Efficient Usage of Collective Classification Algorithms for Collaborative Decision Making

Tomasz Kajdanowicz

Faculty of Computer Science and Management, Institute of Informatics
Wroclaw University of Technology, Wroclaw, Poland
tomasz.kajdanowicz.@pwr.wroc.pl

Abstract. Collective classification algorithms with underlying network structure of related entities are a powerful modelling tool that can address collaborative decision making problems. The paper presents the usage of collective classification algorithms for classification problem in which unknown nodes are assigned with classes based on the classes of known nodes. In such problem the classification decision for particular node is inferred from collaborative knowledge of nodes with known classes and underlying network connections. The paper considers Iterative Classification (ICA) and Loopy Belief Propagation (LBP) algorithms applied in various network configurations for collaborative decision making. The experimental results revealed that greater number of output classes decreases classification accuracy and LBP outperforms ICA for dense network structures while it is worse for sparse networks.

Keywords: Collaborative Decision Making, Collective Classification, Iterative Classification, Loopy Belief Propagation.

1 Introduction

Traditionally, collaborative decision making (CDM) was recognized as a process of debates and negotiations among a group of people in order to make a decision. Therefore, CDM can be considered as problem with collaborative outcome that is a result from argumentative discourse and joint cooperation of human beings. It can be usually observed, that expected consensus emerges through the consideration of all alternative competing interests, priorities and constraints. In order to model CDM formally the underlying approach have to be articulated in a concise and agreed upon manner. One of a very powerful representations of collaborative environment, that can be utilized to undertake the collaborative decision is a network. This representation may be given in a various types of graphs. Depending on the assumed property these can be undirected, oriented or directed graphs organized in multigraphs, hypergraphs or pseudographs [1]. Additionally graphs can be unlabelled, edge-labelled, vertex-labelled or vertex and edge labelled. Using networks allows to represent the humans as vertices and all the relations between them as edges. Additionally, labels assigned to vertices can reflect particular standpoints.

It is addressed in the paper a very specific collaborative decision making problem - collective classification accomplished in network. Solving collective classification task

Y. Luo (Ed.): CDVE 2013, LNCS 8091, pp. 73–80, 2013.

it is expected to obtain the class of unknown network's nodes based on the available classes of known nodes. As the structure of connections is accessible for all nodes (with known and unknown class labels) it is utilized in the label inference process.

In general, there have been proposed several types of collective classification approaches. This paper provides an experimental comparison for Iterative Classification Algorithm (ICA) and Loopy Belief Propagation (LBP) algorithms that can be applied in order to solve the collaborative decision making problem.

The paper provides concise presentation of related work in the field of collective classification in Section 2, short description of considered algorithms in Section 3, experimental results and comparison of the methods together with evaluation of the algorithms' accuracy for different contributions of the known labels in the entire network are gathered in Section 4 and concluded in Section 5.

2 Related Work

There exist a variety of methods for collective classification. However, it can be distinguished two distinct types of them: local and global. The former methods use a collection of local conditional classifiers successively applied to the unknown nodes whereas the latter are defined as optimization of one global objective function [2].

Additionally classification of nodes in network can be solved using two distinct approaches: within-network and across-network inference. Within-network classification [3], for which training nodes are connected directly to other nodes, whose labels are to be classified, stays in contrast to across-network classification [4], where models learnt from one network are applied to another similar network.

There are related several problems with collective classification that have been currently addressed by researchers. One of them is the problem of what features should be used to maximize the classification accuracy. In approaches which use local classifiers the relational domain needs to be transformed to standard notation by application of proper aggregation operator. It has been reported that precise solution strongly depends on the application domain [5]. The previous research showed that new attribute values derived from the graph structure of the network, such as the betweenness centrality, may be beneficial to the accuracy of the classification task [6]. It was also confirmed by other research discussed in [7].

Another interesting problem in collective classification based on iterative algorithms is the ordering strategy that determines, in which order to visit the nodes iteratively to re-label them. The order of visiting the nodes influences the values of input features that are derived from the structure. A variety of sophisticated or very simple algorithms can be used for this purpose. Random ordering that is one of the simplest ordering strategies used with iterative classification algorithms can be quite robust [8].

One of the most popular local collective classification methods is Iterative Classification Algorithm (ICA) introduced by Geman & Geman in the image processing context [9]. It belongs to so called approximate local inference algorithms basing on local conditional classifiers [10]. Another technique is a Loopy Belief Propagation (LBP) [11] that is the global approximate inference method used for collective classification. As in the literature it was not found the comparison of predictive accuracy of mentioned methods across datasets from distinct domains we examine these algorithms using eight distinct datasets and present the comparison in the paper.

3 Collective Classification Techniques

3.1 Iterative Classification

The basic idea behind ICA is quite simple but reasonable. Considering a node $v_i \in V^{UK}$, where V^{UK} is a set of nodes with unknown label, $V^{UK} \subset V$, we aim to discover its label l_i. Having known labels of v_i's neighbourhood ICA utilizes a local classifier Φ that takes the attribute values of nodes with known labels (V^K) and returns the best label value for v_i from the class label set L. If the knowledge of the neighbouring labels is partial the classification process needs to be repeated iteratively. In each iteration labelling of each node v_i is done using current best estimates of local classifier Φ and continues until the label assignments are stabilized. A local classifier might be any function that is able to accomplish the classification task. It can range from a decision tree to an SVM in its place.

Algorithm 1 depicts the ICA algorithm as a pseudo-code where the local classifier is trained using the initially labelled nodes V^K only. It can be observed that the attributes utilized in classification depend on current label assignment (lines 8 and 9 in Algorithm 1). Thus there need to be performed the repetition of classification phase until labels stabilize or maximal number of of iteration is reached.

Algorithm 1. Iterative Classification Algorithm (ICA), the idea based on [10]

1: **for** each node $v_i \in V^{UK}$ **do**
2:     compute x_i, i.e. v_i's attributes using observed nodes V^K
3: **end for**
4: train classifier Φ by Θ optimization using attributes of V^K nodes
5: **repeat**
6:     generate ordering O over nodes in V^{UK}
7:     **for** each node $v_i \in O$ **do**
8:         compute x_i, i.e. v_i's attributes using current assignments
9:         $l_i \leftarrow \Phi(x_i, \Theta)$
10:     **end for**
11: **until** label stabilization or maximal number of iterations

3.2 Loopy Belief Propagation

Loopy Belief Propagation (LBP) is an alternative approach to perform collective classification in comparison to ICA. The main difference is that it defines a global objective function to be optimized, instead of performing local classifiers optimization.

Intuitively, LBP is an iterative message-passing algorithm. The messages are transferred between all connected nodes v_i and v_j, $v_i, v_j \in V$, $(v_i, v_j) \in E$, and might be interpreted as belief of what v_j label should be based on v_i label.

The global objective function that is optimized in the LBP is derived from the idea of pairwise Markov Random Field (pairwise MRF) [12]. In order to calculate the message to be propagated the calculation presented in Equation 1 is performed.

$$m_{i\to j}(l_j) = \alpha \sum_{l_i\in L} \Psi_{ij}(l_i,l_j)\phi(l_i) \prod_{v_k\in V^{UK}\setminus v_j} m_{k\to i}(l_i) \tag{1}$$

where $m_{i\to j}(l_j)$ denotes a message to be sent from v_i to v_j, α is the normalization constant that ensures each message sum to 1, Ψ and ϕ denotes the clique potentials. For further explanation see [10].

The calculation of believe can be concisely expressed as in Equation 2:

$$b_i(l_i) = \alpha\phi(l_i) \prod_{v_j\in V^{UK}} m_{j\to i}(l_i) \tag{2}$$

The LBP algorithm consist of two main phases: message passing that is repeated until the messages are stabilized and believe computation, see Algorithm 2.

Algorithm 2. Loopy Belief Propagation (LBP), the idea based on [10]

for each edge $(v_i,v_j)\in E, v_i,v_j\in V^{UK}$ **do**
    for each class label $l_i inL$ **do**
        $m_{i\to j}(l) \leftarrow 1$
    end for
end for
//perform message passing
repeat
    for each edge $(v_i,v_j)\in E, v_i,v_j\in V^{UK}$ **do**
        for each class label $l_i\in L$ **do**
            $m_{i\to j}(l_j) \leftarrow \alpha\sum_{l_i\in L}\Psi_{ij}(l_i,l_j)\phi(l_i)\prod_{v_k\in V^{UK}\setminus v_j} m_{k\to i}(l_i)$
        end for
    end for
until stop condition
//compute beliefs
for all $v_i\in V^{UK}$ **do**
    for all $l_i\in L$ **do**
        $b_i(l_i) \leftarrow \alpha\phi(l_i)\prod_{v_j\in V^{UK}} m_{j\to i}(l_i)$
    end for
end for

4 Experimental Study

4.1 Experimental Scenarios

In order to evaluate the considered collective classification algorithms in the context of collaborative decision making modelled as collective classification the predictive accuracy of ICA and LBP algorithms was examined. For this purpose an experimental environment has been developed in Java language. The ICA approach was provided with C4.5 decision tree as a base classifier. The experiments were carried out on original dataset with primary prepared splits between nodes with known and unknown labels. Each dataset was split into known and unknown node sets in nine distinct proportions (from 10% to 90% unkown labels). The split was accomplished by node sampling using uniform distribution. In order to assess distinct classification approaches standard measure of classification accuracy was recorded.

4.2 Datasets

The experiments were carried out on six datasets. The AMD_NETWORK graph presents seminary attendance at conference. The dataset was a result of a project that took place during The Last HOPE Conference held in July 18-20, 2008, New York City, USA. At this conference RFID (Radio Frequency Identification) devices were distributed among participants and allowed to uniquely identify them and track in which sessions they attended. The data set is build from information about descriptions of interests of participants, their interactions via instant messages, as well as their location over the course of the conference. Location tracking allowed to extract a list of attendances for each conference talk. In general, the most interesting for experiment information included in dataset are: information about conference participants, conference talks and presence on talks. The genealogy dataset CS_PHD is the network that contains the ties between Ph.D. students and their advisers in theoretical computer science where arcs points from an advisers to a students [13]. The dataset NET_SCIENCE contains a coauthorship network of scientists working on network theory and experiment [14]. It was extracted from the bibliographies of two review articles on networks. The biological dataset YEAST consists of protein-protein interaction network [15]. The PAIRS dataset is a dictionary from The University of South Florida word association, rhyme, and word fragment norms. This graph presents correlation between nouns, verbs and adjectives. Additionally, collective classification approaches were examined on artificially generated graph: CRN. The dataset was created according to simple sampling procedure constructing edges between nodes in accordance to the frequency of given class label in whole dataset. Namely if the the node is of a frequent class it has small degree and if the class is rare it has high degree. For CRN it was used 4 classes with highly skewed distribution. The profiles of the datasets were shortly depicted in Tab. 1.

4.3 Results

The accuracy values for various contribution of known nodes (from 10% to 90%), for both classification algorithms (ICA and LBP) were presented in Fig. 2.

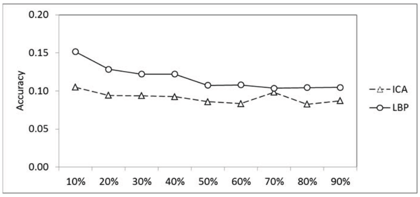
As we can see the average accuracy is at different level for various datasets. For the NET_SCIENCE dataset, it exceeds 10% only once, whereas for PAIRS_FSG, it is regularly above 70%. Overall, better results can be achieved if the problem is simpler, i.e. the greater the number of classes the worse results. It means that the quality of collective classification, like other, regular classification methods, strongly depends on the problem and sometimes it is hardly to obtain very good results.

Table 1. Basic properties of datasets utilized in experiments

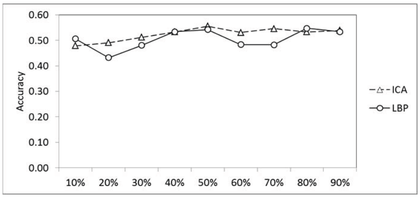
Dataset	Nodes	Edges	Classes	Avg. node degree
AMD_NETWORK	332	69092	16	208.108
CRN	327	324	4	0.990
CS_PHD	1451	924	16	0.636
NET_SCIENCE	1588	2742	26	1.726
PAIRS_FSG	4931	61449	3	12.461
YEAST	2361	2353	13	0.996

In almost all cases Iterative Classification (ICA) outperforms Loopy Belief Propagation (LBP), especially where it works worse for the sparse networks, i.e. with the small average degree value about 1 (CS_PHD, YEAST, CRN, NET_SCIENCE). However, the LBP's results are boosted for dense networks - average degree above 6 as for AMD_NETWORK and PAIRS datasets. These differences were smaller for artificial dataset CRN than for real ones. The difference in accuracy for smaller contribution of unknown nodes (e.g. 10%) and for most nodes unlabelled (90%) is not significant.

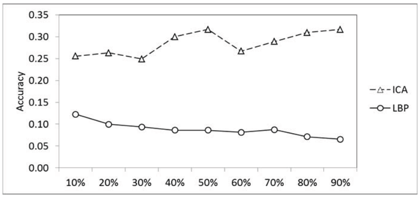
Table 2. Accuracy of collective classification performed by ICA and LBP algorithms for particular datasets



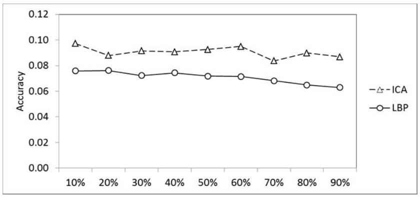
(a) AMD_NETWORK



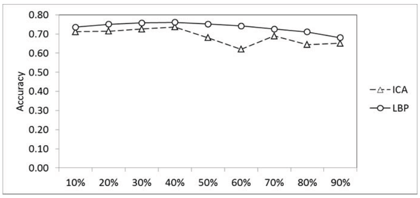
(b) CRN



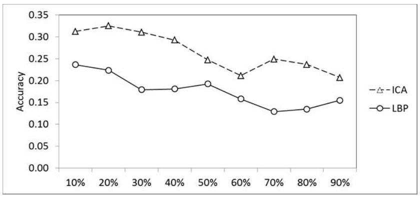
(c) CS_PHD



(d) NET_SCIENCE



(e) PAIRS_FSG



(f) YEAST

5 Conclusions and Future Works

The main goal of the paper was to present and investigate various algorithms for classification of nodes in the network (collective classification algorithms) in the context of collaborative decision making. The selected methods that were considered in the paper represent two distinct approaches to collaborative modelling of classification. Whereas the Iterative Classification (ICA) utilizes local classifiers the Loopy Belief Propagation (LBP) algorithm optimizes global objective function. This makes the latter algorithm more intuitively applicable for collaborative decision making.

The general conclusion derived from the experiments carried out on 6 datasets revealed that LBP outperforms ICA for dense networks and it is worse for sparse structures. Generally, better results can be obtained in case of smaller number of classes.

Summarizing, the usage of collective classification algorithms as well as underlying network representation of collaborative environment is a powerful modelling tool that can address collaborative decision making problems.

The future work will focus on further analysis of collaborative schemes in collective classification problems as well as on the analysis of computation efficiency of considered algorithms.

Acknowledgement. This work was partially supported by The Polish National Center of Science the research project 2011-2014 and European Commission within FP7-ICT-2009-4 under the grant agreement no. 247787.

References

[1] Golumbic, M.C.: Algorithmic Graph Theory and Perfect Graphs (Annals of Discrete Mathematics), vol. 57. North-Holland Publishing Co. (2004)

[2] Kajdanowicz, T., Kazienko, P., Janczak, M.: Collective classification techniques: An experimental study. In: Pechenizkiy, M., Wojciechowski, M. (eds.) New Trends in Databases & Inform. Sys. AISC, vol. 185, pp. 99–108. Springer, Heidelberg (2012)

[3] Desrosiers, C., Karypis, G.: Within-network classification using local structure similarity, vol. 5781, pp. 260–275. Morgan Kaufmann, San Francisco (2009)

[4] Lu, Q., Getoor, L.: Link-based classification. In: Proceedings of 20th International Conference on Machine Learning ICML, San Francisco, pp. 496–503 (2003)

[5] Perlich, C., Provost, F.: Distribution-based aggregation for relational learning with identifier attributes. Machine Learning 62(1-2), 65–105 (2006)

[6] Gallagher, B., Eliassi-Rad, T.: Leveraging label-independent features for classification in sparsely labeled networks: An empirical study. In: Giles, L., Smith, M., Yen, J., Zhang, H. (eds.) SNAKDD 2008. LNCS, vol. 5498, pp. 1–19. Springer, Heidelberg (2010)

[7] Kazienko, P., Kajdanowicz, T.: Label-dependent node classification in the network. Neurocomputing 75(1), 199–209 (2012)

[8] Knobbe, A.J., de Haas, M., Siebes, A.: Propositionalisation and aggregates. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 277–288. Springer, Heidelberg (2001)

[9] Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741 (1984)

[10] Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., Eliassi-Rad, T.: Collective classification in network data. Artificial Intelligence Magazine 29(3), 93–106 (2008)
[11] Pearl, J.: Probabilistic reasoning in intelligent systems. Morgan Kaufmann (1988)
[12] Taskar, B., Abbeel, P., Koller, D.: Discriminative probabilistic models for relational data. In: Proceedings of 18th Conference in Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco (2002)
[13] Nooy, W., Mrvar, A., Batagelj, V.: 11. In: Exploratory Social Network Analysis with Pajek. Cambridge University Press (2004)
[14] Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E 74, 036104+ (2006)
[15] Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G., Chen, R.: Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Research 31(9), 2443–2450 (2003)