Pattern Recognition Letters 37 (2014) 41–53

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Feature enrichment and selection for transductive classification on networked data

Zehra Cataltepe *, Abdullah Sonmez, Baris Senliol

Istanbul Technical University, Faculty of Computer and Informatics, 34469 Maslak, Istanbul, Turkey

ARTICLE INFO

Article history:
Available online 26 July 2013

Keywords:
Feature enrichment
Feature selection
Collective classification
Fast Correlation Based Filter Sharp (FCBF#)
Networked data
Semi-supervised learning

ABSTRACT

Networked data consist of nodes and links between the nodes which indicate their dependencies. Nodes have content features which are available for all the data; on the other hand, the labels are available only for the training data. Given the features for all the nodes and labels for training nodes, in transductive classification, labels for all remaining nodes are predicted. Learning algorithms that use both node content features and links have been developed. For example, collective classification algorithms use aggregated (such as sum or average of) labels of neighbors, in addition to node features, as inputs to a classifier. The classifier is trained using the training data only. When testing, since the neighbors' labels are used as classifier inputs, the labels for the test set need to be determined through an iterative procedure.

While it is usually very difficult to obtain labels on the whole dataset, features are usually easier to obtain. In this paper, we introduce a new method of transductive network classification which can use the test node features when training the classifier. We train our classifier using enriched node features. The enriched node features include, in addition to the node's own features, the aggregated neighbors' features and aggregation of node and neighbor features passed through simple logical operators OR and AND. Enriched features may contain irrelevant or redundant features, which could decrease classifier performance. Therefore, we employ feature selection to determine whether a feature among the set of enriched features should be used for classifier training or not. Our feature selection method, called FCBF#, is a mutual information based, filter type, fast, feature selection method. Experimental results on three different network datasets show that classification accuracies obtained using network enriched and selected features are comparable or better than content only or collective classification.

1. Introduction

Learning problems with network information, where each instance has its content features (attributes) and relations (links) with other instances, have recently become more common. Examples include social, financial, computer, citation, semantic, ecological and gene regulatory networks. Classification of nodes or links in the network, identification of unobserved or essential links or nodes are some of the research areas on networked data.

In content only classification of a dataset, only the content features of an instance are used as inputs to the classifier. On the other hand, when networked data are available, in addition to the features of an instance (node), links between instances are also given. In this paper, we consider the transductive classification problem on a network (Macskassy and Provost, 2007; Sen et al., 2008; Ji et al., 2010). We assume that features and links of all nodes are known. There is a training set of nodes for which labels are also known. The task is the prediction of labels for the remaining (test) nodes as accurately as possible. We are interested in using link information, in addition to content features, as well as using the knowledge of test nodes' links and features, to improve classification accuracy. In order to achieve this goal, we propose using enriched content features, which allows all nodes', including test nodes', link and feature information to be used for classifier training. We also propose using feature selection to determine which enriched features, among many possible, should be included in the classifier.

For example, consider the problem of predicting the age group (0 if the person is below 18 or 1 is the person is 18 or older) of an individual based on the singers/groups s(he) likes, her/his friends and the singers/groups they like. In Fig. 1, top row, each individual (A, B, C, D) is shown as a node. Labels for B (0) and D (1) are known and these nodes are in the training set. We want to train a classifier which will best predict the label for nodes in the test set, namely A and C. Whether each individual likes or dislikes, in order, Adele, Justin Bieber, R.E.M., Katy Perry, U2 and The Beattles are shown in her feature vector as 0's (dislikes) and 1's (likes). So, person A, who has the feature vector [101101] likes Adele, R.E.M., Katy Perry and The Beattles. In content only classification, we train classifiers which

* Corresponding author. Tel./fax: +90 212 285 3551.
E-mail address: cataltepe@itu.edu.tr (Z. Cataltepe).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.07.009

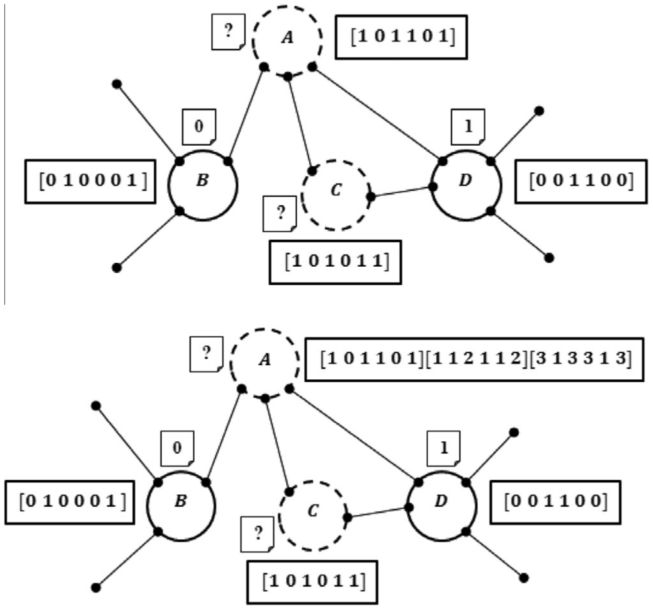

Fig. 1. (Top) An example transductive binary classification problem on a networked data set. (Bottom) Node A's enriched features consisting of content, neighbor and neighbor OR features.

only use these content features for each individual. A link only classifier would only use the labels of the neighbors of node A. Since A's neighbor C is in the test set, whether C is labeled as 0 or 1 effects the label predicted for node A. Therefore, for link only prediction one has to utilize an iterative procedure which lets test nodes be labeled based on their neighbors' actual (if they are in the training set) or current (if they are in the test set) label assignments. In order to use both content and link information, one can train classifiers by simply appending the node features with additional features which reflect properties of the neighbors' labels. For example, if C's current assignment is 1, then for node A, the number of neighbors who are above 18, would be 2 and the feature vector would have been [1011012].

Especially when there are many more test nodes than training nodes, or when unlabeled test nodes have many neighbors who are in the test set, relying on the neighbors predicted labels may not be a very good idea. In this paper, we propose making use of the neighbors features, as opposed to their predicted or actual labels, as additional features. We propose using enriched features, which consist of the sums of neighbors' feature vectors and also sums of outcomes of simple operators, such as AND or OR, between node and neighbor features. For example, in the bottom row of Fig. 1, for node A, we use the sum of neighbors feature vectors, which is [112112] and the sum of the node and neighbors' feature-wise OR'ed vectors, which is [313313]. Since the enriched feature vectors contain more features, training and testing classifiers with them would take more time. Also, some of the enriched features could be less useful for classification than the others. Therefore, we propose using feature selection to identify a set of features which would be the most useful for the classification task at hand.

Homophily (McPherson et al., 2001), i.e. that linked nodes are more likely to have the same label, has been one of the important requirements for link information to be useful for classification. Algorithms have been devised to take into consideration the neighbors' labels. weighted-vote relational classifier (wvRN) (Macskassy and Provost, 2003), is a relational learning algorithm that aggregates the neighbors' labels and uses them as inputs to a classifier. Aggregation methods (Perlich and Provost, 2006; Lu and Getoor, 2003; Sen and Getoor, 2007) which summarize the label information of the neighbors in a constant dimensional vector through taking the sum, average, max or existence of neighbor labels, have been used. By means of training classifiers with node content features, appended with aggregated neighbor labels, (Macskassy and Provost, 2007; Sen et al., 2008) have been able to use both content and link information to train classifiers.

There have been previous methods of feature construction which are related to our work and which aim to take advantage of network information to train better classifiers. The simplest method of feature construction is performed by weighted-vote relational classifier (wvRN) (Macskassy and Provost, 2003). wvRN determines the class of a node based on a weighted average of its neighbors' class probabilities. Chakrabarti et al. (1998) performed experiments on web pages with hyperlinks between them using Naive Bayes classifier and the relaxation labeling method. They showed that, for the datasets that they used, while using the labels of the neighbors in addition to nodes' contents improved performance, using the neighbors' contents did not.

There has also been some previous work on feature construction and then feature selection for networked data. Previously, (Popescul and Ungar, 2004) suggested approaches for feature construction from database tables using refinement graphs. Then they selected features using a statistical model selection criteria. Perlich and Provost's relational learning system ACora (Automated Construction of Relational Attributes) (Perlich and Provost, 2006) investigated many methods of feature construction, such as count, mode, max, using a node and its related entities. They outlined principles of feature aggregation, namely, aggregation should help with classification and various aggregation methods should be considered. So, they considered distances to the class-conditional distributions and used standard aggregates for feature construction. Although Perlich and Provost (2006) suggested that feature selection should be performed on the constructed features, they did not report results with feature selection because it did not improve results for the datasets they used.

In our previous work (Senliol et al., 2009), we used mRMR (Peng et al., 2005) feature selection for classification of networked data using the node features. We showed that content only or collective classification using feature selection can achieve accuracies as high as using all the features. In this paper, we introduce the network enriched features together with the fast and accurate FCBF# feature selection method, and we show that we can achieve better classification accuracies than content only or collective classification.

When test node labels need to be predicted, classifiers that use neighbor label information face a problem. Because the label to be assigned for a test node depends on the labels assigned to its neighboring test nodes. Collective classification algorithms, such as ICA (Iterative Classification Algorithm) (Macskassy and Provost, 2007; Sen et al., 2008), have been used for this purpose. In collective classification, the classifier is first trained using only training data, ignoring the test nodes, because their labels are not known. The trained classifier is used to assign the initial test labels. When a test node has neighboring test nodes, the assignment of neighbors may change the assignment for the node. Therefore, test node labels are reassigned based on each other until a stable solution is obtained. Note that, during collective classification, the classifier does not change, only the test node label assignments change. ICA algorithm can be related to other network diffusion algorithms, such as affinity propagation (Frey and Dueck, 2007), which is used for clustering and nodes propagate a degree of how they see the other node as their examplers. In ICA algorithm, for each test node the classifier output is computed and then propagated to its neighbors. Previously in Cataltepe et al. (2011) we have shown that instead of combining content and link features into a single feature vector and training a single classifier, training separate classifiers