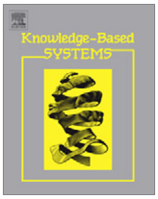
Knowledge-Based Systems 24 (2011) 1203–1213

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



# A multi-instance ensemble learning model based on concept lattice

Xiangping Kang [a,b], Deyu Li [a,b,*], Suge Wang [a,b,c]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China
[c] School of Mathematics Science, Shanxi University, Taiyuan 030006, Shanxi, China

ARTICLE INFO

*Article history:*
Received 23 August 2010
Received in revised form 19 May 2011
Accepted 20 May 2011
Available online 25 May 2011

*Keywords:*
◇-Concept lattice
Multi-instance learning
Local target feature set
Ensemble learning
Content-based image retrieval

ABSTRACT

This paper introduces concept lattice and ensemble learning technique into multi-instance learning, and proposes the multi-instance ensemble learning model based on concept lattice which can be applied to content-based image retrieval, etc. In this model, a ◇-concept lattice is built based on training set firstly. Because bags rather than instances in bags will serve as objects of formal context in the process of building ◇-concept lattice, the corresponding time complexity and space complexity can be effectively descend to a certain extent; Secondly, the multi-instance learning problem is divided into multiple local multi-instance learning problems based on ◇-concept lattice, and local target feature sets are found further in each local multi-instance learning problem. Finally, the whole training set can be classified almost correctly by ensemble of multiple local target feature sets. Through precise theorization and extensive experimentation, it proves that the method is effective. Conclusions of this paper not only help to understand multi-instance learning better from the prospective of concept lattice, but also provide a new theoretical basis for data analysis and processing.

## 1. Introduction

The term multi-instance learning was coined by Dietterich et al. [1] in mid-90s of last century, when they were investigating the problem of drug activity prediction. The training set is composed of labeled bags each consisting of many unlabeled instance. A bag is positively labeled if it contains at least one positive instance and negatively labeled otherwise. Learning system learns from the training set of bags to predict the label of the bag out of the training set as accurately as possible. As a blind spot of machine learning multi-instance learning possesses the promising application and unique feature, which has attract much attention of international machine learning community. And it has been regarded as the fourth instance learning framework juxtaposed with supervised learning, unsupervised learning, and reinforcement learning.

Formal concept analysis (FCA) is an order-theoretic method for the mathematical analysis of scientific data, pioneered by [3] in mid 80s. Over the past 20 years, FCA has been widely studied [2,4–9] and becomes a powerful tool for machine learning [10], software engineering [11] and information retrieval [12].

Since the multi-instance learning model was coined by Dietterich et al., machine learning community pays much attention to it. The study of multi-instance learning is very flourishing, such as [13–18]. In multi-instance learning research some tools of data analysis and processing were introduced to enrich multi-instance learning algorithms greatly. For example, Ruffo [19] presented Relic algorithm which introduced traditional decision tree C4.5 into multi-instance learning; Chevaleyre and Zucker [20] improved the traditional decision tree algorithm ID3 by introducing multiple-entropy, and presented the ID3 algorithm called ID3-MI based on multi-instance learning. Zhou and Zhang [21] applied ensemble learning technique to multi-instance learning, etc. Although there exist some algorithms, some problems cannot be solved effectively. Therefore, better multi-instance learning algorithms are needed further. This paper introduced concept lattice and ensemble learning technique into multi-instance learning, and proposed the multi-instance ensemble learning model based on concept lattice. The model not only helps to understand the multi-instance learning problem better, but also provides a new theoretical basis of analyzing and processing multi-instance learning problem.

This paper is organized as follows. Section 2 proposes a new research background of multi-instance learning. Section 3 briefly recalls basic notions of FCA. Section 4 generates a multi-instance formal context and builds a ◇-concept lattice. Section 5 finds the local target feature set in each local multi-instance learning problem based on ◇-concept lattice. Section 6 discusses the ensemble of multiple local target features sets thoroughly. Conclusions and discussions of further work will close the paper in Section 7.

[*] Corresponding authors at: School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China.

*E-mail addresses:* kangxiangping@yahoo.cn (X. Kang), lidy@sxu.edu.cn (D. Li), wsg@sxu.edu.cn (S. Wang).

0950-7051/$ - see front matter Crown Copyright © 2011 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.05.010

## 2. A new research background of multi-instance learning

In the task of drug activity prediction, the instance of the bag is the abstract of different shape of bag's corresponding molecule. As molecule cannot possess different shape at one time, instances of the bag will not appear simultaneously. In the task of natural scenes classification, instances of the bag only described the different components of the bag, and they must ensemble so as to describe a bag perfectly. So multi-instance learning can solve various practical problems by building different models. The background of the multi-instance ensemble learning model based on concept lattice proposed in this paper is shown as follows.

With the rapid expansion of multimedia information and the widespread popularity of computer network, content-based image retrieval attracts more and more attention. However existing image recognition technique is not mature yet in the task of image retrieval, namely image recognition algorithms cannot distinguish objects or identify implicit relations of objects in the image. So a large number of pictures on the Internet cannot be retrieved by users, which leads to a waste of picture resources. How to make use of picture resources on the Internet effectively is a problem we are faced with. In order to understand scenes in the image preferably we employ human's priori knowledge as the additional information of the image, because human can distinguish objects and identify implicit relations between objects in the image. Introducing human's priori knowledge is a new technology and also an innovation point of this paper.

The additional information of images and implicit relations of objects in images are expatiated as follows.

when users are faced with a picture named picture 1 containing {antelope, lion, grassland, zebra, tree, mountain, sky, ...}, they can easily identify all objects in the picture. But users only care about the objects they are interested in. For example, in the picture user A is only interested in {zebra, antelope}, user B is only interested in {zebra, grassland}, user C is only interested in {lion, zebra, antelope} and so on. Sets of users' interests are collected as additional information of the picture 1, and corresponding additional information is shown as follows:

$$\begin{array}{llll} Picture1: & User\ A & : & zebra, antelope \\ & User\ B & : & zebra, grassland \\ & User\ C & : & lion, zebra, antelope \\ & \cdots & : & \cdots \quad \cdots \end{array}$$

In fact relations of objects in the image are diverse and complex, and they are not sets of individual objects without any connection. How to explore these implicit relations and understand scenes in the image better is the key of enhancing image retrieval capability further. Since human can identify objects and their implicit relations in the image, we introduce human's priori knowledge to overcome above problems. For example, the picture 1 describes the following scenes: in the blue sky, lions are chasing after a group of zebras and antelopes on the grassland, and there are mountains, trees and so on in the distant. Based on scenes in the picture human can identify implicit relations between objects as follows: 1. zebras and antelopes are herbivorous; 2. lions are hunting zebras and antelopes; 3. zebras are grazing on the grassland; 4. antelopes, lions, grassland, zebras, ... belong to the ecosystem of African grassland; ... Objects in the image can be clustered by these relationships, for example, Relationship 1 can cluster antelopes and zebras into one group; Relationship 2 can cluster lions, antelopes and zebras into one group, etc.

Assume objects of the set of users' interests must satisfy certain relations. As different users may have different interests, for example, in the picture user A may be interested in herbivorous; user C may be interested in "lions are hunting zebras and antelopes" and so on, users only care about implicit relations they are interested in. Clearly the additional information cannot only identify objects of the picture but also respectively cluster objects satisfying different relations together.

As there are too many implicit relations in the picture, in order to enhance the speed and the quality of image retrieval we only retain several relations users are most interested in. For example, although there exists the implicit relation "zebras are grazing on the prairie" in the picture 1, but few users are interested in it, so we donot contain it in additional information of the picture.

In real life, content-based image retrieval has been widely applied to application. The multi-instance ensemble learning model based on concept lattice proposed in this paper can be effectively applied to content-based image retrieval, of which the prerequisite conditions is that all pictures on the Internet have the additional information. Taking a practical application as an example the formal description of the learning model is shown as follows.

Assume there is a picture gallery of African grassland on one website, in which pictures are positively labeled or negatively labeled according to their image retrieval rates. The picture possessing high image retrieval rate is positively labeled and negatively labeled otherwise. Because every picture contains many objects, for example, picture 1 containing many objects is positively labeled, the current problem is that we cannot immediately determine which objects influence the classification according to additional information of pictures. If we get the answer, we can label pictures of the Internet positively or negatively. Pictures labeled positively are added to the picture gallery of the website, so that the image retrieval rate can be enhanced further. In this application, we assume that pictures of the Internet have additional information.

We label the picture positively or negatively according to two factors: one is whether existing one or more objects influence classification; the other is the relations of pictures users are interested in. For example, picture 2 also contains lion, antelope and zebra, and it describes the scene: lion is napping in a cage, and zebra, antelope are grazing in the distance. Obviously there does not exist a reasonable implicit relation which can cluster lion, antelope and zebra into one group. So when {lion, zebra} has influence on classification, we regard picture 1 as a positive class. We do not think picture 2 is also a positive class, although picture 2 also contains lion, antelope and zebra. Therefore in every positive class there must exist a reasonable relation which can cluster {lion, zebra} into one group. For example, picture 3 describes the scene: lion, zebra and elephants are performing for audience in the zoo. Obviously there must exist a reasonable relation "lion and zebra are performing together" which can cluster {lion, zebra} into one group, then picture 3 is classified as a positive class.

Based on above application background the problem can be formalized as follows:

The training set is given:

$$training\_set = \{T_i^{\delta} | 1 \leqslant i \leqslant n\},$$

where $T_i^{\delta} = \{g_{ij} | 1 \leqslant j \leqslant m_i\}$ is a bag with $\delta \in \{+, -\}$.

If $\delta = +,\ T_i^{+}$ is a positive bag; Conversely, if $\delta = -,\ T_i^{-}$ is a negative bag. $g_{ij}$ denotes a instance of $T_i^{\delta}$. Here, the bag is equivalent to the additional information of the picture, and the instance is equivalent to one set of users' interests.

For the sake of convenience, we denote *training_set* in simplified form as $\Sigma$ in the following.

Assume $T^{\delta} = \{g_j | 1 \leqslant j \leqslant m\}$ is a bag whose feature set is denoted by $T_{\triangleright}^{\delta} = \bigcup_{j=1}^{m} g_j$. The feature set of training set $\Sigma$ is denoted by $\Sigma_{\triangleright} = \bigcup_{i=1}^{n} T_{i\triangleright}^{\delta}$.