Using Semantic Data Mining for Classification Improvement and Knowledge Extraction

Fernando Benites and Elena Sapozhnikova

University of Konstanz, 78464 Konstanz, Germany.

Abstract. The objective of this position paper is to show that the integration of semantic data mining into the DAMIART data mining system can help further improve classification performance and knowledge extraction. DAMIART performs multi-label classification in the presence of multiple class ontologies, hierarchy extraction from multi-labels and concept relation by association rule mining. Whereas DAMIART combines knowledge from multiple data sources and multiple class ontologies, the proposed extension should also explore available ontologies over attributes. This will allow the system to produce not only more accurate classification results but also improve their interpretability and overcome such problems as data sparseness.

Keywords: Semantic Data Mining, Linked Open Data, Ontology

1 Introduction

Data Mining is defined as the process of discovering implicit, novel, potentially useful and understandable patterns or relationships in large volumes of data [8]. In this context, conventional data mining algorithms treat the data simply as numbers lacking any semantic information and process them independently from the particular domain. Data preprocessing as well as interpretation of the obtained results are though domain-dependent tasks, which are usually solved by human experts possessing required domain knowledge. However, such knowledge can be very useful at any other stage of the data mining process, e.g. for choosing suitable data and proper mining techniques or for the effective pruning of the hypothesis space. So, it has been early realized that the incorporation of available domain knowledge is one of the most important problems in data mining [9]. Now, its importance is growing even more because the data are becoming more and more complex, and a manual approach to obtaining domain knowledge is not sufficiently efficient. With more interconnected data, more possible interpretations can be generated by data mining algorithms, overwhelming any human expert.

The new field of Semantic Data Mining, which has emerged in the past few years, has suggested a possible solution to this problem: Domain knowledge can be derived from semantic data (data which include semantic information, e.g. ontologies or annotated data collections) and directly incorporated in the data mining process. The term Semantic Data Mining was first introduced by [15] in order to designate a data mining approach where domain ontologies are used as background knowledge for data mining (Fig. 1). It includes methods for systematic incorporation of domain knowledge in an intelligent data mining environment [12].

Alternatively, d'Amato et al. proposed the term Ontology Mining for the same research area, reflecting the importance of the role ontologies play in knowledge representation [4]. A domain ontology can be viewed as a model that contains the structural and conceptual information about the domain. It typically consists of all important concepts of the domain, their specific properties, the relationships between the concepts, and possibly additional restrictions on the domain. A common example may be an ontology for the tourism domain containing concepts such as accommodation, attractions and transport (where, for example, "hotel" and "youth hostel" are subconcepts of "accommodation"). Due to the rapid growth of the number of ontologies becoming available on the Web in a wide range of domains, semantic data mining has great potential in many application areas such as biology, sociology, and finance.

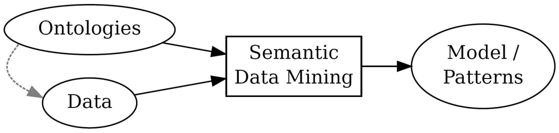

Fig. 1. Semantic Data Mining.

The main advantage of ontology-based systems is their ability of sharing knowledge among people as well as among computer systems. This was the motivation behind the development of the semantic Web [7] when companies and large institutions intended to automatically exchange data over the Internet. Nowadays, semantic Web ontologies have been established as a key technology for intelligent knowledge processing. This has resulted in a paradigm shift within the data mining community: instead of mining the large volumes of numerical data supported by scarce domain knowledge, the new challenge is to mine the abundance of knowledge encoded in domain ontologies, constrained by the heuristics computed from the data [11]. Recently developed semantic Web technologies, such as RDF (Resource Description Framework) and OWL (Ontology Web Language), enable domain knowledge to be captured automatically with minimum manual effort. The first attempts to utilize Linked Open Data (LOD) in data mining process have been shown to be successful in many application areas, including user recommendation systems [18], medical domain [13], Web search [17] and cross lingual text classification [14].

Another challenge of modern data mining is the connecting of different data sources. As a result of increasing data complexity, there is often a variety of interrelated data sources, which can all be used to describe the same problem. Classical data mining utilizing just one data set at a time is insufficient in such a case. It is especially important in biological applications as, for example, in functional genomics where a single data source can often reveal only a certain perspective of the underlying complex biological mechanism. By integrating evidence from multiple data sources, it is possible to obtain more accurate predictions of unknown gene functions. Generally, combining information from the ontologies providing different insights into a problem domain can be very helpful and can lead to the discovery of new knowledge. So, the integration of available evidence from multiple data sources may significantly decrease human efforts in creating useful knowledge representations. This was the goal of the project "DAMIART – Data Mining of heterogeneous data with an Adaptive-Resonance-Theory-based neural network" which intended to solve the latter problem by integrating available data sources and class ontologies.

This paper starts by reporting on the data mining system developed in the project DAMIART. Then we propose a natural extension to this system which should exploit LOD for performance improvement and knowledge extraction. We also discuss a set of objectives this extension should deal with. The Conclusion and Future Work section closes the paper.

2 DAMIART System

The DAMIART project combined multiple data source and multiple class ontology approaches into a single data mining system performing multi-label classification by a neuro-fuzzy classifier. It should lead to the improvement of classification performance and result interpretation because of using complementary domain knowledge extracted from different data sources. The most important tasks of the developed system are hierarchy extraction from multi-labels [3] and concept relation [1], both solved by association analysis. Concept relation implies that relations found between the classes of multiple class ontologies can assist experts in extracting new knowledge from data. For example, if a film can be classified either by its genre into a genre ontology or by the producing company in an ontology of producers, one can find a possible interesting connection between a certain genre and a producing company, specializing in this genre. This potentially useful information can be utilized in many ways, for example, to narrow the huge search space for data mining algorithms or to better interpret the results presented to the user. It was shown that the system was able to discover valuable relationships between class ontologies [2]. Additionally, fuzzy rules extracted from the trained classifier can be used for the plausibility check of discovered association rules. When experts subsequently interact with the system (see Fig. 2), it should be possible to reveal conflicts in the classification rules and to correct them. Finding relations between concepts in our system is instance-based, which means that they are determined by data only and may change accordingly when the data change.

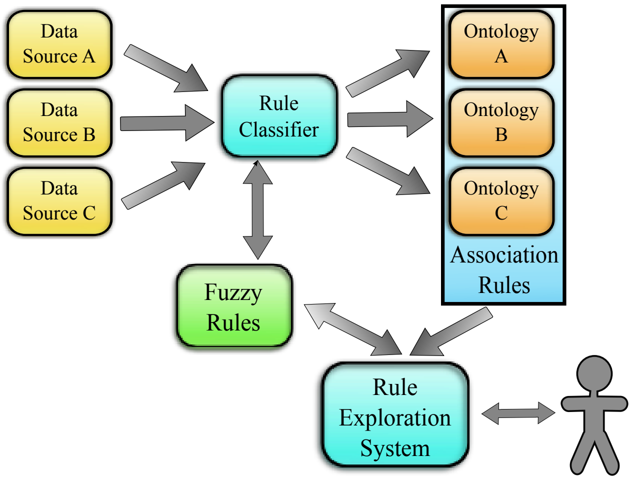

Fig. 2. DAMIART System.

3 Motivation and Objectives

In the focus of the project DAMIART were only multiple class ontologies providing additional information about relationships between the labels found in a training set. Considering a large number of ontologies available also for data attributes, such restriction to class ontologies seems to be inappropriate. To cope with the challenges discussed in Introduction, the DAMIART system can be naturally extended to utilize not only class ontologies, but also ontologies available for data attributes (see Fig. 3). In the above example of the film classification, the data contain short film descriptions, which are then transformed into feature vectors by standard text mining methods. The use of an ontology like Wordnet [20] enables interesting connections between some subsets of words and a certain genre of films, such as e.g. thriller or comedy, to be extracted. We therefore propose to extend the DAMIART system by integrating existing ontological knowledge about attributes. The analysis of state-of-the-art approaches revealed that the proposed extension can be used to solve at least the following tasks:

1. to enrich training data with additional features derived from LOD;
2. to perform feature selection effectively;
3. to further improve classification performance;
4. to enhance the interpretation of fuzzy rules extracted from the classifier;
5. to facilitate understanding of obtained classification results.

It has been already shown in different applications (e.g. [16]) that significant improvement of classification performance can be achieved by the data enrichment through large Web ontologies like DBpedia [6]. It is important to note that the methods can be diverse: one can either directly incorporate additional features in a dataset or exploit high-level knowledge in order to avoid overfitting by replacing specialized features with more general concepts, e.g. names of certain streets can be replaced with the concept "Street". Obviously, the use of additional information facilitates the feature selection [10]. It has also been shown in [19] how LOD can be successfully applied to enhancing the interpretation of data mining results.

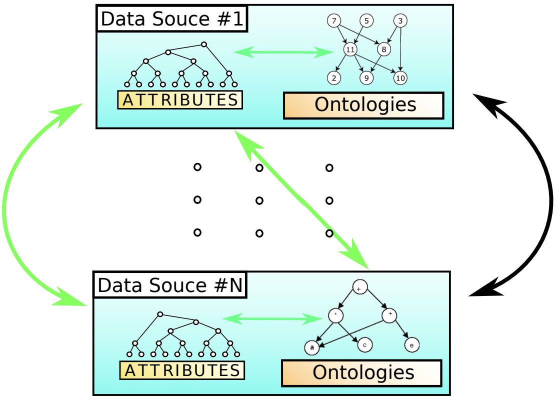

Fig. 3. Extension of the DAMIART System.

4 Conclusion and Future Work

In this paper we have proposed the semantic data mining extension to the DAMIART system. Its implementation will be the subject of future work. Among the benefits expected from the extension is the reduction of data sparseness if a dataset has only few features: Including new features from LOD helps solve this problem. However, the danger of overfitting arises if a dataset already has a lot of features. In this case it is important to select the features that are sufficiently general to represent more specific features of the training data. For this purpose ontological structures of LOD are very useful. It is also expected that the system will be able to produce more accurate results. Additionally, we expect an improvement of interpretability and understandability of the classification results due to better representation of the fuzzy rules extracted from the trained classifier. Moreover, possible new knowledge found by combining different data sources could be used to further update the ontologies, generating a feedback cycle in the data mining process similarly to [5]. The system will have many potential applications such as politics (analysis of the election results), medicine (patient-report analysis), genetics (functional gene classification), and machine translation. An additional point of the future work is evaluation of the proposed extension in one or several application fields.

References

1. Benites, F., Sapozhnikova, E.: Learning different concept hierarchies and the relations between them from classified data. In: Intel. Data Analysis for Real-Life Applications: Theory and Practice, pp. 18–34. IGI Global, Hershey (2012)
2. Benites, F., Simon, S., Sapozhnikova, E.: Mining rare associations between biological ontologies. PLOS ONE 9(1), e84475 (2014)
3. Brucker, F., Benites, F., Sapozhnikova, E.P.: Multi-label classification and extracting predicted class hierarchies. Pattern Recognition 44, 724–738 (2011)
4. d'Amato, C., Fanizzi, N., Esposito, F.: Inductive learning for the semantic web: What does it buy? Semantic Web 1, 53–59 (2010)
5. d'Aquin, M., Kronberger, G., Surez-Figueroa, M.: Combining data mining and ontology engineering to enrich ontologies and linked data. In: Workshop: Knowledge Discovery and Data Mining Meets Linked Open Data (2012)
6. DBpedia, `http://dbpedia.org`
7. Domingue, J., Fensel, D., Hendler, J.A. (eds.): Handbook of Semantic Web Technologies. Springer, Heidelberg (2011)
8. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: An overview. In: LECT NOTES ARTIF INT, pp. 1–34. LNCS (1996)
9. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine 17, 37–54 (1996)
10. Jeong, Y., Myaeng, S.H.: Feature selection using a semantic hierarchy for event recognition and type classification. In: Sixth Int. Joint Conf. on Natural Language Processing, pp. 136–144. Asian Federation of Natural Language Processing, Nagoya, Japan (October 2013)
11. Lavrač, N., Vavpetič, A., Soldatova, L., Trajkovski, I., Novak, P.K.: Using ontologies in semantic data mining with SEGS and g-SEGS. In: 14th Int. Conf. on Discovery science, pp. 165–178. DS'11, Springer, Heidelberg (2011)
12. Liu, H.: Towards semantic data mining. In: 9th Int. Semantic Web Conf. (2010)
13. Moss, L., Sleeman, D.H., Sim, M., Booth, M., Daniel, M., Donaldson, L., Gilhooly, C.J., Hughes, M., Kinsella, J.: Ontology-driven hypothesis generation to explain anomalous patient responses to treatment. Knowl.-Based Syst. 23, 309–315 (2010)
14. Ni, X., Sun, J.T., Hu, J., Chen, Z.: Cross lingual text classification by mining multilingual topics from wikipedia. In: WSDM '11 fourth ACM international conference on Web search and data mining, pp. 375–384. ACM, New York (2011)
15. Novak, P.K., Vavpetič, A., Trajkovski, I., Lavrač, N.: Towards semantic data mining with g-SEGS. In: 13th International Multiconference Information Society (IS 2010), pp. 173–176 (2010)
16. Paulheim, H.: Exploiting linked open data as background knowledge in data mining. In: Int. Workshop on Data Mining on Linked Data, with Linked Data Mining Challenge at ECMLPKDD 2013, pp. 1–10 (2013)
17. Phan, X.H., Nguyen, L.M., Horiguchi, S.: Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: 17th International Conference on World Wide Web, pp. 91–100. ACM, New York (2008)
18. Singhal, A., Kasturi, R., Sivakumar, V., Srivastava, J.: Leveraging web intelligence for finding interesting research datasets. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 1, 321–328 (2013)
19. Tiddi, I.: Explaining data patterns using background knowledge from linked data. In: ISWC-DC, pp. 56–63 (2013)
20. WordNet, `http://wordnet.princeton.edu/`