International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:3, 2008

DIFFER: A Propositionalization approach for Learning from Structured Data

Thashmee Karunaratne, and Henrik Böstrom

***Abstract*—**Logic based methods for learning from structured data is limited w.r.t. handling large search spaces, preventing large-sized substructures from being considered by the resulting classifiers. A novel approach to learning from structured data is introduced that employs a structure transformation method, called finger printing, for addressing these limitations. The method, which generates features corresponding to arbitrarily complex substructures, is implemented in a system, called DIFFER. The method is demonstrated to perform comparably to an existing state-of-art method on some benchmark data sets without requiring restrictions on the search space. Furthermore, learning from the union of features generated by finger printing and the previous method outperforms learning from each individual set of features on all benchmark data sets, demonstrating the benefit of developing complementary, rather than competing, methods for structure classification.

***Keywords*—**Machine learning, Structure classification, Propositionalization.

I. Introduction

SEVERAL machine learning algorithms have been introduced to extract relevant information from structures for classification tasks. Inductive logic programming approaches [2], [3], [4] are quite popular for this purpose due to their relative simplicity and efficiency in representing data and due to the comprehensibility of resulting models. Another logic based approach to learning from structured data is to transform structured data into feature vectors by propositionalization [5], [6]. However, these, as well as the standard ILP methods, often encounter huge search spaces, for which constraints have to be imposed [7]. The limits on search depth and clause length, commonly referred to as search bias, typically result in that the substructures discovered by ILP methods are quite small and usually are limited to 5-6 structural relations [8]. This limitation prevents the discovery of large discriminatory substructures. Therefore new search methods and reasoning methods needs to be investigated, as suggested by Page and Srinivasan [1], or different data representation methods should be explored [8]. The method presented in this paper, which extracts features from structures by a method called *finger printing,* is motivated exactly by this need and follows the second line of research.

The purpose of the work reported in this paper is to develop a method for classification of structured data that successfully overcome the limitations of existing classification algorithms such as search bias.

The remaining of the paper is organized as follows. The finger printing method is introduced in the next section. In section 3, an empirical evaluation of the method on some standard benchmark datasets is presented. Finally, in section 4, we give concluding remarks and outline future work.

II. Finger Printing

Motivated by the limitations of current logic based methods for learning from structured data, our approach to structure classification employs a method of structure transformation into feature vectors by a method called *finger printing*, which does not require any constraints to be imposed on the search space. The method follows a data to model (bottom – up) search strategy and digs down any potential substructures irrespective of its length, in the following way.

Structured data is assumed to be represented by nodes (e.g., an atom in a molecule) and edges (e.g. a bond connecting two atoms). Furthermore, it is assumed that all nodes have been given labels, allowing similar nodes in different graphs to be handled in a similar way (e.g., an atom could be given the label ‘carbon’). Each example is represented by the set of all pairs *(Li,Lj)*, such that there is an edge in the graph of the example between nodes *Ni* and *Nj* that are labeled *Li* and *Lj* respectively. We refer to such a set as a *finger print*.

The finger prints are used for substructure search in the following way. For all pairs of examples, the intersection of their finger prints, which is referred to as *the maximal common substructure*, is formed, and ranked according to their frequency in the entire set of examples (i.e., the number of finger prints for which the maximal common substructure is a subset). It should be noted that no constraints are applied on the length of the substructures considered during this process. Therefore the discovered substructures are not subjected to pruning the search space beforehand in any manner.

III. Empirical Evaluation

A feature construction and classifier system called DIFFER (**DI**scovery of **F**eatures using **F**ing**ER** prints) has been developed that incorporates the finger print method. The input to DIFFER consists of examples, containing structured data. From this, DIFFER produces a set of features together with an encoding of the examples using these features, in the form of

T. Karunaratne and H. Boström are with the Department of Computer & Systems Sciences (DSV), Stockholm University and Royal Institute of Technology, Forum 100, SE 164 40, Kista, Sweden (e-mail: {si-thk, henke}@dsv.su.se).

a text file on .arff format which is the recognizable format for WEKA data mining toolkit [13].

A. *Data Sets*

We have used four benchmark datasets to compare the performance of DIFFER with other available methods for learning from structures. The first benchmark dataset concerns predicting mutagenicity on *Salmonella typhimurium* [9]. The second benchmark data set concerns the popular east-west train problem [10]. The third dataset, carcinogenesis is also, like the first, from the domain of chemo-informatics and was originally developed within the US national toxicology program [11]. The fourth dataset we selected for this study concerns predicting Satellite faults [12].

B. *Experimental Setup and Results*

Graphs were generated from the four datasets in the following way: For the first and third datasets, the graphs represent molecules and consist of nodes that correspond to atoms, which are labeled with the atom type and the properties of bonds attached to the atom. For example, a carbon atom with two single bonds, one double bond and no aromatic bonds is labeled by atom*(c, 2, 1, 0).* In the second dataset, the structures represent trains, and each train has a set of cars which are labeled by a set of properties, e.g., shape, no. of wheels etc. A node in this domain is represented by a tuple *car(<properties>).* For example, a car with a long rectangular shape, a flat roof, sides that are not double and 3 wheels, is labeled by *car(rectangle, long, not_double, flat, 3).* Structures of the last dataset represent the states of 29 sensors at a give time instant as and is on the form of a tuple *fault*(S_1, ..., S_{29}), where each S_i is a boolean feature. For each benchmark dataset, all examples were given as input to DIFFER in order to generate a feature set. The training and test examples were expressed using these features, and classification models were generated from the training examples in this experiment by random forest, as implemented in the WEKA data mining toolkit [13], with 50 trees in each model and 10 random features evaluated at each node. Due to the limited number of examples in the data sets, we used 10 fold cross validation as the evaluation method. It should be noted that the accuracy estimates will not be biased even though the feature generation method is given all examples as input, since class labels are not taken into account[1]. The results we obtained with DIFFER were compared to that of RSD [5], a state-of-the-art logic-based propositionalization method, which similarly to DIFFER, transforms structured examples into feature vectors, and for which the same learning method (random forest) and experimental setup was used. The results are summarized in Table I.

TABLE I
ACCURACIES FOR DIFFER RSD AND DIFFER + RSD

Dataset	Accuracy		
	DIFFER	RSD	DIFFER + RSD
Trains	80%	75%	85%
Mutagenesis	80.61%	88.86	92.76%
Carcinogenesis	65.25%	54.37	65.33%
Satellite faults	71.43%	71.43	80.95%

The results show that DIFFER performs comparably to RSD when used on its own (two wins, one loss and one tie) without having to impose constraints on the search space[2]. Since RSD generates features in an orthogonal way (by using background knowledge expressed in first-order logic together with constraints on the search space), these do not necessarily overlap with the features constructed by DIFFER. By inspecting the features generated by the two methods in the mutagenesis domain, it was noticed that RSD's features concern global properties of molecules, which are not directly related to the structures, while DIFFER's features express relationships between atoms. We investigated the effect of merging these complementary feature sets, and it was observed that for all four benchmark data sets, the models generated from each individual feature set were outperformed by the models generated from the merged feature set (see third column in Table I). The experiment hence demonstrates the benefit of treating these methods as complementary, rather than competing.

The obtained results also seem to compare well with those reported for state-of-the art graph based structure classification methods [14], [15]. Tree$^2\chi^2$ was reported to obtain an accuracy of 80.26% on the mutagenesis dataset [15] and SUBDUE-CL was reported to obtain an accuracy of 61.54% on carcinogenesis [14]. The results are not surprising since the graph based methods suffer from the constraint of requiring that discovered sub-graphs are connected, while models generated by DIFFER and RSD may include non-connected sub-graphs as well.

IV. CONCLUDING REMARKS

Logic based techniques for learning from structured data is limited w.r.t. searching for large substructures. In order to overcome these limitations, a novel method that transforms structured data into a form called finger prints, has been presented. The new method, which has been implemented in a system, called DIFFER, was shown to perform comparably to an existing state-of-the-art method on four standard benchmark data sets, without having to impose constraints on the search space. The reason for its effectiveness can be explained by its ability to mine large–sized substructures by searching bottom-up. A very interesting observation is that the classification performance can be improved by merging the features generated by DIFFER with features generated by

[1] Having access to unlabeled test examples during training is sometimes referred to as transductive learning.

[2] We did not achieve the same accuracy for RSD as reported in [5] for the mutagenesis dataset although the same code and files were used in reconstruction of features.

other methods and thereby integrating the different qualities of several methods.

There are a number of possible directions for future work. At present DIFFER's substructure search is a pair-wise approach, for which the computational cost grows quadratically with the number of examples. A more efficient procedure could be obtained by incrementally searching for the substructures, or by sampling of the pairs to consider (cf. [3]). Alternatives to the ranking criterion for generated features could be investigated. Candidates for this include model driven approaches such as voting by the convex hull or a coverage measure. Currently the edges in the graphs are assumed to be unlabeled, and any relationship between the objects in a graph has to be encapsulated in the node labels (e.g., properties of the bonds are used when labeling the nodes of molecules). By also considering labels of the edges (e.g. whether it is a single or double bond), more general node definitions could be obtained allowing further generalization from the examples.

The promising result of combining the features generated by DIFFER and RSD also points to considering merging the features of DIFFER with those of other methods, perhaps improving the predictive performance even further.

References

[1] Page, D. and Srinivasan, A., (2003), *ILP: A Short Look back and a Longer Look Forward*, Journal of machine learning research, 4(Aug):415-430.

[2] Quinlan, J. R., Cameron-Jones, R. M., (1993), *FOIL*, Proceedings of the 6th ECML, Lecture Notes in AI, Vol. 667, 3-20. Springer-Verlag

[3] Muggleton S.H. and Feng C. (1990). *Efficient induction of logic programs*, Proceedings of the First Conference on Algorithmic Learning Theory, Tokyo.

[4] Srinivasan A,. King, R.D, and Muggleton S, (1999), *The role of background knowledge: using a problem from chemistry to examine the performance of an ILP program*, Technical Report PRG-TR-08-99, Oxford University.

[5] Krogel, M-A., Rawles, S., Železný, F., Flach, P. A., Lavrač, N., and Wrobel, S., (2003), *Comparative evaluation of approaches to propositionalization*, Proc.of the 13th International Conference on ILP, Lecture Notes in CS, 197-214.

[6] Lavrac, N. and Flach P., (2000), "*An extended transformation approach to Inductive Logic Programming*", University publication, University of Bristol.

[7] Nattee, C., Sinthupinyo, S., Numao, M., Okada, T., (2005), *Inductive Logic Programming for Structure-Activity Relationship Studies on Large Scale Data*, SAINT Workshops 2005: 332-335.

[8] Inokuchi, A., Washio, T., and Motoda, H., (2003), *Complete mining of frequent patterns from graphs*, Mining graph data, Machine Learning, 50:321-354.

[9] Debnath, A.K. Lopez de Compadre, R.L., Debnath, G., Shusterman, A.J., and Hansch, C. (1991), *Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds: Correlation with molecular orbital energies and hydrophobicity*, Journal Med. Chem. 34:786-797.

[10] Michie,D., Muggleton,S., Page,D., and Srinivasan,A., (1994), "*To the international computing community: A new East-West challenge*" Oxford University Computing laboratory, Oxford, UK.

[11] US National Toxicology program, http://ntp.niehs.nih.gov/index.cfm?objectid=3 2BA9724-F1F6-975E-7FCE50709CB4C932

[12] Pearce D.A. (1988). *The induction of fault diagnosis systems from qualitative models*, Proceedings 7th National Conference on AI, Saint Paul, Minnesota.

[13] Ian H. Witten and Eibe Frank (2005) "*Data Mining: Practical machine learning tools and techniques*", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[14] Gonzalez, J., Holder, L. B. and Cook, D. J. (2001), "*Application of Graph-Based Concept Learning to the Predictive Toxicology Domain",* Proceedings of the Predictive Toxicology Challenge Workshop.

[15] Bringmann, B., and Zimmermann, A., (2005), "*Tree - Decision Trees for Tree Structured Data*", Proceedings of PKDD 2005, LNAI 3721, pp. 46-58, Springer.