Neurocomputing 174 (2016) 375–382

Contents lists available at ScienceDirect

Neurocomputing

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

Uncertain XML documents classification using Extreme Learning Machine

Xiangguo Zhao *, Xin Bi, Guoren Wang, Zhen Zhang, Hongbo Yang

College of Information Science and Engineering, Northeastern University, Liaoning, Shenyang 110819, China

ARTICLE INFO

Article history:
Received 18 September 2014
Received in revised form
12 January 2015
Accepted 19 February 2015
Available online 12 August 2015

Keywords:
Extreme Learning Machine
Classification
XML
Uncertain Data

ABSTRACT

Driven by the emerging network data exchange and storage, XML documents classification has become increasingly important. Most existing representation model and conventional learning algorithm are defined on certain XML documents. However, in many real-world applications, XML datasets contain inherent uncertainty, which brings greater challenges to classification problem. In this paper, we propose a novel solution to classify uncertain XML documents, including uncertain XML documents representation and two uncertain learning algorithms based on Extreme Learning Machine. Experimental results show that our approaches exhibit prominent performance for uncertain XML documents classification problem.

1. Introduction

Classification over XML documents is a challenging and important task in XML data mining and management. In general XML documents' classification problems, XML documents have to be transformed into a specific representation model, and taken as the input to classifiers. *Classifiers* can be trained using various learning algorithms, among which Extreme Learning Machine (ELM) [1,2] shows good generalization performance and extreme learning speed in a variety of applications, including multimedia recognition [3,4], industry process control [5,6], financial prediction [7], bioinformatics [8,9], mobile objects [10], etc.

On the other hand, since *representation models* of plain text are unable to express both semantic and structure information of XML documents, Structured Link Vector Model (SLVM) was proposed in [11] to take advantage of the structure and link information. Each element of SLVM is determined by term weights, XML structure and neighboring documents. Based on SLVM, improved XML representation models RS-VSM [12] and DSVM [13] were designed to enhance both semantic and structure representation abilities.

However, in real-world applications, due to the occurrence of inaccurate measurement, noises and incompleteness in data generation and collection, semantic and structural information of XML documents are usually uncertain. Probabilistic XML data model was first studied in [14] to represent uncertainty in XML documents. Different from certain XML data model, two major features were introduced. Probability attribute denotes the probability of a particular element existing in the XML database at the specified location, while distribution construct, including mutually-exclusive and independent, records dependencies between element value probabilities. Based on the expressiveness analysis of probabilistic XML models in [15], a full complexity analysis of queries and updates on probabilistic tree model *p-document* was given in [16,17], along with the probable possible worlds removing and probabilistic tree validating. The expressive power of p-documents was further studied in [18], drawing the conclusion that various known models of probabilistic XML can be represented as instantiations of the abstract notion of p-document. Based on those probabilistic XML models, keyword queries were studied in [19,20] and twig queries in [21–23]. In order to prune useless intermediate results with small probabilities, top-*k* ranking was also introduced into uncertain XML query processing in [24–28].

To our best knowledge, this is the *first* paper addressing the problem of uncertain XML documents classification. When preprocessing the raw datasets of uncertain XML documents, uncertain XML documents with appearance probabilities of elements will derive exponential number of possible world trees. A naïve method is to enumerate all the instances in the whole possible worlds, calculate the appearance probability of each instance and treat it as a training sample. But this enumeration method within such extremely large search space is far too inefficient. Therefore, it is necessary to propose a solution to efficient uncertain XML documents classification.

To address the above challenges, our contributions in this paper can be summarized as follows.

* Corresponding author.
E-mail address: zhaoxiangguo@mail.neu.edu.cn (X. Zhao).

http://dx.doi.org/10.1016/j.neucom.2015.02.095
0925-2312/

- *First* paper to discuss the problem of uncertain XML documents classification, including the uncertain XML representation and training strategies.
- An enumerated instance appearance probability based uncertain ELM is proposed to demonstrate the problem definition.
- A sampling method based uncertain ELM is proposed to further improve the uncertain learning procedure.
- Extensive experiments are conducted to verify the effectiveness and efficiency of our approaches.

The rest of this paper is organized as follows. Section 2 introduces uncertain XML data model and the problem definition of uncertain XML documents classification. Section 3 gives a brief introduction of ELM. Section 4 presents the appearance probability based uncertain learning algorithm. A sampling based uncertain learning algorithm is proposed in Section 5. Experimental results are analyzed in Section 6, followed by the conclusion of this paper in Section 7.

2. Problem definition

The representation of uncertain XML documents in this paper is based on the uncertain data model and probabilistic XML model described in this section. In order to be contrasted, an example of uncertain XML tree will be presented. The problem of uncertain XML documents classification is also formally defined.

2.1. Uncertain XML

A deterministic XML document is often represented in a certain tree model, which is defined as follows.

Definition 1 (*XML tree*). An *XML tree* t is an unranked and unordered tree, in which $V(t)$ is the set of nodes, $E(t) \subseteq V(t) \times V(t)$ is the set of edges, the *root* and *leaves* of the tree are denoted as $R(t) \in V$ and *leaves*(t), function $\varphi(v)$ associates a specific *label* to each node v of V. If $(n_1, n_2) \in E(t)$, $n_1 \in V(t)$ and $n_2 \in E(t)$, then we say that n_1 is the *parent* of n_2 and n_2 is the *child* of n_1. If there is a path from n_1 to n_2 in t, then n_1 is an *ancestor* of n_2 and in turn, n_2 is a *descendant* of n_1.

An uncertain XML document is modeled as a probabilistic XML tree, i.e., p-document, which is a probability distribution over a space of ordinary documents. A specific model is a mechanism for defining that distribution in terms of a probabilistic process that generates a random possible world, which is an ordinary XML document [17].

Definition 2 (*Probabilistic XML tree*). Based on p-document, a *probabilistic XML tree* t_p is a tree whose $V(t_p)$ consists of two types of nodes, i.e., *ordinary nodes* $V_{ord}(t_p)$ consistent with V in t and *distributional nodes* $V_{dst}(t_p)$ indicating a distribution (e.g. *IND* for probabilistically independent and *MUX* for probabilistically mutually exclusive) over the subsets of its children. The root and leaves are required to be ordinary nodes, that is $R(t_p) \in V_{ord}(t_p)$ and *leaves*$(t_p) \subseteq V_{ord}(t_p)$.

In p-document, the children of the distributional nodes are attached with appearance probabilities. According to the distributional node type, an uncertain XML document presented in p-document can derive a number of possible worlds. The appearance probability of a possible world is calculated according to the probability distribution of the p-document. Example 1 shows a p-document model of an uncertain XML document and one of its possible worlds.

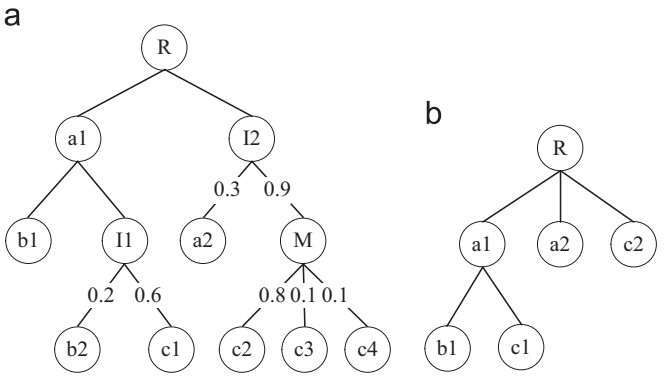

Fig. 1. An example of an uncertain XML document. (a) An uncertain XML document. (b) A possible world.

Example 1. An example of an uncertain XML document modeled in p-document is presented in Fig. 1(a). Node R is the root of the XML document; a1 and b1 are ordinary nodes. Since the parent of b2 and c1 is a distributional node of type IND, b2 with an appearance probability of 0.2 and c1 with 0.6 are independent with each other. I2 is another IND distributional node with two child nodes, i.e., a2 and M. M is a distributional node of type MUX, that is, any two of these three nodes c2, c3 and c4 are mutually exclusive with each other. Therefore in any possible world, either c2 appears with a probability of 0.8 or c3 with 0.1 or c4 with 0.1. Fig. 1(b) shows a possible world derived from Fig. 1(a). Since nodes R is the root node, a1 and b1 are original nodes, they appear in all the possible worlds. We assume that b2 does not appear and c1 appears, the probability of this situation is $(1-0.2) \times 0.6$; we also assume that both a2 and M exist, in which case the probability is 0.3×0.9; for the node M, only one child at most is allowed to appear in a possible world, which in this case we assume is c2 with its appearance probability of 0.8. Therefore, the probability of the possible world in Fig. 1(b) is $(1-0.2) \times 0.6 \times 0.3 \times 0.9 \times 0.8 = 0.10368$.

2.2. XML representation model

In the problem of document classification, a document should be transformed into the representation of a vector, and then taken as an input to learning algorithms. Vector Space Model (VSM) [29] is often used to represent plain text documents, which takes term occurrence statistics as feature vectors.

However, representing an XML document in VSM directly will lose the structural information. Structured Link Vector Model (SLVM) is proposed in [11] based on VSM to represent semi-structured documents, which contains both semantic and structural information. SLVM is defined as

$$\mathbf{d}_{slvm} = \langle \mathbf{d}_1, \ldots, \mathbf{d}_n \rangle \tag{1}$$

where $\mathbf{d}_i$ is a feature vector of the ith XML element calculated as

$$\mathbf{d}_i = \sum_i (TF(w_i, doc.e_j) \times \varepsilon_j) IDF(w_i) \tag{2}$$

where w_i is the ith term, ε_j is a unit vector corresponding to the element e_j.

In SLVM, each $\mathbf{d}_{slvm}$ is a feature matrix $\mathbf{R}^{n \times m}$, which is viewed as an array of VSMs. $\mathbf{d}_i$ consists of the feature terms corresponding to the same XML element, which is an m dimensional feature vector in each element unit.

Based on SLVM, in [12], we proposed Reduced Structured Vector Space Model (RS-VSM), which not only inherits the advantages of representing structural information of SVLM, but also achieves a better performance due to the feature subset selection based on information gain. We also proposed Distribution based Structured