Within-network classification using local structure similarity

Christian Desrosiers DESROS@CS.UMN.EDU
George Karypis KARYPIS@CS.UMN.EDU
University of Minnesota (DTC), 117 Pleasant Street SE, Minneapolis MN 55455 USA

Abstract

This paper presents a novel collective classification approach that classifies entities of a network using their local structure. Moreover, a new node similarity measure based on random walks, which takes into account label uncertainty and the degree of nodes, is introduced. Through experimentation on real-life datasets from different domains, we show our method to outperform several state-of-the-art approaches for this problem.

Within-network classification, where the goal is to classify the nodes of a partly labeled network, is a semi-supervised learning problem that has applications in several important domains like image processing, the classification of documents and web pages, classifying gene expression data, part-of-speech tagging, detecting malicious activities, and recommending items.

Because of the interdependence of labels in network data, approaches for this problem normally infer the class labels simultaneously, a technique known as collective classification (Macskassy & Provost, 2007). Methods based on this principle can generally be divided in the groups of exact and approximation inference methods. Exact inference methods, such as Markov Random Fields (Lafferty et al., 2001), Markov Logic Networks (Domingos & Richardson, 2004) and Relational Bayesian Networks (Taskar et al., 2001), learn the joint probability distribution of class labels. However, due to the high complexity of exact inference, approximation methods, like Loopy Belief Propagation (Yedidia et al., 2005), Relaxation Labeling (RL) (Chakrabarti et al., 1998), and Iterative Classification (IC) (Lu & Getoor, 2003; Neville & Jensen, 2005) are usually used for larger networks.

The methods proposed for within-network classification are generally based on the *homophily* hypothesis that linked or nearby nodes are likely to have the same labels. While evidence suggests this to apply to several types of networks, such as social networks (Barabasi et al., 2002), there are many other types of network, like molecules and biological networks, for which this assumption fails. Furthermore, while a few of these methods use information on nearby nodes to infer the class labels, e.g., (Lu & Getoor, 2003; Perlich & Provost, 2006), only the distribution of labels in this neighborhood is considered, not its structure.

Contributions

This paper makes two contributions to the problem of within-network classification. First, it introduces a novel collective classification framework that extends the relaxation approach described in (Macskassy & Provost, 2007). While our method also uses similarity between nodes to define the class membership probabilities, it is more general in the sense that it allows the use of complex similarity measures that are not based on a vectorial representation of the neighborhood.

Secondly, although methods based on random walks have recently been proposed for this problem (Callut et al., 2008; Zhu et al., 2003), such methods evaluate the proximity between nodes (indirectly) based on the number and length of paths connecting them, thus subscribing to the homophily assumption. Following the success of structural kernels on the problem of graph classification (Borgwardt et al., 2005; Li et al., 2007), we present a novel relational classifier that extends marginalized graph kernels (Kashima et al., 2003) by including label probabilities and node degrees in the computation of structural similarity. As we will show in the experimental section of this paper, considering the local structure of a node in the network can yield better classification results than simply considering the label distribution of neighbor nodes.

1. A novel classification approach

We model networked data as a partially labeled graph $G = (V, E, W, L_V, L_E, l)$ where V is a set of nodes, E a set of edges between the nodes of V, $W \subset V$ is the set of nodes for which the true labels are known,

Presented at *ILP-MLG-SRL*, Leuven, Belgium, 2009.

L_V and L_E are respectively the sets of node and edge labels, and l is a function that maps each node and edge to a label of the corresponding set. We write l_v the label of a node $v \in V$ and $l_{u,v}$ the label of an edge $(u, v) \in E$. Denoting U the set of unlabeled nodes of G, i.e. $U = V \setminus W$, the within-network classification problem consists in assigning to each $u \in U$ a label in L_V based on the labels of nodes in W.

Our classification approach is comprised of two elements: a collective inference algorithm based on RL, and a structure similarity measure based on random walks.

1.1. Relaxation labeling framework

As other RL methods, our approach works by iteratively updating the label probabilities of each unlabeled node until convergence. For any node $v \in V$ and any label $k \in L_V$, we denote $\pi_{v,k}$ the probability of v to have label k. If the true label of a node w is known, i.e. $w \in W$, then this value is binary: $\pi_{w,k} = \delta\left(l_w = k\right)$, where δ is the Kronecker delta. Furthermore, let $\mathrm{sim} : V^2 \to \mathbb{R}$ be a function that evaluates the similarity between two nodes, the probability of an unlabeled node $u \in U$ of having label $k \in L_V$ is computed from the other nodes as

$$\pi_{u,k} \;=\; \frac{\sum\limits_{v \in V} \pi_{v,k}^{\alpha}\, \mathrm{sim}(u, v)}{\sum\limits_{v \in V} \pi_{v,k}^{\alpha}}, \qquad (1)$$

where $\alpha \geq 0$ is a user-supplied parameter that controls how label uncertainty influences the computation of $\pi_{u,k}$. Thus, by increasing the value of α, one can give more importance to nodes for which the true label is known.

1.2. Random walk structure similarity

Our approach to evaluate the local structure similarity of two nodes is based on marginalized graph kernels (Kashima et al., 2003), which compute similarities as the probability of generating the same sequence of labels in two parallel random walks. While a more general approach, using product graphs, has been proposed to compute the structural similarity between graphs (Gaertner et al., 2003), the probabilistic framework of marginalized kernels is better suited to cope with the label uncertainties of our RL method. We should also mention that other types of kernels have been proposed to measure the similarity between nodes, such as exponential and diffusion kernels (Kondor & Lafferty, 2002), kernels using regularization operators (Smola & Kondor, 2003), and kernels based on random walks (Callut et al., 2008; Zhu et al., 2003). However, these kernels are mostly based on the physical proximity of the nodes in the graph, not their structural similarity.

Our similarity measure differs from marginalized kernels in two respect. First, it evaluates the similarity between two nodes of a same graph, instead of between two different graphs. Accordingly, the similarity between two nodes u and u' is defined as the probability of generating the same sequence with random walks starting at u and u'. Secondly, the labels of some nodes are only known as a probability. To cope with this problem, we make the label generation stochastic such that label k is generated at node v with probability $\pi_{v,k}$.

Using a constant walk termination probability γ, a node transition probability uniformly distributed over the edges leaving a node, as shown in (Desrosiers & Karypis, 2009), the probability $R_{u,u'}^{(N)}$ of generating the same sequences of at most N labels starting from nodes u and u' can be expressed recursively as

$$R_{u,u'}^{(N)} = \frac{(1-\gamma)^2}{d_u d_{u'}} \sum_{v \in N_u} \sum_{v' \in N_{u'}} \sum_{k \in L_V} \delta\left(l_{u,v} = l_{u',v'}\right) \pi_{v,k}\pi_{v',k}\left(\gamma^2 + R_{v,v'}^{(N-1)}\right), \qquad (2)$$

where N_u is a set containing the neighbors of a node u and $d_u = |N_u|$ is the degree of u. To compute the kernel, we use an bottom-up iterative approach, where (2) is used to compute $R^{(N)}$ based on $R^{(N-1)}$. We repeat this process for increasing values of N, until the similarity values converge, i.e. the average change is smaller than a given ϵ, or N reaches a given limit $N_{\max}$.

Exploiting node degrees

A problem with this definition is that it does not consider the difference between the degrees of two nodes u and v, while evaluating their similarity. To illustrate this, suppose we limit the walk length in (2) to $N_{\max} = 1$, i.e. we consider only the direct neighbors. Moreover, suppose that the label of every node is known, i.e. $\pi_{u,k} = \delta\left(l_u = k\right)$. Under these constraints, the similarity kernel becomes

$$\mathrm{sim}(u, v) \;=\; \frac{(1-\gamma)^2\gamma^2}{d_u d_v} \sum_{k \in L_V} n_{u,k}\, n_{v,k},$$

where $n_{u,k} \leq d_u$ denotes the number of neighbors of u that have label k. Using this formulation, the local structure similarity between the nodes u,v of Figure 1 (a)-(b) is equal to their "self-similarity": $\mathrm{sim}(u, u) = \mathrm{sim}(v, v) = \mathrm{sim}(u, v) = \frac{1}{2}(1-\gamma)^2\gamma^2$.

In order to consider the difference in the node degrees, we add temporary edges to a dummy node of label

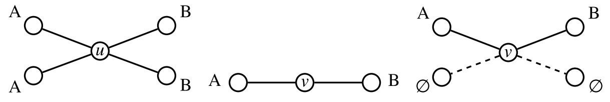

Figure 1. (a)-(b) The neighborhood of two nodes u, v and (c) the transformed neighborhood of v.

$\varnothing \notin L_V$ such that both nodes have the same degree, as shown in Figure 1(c). With the same probability as the true neighbors, the random walk can jump to this dummy node, after which the probability of generating the same sequence becomes null. Under this modification, (2) becomes:

$$R^{(N)}_{u,u'} = \frac{(1-\gamma)^2}{\max\{d_u, d_{u'}\}^2} \sum_{v \in N_u} \sum_{v' \in N'_u} \sum_{k \in L_V} \delta\left(l_{u,v} = l_{u',v'}\right) \pi_{v,k}\pi_{v',k}\left(\gamma^2 + R^{(N-1)}_{v,v'}\right). \qquad (3)$$

Using this new formulation, the similarity values for nodes u and v, again limiting the walk length to $N_{\max} = 1$, are $\mathrm{sim}(u,u) = \mathrm{sim}(v,v) = \frac{1}{2}(1-\gamma)^2\gamma^2 \geq \frac{1}{4}(1-\gamma)^2\gamma^2 = \mathrm{sim}(u,v)$.

1.3. Convergence and complexity

While the convergence of the similarity kernels defined above can be shown (Desrosiers & Karypis, 2009), the collective classification method presented in this paper, as most RL methods, is not guaranteed to converge. However, by limiting the number of allowed iterations to $T_{\max}$, we can still obtain a solution in the non-converging case. Furthermore, while the classification process can be expensive in the worst-case, i.e. $O\left(T_{\max} N_{\max} d^2_{\max} |L_V||V|^2\right)$, its complexity is closer to $O(|V|^2)$ in practice due to four reasons: 1) there are much less node labels than nodes, 2) the nodes of many real-life graphs have a low bounded degree (e.g., molecular graphs), 3) the relevant structural information of a node is contained within a short distance, and 4) the RL algorithm normally converges in a few iterations, regardless of $|V|$.

2. Experimental evaluation

2.1. Experimental setting

We tested our classification approach on five datasets. The first three datasets, which are available online at the IAM Graph Database Repository[1], were originally used for the prediction of mutagenicity, AIDS antiviral activity, and protein function. The first two model chemical compounds as undirected graphs where the nodes represent atoms, node labels are the chemical symbols of these atoms, and edges are covalent bonds between atoms. Edge labels give the valency of these bonds. The third dataset models proteins into undirected graphs using their secondary structure, such that nodes are secondary structure elements (SSE) labeled as helix, sheet, or turn. Every node is connected with an edge to its three nearest neighbors[2] in space, and edges are labeled with their structural type.

Finally, the last two datasets, which were created for the WebKB project, contain graphs modeling the links between Web pages collected from computer science departments of the Cornell and Texas Universities. These two datasets, available online[3], have often been used to benchmark within-network classification methods, as in (Macskassy & Provost, 2007). While the link information is sometimes converted into a co-citation graph, we evaluate our approach directly on the original Web page link graph. Furthermore, we consider the multi-class classification problem where pages can have one of six types: *student, faculty, staff, department, course* and *project*. Finally, while they are used in the evaluation of other methods, the edges weights representing the number of links between two Web pages, are ignored by our methods.

Table 1. Properties of the datasets.

Property	Mutagen.	AIDS	Protein	Cornell	Texas
Nb. graphs	4,337	2,000	600	1	1
Avg. nodes	30.3	15.7	32.6	351	338
Avg. edges	30.8	16.2	62.1	1392	986
Node labels	14	38	3	6	6
Edge labels	3	3	5	1	1
Freq. class	44.3%	59.3%	49.4%	41.5%	48.1%

Table 1 gives some properties of these datasets: the number of graphs, the average number of nodes and edges of these graphs, their number of node and edge labels, and the percentage of nodes having the most frequent class label.

The five datasets were used differently in our experiments. For the first three ones, which contain many small graphs, we randomly sampled six sets of 100 graphs and then merged the graphs of each of these sets into larger test graphs, considering the small graphs as individual components of the larger ones. These test graphs have 1500 to 3500 nodes, depending on the dataset. We then randomly selected one of these test graphs to tune the parameters of the tested methods and used the five others to evaluate their performance. For each of these five test graphs, 10 runs were performed, where we randomly selected a subset of nodes from which we removed the labels. We then computed the F1-score using the precision and

[1] http://www.iam.unibe.ch/fki/databases/iam-graph-database

[2] Note that a node can have more than three neighbors since the relation "nearest-neighbor" is not symmetric.

[3] http://netkit-srl.sourceforge.net/data.html

recall obtained for each class, weighted by the number of nodes in these classes, and averaged this value over the 5 × 10 classification runs. For the graphs of the last two datasets, parameters were tuned using another WebKB dataset modeling the links between Web pages of the University of Washington. As with the other datasets, 10 runs were performed on each of these two graphs and the F1-scores were averaged over these runs.

As suggested in (Macskassy & Provost, 2007), we compared our approaches using the structural similarity of (2) and (3), respectively named RL-RW and RL-RW-deg, with five classification methods implemented in NetKit-SRL[4]: IC with the Bayesian classifier of (Chakrabarti et al., 1998) (called IC-NOB), IC with the logistic regression using the [raw/normalized] number of neighbor labels (Lu & Getoor, 2003) (called IC-NOLB-[count/norm]), RL using a weighted average of the neighbor labels (Macskassy & Provost, 2003) (named RL-WVRN), and RL using similarity with reference label distribution vectors (Macskassy & Provost, 2007; Perlich & Provost, 2006) (named RL-CDRN). Note that we have tried IC *as well as* RL for all these classification approaches but only report the one giving the best results.

2.2. Results

Figures 2 give the F1-scores obtained by the seven tested methods on the five datasets, for decreasing percentages of labeled nodes. From these results, we can see that our structural similarity considering node degrees, i.e. RL-RW-deg, largely outperforms the other classification methods for datasets where the homophily assumption does not hold (i.e. molecular graphs), especially when a small portion of nodes are labeled. Moreover, our classification approach considering node degrees also works well on types of data where the local structure is not so correlated with the type of a node, such as Web page link graphs, where it is as good as the best NetKit-SRL methods.

References

Barabasi, A., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*, *311*, 590–614.

Borgwardt, K., Ong, C., Schönauer, S., Vishwanathan, S., Smola, A., & Kriegel, H.-P. (2005). Protein function prediction via graph kernels. *Bioinformatics*, *21*, 47–56.

Callut, J., Francoisse, K., Saerens, M., & Dupont, P. (2008). Semi-supervised classification from discriminative random walks. *Lecture Notes in Artificial In-*

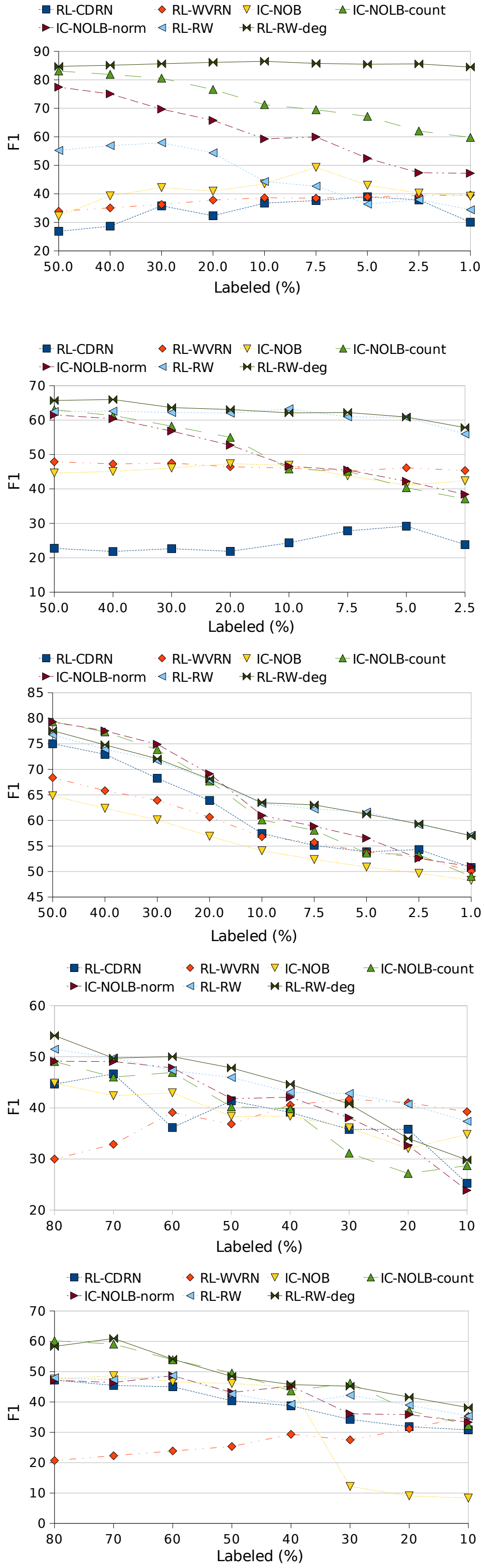

Figure 2. F1-scores obtained for (top to bottom): the Mutagenicity, AIDS, Protein, Cornell and Texas datasets.

[4] http://netkit-srl.sourceforge.net/.

telligence No. 5211, ECML PKDD 08 (pp. 162–177). Springer.

Chakrabarti, S., Dom, B., & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. *SIGMOD '98: Proc. of the 1998 ACM SIGMOD Int. Conf. on Management of data* (pp. 307–318). New York, NY, USA: ACM.

Desrosiers, C., & Karypis, G. (2009). *Within-network classification using local structure similarity* (Technical Report TR# 09-010). Dept. of Computer Science, University of Minnesota.

Domingos, P., & Richardson, M. (2004). Markov logic: A unifying framework for statistical relational learning. *Proc. of the ICML-2004 Workshop on Statistical Relational Learning and its Connections to Other Fields* (pp. 49–54).

Gaertner, T., Flach, P., & Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. *Proc. of the 16th Annual Conf. on Computational Learning Theory* (pp. 129–143). Springer-Verlag.

Kashima, H., Tsuda, K., & Inokuchi, A. (2003). Marginalized kernels between labeled graphs. *Proc. of the 12th In. Conf. on Machine Learning* (pp. 321–328). AAAI Press.

Kondor, R. I., & Lafferty, J. D. (2002). Diffusion kernels on graphs and other discrete input spaces. *ICML '02: Proc. of the 19th Int. Conf. on Machine Learning* (pp. 315–322). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML '01: Proc. of the 18th Int. Conf. on Machine Learning* (pp. 282–289). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Li, X., Zhang, Z., Chen, H., & Li, J. (2007). Graph kernel-based learning for gene function prediction from gene interaction network. *BIBM '07: Proc. of the 2007 IEEE Int. Conf. on Bioinformatics and Biomedicine* (pp. 368–373). Washington, DC, USA: IEEE Computer Society.

Lu, Q., & Getoor, L. (2003). Link-based classification. *Proc. 12th Int'l Conf. Machine Learning (ICML)* (pp. 496–503). AAAI Press.

Macskassy, S. A., & Provost, F. (2003). A simple relational classifier. *Proc. of the 2nd Workshop on Multi-Relational Data Mining (MRDM 03)* (pp. 64–76).

Macskassy, S. A., & Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, *8*, 935–983.

Neville, J., & Jensen, D. (2005). Leveraging relational autocorrelation with latent group models. *MRDM '05: Proc. of the 4th Int. workshop on Multi-relational mining* (pp. 49–55). New York, NY, USA: ACM.

Perlich, C., & Provost, F. (2006). Distribution-based aggregation for relational learning with identifier attributes. *Machine Learning*, *62*, 65–105.

Smola, A., & Kondor, R. (2003). Kernels and regularization on graphs. *Proc. of the 2003 Conf. on Computational Learning Theory (COLT) and Kernels Workshop* (pp. 144–158). M.Warmuth and B. Schölkopf.

Taskar, B., Segal, E., & Koller, D. (2001). Probabilistic classification and clustering in relational data. *In Proc. of the Seventeenth Int. Joint Conf. on Artificial Intelligence* (pp. 870–878).

Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, *51*, 2282–2312.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. *In Proc. of the 12th Int. Conf. on Machine Learning (ICML)* (pp. 912–919).