Expert Systems with Applications 42 (2015) 8283–8293

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An enhanced noise resilient K-associated graph classifier

Mahdi Mohammadi [a,*], Bijan Raahemi [a], Saeed Adel Mehraban [b], Elnaz Bigdeli [a], Ahmad Akbari [b]

[a] Knowledge Discovery and Data Mining Lab, University of Ottawa, Canada
[b] Iran University of Science and Technology, Computer Engineering Department, Iran

ARTICLE INFO

Article history:
Available online 29 June 2015

Keywords:
Graph-based classifier
Noisy samples
K-associated graph

ABSTRACT

In this paper, we propose a non-parametric, noise resilient, graph-based classification algorithm. By modifying the training phase of the k-associated optimal graph algorithm, and proposing a new labeling algorithm in the testing phase, we introduce a novel approach that is robust in the presence of different level of noise. In designing the proposed classification method, each class of dataset is represented by a set of sub-graphs (components), and a new extension of the k-associated optimal graph algorithm is introduced in the training phase to combine the smaller components. With this enhancement, we demonstrate that our algorithm distinguishes between noisy and non-noisy sub-graphs. Moreover, in the testing phase, we combine relational data, such as the degree of relevancy, with non-relational attributes, such as distance, for each sample in a graph to make the proposed algorithm less sensitive to noise. Gravity formula is the main concept behind the proposed test sample with various modifications to tailor it to the arbitrary shape and non-uniform sample scattering of the graph structure. We compare the proposed method with a graph-based classifier, as well as two other well-known classifiers, namely, Decision Tree and Multi-Class Support Vector Machine. Confirmed by the *t*-Test score, our proposed method shows a superior performance in the presence of different levels of noise on various datasets from the UCI repository. At a noise level of 5% or higher, the proposed algorithm performs, in average, 7% better than the graph-based classification algorithm. At a noise level of 20%, the proposed method performs, in average, 5% better than Decision Tree and multi-class SVM.

1. Introduction

Graph structure has been studied in machine learning areas in tasks such as clustering, classification and feature reduction (Belkin & Niyogi, 2003; Dhanjala, Gaudelb, & Clémençonc, 2014; Vathy-Fogarassy & Abonyi, 2009). Graph representation has specific characteristics. It can present the topological structure of the data. It can also propose a hierarchical structure by considering a graph as a set of sub-graphs representing an arbitrary shape for a class or a cluster. This is the reason the graph structure has attracted much attentions in machine learning area. Perhaps, graph clustering is the most important application of the graph structure since it is capable of extracting the arbitrary shape of clusters (Chen & Jian, 2014; Dhanjala et al., 2014; Ye & Jin, 2014). Semi-supervised learning is another application of the graph structure (Chen, Li, & Peng, 2009; Ozaki, Shimbo, Komachi, & Matsumoto, 2011; Zhu, 2008) in which only a small portion of the data is labeled. Based on the labeled data, a graph-based classifier is trained to predict the label of unlabeled data. The newly labeled data can be added to the formerly labeled data to retrain the classifier and improve the model accuracy. Classification is an application of the graph structure which has not been receiving much attention in comparison with graph clustering and semi-supervised learning. The graph classification problem can be discussed in two different ways. The first one is how to classify separate and individual graphs in a graph database into two or more categories (Ketkar, Holder, & Cook, 2009). The other application is how to represent a vector dataset as a set of sub-graph, each of which illustrates a class of training dataset. In other words, a graph-based classifier consists of some subgraphs representing the training dataset. A number of algorithms have been introduced for building the set of sub-graphs for an input training dataset (Bertini, Zhao, Motta, & Lopes, 2011; Chatterjee & Raghava, 2012; Iliofotou et al., 2011). In this paper, we focus on the latter application of graph classification for machine learning.

In Bertini et al. (2011), the authors propose a graph-based classification method based on K-associated graph which presents each class of data as a set of sub-graphs (components). Their

* Corresponding author at: Knowledge Discovery and Data mining Lab, University of Ottawa, 55 Laurier Ave, E., Ottawa, ON K1N 6N5, Canada. Tel.: +1 613 5625800.
E-mail addresses: mmohamm6@uottawa.ca (M. Mohammadi), braahemi@uottawa.ca (B. Raahemi), samehraban@comp.iust.ac.ir (S.A. Mehraban), ebigd008@uottawa.ca (E. Bigdeli), akbari@iust.ac.ir (A. Akbari).

http://dx.doi.org/10.1016/j.eswa.2015.06.044
0957-4174/

proposed method is a non-parametric algorithm contrasting to *K* nearest neighbor classifiers, does not need model selection, does not consider relational data and does not make use of kernel nor Laplacian. They also introduced a new concept, called *purity* which measures the connectivity level of samples in each component. The algorithm builds a K-Associated Optimal Graph (KAOG) which is based on K-associated graph algorithm. The detail of the K-Associated Optimal Graph and K-associated graph steps is presented in Sections 2.1 and 2.2. The experimental results in their work indicated that the KAOG algorithm can compete with other well-known and non-graph -based algorithm such as Decision Tree, support vector machine and *K* nearest neighbor in terms of correct detection rate with and without the presence of the noisy data.

In this paper, we modify the training phase of KAOG algorithm, and also, propose a new testing phase. Accordingly, we propose a new graph-based classifier which is more noise resilient than the KAOG. In the testing phase, we combine the relational data with purity measure to propose a new testing formula. Using relational data, the label of each sample point does not only depend on its own attributes, it might also be affected by the label of its neighbors. Relational data can be presented in graph form, in which, some samples are more correlated to their own graph rather than the other samples in the same graph. The samples with more correlation or membership value can be more effective than the other samples in labeling a new test sample.

Although there exist graph-based classification algorithms (Chatterjee & Raghava, 2012; Iliofotou et al., 2011), to the best of our knowledge, there are no methods focusing on robust graph-based classification algorithm using relational data and density to achieve acceptable accuracy in presence of noise. The main contributions of the proposed algorithm are as follows:

- We introduce a graph-based classification with higher accuracy on noisy dataset in comparison with published well-known classification algorithms.
- The main contribution of the training phase of the MKAOG is to merge small components into larger ones which, consequently, makes the proposed algorithm less sensitive to noise.
- We also propose a new testing method to label test samples in which, the number of samples in each graph and the level of connectivity play a distinctive role in labeling process. We illustrate that the noisy sub-graphs have a lower chance to win the competition against non-noisy sub-graphs.

The rest of the paper is organized as follows: an overview of graph-based classifiers and its applications are presented in Section 2, followed by explaining the KAOG algorithm. In Section 3, we introduce the proposed algorithm. Analysis and discussion of the proposed method are presented Section 4. The experimental results are explained in Section 5, followed by the conclusions in Section 6.

2. An overview of graph-based classification

Graph theory has different applications in machine learning and data mining mainly on clustering (Chen & Jian, 2014; Dhanjala et al., 2014; Ye & Jin, 2014) and association rules mining (Al-Khassawneh, Bakar, & Zainudin, 2012; Gross, Khopkar, & Nagi, 2013; Yiyuan, Yuejia, Shijie, & Dapeng, 2015). The problem of graph classification was first studied by Gonzalez, Holder, and Cook (2002) as a greedy method to find sub-graphs in a dataset. In Deshpande, Kuramochi, and Karypis (2005), the authors presented a graph classification algorithm which uses frequent subgraph discovery algorithms to find all topological substructures in a dataset. By using highly efficient frequent subgraph discovery algorithms, they reduced the computational complexity of the proposed algorithm, based on which they were able to select the most discriminative sub-graph candidate to improve the accuracy of the classifier. Chatterjee and Raghava (2012) proposed a data transformation algorithm to improve the accuracy of two classifiers (LD and SVM). First they employed a similarity graph neighborhood (SGN) in the training feature subspace and mapped the input dataset by determining displacements for each entity and then trained a classifier on the transferred data.

The graph structure has been also employed as a classifier to predict the label of the test data. In Benso, Carlo, and Politano (2011) the authors proposed a microarray data classifier by using the graph theory which is comparable with other classification algorithms. In their proposed method, first a graph is built where each vertex represents a gene, and the edges explain the relation between the genes. In the case of microarray technology, there are two main limitations: the reliability of the training data sets which is used to build the classifiers and the classifiers' performances. The authors illustrated that their proposed graph-based method could cope with most of the limitations of known classification methodologies.

The graph-based classification illustrates promising results in many applications such as wireless networks. In Mahmood, Muhamad, and Khan (2008), the authors proposed a combinational method consisting of neural network and graph theory. In their work a new type of neural network, called DHNG, is proposed which uses a hierarchical graph-based structure of the input patterns adopting a one-cycle learning process. In wireless networks, the rate of input data which needs to be classified is very high. In this case, time and memory complexity the classifier is very important. This is why the authors examined the computational complexity of their proposed method with the computational complexity of a self-organizing map (SOM) classifier in a supervised environment. The proposed method has lower computational complexity than that of SOM, and also, outperforms SOM in terms of correct detection rate.

In Bertini et al. (2011), the authors proposed a multi-class classification (KAOG) which benefits from the local attributes, as well as global statistical properties of the graph structure. Although the KAOG method can outperform some other classification algorithms, it still has some drawbacks in terms of building the graph model based on the given training data. In our proposed method, we improve the accuracy of KAOG algorithm in presence of different level of noise. We introduce a new testing phase for the algorithm to make it robust to distinguish between noisy and non-noisy samples. We also propose a solution to correctly classify the noisy samples. The next section explains the KOAG algorithm in details. The KAOG algorithm works based on the K-associated graph which is explained in Section 2.1.

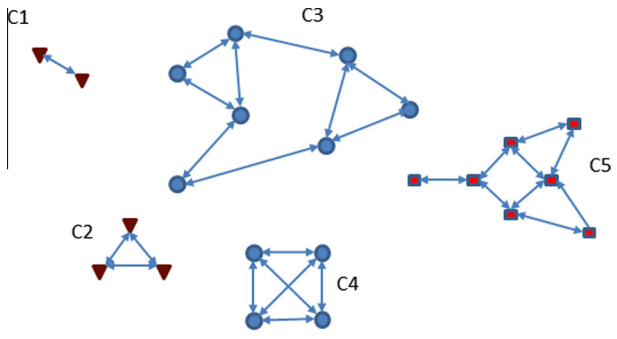

Fig. 1. An example of the clusters generated by the KAG algorithm with 5 components ($K = 3$). (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.)