaoning, Shenyang 110819, China
[c] Key Laboratory of Medical Image Computing (Northeastern University), Ministry of Education, China

ARTICLE INFO

Article history:
Received 29 September 2014
Received in revised form
16 May 2015
Accepted 17 May 2015
Available online 10 August 2015

Keywords:
Extreme learning machine
Classification
Uncertain data
Single hidden layer feedforward neural networks

ABSTRACT

In recent years, along with the generation of uncertain data, more and more attention is paid to the mining of uncertain data. In this paper, we study the problem of classifying uncertain data using Extreme Learning Machine (ELM). We first propose the UU-ELM algorithm for classification of uncertain data which is uniformly distributed. Furthermore, the NU-ELM algorithm is proposed for classifying uncertain data which are non-uniformly distributed. By calculating bounds of the probability, the efficiency of the algorithm can be improved. Finally, the performances of our methods are verified through a large number of simulated experiments. The experimental results show that our methods are effective ways to solve the problem of uncertain data classification, reduce the execution time and improve the efficiency.

1. Introduction

In recent years, a large amount of uncertain data is generated and collected due to new techniques of data acquisition, which are widely used in many real-world applications, such as wireless sensor networks [1,2], moving object detection [8,34,36], meteorology and mobile telecommunication. However, since the intrinsic differences between uncertain and deterministic data, it is difficult to deal with uncertain data using traditional data mining algorithms for deterministic data. Therefore, many researchers put efforts in developing new techniques of data processing and mining on uncertain data [6,7,32,33].

Uncertain data model can be loosely classified into the following three categories [39]: (1) the most rigorous assumption of uncertainty is conceptually described by a continuous probability density function (pdf) in the data space D. Given any uncertain tuple x in D and its pdf $f(x)$, $\int_{x\in D} f(x)\,dx = 1$; (2) an uncertain tuple x_i consists of a set of possible values in the data space as its instances, denoted by $x_i = x_i^1, x_i^2, \ldots, x_i^j, \ldots, x_i^n$. The number of instances for an uncertain data x_i is denote by $|x_i| = n$. It can be regarded as the discrete case of probability density function. Let $p(x_i^j)$ denote the exist probability of instance x_i^j, then $p(x_i^j) > 0$, and $\sum_{j=1}^{j=n} p(x_i^j) = 1$; (3) it assumes that the standard deviation of each tuple is available [5], although such assumption of uncertainty is fairly simple and modest, but it is not a mainstream model in the uncertain data management field. Note that in this paper, we mainly focus on the pdf model as it has been widely used in the setting of uncertain data model. In this pdf model, there exist two different distribution: (a) uniformly distributed, the probabilities of the instances of the same uncertain data are equal, as shown in Fig. 1; (b) non-uniformly distributed, the probability of instance in accordance with same distribution, as shown in Fig. 1(b).

Classification is one of the key problems in data mining area which can find interesting patterns, and has significant application merits in many fields. There are many published works on classification method [3,4,11,12,31,37,38]. Inducing uncertainty to the data makes the problem far more difficult to tackle, as this will further limit the accuracy of subsequent classification. Therefore, how to effectively classify uncertain data is great importance. There are many challenges which will affect the uncertain data classification.

Challenge 1: What is the classification over uncertain data? When considering deterministic data, it is deterministic which classes the certain object belongs to. However, over uncertain data, it is uncertain that which classes the uncertain object belongs to. Thus, classification result over uncertain data cannot be de fined just based on the definition of classification over deterministic data.

☆This research is supported by the NSFC (Grant nos. 61173029, 61472069, 61332006, 60933001, 75105487 and 61100024), National Basic Research Program of China (973, Grant no. 2011CB302200-G), National High Technology Research and Development 863 Program of China (Grant no. 2012AA011004) and the Fundamental Research Funds for the Central Universities (Grant no. N110404011).

http://dx.doi.org/10.1016/j.neucom.2015.05.121
0925-2312/

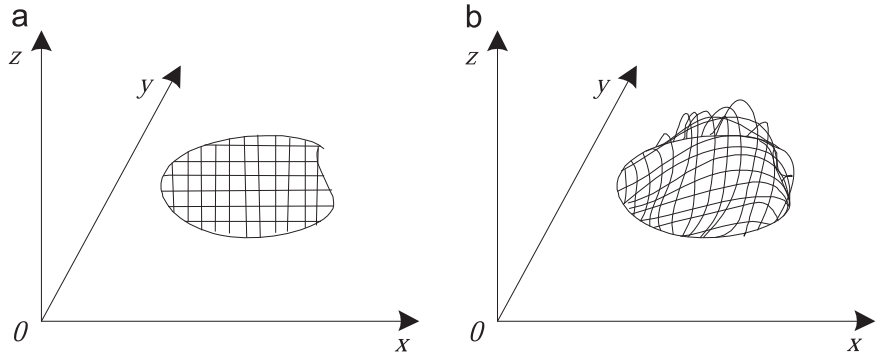

Fig. 1. Uncertain object: (a) uniformly distributed and (b)non-uniformly distributed.

Our contribution. We present a new definition of classification over uncertain data. Keeping the basic idea of the traditional definition of classification, we employ probability in this new definition. We can get the probability of the uncertain object belonging to any class, then it will be assigned to the class with the maximum probability.

Challenge 2: How can the uncertain data be classified efficiently? Each uncertain object contains some possible instances, and any instance has an exist probability. A naive approach is to process all instances of uncertain object, and the sum up the probabilities of instances that belong to the same class, then the uncertain object is assigned the class which the maximum probability belongs to. This naive approach is infeasible in realistic due to the reason, it costs too much time to process all the instances contained in an uncertain object. In order to classify the uncertain data, a more effective approach is expected.

Our contribution. We propose a pruning-based approach to effectively and efficiently reduce the processing amount of instances and save the cost. First, based on ELM method, we propose the Uniformly distributed Uncertain data classification based on ELM (UU-ELM) algorithm for classification over uncertain data which are uniformly distributed, which can quickly get the bounds of probability of objects belong to each class, to improve the efficiency. Second, we propose NU-ELM algorithm for classification over uncertain data which are in non-uniformly distributed. By calculating the upper bounds and lower bounds, we can reduce the amount of calculation.

Motivation (sensor data): Sensor networks are frequently used to monitor the surrounding environment, in which each sensor reports its measurements to a central location. In the case of environmental monitoring sensor net, measurements may include air pressure, temperature and humidity. The true measured values cannot be accurately obtained due to limitations of the measuring equipments. Instead, sensor readings are sent in order to approximate the true value, leading to uncertain objects.

Traditional classification algorithms are unable to deal with such challenges. In this paper, we investigate uncertain data classification based on ELM [13,15–17,19–22,26,27]. In the remainder of this paper, we first introduce the ELM in Section 2. After that, we formally define our problem in Section 3. We analyze the challenges of classification over uncertain data, and develop the two algorithms respectively in Section 4. Section 5 presents an extensive empirical study. In Section 6, we conclude this paper with directions for future work.

2. Brief of extreme learning machine

In this section, we present a brief overview of ELM, developed by Huang et al. [18,23,26,28,29]. ELM is based on a generalized Single Hidden-layer Feedforward Network (SLFN). The interpolation capability and universal approximation capability of ELMs have been investigated [24]. In ELM, the hidden-layer node parameters are mathematically calculated instead of being iteratively tuned, providing good generalization performance at thousands of times higher speeds than traditional popular learning algorithms for feedforward neural networks [24]. The output function of ELM for generalized *SLFNs* is represented by

$$f_L(x)=\sum_{i=1}^{L}\beta_i g_i(x)=\sum_{i=1}^{L}\beta_i G(a_i,b_i,x),\quad x\in R^d,\ \ \beta_i\in R^m \tag{1}$$

where $\beta=[\beta_1,\ldots,\beta_L]^T$ is the vector of the output weights between the hidden layer of L nodes and the output node, and g_i denotes the output function $G(a_i,b_i,x)$ of the ith hidden node. For additive nodes with activation function g, g_i is defined as

$$g_i=G(a_i,b_i,x)=g(a_i\cdot x+b_i),\quad a_i\in R^d,\ \ b_i\in R \tag{2}$$

For Radial Basis Function (RBF) nodes with activation function g, g_i is defined as

$$g_i=G(a_i,b_i,x)=g(b_i\|x-a_i\|),\quad a_i\in R^d,\ \ b_i\in R^+ \tag{3}$$

The above equations can be written compactly as

$$H\beta=T \tag{4}$$

where

$$H=\begin{bmatrix}h(x_1)\\ \vdots\\ h(x_N)\end{bmatrix}=\begin{bmatrix}G(a_1,b_1,x_1)\cdots G(a_L,b_L,x_1)\\ \vdots\\ G(a_1,b_1,x_N)\cdots G(a_L,b_L,x_N)\end{bmatrix}_{N\times L} \tag{5}$$

$$\beta=\begin{bmatrix}\beta_1^T\\ \vdots\\ \beta_L^T\end{bmatrix}_{L\times m}\quad\text{and}\quad T=\begin{bmatrix}t_1^T\\ \vdots\\ t_N^T\end{bmatrix}_{N\times m} \tag{6}$$

H is the hidden layer output matrix of the *SLFN* [13,14,24]; the i th column of H is the i th hidden node output with respect to inputs $x_1,x_2,\ldots,x_N\cdot h(x)=G(a_1,b_1,x),\ldots,g(a_L,b_L,x)$ is called the hidden layer feature mapping. The i th row of H is the hidden layer feature mapping with respect to the i th, input $x_i: h(x_i)$. It has been proved [24,28] that from the interpolation capability point of view, if the activation function g is infinitely differentiable in any interval, the hidden layer parameters can be randomly generated.

For the problem of multiclass classifier with single output, ELM can approximate any target continuous function and the output of the ELM classifier $h(x)\beta$ can be as close as possible to the class labels in the corresponding regions [25]. To maximize the training errors, ξ_i and to minimize the norm of the output weights, the classification problem for the proposed constrained-optimization-based ELM with a single-output node can be formulated as [25]

$$\texttt{Minimize}: L_{P_{ELM}}=\tfrac{1}{2}\|\beta\|^2+C\tfrac{1}{2}\sum_{i=1}^{N}\xi_i^2$$