Chapter 1
Moderated Class–membership Interchange in Iterative Multi–relational Graph Classifier

Peter Vojtek and Mária Bieliková

Abstract Organizing information resources into classes helps significantly in searching in massive volumes of on line documents available through the Web or other information sources such as electronic mail, digital libraries, corporate databases. Existing classification methods are often based only on own content of document, i.e. its attributes. Considering relations in the web document space brings better results. We adopt multi–relational classification that interconnects attribute–based classifiers with iterative optimization based on relational heterogeneous graph structures, while different types of instances and various relation types can be classified together. We establish moderated class–membership spreading mechanism in multi–relational graphs and compare the impact of various levels of regulation in collective inference classifier. The experiments based on large–scale graphs originated in MAPEKUS research project data set (web portals of scientific libraries) demonstrate that moderated class–membership spreading significantly increases accuracy of the relational classifier (up to 10%) and protects instances with heterophilic neighborhood to be misclassified.

Key words: relational classification, graph, homophily

Peter Vojtek
Faculty of Informatics and Information Technologies, Slovak University of Technology
e-mail: pvojtek@fiit.stuba.sk

Mária Bieliková
Faculty of Informatics and Information Technologies, Slovak University of Technology
e-mail: bielik@fiit.stuba.sk

1.1 Introduction

Classification is an established data mining method useful in automated document grouping and specifically populating directories of web pages, e.g., Google Directory[1]. Increasing complexity and structure of data on the Web revealed limitations of the *traditional* attribute–based (content) classification based solely on own content of data objects. In search for advanced methods capable to exploit structure of interconnected data instances more intensively, single–relational classification [7, 5] originated as more efficient alternative to content classification.

Multi–relational classifiers are a successor of single–relational methods, designed to uncover and take advantage of broader dependencies present in the data. Multiplicity of the classifier is both in the nature of data instances and relations between them. Direct classification of heterogeneous web objects as search queries and web pages or classification of scientific publications and associated authors and keywords presents areas of interest where multi–relational approach takes advantage over single–relational and content–based methods. Similarly, domains with social interaction between instances (usually people) are good candidates for multi–relational classifiers, e.g., brokers fraud detection [9], tax frauds or user preferences gathering.

Our work is focused on graph–based classifier, Fig. 1.1 illustrates an example of multi–relational graph representation of data from the domain of scientific web portals. Data with similar nature are used in experimental evaluation presented in this paper, using MAPEKUS dataset[2] created within a research project on personalization of large information spaces in domain of digital libraries [1]. Three object types are covered: *Publications*, *Authors* and *Keywords*. Vertices *P1* and *P2* are instances of object type *Publication* connected with intra–relation *references*. Similarly, instances of *Authors* are intra–related through *isCollaboratorOf* relation type. Two types of inter–relations are present in the graph; *isAuthorOf* connects *Authors* and *Publications*, and *hasKeyword* associates instances of *Publications* and *Keywords*.

Typical classification task suitable for presented data network is to determine publications interested in class *Hardware*. Concerning attributes of publication exclusively can induce some results, however augmenting the classifier by neighboring publications, authors or keywords can provide the classifier more useful information, assuming homophily[3] between related instances. Additionally, it is feasible to determine authors interested in *Hardware* as well, without the necessity of supplementary classification or subgraph extraction.

Iterative Reinforcement Categorization (IRC) introduced by Xue et al. [11] is one of the first methods competent to perform classification in such multi-

[1] `http://www.google.com/dirhp`

[2] MAPEKUS dataset: `http://mapekus.fiit.stuba.sk/`

[3] Assumption of homophily – related (neighboring) instances are more likely to share similarities (e.g., same class) as non–related instances [7].

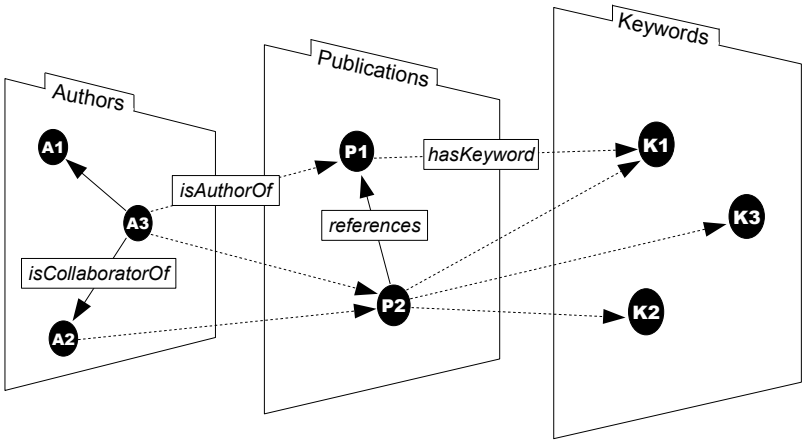

Fig. 1.1 Graph with multiple object types and various relations between them, domain of scientific publications.

relational graph structure directly, without the need for subgraph extraction, preserving and providing whole context the data is situated in to a classifier. Beside the advances of such a non–weakening approach our initial experiments exhibited following handicap: performance of the multi–relational classifier is heavily affected by structure of the graph and its accuracy gain is not always positive (Section 1.3.1 provides empirical evidence of even negative influence of non–moderated multi–relational classifier when compared to content–based classifier). To deal with this problem, the graph structure can be investigated and readjusted in some way. However, such inspection is domain specific and can be time and resource demanding (we discuss further this issue in Section 1.3.2). Better solution is to use a mechanism to automatically analyze quality of relations in the network, sustain helpful connections between the data instances and inhibit influence of non-beneficial relations, i.e. establish a technique to *moderate* information spreading in the graph.

We propose universal and effective method to moderate information exchange between classified instances based on *class–membership* of each classified instance. Multi–relational IRC method adopted in this work encapsulates domain specific knowledge into statistical parameter *class–membership*, which refers to the probability that an instance belongs to a certain class. Such a domain independent moderation of information spreading adjusts classifier performance, decreases the risk of improper object re-classification, provides a mechanism to deal with and take advance of even very weak homophily.

The rest of this paper has following structure: Section 1.2 describes the core of multi–relational classifier with class–membership spreading moderation. Experimental evaluation aimed at finding optimal parameters of the classifier, comparison of content classifier, original IRC classifier and moderated method and analysis of dataset homophily is in Section 1.3, using freely available MAPEKUS data set incorporating networks of electronic publica-

tions, authors, keywords, etc., from electronic publication web portals. Next, Section 1.4 contains overview of related work and Section 1.5 concludes the paper and points out some issues requiring further work.

1.2 Principles of Moderated IRC Method

This section describes our proposal of the moderated IRC classifier in a details. The method consist of following steps:

1. *class–membership initialization* using content–based classifier;
2. *single iteration step of class–membership optimization* exploiting the graph structure of interconnected classified instances. Moderation of information interchange is applied in this step; class-membership of each vertex is inspected and only when the class-membership is evaluated as beneficial to the overall classifier performance, the vertex can provide class-membership information to its neighbors.
3. *sequential iteration steps* converging into fixed graph state succeeded by final assignment of classes to instances.

In the pre–classification step only local features of each instance (object in the graph) are taken into account (e.g., each publication is pre–classified according to text of the publication), this step is $defacto$ content classification where each instance is assigned a fuzzy class–membership. The method to be used can vary (e.g., Naive Bayes, decision trees [6]). If only one object type is assigned a training class–membership and other object types are subsidiary, only the leading object type instances $x_1, x_2, \ldots, x_n \in X$ are pre–classified.

Class–membership absorption

Following preconditions are arranged already: real class–membership of X_{train} instances, initial class–membership of each instance in X_{test}[4], auxiliary instances of remainder types (denoted as belonging to set Y disregarding their type) and relations between all instances. In current step each object from X_{test} and Y absorbs class–memberships of neighboring objects and recomputes its own membership. Two types of neighborhood are used; $trainNeigh(n_i)$ returns set of neighboring instances from X_{train} of an instance n_i (n_i can be either from X or Y) and $testNeigh(n_i)$ refers to neighbors from X_{test} and Y. Usually only closest instance neighborhood is taken into account (i.e. only instances directly connected via edges).

For each object $n_i \in X_{test} \cup Y$ and each class $c_j \in C$ a class–membership $p(c_j|n_i)$ determines odds that n_i will be labeled with class c_j.

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

$$p(c_j|n_i) = \underbrace{\lambda_1 p(c_j|n_i)}_{self} + \underbrace{\lambda_2 \frac{\sum_{x_z \in trainNeigh(n_i)} w(n_i, x_z) p(c_j|x_z)}{sizeOf(trainNeigh(n_i))}}_{X_{train}} + \underbrace{\lambda_3 \frac{\sum_{n_z \in testNeigh(n_i)} w(n_i, n_z) p(c_j|n_z)}{sizeOf(testNeigh(n_i))}}_{X_{test} \cup Y} \quad (1.1)$$

Moderation of class–membership spreading

Membership computed in Eq. 1.1 can be harmful; an instance n_i affiliated to each class with the same probability (e.g., binary classification with $p(c_+|n_i) = 0.5$ and $p(c_-|n_i) = 0.5$) can provide meaningless information to neighboring instances, or even worse, can affect their class–membership negatively. Our assumption is that instance provides to its neighbors most useful information when class-membership of this instance belongs with high probability to positive or negative class, i.e. $p(c_+|n_i) \to 1.0$ or $p(c_-|n_i) \to 0.0$.

An eligible solution to improve Eq. 1.1 is to accept information only from instances with well–formed membership, i.e. to compute entropy based on node's classmembership $H(n) = -\sum_{i=1}^{z} p(c_i|n) \log p(c_i|n)$ and if the value exceeds specified moderation threshold, information from the node is either accepted or ignored in information exchange.

Cycles of iteration and final assignment

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 objects $n \in X_{test} \cup Y$ 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 $argmax_{c_j} p(c_j|n_i)$.

1.3 Experimental Evaluation and Discussion

In following experiments, designed to evaluate effect of the moderation mechanism, we use MAPEKUS dataset with instances obtained from ACM (Association for Computing Machinery) portal[5] similar to the graph present in Fig. 1.1. Three instance types are treated: leading type *Publication*, which is primary classified according to ACM classification, and two subsidiary instance types, *Author* and *Keyword*.

Two inter–relation types occur in the data: *isAuthorOf* and *hasKeyword*. Weight of each relation edge is set to $w(n_i, n_j) = 1.0$. Size of the graph used in our experiments is following: 4 000 publication instances, 7 600 keywords and 9 700 authors, totally 21 300 unique instances with 35 000 edges.

Accuracy gain is observed and evaluated as an indicator of classifier quality. The term *accuracy gain* expresses contrast between accuracy of content–based classifier and multi–relational classifier on the same data sample, e.g., when content classifier achieves *accuracy*= 80% and multi–relational classifier attains *accuracy*= 90%, the accuracy gain is +10%. We adopt Naive Bayes method as basal content–based classifier, preclassification is based on text of publications' abstracts. Vectorization of abstract text is preceded by stemming and stop–word removal. Provided statistics are averaged from 200 runs.

1.3.1 Moderation Threshold and Accuracy Gain

Parameter of moderation established in Section 1.2 is introduced with the aim to boost classifier accuracy. We performed series of experiments where the moderation threshold (labeled as *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 objects 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 $p(c_-|n_i) = 1.0$).

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 1.1. Fig. 1.2 refers to results of the experiment. X–axis displays various values of moderation threshold, y-axis indicates accuracy gain.

[5] ACM: `http://www.acm.org/dl`

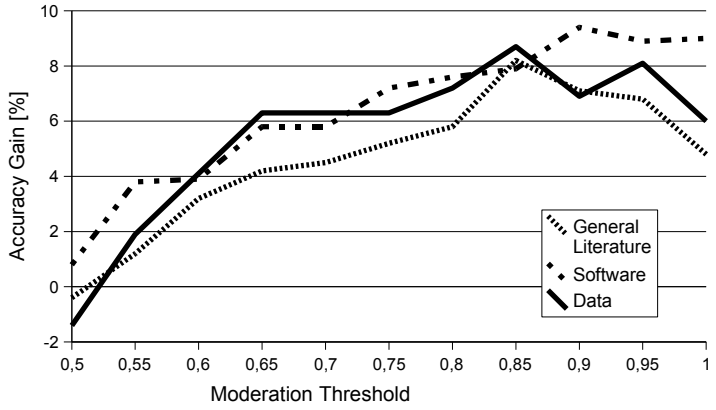

Fig. 1.2 Influence of moderation threshold on accuracy gain, different classes of ACM.

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 0.85 and 0.95. Decrease of accuracy gain in $mod = 1.0$ demonstrates importance of instances of the testing set to overall accuracy gain (these instances are eliminated from class–membership spreading in the strong moderated case when $mod = 1.0$). This experiment successfully demonstrated importance of moderation threshold. Non–moderated multi–relational classifier (corresponds to $mod = 0.5$ in Fig. 1.2) achieves inadequate, or even negative accuracy gain (-1.4% for class *Data*).

1.3.2 Analysis of Relation Quality

In previous experiment entire graph with *Publication*, *Author* and *Keyword* instances and *isAuthorOf* and *hasKeyword* relation types was employed. However, impact of these two relation types on the classifier performance can be different when considered independently, as they can exhibit different degree of homophily between instances they connect. In the current experiment we investigate quality of these relations types, comparing accuracy gain for three networks:

- graph with publications and keywords (*hasKeyword* relation type);
- graph with publications and authors (*isAuthorOf*);
- join graph with publications, authors and keywords (*isAuthorOf*+*hasKeyword*).

Initial conditions are following: moderation threshold is fixed to $mod = 0.9$ and class *Software* is considered. Pre–experimental hypothesis is that classifier performance will be highest when both relation types are included in

the graph as most information is provided to the classifier in this configuration. Fig. 1.3 displays result of the experiment. Curve of accuracy gain for $mod \in \langle 0.5, 0.9\rangle$ indicates our initial hypothesis is wrong, accuracy gain in this interval for graph with relation type *isAuthorOf* over–performs graph with both types of relation (*isAuthorOf* and *hasKeyword*). The hypothesis is satisfied only for graph with relation type *hasKeyword*, which is over–performed by richer graph with both *isAuthorOf* and *hasKeyword* relation types.

Loss of the accuracy gain for *isAuthorOf*+*hasKeyword* graph is induced by different character of the relation types, mainly neighborhood quality of associated vertices (*Author* associated with *isAuthors* and *Keyword* associated with *hasKeyword*). *Authors* are more likely to have neighboring publications holding the same class–orientation (i.e. positive or negative examples of a class).

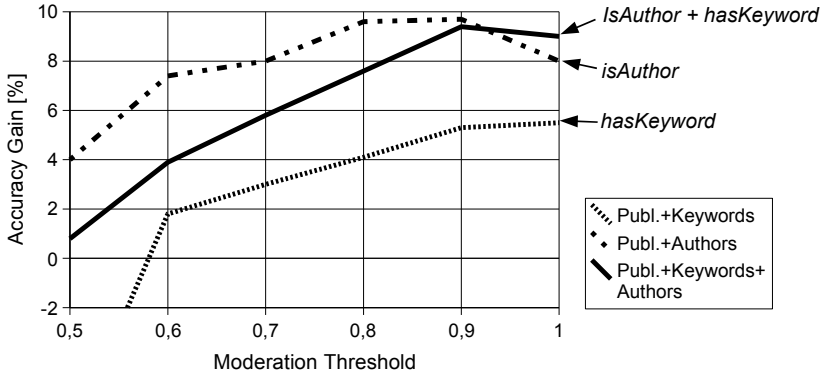

Fig. 1.3 Classifier performance influenced by relation type.

Well–formed neighborhood (when all neighboring vertices of a vertex are exclusively positive or negative) corresponds to positive–to–negative ratio 1 : 0 or 0 : 1. As much as 89% of *Author* vertices and only 42% *Keyword* vertices fall into this range. Positive–to–negative ratio 1 : 1 (heterophilic neighborhood) is present in 3% of *Author* vertices and 43% of *Keyword* vertices. These statistics exhibit significant difference between concerned relation types.

Revisiting Fig. 1.3 shows our original hypothesis[6] is correct for $mod \in (0.9, 1.0\rangle$, when accuracy gain for graph with both relation types predominates – this improvement is stimulated by positive influence of strong moderation, eliminating most of heterophilic vertices, either from *Keyword* and *Author* set.

[6] Classifier performance will be highest when both relation types are included.

1.4 Related Work

Users searching the Web commonly deal with information overload. Classifiers are frequently employed in Web search as they can automatically conceptualize and schematize concerned information, which is the base of data indexing. One of the first methods which applied single–relational classifier to organize hypertext documents connected via hyperlinks is designed by Chakrabarti et al. [3]. Classifiers designed to uncover majority of information present in multi–relational data structures appeared in past few years. Moderated iterative multi–relational classification described in this work is an extension of IRC (Iterative Reinforcement Categorization) designed by Xue et al. [11], experimental evaluation of this method was based on classification of web pages together with user sessions and their search queries. This branch of multi–relational classifiers uses graph representation of data.

Similar to IRC is Relational Ensemble Classification (REC) [10]. The main difference between IRC and REC is in the graph processing phase; IRC method iteratively spreads class–membership between intra– and inter–related objects while REC method requires construction of homogeneous subgraphs (each subgraph has single object– and single relation– type). After the iterative class–membership spreading ends the results are compiled together using ensemble classification.

Moderation of class-membership spreading is the task of determining the proper amount of disseminated information. Similar problem is in the scope of Galstyan et al. [4] where single–relational binary classifier with three–state epidemic model is utilized. In a broader sense, information dissematination in graphs is not limited to classification tasks; one of the universal information diffusion methods employed in web search is activation spreading [2].

1.5 Conclusions and Further Work

Multi–relational classification is recently established but powerful data mining technique gaining attention in hard classification problems as classification of instances with sparse or missing attributes where attribute–based classification cannot take advance. Employing multi–relational data structures and corresponding methods brings satisfactory results in such circumstances. Collective inference method called Iterative Reinforcement Categorization (IRC) is enhanced with moderated class–membership spreading mechanism in this paper in order to efficiently deal with varying homophily of related data instances.

Experimental evaluation based on data from scientific web portals validates assumption that the class–membership spreading requires moderation of the diffused amount of information, adjusting classifier accuracy up to 10%.

In addition, moderated class–membership spreading provides an efficient, robust and universal mechanism to deal with different quality of relation types.

Evaluation and comparison of different classifiers is usually performed using a dataset originated from an existing information space, e.g., WebKB[7] is a mirror of the Web. Our future work will focus on involving class–membership moderation in other relational classifiers employing collective inference mechanism [8]. We plan also to investigate influence of different shapes of the moderation function.

This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0391-06, the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 3/5187/07 and by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

1. Bieliková, M., Frivolt, G., Suchal, J., Veselý, R., Vojtek, P., Vozár, O.: Creation, population and preprocessing of experimental data sets for evaluation of applications for the semantic web. In: SOFSEM '08: Current Trends in Theory and Practice of Computer Science, pp. 684–695. Springer Verlag, Heidelberg, DE (2008)
2. Ceglowski, M., Coburn, A., Cuadrado, J.: Semantic search of unstructured data using contextual network graphs (2003)
3. Chakrabarti, S., Dom, B.E., Indyk, P.: Enhanced hypertext categorization using hyperlinks. In: L.M. Haas, A. Tiwary (eds.) Proceedings of SIGMOD-98, ACM International Conference on Management of Data, pp. 307–318. ACM Press, New York, US (1998)
4. Galstyan, A., Cohen, P.R.: Iterative relational classification through three-state epidemic dynamics. In: S. Mehrotra, D.D. Zeng, H. Chen, B.M. Thuraisingham, F.Y. Wang (eds.) ISI, *LNCS*, vol. 3975, pp. 83–92. Springer (2006)
5. 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, pp. 593–598. ACM Press (2004)
6. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, pp. 55–115. Springer (2006)
7. Macskassy, S., Provost, F.: A simple relational classifier. In: Workshop Multi-Relational Data Mining in conjunction with KDD-2003. ACM Press (2003)
8. Macskassy, S.A., Provost, F.: Classification in networked data: A toolkit and a univariate case study. J. Mach. Learn. Res. **8**, 935–983 (2007)
9. Neville, J.: Statistical models and analysis techniques for learning in relational data. Ph.D. thesis, University of Massachusetts Amherst (2006)
10. Preisach, C., Schmidt-Thieme, L.: Relational ensemble classification. In: ICDM '06: Proceedings of the Sixth International Conference on Data Mining, pp. 499–509. IEEE Computer Society, Washington, DC, USA (2006)
11. 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), 229–248 (2006)

[7] WebKB: `http://www.cs.cmu.edu/~webkb/`