Contribution to Relational Classification with Homophily Assumption

Peter Vojtek*
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
pvojtek@fiit.stuba.sk

Abstract

Relational classification is a set of methods employing relations between instances in a dataset as well as their attributes. Homophily is a phenomenon present in graphs which capture real-world data, e.g., social connections between humans. Homophily is defined as following: related (neighbouring) vertices are more likely to share similarities (e.g., the same class, attribute value) as non-related instances. Contemporary relational classifiers implicitly require homophily to be present in a graph (so called homophily assumption), however these methods are unable to determine the homophily of each node and take benefit of this information. Our work is at first dedicated to classification of relational classifiers. Next, impact of homophily assumption on particular branches of relational classifiers is analyses and then homophily measures are defined. According to this analysis, two new relational classifiers are designed. First method belongs to simple relational methods and employs local graph ranking in order to redefine neighbourhood function, second method is collective inference based and involves information exchange moderation. Both methods are capable to increase the quality of class assignment in networked data due to their capability to employ and measure homophily in a graph.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; G.2.2 [**Graph Theory**]: Graph algorithms

Keywords

classification, relational data mining, homophily, graph ranking, relational datasets, inferencing

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 could classify web pages according to their content exclusively, however incorporating the content or class-membership of neighbouring web pages[1] provides better results [6],[19].

Methods which utilise the relations between classified instances are well suited in domains where instances have variable number of attributes (e.g., actors of 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 only by its nickname but connected to many other people via friendship relation).

This work is dedicated to relational classification methods and their interconnection with so called *homophily assumption*, which is the core principle of correct classification in relational data. Homophily originated as a sociological phenomenon and is present mostly in graphs which structure[2] is induced by human activity. Homophily is defined as following – contact between similar[3] people occurs at a higher rate than among dissimilar people. This phenomenon is typically observed in relations between humans, but is indirectly transferred into graphs, where humans do not explicitly act as vertices. For example, web pages can be classified into two classes – those which concern sport and those which do not (binary classification). Here, relational classifiers assume that hyperlinks from web pages which are dedicated to sport lead to web pages with this same class with higher probability than to other web pages.

2. Goals

The ability of relational classifiers to correctly assign a class is based exactly on the assumption of homophily.

*Recommended by thesis supervisor:
Prof. Mária Bieliková
Defended at Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava on May 6, 2010.

© Copyright 2010. All rights reserved. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from STU Press, Vazovova 5, 811 07 Bratislava, Slovakia.

Vojtek, P. Contribution to Relational Classification with Homophily Assumption. Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 2, No. 1 (2010) 26-33

[1] Neighbouring web pages – connected via hyperlinks.

[2] Graph structure – the way how vertices are interconnected via edges.

[3] Similarity in context of classification is the class-membership of an instance.

If the instances were interconnected with distribution independent of their class-membership, relational classifier is unable to classify and achieves same results as random generator. Our work is dedicated to the implications which are brought by such a *naive* assumption of homophily in a classifier design.

Current relational classifiers cannot effectively adapt on varying homophily in a graph. Homophily is varying and can be different for each vertex in a graph due to its dependency on own class-membership of a vertex as well as its dependency on vertex neighbourhood. Two effects are induced by this unstable nature of homophily, both of them can increase misclassification rate:

- vertices share between themselves class-membership information disregarding the quality of this shared information,
- the approach of acquiring neighbouring nodes is not flexible enough and cannot capture local structure of a graph.

If we could consciously measure homophily in a classified graph, this would prevent us against the interchange of worthless and confusing information between the vertices during classification. Moreover, we will be able to gain such a neighbourhood of a vertex, which would bring enough homophilic vertices, despite varying local structure of a graph. In both cases, we will be able to decrease the misclassification rate, because robustness of a relational classifier will increase.

The main contribution of our work is the design of new classification methods which treat both enlisted problems: to prevent sharing confusing information between vertices and to take advantage of local structure of a graph.

Structure of this work is following: Section 3 introduces the classification of classifiers according to their capability to incorporate relational knowledge, Section 4 discusses homophily in nature and introduces how to measure this graph feature. Next, Section 5 describes a new classification method designed to moderate class–membership interchange in a graph and Section 6 is dedicated to our second major contribution – enhancing the relational classifier by local graph ranking. Finally, Section 7 concludes our work.

3. Classification of Classifiers

One of the first methods which drawn attention to relational classification was proposed in [2]. After that, many other relational methods appeared. Summarization of these methods was introduced in [8, 13]. Following three classes of classifiers refers to three different paradigms, each with a different ability of a classifier to capture relations in a dataset.

3.1 Attribute-based Model ($\hat{c}_a$)

Attribute-based, refereed as $\hat{c}_a$, depicted in Figure 1(a), estimates joint distribution of the class label and attributes of an instance. The inner box, along with the edge connecting A and C, indicates m different versions of node A (i.e. m attributes of A_i) each depend on C. The outer box indicates that the model creates N different versions of the network, each containing a single node C.

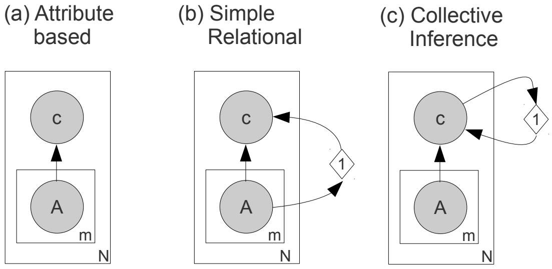

Figure 1: Three basic relational models depicted in Jensen's notation

For example, this model would indicate that the class-membership of a web page (C) depend only on the content of that page (the attributes A_i) and is independent of the topic and words on any other page. The expressiveness power of this model is equal to attribute-based classification.

This branch of classifiers contains all attribute-based methods, e.g., k-neareast neighbors [10], decision trees [9], Support Vector Machines [3], Bayesian networks [15].

3.2 Simple Relational Model ($\hat{c}_r$)

Simple relational model indicates that the attributes of an instance depends on the class label of that instance as well as the class labels of instances one link away. Figure 1(b) shows this model using a modified plate notation in which the integer within the diamond shaped annotation ("1") indicates the graph distance of neighbouring instances. The path of the edge outside the outer box emphasises the dependence on the class labels of adjoining instances. Here, the value of each A_i directly influences class-membership of neighbouring instances.

For example, this model would indicate that the class-membership a web page depends on attributes of that page and attributes of related web pages.

Simple relational model is included in more relational classifiers, e.g., Simple Relational Classifier [12], Iterative Reinforcement Categorization [19], Relational Ensemble Classification [16], Class-distribution Relational Neighbour Classifier and Network-only Bayes Classifier [13]. these methods can be used in conjunction with attribute-based methods, so that $\hat{c}_r \leftarrow \hat{c}_a$.

If we analyse for example Simple Relational Classifier, [12], this method estimates class-membership of a classified instance according to its neighbourhood, 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 (1).

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

where V_k is the set of neighboring vertices of vertex v_k, $w(v_k, v_j)$ is weight of the edge between vertices v_j and v_k, and $W = \sum_{v_j \in V_k} w(v_k, v_j)$ normalises the results.

The set of neighbours 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 si the set of all classes), the final class assigned to v_k is $class(v_k) = argmax_{c_m}[p(c_m|v_k)]$. It is obvious that the *winning* class is determined thanks to the homophily assumption, because the method *assumes* that neighbouring vertices are more likely to share the same class as non-related instances. All other enlisted simple relational methods include the same homophily assumption.

3.3 Collective Inference Model ($\hat{c}_{ci}$)

The model in Figure 1(c) adds dependence between the class label of an instance and the class label of neighbouring instances. Collective inferencing is present in following methods: Iterative Classification [11], Relaxation Labelling [2], Gibbs Sampling [5], Iterative Reinforcement Categorization [19], Relational Ensemble Classification [16].

In contrast to $\hat{c}_r$, $\hat{c}_{ci}$ model itself cannot take benefits from the attributes of related instances. However, usual approach how $\hat{c}_{ci}$ is applied is that it incorporates $\hat{c}_r$ model: $\hat{c}_{ci} \leftarrow \hat{c}_r$. This means that $\hat{c}_{ci}$ wraps the iterative character of the classification process and $\hat{c}_r$ deals with the class-membership exchange in single iteration, which implies that homophily assumption is weaved into $\hat{c}_{ci}$ as well.

4. Measuring the Homophily

Mathematical graphs are a useful metaphor to capture the networked substance of our world. If we consider graphs which capture social links among a set of individuals, we discover a phenomenon called *homophily*. In sociology, homophily is determined as following [14]:

> Contact between similar people occurs at a higher rate than among dissimilar people.

Our society exhibits homophily in many types of relations between individuals, e.g., religion, education, gender, ethnicity, race, age.

The sociological definition can be translated in context of relational classification in a following way:

Definition of Homophily. Each vertex in a graph $G(V, E)$ has assigned a distribution of class-membership $p(c \in C|v \in V)$, where C is the set of all classes. We state that similarity between distributions $p(c \in C|v_i)$ and $p(c \in C|v_j)$ is equal to the probability that there exists an edge $v_iv_j \in E$ between vertices $v_i, v_j \in V$.

If we consider for example binary classification (classes $C = \{c_+, c_-\}$) and following three vertices and their class-membership:

- $p(c_+|v_1) = 1.0$, $p(c_-|v_1) = 0.0$,
- $p(c_+|v_2) = 1.0$, $p(c_-|v_2) = 0.0$,
- $p(c_+|v_3) = 0.0$, $p(c_-|v_3) = 1.0$,

then the probability of edge occurring between vertices v_1 and v_2 is higher than between vertices v_1 and v_3, because v_1 and v_2 belong to the same class, in contrast to v_3. In real-world classification we often face fuzzy class-membership so that $p(c \in C|v \in V) \in \mathbb{R}$, typically normalized to $\langle 0, 1 \rangle$.

4.1 Degree of Homophily

In order to determine the degree of homophily between two vertices it is sufficient to compare their class-membership distributions. If we would bound ourselves to binary classification only, it is sufficient to compute the square difference between these class-membership distributions of v_i and v_j:

$$\begin{aligned} &homophily(v_i, v_j) = \\ &= \sqrt{[p(c_+|v_i) - p(c_+|v_j)]^2 + [p(c_-|v_i) - p(c_-|v_j)]^2} \end{aligned} \quad (2)$$

In reality we face more that two classes of classification, which requires (2) to be generalized:

$$homophily(v_i, v_j) = \sqrt{\sum_{c \in C} [p(c|v_i) - p(c|v_j)]^2} \quad (3)$$

Homophily of v_k is then average of values from (3):

$$homophily_n(v_k) = \frac{1}{|V_k|} \sum_{v_j \in V_k} homophily(v_k, v_j) \quad (4)$$

We call this approach – $homophily_n$ – as node oriented, as there exist joint relation of a central vertex to all other nodes in neighbourhood.

Another view on homophily is set based. Vertex v_k as well as its surrounding vertices in V_k are treated equally, the extra-ordinariness of v_k is ignored. From this point of a view we treat homophily as $homophily_s(V_k \cup \{v_k\})$ (s as set). We can aggregate distribution of class-membership for each vertex from $V_k \cup \{v_k\}$ for example using entropy (5) [17].

$$\begin{aligned} &homophily_s(V_k \cup \{v_k\}) = \\ &\quad = 1.0 + \sum_{v_i \in V_k \cup \{v_k\}, c \in C} p(c|v_i) \log_{base} p(c|v_i) \end{aligned} \quad (5)$$

Special case of $homophily_s$ is its value for a single vertex:

$$homophily_s(v_k) = 1.0 + \sum_{c \in C} p(c|v_k) \log_{base} p(c|v_k) \quad (6)$$

where the value $homophily_s(v_k)$ exhibits the degree of coherence of its class-membership distribution. The more a vertex is assigned to one class (e.g., $p(c_+|v_k) = 0.99$,

$p(c_-|v_k) = 0.01$), the more is its value of *homophily*$_s$ closer to 1.0 (if we adopt logarithm with base 2). On the other side, with decreasing class-membership of a vertex decreases the value of *homophily*$_s(v_k)$ (the worst case is $p(c_+|v_k) = p(c_-|v_k) = 0.5$).

Figure 2 displays the difference between node and set based homophily on an example.

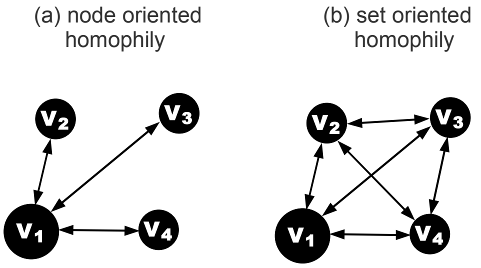

Figure 2: Two views on homophily of a vertex v_1.

5. Moderated Collective Inference-based Classifier

In current section we introduce our contribution to the first problem listed in the beginning of this work: vertices share between themselves class-membership information disregarding the quality of this shared information. Our contribution is based on moderating the information interchange of a collective-based classifier. We adopted the Iterative Reinforcement Categorization (IRC) method [19] due to its capability to assign different weight to training and testing set vertices in the iterative class-membership interchange process.

W describes the IRC classifier in a detailed way in following steps, which in general follow schema of a typical collective inference classifier ($\hat{c}_{ci}$).

Step 1: Class–membership initialization

In the pre–classification step only local features of each instance are taken into account (e.g., each publication is pre–classified according to text of the publication), this step is $defacto$ attribute-based classification ($\hat{c}_a$), however each instance is assigned a fuzzy class–membership rather than a class. The method to be used can vary (e.g., Naïve Bayes, decision trees [10]). If only one instance type is assigned a training class–membership and other instance types are subsidiary, only the leading instance type instances $x_1, x_2, \ldots, x_n \in X$ are pre–classified[4].

Step 2: Class–membership absorption

Following preconditions are arranged already: real class–membership of X_{train} instances, preclassified membership of each instance in X_{test}[5], auxiliary instances of remainder types (denoted as belonging to set Y disregarding their type) and relations between all instances.

In current step each instance from X_{test} and Y absorbs class-memberships of neighbouring instances and recomputes its own membership. Two types of neighbourhood are used, $trainNeigh(n_i)$ returns set of neighbouring instances from X_{train} of an instance n_i (n_i can be either from X or Y) and $testNeigh(n_i)$ refers to neighbours from X_{test} and Y. Usually only closest instance neighbourhood is taken into account (i.e. only instances directly connected via edges).

For each instance $n_i \in X_{test} \cup Y$ and each class $c_j \in C$ a class–membership $p(c_j|n_i)$ determines odds that n_i will be labelled with class c_j. Class-membership of each instance is recomputed using Eq. 7 where $w(n_i, n_j)$ is weight of the edge between the instances n_i and n_j. Parameters λ_1, λ_2 and λ_3 determine the relative importance of each component of class–membership, $\lambda_1 + \lambda_2 + \lambda_3 = 1$. For example, if $\lambda_1 = 0.5$, $\lambda_2 = 0.5$ and $\lambda_3 = 0$, classes of neighboring testing instances are not taken into account.

$$p(c_j|n_i) = \underbrace{\lambda_1 p(c_j|n_i)}_{self} + \underbrace{\lambda_2 \frac{\displaystyle\sum_{x_z \in trainNeigh(n_i)} w(n_i, x_z) p(c_j|x_z)}{\displaystyle\sum_{x_z \in trainNeigh(n_i)} w(n_i, x_z)} +}_{X_{train}} \underbrace{+ \lambda_3 \frac{\displaystyle\sum_{n_z \in testNeigh(n_i)} w(n_i, n_z) p(c_j|n_z)}{\displaystyle\sum_{n_z \in testNeigh(n_i)} w(n_i, n_z)}}_{X_{test} \cup Y} \quad (7)$$

Moderation of class–membership spreading

Membership computed in Eq. 7 can be harmful; an instance n_i affiliated to each class with the same probability (e.g., binary classification with $p(c_1|n_i) = 0.5$ and $p(c_2|n_i) = 0.5$) can provide meaningless information to neighbouring instances, or even worse, can affect their class–membership negatively. An eligible solution is to accept information only from instances with well–formed membership.

We employ *homophily*$_s(v_k)$ (6) here in order measure the degree of homophily of a classified vertex and according to this value, we allow the instance to participate on information exchange, or prevent (moderate) this vertex from class-membership spreading.

If we consider binary classification in some iteration t (any $\hat{c}_{ci}$ model) where $p(c_+|v_i) = (c_-|v_i) = 0.5$, it is obvious that class-membership of a vertex v_i is same for both poles of binary classification and its value *homophily*$_s(v_i) = 0.0$ is the lowest possible. According to our hypothesis we should temporarily prevent this instance from spreading its content to its neighbours and wait until its value of *homophily*$_s(v_i)$ increases.

Inverse situation occurs when an instance is assigned class label $p(c_+|v_j) = 1.0$, $(c_-|v_j) = 0.0$, which stands for the highest possible value of *homophily*$_s(v_j) = 1.0$. We want to preserve the ability of this instance to forward its class-membership information.

[4] We follow this prerequisite in consecutive sections as the dataset used in experiments meets this condition, however, in general the method can classify more leading instance types at once.

[5] Note that the real class-membership of the testing instances is also known and is stored in order to compute performance of the classifier.

These two examples are on the lower and upper bound of homophily. Most of the classified vertices are *inside* this interval and our goal is to experimentally detect the best value of $homophily_s(v_j)$, which should divide the instances to those, which should be allowed to spread their class-membership and to those, which would be temporarily disallowed. We refer to this level as to *the degree of moderation*.

Step 3: Cycles of iteration and final assignment

Remaining steps follow the original IRC method. Class–membership adjustment is an iterative process, probabilities $p_t(c_j|n_i)$ gathered in iteration t are utilized to compute class–membership in iteration $t+1$. If Q_t is membership probability matrix between all instances $n \in X_{test} \cup Y$ and all classes $c_i \in C$ in iteration t, the absorption and spreading of information ends when the difference $||Q_{t+1} - Q_t||$ is smaller than some predefined δ. After the iterative spreading is terminated, final class of each instance n_i is simply taken to be $argmax_{c_j} p(c_j|n_i)$.

5.1 Experimental Evaluation and Discussion

According to the previous section, our goal is to determine the proper degree of moderation. In following experiments we use MAPEKUS dataset with instances obtained from ACM (Association for Computing Machinery) portal[6]. Three instance types are treated: leading type *Publication*, which is primary classified and two subsidiary instance types, *Author* and *Keyword*. Two inter–relation types occurs in the data: *isAuthorOf* and *hasKeyword*. Despite the relation orientation expressed by their name (*isAuthorOf:Author* → *Publication* and *hasKeyword:Publication* → *Keyword*), these relations are considered as unoriented in our experiments, thus instances both from domain and range of the relation can exploit its benefits. Weight of each relation edge is set to $w(n_i, n_j) = 1.0$. Intra–relations are not included in the data sample due to easier post–experimental analysis of each relation type influence, focusing only on relations composing the true nature of multi–relational graphs.

Leading vertex type *Publication* is assigned one or more classes according to the ACM classification[7]. However, the classifier is not designed to direct multi–label classification [7], thus the task is divided into n binary classifications, where each publication instance in the graph is assigned either positive (c_+) or negative (c_-) label according to the class.

Size of the graph used in our experiments is following: 4000 publication instances, 7600 keywords and 9700 authors, totally 21300 unique instances, with 35000 edges (relations).

Accuracy gain

In most of following experiments accuracy gain is observed and evaluated as an important indicator of classifier quality. The term *accuracy gain* expresses the contrast between accuracy of attribute-based classifier and relational classifier on the same data sample, e.g., when attribute-based classifier achieves *accuracy* = 80% and relational classifier attains *accuracy* = 90%, the accuracy gain is +10%. Despite the fact we refer to this indicator as *gain*, its value can be also negative, i.e. in some unfavourable conditions the relational classifier can be outperformed by attribute-based classifier. We identified accuracy as the most proper indicator due to its capability to capture correctness of classification both for true positive and true negative instances.

Initial conditions

In following experiments Naïve Bayes method is adopted as basal content–based classifier, preclassification is based on text of publications' abstracts present in the dataset. Vectorization of abstract text is preceded by stemming and stop–word removal. Usually 10 iterations are used and provided statistics are averaged from 200 runs.

5.2 Influence of Moderation Threshold on Accuracy Gain

Parameter of moderation established in Step 2 is introduced with the aim to boost classifier accuracy. We perform series of experiments where the degree of moderation (*mod*) is set to values between 0.5 and 1.0. *mod* = 0.5 corresponds to original non–moderated IRC classifier and class–membership spreading is without constrains. Increasing the value of moderation refers to stronger control of class–membership interchange between neighboring instances. Setting the threshold to *mod* = 1.0 implies that only instances with well–formed class–membership can spread their values, such a condition is satisfied only by instances from the training set X_{train} as only these are exclusively truly positive (i.e. $p(c_+|n_i) = 1.0$ and $p(c_-|n_i) = 0.0$) or truly negative ($p(c_+|n_i) = 0.0$ and $p(c_-|n_i) = 1.0$).

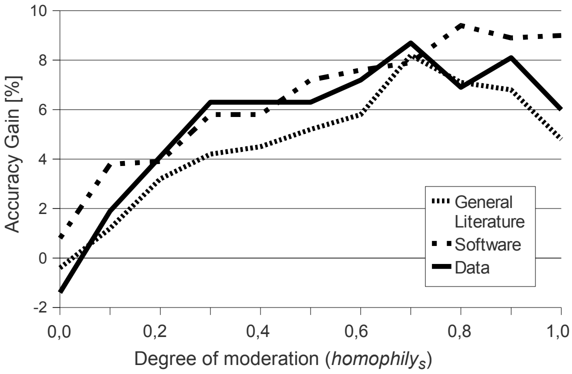

Figure 3: How degree of moderation influences the accuracy gain, different classes of ACM.

The experiment is accomplished with three different top–level classes from ACM (*General literature*, *Software* and *Data*), for each value of *mod* and each class all relations present in the dataset are involved. Parameters λ_1, λ_2 and λ_3 were set equally to $\frac{1}{3}$, denoting same weight of all components in Formula 7. Figure 3 refers to results of the experiment. X-axis displays various values of moderation threshold, y-axis indicates corresponding accuracy gain.

Both three classes exhibit similar behaviour of the classifier. The stronger the moderation is, the higher is the accuracy gain. This trend reaches maximum when *mod* is between 0.7 and 0.85. Decrease of accuracy gain in *mod* = 1.0 demonstrates importance of instances of the

[6] ACM: `http://www.acm.org/dl`
[7] ACM classification system: `http://www.acm.org/class/`

testing set to overall accuracy gain (these instances are eliminated from class-membership spreading in the strong moderated case when $mod = 1.0$).

Our experiment successfully demonstrated importance of homophily in the $\hat{c}_{ci}$ model during moderated class–membership interchange. Non-moderated classifier (corresponds to $mod = 0.0$ in Fig. 3) achieves inadequate, or even negative accuracy gain (-1.4 % for class *Data*). On the other side, moderated classifier exhibited accuracy increase $+9.4$ % (class *Software*, $mod = 0.8$), when accuracy of attribute-based classifier achieved accuracy 80.1 % and relational classifier attained 89.5 % accuracy.

6. Redefining the Neighbourhood of Relational Classifier

In current section we describe our contribution to the following problem enlisted in the beginning: the approach of acquiring neighbouring nodes is not flexible enough and cannot capture local structure of a graph.

In original Simple Relational Classifier as well as in other $\hat{c}_r$ models [19, 16] the 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 neighborhood along with weights indicating degree of proximity of vertices in neighborhood of v_k. Due to this reason, we propose adoption of activation spreading algorithm [1, 18], which is a local graph ranking method with following pseudocode[8]:

```
activate(energy $E$, vertex $v_k$) {
   energy($v_k$) = energy($v_k$) + $E$
   $E$' = $E$ / |V$_k$|
   if ($E$' > $T$) {
      for each vertex $v_j$ $\in$ $V_k$ {
         activate($E$', $v_j$)
      }
   }
}
```

Activate is a recursive algorithm which returns a set of vertices along with their weights (energy), indicating degree of affinity between v_k and ranked vertices. T is a minimum energy threshold which provides quick convergence of algorithm and $|V_k|$ is number of neighbouring vertices. Spreading activation assigns energy values to the vertices, not to the edges – in order to be consistent with (1) we establish $w(v_k, v_j) = \frac{energy(v_k)}{energy(v_j)}$.

Our goal is to compare basic direct neighbourhood with neighbours 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 neighbourhood method should be included into Simple Relational Classifier with the aim to decrease its misclassification rate.

6.1 Simple Relational Classifier and Homophily

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:

$$p(c_m|v_k) = \frac{\sum_{v_j \in V_k | class(v_j) = c_m} w(v_k, v_j)}{\sum_{v_j \in V_k} w(v_k, v_j)} = \frac{W_{k_{c_m}}}{W_k} \qquad (8)$$

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 is $C = \{c_+, c_-\}$ we get $W_k = W_{k_{c_+}} + W_{k_{c_-}}$. If we consider this adjustment within the final class assignment process, in order to determine impact of various neighbourhood 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.

6.2 Experimental Evaluation

We employ dataset based on social network of Slovak Companies register (`http://foaf.sk/`). 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 consists of 350 000 people, 168 000 companies and 460 000 edges between them and 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 city Bratislava, otherwise $class(v_k) = c_-$. The distribution of $c_+ : c_-$ is $27 : 73$.

The results are in Fig. 4, 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[9].

In Fig. 4 we compare three curves: the optimal homophily function is put into contrast with the two observed homophily rates: basic neighbourhood and spreading activation. We see that spreading activation fits optimal homophily much better than basic neighbourhood. In terms of root mean square error (RMSE) we gain:

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

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

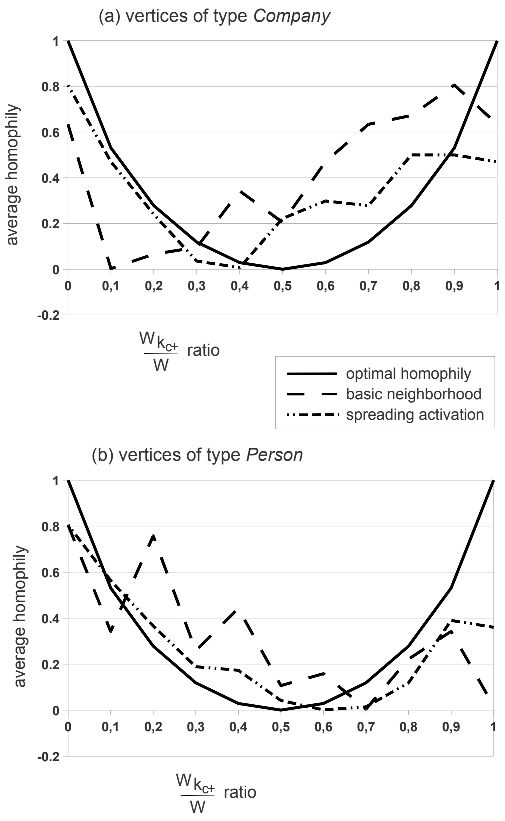

Figure 4: Homophily comparison for basic neighbourhood and spreading activation.

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

This result is supported by the contingency table (Tab. 1) for all measures except Recall of Person (this is due to imbalance between $c_+ : c_- = 27 : 73$).

Table 1: Contingency table measures (*direct* is direct neighbourhood, *spread.* is spreading activation).

	Company		Person	
	direct	spread.	direct	spread.
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

7. Conclusions and Thesis Contributions

Relational classification is currently still considered as an evolving branch of classifiers, currently without adoption in large-scale enterprise. For example, well known data mining software PASW Statistics 18[10] (former SPSS) contains several attribute-based classifiers (neural networks, decision trees, logistic regression), but no relational methods. One of the reasons is initial scepticism as well as low awareness. Next, many people are not used to view on the data as representable by mathematical graphs.

In our work we extended graph relational approach to data organization and analysis. We identified those features of graphs which are the most influential on classifier performance and we pointed out, that we should not rely on this features as given in each graph.

More concretely, the goal of this work was to consider and take into account homophily of the dataset in classifier design. Our work was based on existing overview of relational methods [8, 13]. According to our past analysis and experiments with relational methods we revealed strong connection between these methods and homophily, this remark was inspired by [4, 14].

The main contributions of our thesis are following:

- we designed a method based on information exchange moderation, which enhances the simple relational branch of classifiers ($\hat{c}_r$),
- we designed a method based on local graph ranking which extends the collective inferencing based classifiers ($\hat{c}_{ci}$).

Both methods employ the homophily measures we established and according to them, the classification model provides more robust results with lower misclassification rate. In order to evaluate our methods, we co-authored following two large-scale datasets:

- MAPEKUS, digital libraries based dataset (`http:/mapekus.fiit.stuba.sk/`),
- foaf.sk, social network of Slovak Companies (`http:/foaf.sk/`).

In broader context, our contribution is dedicated to supersede the *naive* homophily assumption in relational classifier design and substitute this inept assumption by an approach which locally considers the degree of homophily and adapts the classifier according to it.

Acknowledgements.

This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09 and VG1/3102/06, the Slovak Research and Development Agency under the contract No. APVT-20-007104, the State programme of research and development "Establishing of Information Society" under the contract No. 1025/04 and by the Cultural and Educational Grant Agency of Slovak Republic, grant No. 3/5187/07. This work is the partial result of the OPVaV - 2008/4.1/01 project ITMS 26240120005 implementation: SMART – Support of Center of Excellence for Smart Technologies, Systems and Services supported by the Research & Development Operational Programme funded by the ERDF.

[10] `http://www.spss.com/software/statistics/`

References

[1] M. Ceglowski, A. Coburn, and J. Cuadrado. Semantic search of unstructured data using contextual network graphs. Technical report, National Institute for Technology and Liberal Education, Middlebury College, Middlebury, Vermont, 05753 USA, 2003.

[2] S. Chakrabarti, B. E. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In L. M. Haas and A. Tiwary, editors, *Proceedings of SIGMOD-98, ACM International Conference on Management of Data*, pages 307–318, Seattle, US, 1998. ACM Press, New York, USA.

[3] H. Drucker, D. Wu, and V. Vapnik. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, 10(5):1048–1054, 1999.

[4] B. Gallagher, H. Tong, T. Eliassi-Rad, and C. Faloutsos. Using ghost edges for classification in sparsely labeled networks. In Y. Li, B. Liu, and S. Sarawagi, editors, *KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 256–264, New York, USA, 2008. ACM Press.

[5] S. Geman, D. Geman, K. Abend, T. J. Harley, and L. N. Kanal. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *Journal of Applied Statistics*, 20(5):25–62, 1993.

[6] L. Getoor, E. Segal, B. Taskar, and D. Koller. Probabilistic models of text and link structure for hypertext classification, 2001.

[7] N. Ghamrawi and A. McCallum. Collective multi-label classification. In *CIKM '05: Proceedings of the 14th ACM International Conference on Information And Knowledge Management*, pages 195–200, New York, USA, 2005. ACM Press.

[8] D. Jensen, J. Neville, and B. Gallagher. Why collective inference improves relational classification. In W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, editors, *KDD '04: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 593–598, New York, USA, 2004. ACM Press.

[9] C. X. Ling, Q. Yang, J. Wang, and S. Zhang. Decision trees with minimal costs. In S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, editors, *ICML '04: Proceedings of the 21st International Conference on Machine Learning*, page 69, New York, USA, 2004. ACM Press.

[10] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2006.

[11] Q. Lu and L. Getoor. Link-based classification. In T. Fawcett and N. Mishra, editors, *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pages 496–503. AAAI Press, 2003.

[12] S. Macskassy and F. Provost. A simple relational classifier. In S. Dzeroski, L. D. Raedt, and S. Wrobel, editors, *Proceedings of the 2nd Workshop on Multi-Relational Data Mining, 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 64–76, New York, USA, 2003. ACM Press.

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

[14] M. Mcpherson, L. S. Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[15] J. Pearl. *The handbook of brain theory and neural networks*, pages 149–153. MIT Press, Cambridge, MA, USA, 1998.

[16] C. Preisach and L. Schmidt-Thieme. Relational ensemble classification. In *ICDM '06: Proceedings of the 6th International Conference on Data Mining*, pages 499–509, Washington, DC, USA, 2006. IEEE Computer Society.

[17] C. E. Shannon, W. Weaver, and Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, September 1998.

[18] J. Suchal. On finding power method in spreading activation search. In V. Geffert, J. Karhumäki, A. Bertoni, B. Preneel, P. Návrat, and M. Bieliková, editors, *SOFSEM 2008: Theory and Practice of Computer Science, 34th Conference on Current Trends in Theory and Practice of Computer Science, Nový Smokovec, Slovakia, January 19-25, 2008, Student Research Forum Proceedings*, pages 124–130. Safarik University, Kosice, Slovakia, 2008.

[19] G. Xue, Y. Yu, D. Shen, Q. Yang, H. Zeng, and Z. Chen. Reinforcing web-object categorization through interrelationships. *Data Min. Knowl. Discov.*, 12(2-3):229–248, 2006.

Selected Papers by the Author

P. Vojtek, & M. Bieliková. Homophily of Neighborhood in Graph Relational Classifier. *Pages 721–730 of:* Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., & Bieliková, M. (eds), *SOFSEM 2010: 36th Conf. on Current Trends in Theory and Practice of Computer Science, Spindleruv Mlyn, Czech Republic, 2010, Proc.*. LNCS, vol. 5901. Springer.

P. Vojtek, & M. Bieliková. Moderated Class-membership Interchange in Iterative Multi-relational Graph Classifier. *Pages 229–238 of:* Snášel, V., Szczepaniak, P.S., Abraham, A., & Kacprzyk, J. (eds), *AWIC 2009: Proc. of the 6th Atlantic Web Intelligence Conf.*. AISC, vol. 67. Springer.

G. Frivolt, J. Suchal, R. Vesely, P. Vojtek, O. Vozár, & M. Bieliková. Creation, Population and Preprocessing of Experimental Data Sets for Evaluation of Applications for the Semantic Web. *Pages 684–695 of:* Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., & Bieliková, M. (eds), *SOFSEM 2008: 34th Conf. on Current Trends in Theory and Practice of Computer Science, Nový Smokovec, Slovakia, 2008, Proc.*. LNCS, vol. 4910. Springer.

M. Barla, M. Kompan, J. Suchal, P. Vojtek, D. Zeleník, & M. Bieliková. Recommendation of News (in Slovak). *In: Znalosti 2010*. Fakulta managementu Vysoké školy ekonomické, Jindřichuv Hradec.

P. Vojtek, & M. Bieliková. Local Graph Ranking in Social Network of Slovak Companies (in Slovak). *In: Proc. of 4th Workshop on Intelligent and Knowledge Oriented Technologies (WIKT 2009).*

J. Suchal, & P. Vojtek. Navigation is Social Network of Slovak Companies Register (in Slovak). *Pages 145–151 of: DATAKON 2009, Proc. of the Annual Database Conf. Srní, Czech Republic.*

P. Vojtek, & M. Bieliková. Increasing the Robustness of Relational Classifier in Datasets with Low Homophily (in Slovak). *In:* Návrat, P., & Vranić, V. (eds), *Proc. of 3rd Workshop on Intelligent and Knowledge Oriented Technologies (WIKT 2008).*

E. Gatial, Z. Balogh, L. Hluchý, & P. Vojtek. Identification and Acquisition of Domain dependent Internet Resources (in Slovak). In: Návrat, P., Bartoš, P., Bieliková, M., Hluchý, L., and Vojtáš, P., (eds),*Tools for Acguisition, Organisation and Presenting of Information and Knowledge: Research Project Workshop*, Poľana, 2007.

J. Suchal, P. Vojtek, & G. Frivolt. Interactive Navigation in Large Graphs based on Clustering (in Slovak). In: Návrat, P., Bartoš, P., Bieliková, M., Hluchý, L., and Vojtáš, P., (eds), *Tools for Acguisition, Organisation and Presenting of Information and Knowledge: Research Project Workshop*, Poľana, 2007.

M. Laclavík, M. Ciglan, M. Šeleng, S. Krajčí, P. Vojtek, & L. Hluchý. Semi-automatic Semantic Annotation of Slovak Texts. *Pages 126–137 of: Proc. of 4th International Seminar on NLP, Computational Lexicography and Terminology (SLOVKO 2007).*

P. Vojtek, & M. Bieliková. Comparing Natural Language Identification Methods based on Markov Processes. *Pages 271–281 of: Proceedings of 4th International Seminar on NLP, Computational Lexicography and Terminology (SLOVKO 2007).*

P. Vojtek. How Graph Generated from User Logs Extends Collective Classifier. *Pages 224–231 of:* Bieliková, M. (ed.), *IIT.SRC 2009: Student Research Conf.* Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.

P. Vojtek. Moderate Iterative Multi-Relational Classification. *Pages 269–276 of:* Bieliková, M. (ed), *IIT.SRC 2008: Student Research Conf.* Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava.