Homophily of Neighborhood in Graph Relational Classifier

Peter Vojtek and Mária Bieliková

Institute of Informatics and Software Engineering,
Faculty of Informatics and Information Technologies,
Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
{pvojtek, bielik}@fiit.stuba.sk

Abstract. Quality of collective inference relational graph classifier depends on a degree of homophily in a classified graph. If we increase homophily in the graph, the classifier would assign class-membership to the instances with reduced error rate. We propose to substitute traditionally used graph neighborhood method (based on direct neighborhood of vertex) with local graph ranking algorithm (activation spreading), which provides wider set of neighboring vertices and their weights. We demonstrate that our approach increases homophily in the graph by inferring optimal homophily distribution of a binary Simple Relational Classifier in an unweighted graph. We validate this ability also experimentally using the Social Network of the Slovak Companies dataset.

1 Introduction

Relational classifiers extend the attribute-based classifiers by adopting relations between classified instances, treating the dataset as a mathematical graph. For example, we can classify web pages according to their content only, however incorporating the content or class-membership of neighboring web pages[1] provides better results [1,2].

Methods which utilize the relations between classified instances are well suited for domains where instances have variable number of attributes (e.g., actors in a movie), attribute values are very sparsely distributed and inadequately correlate with classes, or instances have very few attributes but many relations (e.g., person in a social network identified by its nickname only but connected to many other people via friendship relation).

Univariate relational classifiers with collective inference [3,4] compose an interesting branch of classification methods where classified instances share only their class-membership between themselves via their relations (edges in a graph). The mechanism of final resolution of instance's class-membership is based on assumption of homophily – the classifier assumes that related (neighboring) instances are more likely to share similarities (e.g., the same class) as nonrelated

[1] neighboring web pages = connected via hyperlinks

instances [5]. This phenomenon is present in many graphs and mostly in social networks – people tend to group according to their race or ethnicity very strongly [6], and similarly it is with other person attributes (i.e., class-membership). Homophily is also induced in graphs where vertices are somewhat more abstract, like web pages or avatars, but generally they are created by humans and so they contain homophily tendencies.

In our work we analyze a Simple Relational Classifier [7] and discuss its dependency on homophily (Section 2). We draw the attention to the basic method of neighborhood acquisition applied in the Simple Relational Classifier and put it into contrast with a local graph ranking algorithm named spreading activation as an alternative in order to increase homophily. Next, we define how to measure homophily in a classified graph utilizing information entropy and we derive relationship between homophily and Simple Relational Classifier class assignment mechanism (Section 3). In Section 4 we provide an experimental evaluation that neighbors acquired via spreading activation outperform simple direct neighborhood in terms of homophily, using the Social Network of the Slovak Companies dataset. Section 5 contains related work and Section 6 concludes the paper and points out some issues requiring further work.

The goal of our work is to bring following contributions:

- we propose to use spreading activation as a better method to neighborhood acquisition in order to increase performance of the classifier,
- point out the close relation between classifier performance and dataset homophily,
- define how to measure the homophily in a classified graph,
- utilize homophily as a measure of classifier quality as an alternative to traditionally used supervised learning schema.

2 Neighborhood in Simple Relational Classifier

Simple Relational Classifier [7] estimates class-membership of the classified instance according to its neighborhood, exploiting a graph based data set $G = (V, E)$. If $p(c_m|v_k)$ is defined as a class-membership probability that vertex v_k belongs to class c_m then the Simple Relational Classifier assumes class-membership of v_k using Formula 1.

$$p(c_m|v_k) = \frac{1}{W} \sum_{v_j \in V_k | class(v_j) = c_m} w(v_k, v_j) \tag{1}$$

where V_k is the set of neighboring vertices of vertex v_k, $w(v_k, v_j)$ is a weight of the edge between vertices v_k and v_j, and $W = \sum_{v_j \in V_k} w(v_k, v_j)$ normalizes the result. The set of neighbors V_k contains all vertices directly connected to the classified vertex v_k via edges. If the class-membership consists of classes $c_m \in C$ (C is the set of all classes), the final class assigned to v_k is in Formula 2.

$$class(v_k) = argmax_{c_m}[p(c_m|v_k)] \tag{2}$$

Neighborhood Acquisition

In original Simple Relational Classifier as well as in other relational classifiers [2,8] neighborhood of a vertex v_k is designed as a set of vertices directly connected via edges, so that $V_k = \{v_j : v_j \in V, exists(e_{kj})\}$, where $exists(e_{kj})$ denotes an event that the graph contains an edge between vertices v_k and v_j.

Our hypothesis is that the neighborhood method should be more robust in order to absorb broader set of vertices along with weights indicating degree of vertex proximity. Due to this reason, we propose to adopt activation spreading algorithm [9,10], which is a local graph ranking method with following pseudocode[2]:

```
activate(energy E, vertex v_k) {
   energy(v_k) = energy(v_k) + E
   E' = E / |V_k|
   if (E' > T) {
      for each vertex v_j ∈ V_k {
         activate(E', v_j)
      }
   }
}
```

Activate is a recursive algorithm, its output is a set of vertices along with their weights (energy), indicating degree of affinity between v_k and ranked vertices. Minimum energy threshold T provides quick convergence of algorithm and $|V_k|$ is a number of neighboring vertices. Spreading activation assigns energy values to the vertices, not to the edges – in order to be consistent with (1) we establish $w(v_k, v_j)$ in Formula 3.

$$w(v_k, v_j) = \frac{energy(v_k)}{energy(v_j)} \tag{3}$$

Fig. 1(a) depicts example of a graph with unweighted edges. If we would classify v_1, then $V_1 = \{v_2, v_3, v_4\}$. However, if we adopt spreading activation (Fig. 1(b)), we get $V_1 = \{v_2, v_3, v_4, v_5, v_6\}$ along with weights indicating affinity between vertices, so that $w(v_1, v_2) = \frac{energy(v_1)}{energy(v_2)}$, $w(v_1, v_3) = \frac{energy(v_1)}{energy(v_3)}$, etc.

Discussion on alternative graph ranking methods is in Section 5.

3 Measuring Homophily

Evaluating the difference between original and proposed neighborhood acquisition method in terms of classifier error rate draws our attention to an observation that quality of Simple Relational Classifier class assignment depends on degree of

[2] In order to maintain simplicity and to be coherent with graph used in experimental evaluation, following algorithm is designed for unweighted graph, the original one can deal with weighted graphs.

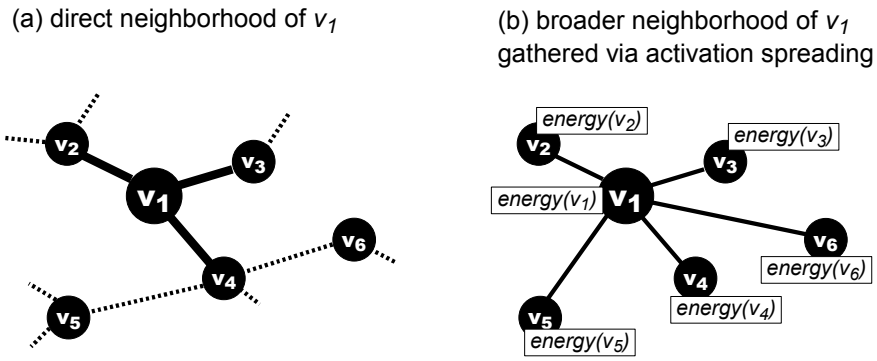

Fig. 1. Two methods of neighborhood acquisition.

homophily in the classified graph. First we introduce homophily and its measure and then point out its relation to the Simple Relational Classifier.

Assumption of homophily is informally defined as following [6]:

Related instances are more likely to share same class as nonrelated instances.

We can rewrite this sentence in terms of probability theory in following way:

$$\begin{aligned} &p(exists(e_{kj})|class(v_k) = class(v_j)) > \\ &> p(exists(e_{kj})|class(v_k) \neq class(v_j)) \end{aligned} \tag{4}$$

where $class(v_k) \in C$ is class-membership of vertex v_k. We define the degree of homophily of vertex v_k as a measure based on class-membership distribution of its neighboring vertices by adopting information entropy:

$$homophily(v_k) = 1.0 + \sum_{c_m \in C} p(c_m|v_k) \log_{base} p(c_m|v_k) \tag{5}$$

This measure is designed to deal with unlimited number of classes and the homophily is in range $\langle 0, 1 \rangle$, so that 0 is the lowest homophily and 1.0 is the highest homophily.

If we consider binary classification with classes $C = \{c_+, c_-\}$, $base = 2$ and weights of all edges are set to 1.0 (i.e. unweighted graph), we gain following boundary states:

- the highest $homophily(v_k) = 1.0$ if all neighboring vertices are assigned to c_- or c_+ exclusively (Fig. 2(a));
- lowest $homophily(v_k) = 0.0$ occurs if 50% of neighboring vertices are assigned to class c_-, the rest belongs to c_+ (Fig. 2(b)).

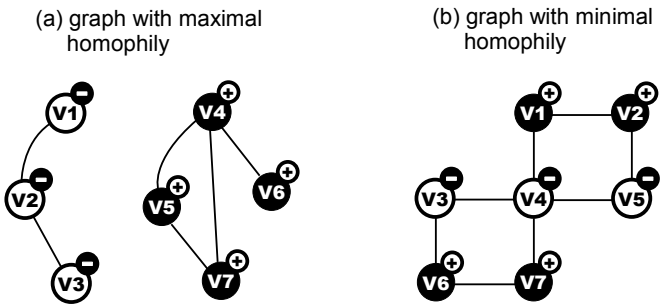

Fig. 2. Examples of homophily in a graph, each vertex in (a) has the same level of homophily (similarly (b)).

If we include substitution $W = \sum_{v_j \in V_k} w(v_k, v_j)$ into (1) we can rewrite the general Simple Relational Classifier formula as following:

$$p(c_m|v_k) = \frac{\sum_{v_j \in V_k | class(v_j)=c_m} w(v_k, v_j)}{\sum_{v_j \in V_k} w(v_k, v_j)} = \frac{W_{k_{c_m}}}{W_k} \tag{6}$$

It is obvious that $W_k = \sum_{c_m \in C} W_{k_{c_m}}$.

Because our experiments are based on binary classification, with set of classes $C = \{c_+, c_-\}$, we get $W_k = W_{k_{c_+}} + W_{k_{c_-}}$. If we consider this adjustment within (2), in order to determine impact of various neighborhood acquisition methods we only need to observe the ratio $W_{k_{c_+}} : W_k$. If $\frac{W_{k_{c_+}}}{W_k} > 0.5$, classified vertex v_k is assigned to positive class, if $\frac{W_{k_{c_+}}}{W_k} < 0.5$ then $class(v_k) = c_-$, otherwise $class(v_k)$ is left unassigned.

4 Experimental Evaluation

If we return to our hypothesis presented Section 2, our goal is to compare basic direct neighborhood with neighbors acquired with spreading activation and determine how these two approaches influence homophily in a graph (which in turn influences classifier performance). With this knowledge we will be able to distinguish which neighborhood method should be included into Simple Relational Classifier with the aim to decrease its misclassification rate.

We employ dataset based on social network of Slovak Companies register (`http://foaf.sk/`) [11]. A bipartite graph consist of two vertex types, *Company* and *Person* and a relation between them (*is_in*), which indicates that person P

plays a role in company C as a shareholder, director, etc. The dataset contains 350 000 persons, 168 000 companies and 460 000 edges between them. It is a typical social network with exponential distribution of vertex degree and graph component size.

Vertices in the graph hold several attributes – name, address, basic capital, scope of business activity, etc. A vertex class-membership is then derived from one of these attributes. We use class-membership named *is_in_Bratislava* which defines that $class(v_k) = c_+$ if person or company is located in the city Bratislava (capital city of Slovakia), otherwise $class(v_k) = c_-$. The distribution of $c_+ : c_-$ is $27 : 73$.

In practice, such a classification task is useful for two reasons: derive (at least at the regional level) addresses of people and companies with unknown location and validate address of instances affected by noise of the data acquisition method[3].

In our experiment we put into contrast ratios from (6). The results are summarized in Fig. 3, x-axis represents the ratio of $\frac{W_{k_{c_+}}}{W_k}$ and y-axis is average vertex homophily, where vertices are grouped according to x-axis[4].

In Fig. 3 we compare three curves: the optimal homophily function (as defined in (5)) is put into contrast with the two observed homophily rates: basic neighborhood and spreading activation. We see that spreading activation fits optimal homophily much better than basic neighborhood. In terms of root mean square error (RMSE) we gain:

- *Company*: $RMSE_{basic\_neigh} = 0.360$ and $RMSE_{act\_spread} = 0.219$
- *Person*: $RMSE_{basic\_neigh} = 0.374$ and $RMSE_{act\_spread} = 0.222$

For a comparison a list of contingency table derived measures is in Table 1. We see that spreading activation clearly outperforms basic neighborhood in all measures except the recall of *Person* vertex type. Imbalance of recall is induced by imbalance between $c_+ : c_-$ ratio in the dataset, where c_- is assigned to 73% of vertices, but recall is computed on the subset of c_+ vertices.

Table 1. Contingency table derived measures.

	Company		*Person*	
	basic neigh.	spread. act.	basic neigh.	spread. act.
recall [%]	85.8	90.8	71.0	56.8
precision [%]	18.2	59.5	24.3	89.1
f1 [%]	74.7	86.1	77.5	79.4
accuracy [%]	30.0	71.9	36.2	69.4
RMSE	0.360	0.219	0.374	0.222

[3] `http://foaf.sk/` dataset is gathered via wrapping the Slovak Companies register `http://orsr.sk/` administrated by the Ministry of Justice of the Slovak Republic.

[4] x-axis is sampled with $step = 0.1$, e.g., when a vertex v_k has three neighbors with positive class and one neighbor with negative class, $\frac{W_{k_{c_+}}}{W_k} = \frac{3}{4}$

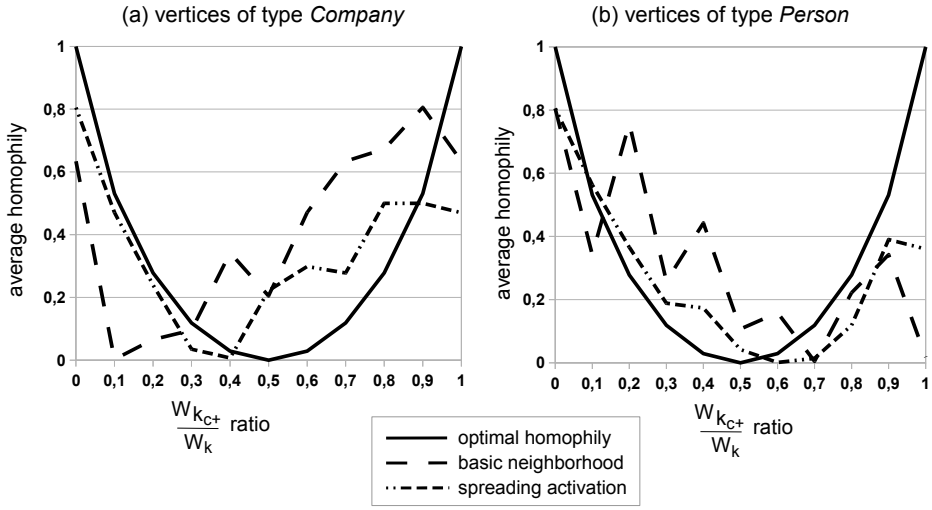

Fig. 3. Homophily comparison for basic neighborhood and spreading activation.

There are more reasons why spreading activation outperforms basic neighborhood method and provides smoother and more robust homophily lapse. Consider a graph in Fig. 4. If we use basic neighborhood method, vertex v_1 is surrounded by vertices $V_1 = \{v_2, v_3\}$. This constellation implies very disadvantageous homophily; $class(v_2) = c_-$ and $class(v_3) = c_+$, so that $\frac{W_{1_{c+}}}{W_1} = 0.5$ and $homophily(v_1) = 0.0$. However, if we consider neighborhood computed with spreading activation (starting with energy $E = 1.0$ and threshold $T = 0.15$), we get neighbors with weights as depicted in Fig. 4. If we compute homophily for this kind of neighborhood we get $\frac{W_{1_{c+}}}{W_1} = 0.625$ and $homophily(v_1) = 0.045$.

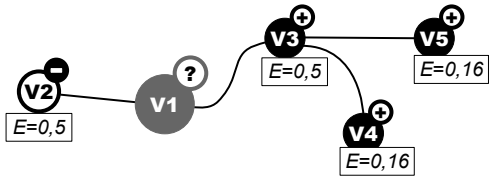

Fig. 4. Example of a graph with varying homophily according to the neighborhood acquisition method.

The spreading activation energy was set to 300.0 in the experiment and threshold $T = 1.0$ so that the neighborhood usually contains between 10 and 100 vertices and the ranking converged very quickly. Increasing the energy or

decreasing the threshold would provide us broader neighborhood of a vertex, however the computational time will heighten. Decreasing the activation energy should not be beneficial as then the energy would spread to direct neighbors only (the flow will be then stopped by the threshold limit), providing the same information about vertex' neighborhood as the basic method.

5 Related Work

Relational classifiers (also called 'collective') are new and a developing branch of predictive methods. Overview and *classification* of relational classifiers is available in [3,4]. More complex alternatives to Simple Relational Classifier are Iterative Reinforcement Categorization Algorithm [2] and Relational Ensemble Classifier [8], both capable to deal with more types of classified instances as well as handle more than one relation in a graph.

Graph ranking algorithms as spreading activation are well analyzed, mainly due to popularity of global ranking algorithms as PageRank and HITS in web search, an overview of these methods is in [10]. Spreading activation is a local ranking method similar to Random Walks with Restart [12]. We decided to employ spreading activation due to its simple understandability and effective runtime execution – we use the same method in real time on `http://foaf.sk` portal when searching for related people and companies, serving more than 500 000 page views per month.

There exist few proposals of alternative neighborhood acquisition methods to direct vertex neighborhood composed of directly connected vertices. Gallagher et al. [12] employs Random Walks with Restart method in order to improve classifier performance in graphs with weakly connected nodes, however without deeper homophily phenomenon analysis. An overview work by Jensen et al. [4] contains a neighborhood method concerning distance of neighboring objects but its impact on classifier performance is not provided.

Homophily in the task of classification is referenced in several works [5,12], using synonyms as 'auto-correlation' or 'local consistency'. A discussion of homophily measurement methods is in [13], however the degree of homophily is set-based (a homophily of chosen attribute in a set of vertices), while we are focused on homophily from a single vertex' point of view.

According to our contribution in previous sections we can refer to homophily as a quality metric of a relational classification. Classical measures as accuracy, recall, precision or F1 can be only derived from true and false positives/negatives from the contingency table, which subsequently requires the data set to be divided into a training and testing set, usually using some cross validation method [14]. Quality of relational classifiers evaluated via these contingency table measures is a subject of bias induced by relations between vertices in the training and the testing set [13,15]. On the other side, homophily explicitly requires these relations, being capable for relational classifier only (excluding attribute-based methods).

6 Conclusion and Further Work

We analyzed quality of class assignment in a relational classifier and its correlation with homophily in the classified data set represented as a graph. We proposed to adopt spreading activation as an alternative to traditionally used direct neighborhood in the classification of graph vertices using Simple Relational Classifier. We demonstrated that to determine the positive impact of spreading activation on the misclassification rate it is sustainable to simply observe the homophily induced by this neighborhood method rather than set up an experiment with training and test set and calculate contingency table metrics, which acquits us from the bias induces by relational component in the dataset.

In further work we derive the relation between homophily and classifier quality of other relational classifiers, mainly Iterative Reinforcement Categorization Algorithm [2] and Relational Ensemble Classifier [8]. It is an interesting notice that Simple Relational Classifier is in fact a kind of Iterative Reinforcement Categorization method under specific conditions.

Acknowledgments. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09 and Slovak Research and Development Agency under the contract No. APVV-0391-06.

References

1. Getoor, L., Segal, E., Taskar, B., Koller, D.: Probabilistic models of text and link structure for hypertext classification (2001)
2. Xue, G., Yu, Y., Shen, D., Yang, Q., Zeng, H., Chen, Z.: Reinforcing web-object categorization through interrelationships. Data Min. Knowl. Discov. **12**(2-3) (2006) 229–248
3. Macskassy, S.A., Provost, F.: Classification in networked data: A toolkit and a univariate case study. J. Mach. Learn. Res. **8** (2007) 935–983
4. Jensen, D., Neville, J., Gallagher, B.: Why collective inference improves relational classification. In: KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM Press (2004) 593–598
5. Jackson, M.O.: Average distance, diameter, and clustering in social networks with homophily. In: WINE '08: Proceedings of the 4th International Workshop on Internet and Network Economics, Berlin, Heidelberg, Springer-Verlag (2008) 4–11
6. Mcpherson, M., Lovin, L.S., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology **27**(1) (2001) 415–444
7. Macskassy, S., Provost, F.: A simple relational classifier. In: Workshop Multi-Relational Data Mining in conjunction with KDD-2003, ACM Press (2003)
8. Preisach, C., Schmidt-Thieme, L.: Relational ensemble classification. In: ICDM '06: Proceedings of the Sixth International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society (2006) 499–509
9. Ceglowski, M., Coburn, A., Cuadrado, J.: Semantic search of unstructured data using contextual network graphs (2003)

10. Suchal, J.: On finding power method in spreading activation search. In Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., Bieliková, M., eds.: SOFSEM, Safarik University, Kosice, Slovakia (2008) 124–130
11. Suchal, J., Vojtek, P.: Navigácia v sociálnej sieti obchodného registra SR. In: DATAKON, Srní, Czech Republic, (in Slovak). (2009)
12. Gallagher, B., Tong, H., Eliassi-Rad, T., Faloutsos, C.: Using ghost edges for classification in sparsely labeled networks. In: KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM (2008) 256–264
13. Jensen, D., Neville, J.: Linkage and autocorrelation cause feature selection bias in relational learning. In: ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2002) 259–266
14. Liu, B. In: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer (2006) 55–115
15. Korner, C., Wrobel, S.: Bias-free hypothesis evaluation in multirelational domains. In: MRDM '05: Proceedings of the 4th international workshop on Multi-relational mining, New York, NY, USA, ACM Press (2005) 33–38