New Trends in Knowledge Driven Data Mining

Cláudia Antunes and Andreia Silva
Instituto Superior Técnico, Universidade de Lisboa, Av Rovisco Pais, Lisboa, Portugal

Keywords: Domain Knowledge, Semantic Aspects of Data Mining, Domain Driven Data Mining, Constrained Data Mining, Knowledge Representation, Domain Ontologies.

Abstract: Existing mining algorithms, from classification to pattern mining, reached considerable levels of efficiency, and their extension to deal with more demanding data, such as data streams and big data, show their incontestable quality and adequacy to the problem. Despite their efficiency, their effectiveness on identifying useful information is somehow impaired, not allowing for making use of existing domain knowledge to focus the discovery. The use of this knowledge can bring significant benefits to data mining applications, by resulting in simpler and more interesting and usable models. However, most of existing approaches are concerned with being able to mine specific domains, and therefore are not easily reusable, instead of building general algorithms that are able to incorporate domain knowledge, independently of the domain. In our opinion, this requires a drift in the focus of the research in data mining, and we argue this change should be from domain-driven to knowledge-driven data mining, aiming for a stronger emphasis on the exploration of existing domain knowledge for guiding existing algorithms.

1 INTRODUCTION

The rise of information society and its increasing maturity, allied to the more recent explosion of *big data*, made even clearer the need of efficient techniques for mining unknown information from data. In the last two decades, the field of data mining (DM) proved to be effective when applied to almost all domains and a large range of kinds of data, from structured sources, as databases, to non-structured ones, like social networks and text. However, this success is mainly reached in classification tasks, where the goal is clearly defined and is possible to make use of past records, in order to predict new outcomes. Indeed, when this is not the case, the results are far from being useful *per se* in the majority of situations.

Pattern mining is a paradigmatic mining task, where this phenomenon occurs – existing approaches discover either a small number of irrelevant patterns or a very large number (usual thousands) of possibly interesting ones. The difficulty is just on choosing the best ones to analyze, but the great variety of interestingness measures available do not help on choosing the right ones in accordance to user interests or expectations.

Actually, the advances in the area of data mining are mainly centered on dealing with a wider range of types of data and domains, and much less on the quality of the models discovered, at least in terms of their simplicity and easyness of interpretation. The identified problem is not new, and was addressed by different techniques, through ILP (*Inductive Logic Programming*) or D3M (*Domain-Driven Data Mining)* fields. In this paper, we discuss the reasons for the failure of those approaches. Despite all the progress made, most of them are focused on mining specific domains, and therefore cannot be easily reused in different domains, even if the domain knowledge is represented with the same formalisms. In this sense, we argue that it is necessary a drift from domain-driven to knowledge-driven data mining, and focus should be given to the definition of techniques that are able to introduce domain knowledge deep into existing and more general algorithms.

Moreover, we argue that the best way to approach this problem is to use more general techniques, guided by domain knowledge, represented through domain ontologies. We base our argument on a clear problem statement and discuss how algorithms may be adapted to be driven by existing knowledge. Beside referring some preliminar results, we discuss how the improvements may be measured, in order to validate the real gains.

Antunes C. and Silva A..
New Trends in Knowledge Driven Data Mining.
DOI: 10.5220/0004974003460351
In *Proceedings of the 16th International Conference on Enterprise Information Systems* (ICEIS-2014), pages 346-351
ISBN: 978-989-758-027-7

2 THE LANDSCAPE OF KNOWLEDGE IN DM

The use of domain knowledge has been explored in data mining since its early years, in a somehow independent manner among different areas.

The first effort was undoubtedly made through *Inductive Logic Programming* (ILP for short), which is a paradigm of machine learning that is concerned with inducing classification rules from examples and background knowledge, all of which expressed as Prolog programs (Lavrac et al. 2011) (Nienhuys-Cheng and Wolf 1997), (Lisi and Malerba 2004), (Lisi and Esposito 2009). It was born from the interception of Concept Learning and Logic Programming, with the goal of prediction within the representation framework of Horn Clausal Logic.

The fact that all information must be written in declarative languages (like Prolog and Datalog) is one of the drawbacks of ILP approaches, and one of the reasons for not being widely used. Nevertheless, its structure promotes the representation and use of domain knowledge. There are many ILP algorithms that are able to introduce this knowledge into the discovery process (see, for example, (Raedt and Ramon 2004), (Malerba and Lisi 2001), (Levy and Rousset 1998), (Rouveirol and Ventos 2000)).

ILP techniques must also deal with the tradeoff between expressiveness and efficiency of the used representations. Studies show that current algorithms would scale relatively well as the amount of background knowledge increases. But they would not scale, at all, with the number of relations involved, and in some cases, with the complexity of the patterns being searched (Dzeroski 1996), (Lisi and Malerba 2004).

A second strategy, was relying in the shoulders of users / experts to guide the discovery, by choosing the most promising search paths. These interactive approaches (Nag, Deshpande and DeWitt 1999), (Goethals and Bussche 2000), (Goethals, Moens and Vreeken 2011), (Druck and McCallum 2011) use users feedback iteratively and incrementally, allowing them not only to view the intermediate results, but also to chose the best steps and measures, and even to change parameters. In this sense, users understand more easily what decisions lead to what results, therefore leading to more interesting results, in their perspective.

However, one on the problem of these systems is that users do not always know what they want, or what are the best choices. And besides being hard to define and implement, more elaborated domains may make this process too labor intensive and error prone. Furthermore, it is not straightforward the reuse of the knowledge and efforts applied before.

More recently, the methodology of Domain Driven Data Mining, D3M, was proposed (Cao and Zhang 2006), (Cao 2008), (Cao 2010), defending an urgent need for Actionable Knowledge Discovery to support businesses and applications.

The motivation behind D3M is the gap between academic objectives (innovation, performance and generalization) and business goals (problem solving), and between academic outputs and business expectations (Cao et al. 2010). So that this new data mining paradigm can be better accepted and advantageously applied in real businesses and applications, it is necessary to create methods and tools capable of analyzing real world data and extracting actionable knowledge, i.e. useful information that can be (as far as possible) directly converted into decision-making actions. The term “actionability” measures the ability of a pattern to prompt a user to take concrete actions to his advantage in the real world (Cao, Luo and Zhang 2007) .

To achieve that, data mining must involve the ubiquitous intelligence surrounding the business problem, such as human intelligence, domain intelligence, network and organizational/social intelligence (Cao et al. 2010). D3M proposes, therefore, a paradigm shift from data-centered knowledge discovery to domain-driven actionable knowledge discovery.

Research included in this area of D3M has been centered on the proposal of methods dedicated to specific domains, with a special emphasis on the actionability of the results. The specificity of those methods difficult their application to other domains, and the need for a standard methodology that is able to incorporate the existing knowledge of any domain into the mining process remains an open issue. In our opinion, existing work in D3M is more centered in the actionability of results in some domain, than on the reuse of the proposed strategies.

Along with the efforts in D3M, the use of existing domain knowledge to enrich the mining process has been explored under the umbrella of *Semantic Aspects of Data Mining*, by trying to add semantics to data under analysis.

The simplest approach, usually known as *semantic annotation* (Diamantini and Potena 2008), (Liu 2010), is just to use existing knowledge to annotate data, in order to help users understanding the data, and use it to get better results. This approach is gaining more adepts with the development of the Semantic Web, making it more plausible.

A second alternative is to use the semantics of some specific domain, represented through sound knowledge representation formalisms, to guide the mining algorithms, as proposed in (Antunes 2009), (Novak et al. 2009), (Jozefowska, Lawrynowicz and Lukaszewski 2010).

In this paper, we argue in favor of this last approach, and discuss the use of domain knowledge to guide DM algorithms in the search for more focused results.

3 KNOWLEDGE DRIVEN DATA MINING

The use of domain knowledge to improve the mining process tries to accomplish two main goals: to find more accurate and more easily understandable models.

A paradigmatic technique, that pursues these two goals are Bayesian networks (Pearl 1988), where a directed acyclic graph represents the known dependencies among variables, and algorithms are able to estimate a classification model. However, the difficulties on automatically designing these networks are well known, being this problem NP-hard (Heckerman, Geiger and Chickering 1995). Indeed, these networks are one of the preferred models, for example among physicians (see for example (Kononenko 1997)), since they are easily understandable reflecting cause-effects dependencies, usually known and described by them. Other techniques, like support vector machines or ensemble, whose discovered models are too hard to interpret, are actually put away, despite their major accuracy.

In our opinion, the great advantage of these models is their graphical representation, being completely clear to any informed user. When thinking about graphical knowledge representation formalisms, taxonomies and ontologies are the counterparts of Bayesian networks.

Before proceeding, we briefly overview the meaning of taxonomies and domain ontologies.

3.1 Knowledge Representation

Modelling has been one of the core parts of information science, either in information systems or in artificial intelligence. In both ones, it is generally accepted that without a good model, no system works adequately.

The advances in the areas of modeling and knowledge representation allow for using the developed mature formalisms to represent existing knowledge, and therefore making possible the exploration of those models to guide the discovery process. In particular, ontologies have gained a central role, for example in the semantic web, and begun to be used in many other contexts.

Ontologies are content theories about the objects, their properties and relations, that are possible in a specified domain of knowledge (Chandrasekaran, Josephson and Benjamins 1999), along with a set of explicit assumptions regarding the intended meaning of the vocabulary words (Lisi and Esposito 2009).

An ontology captures the intrinsic conceptual structure of the domain. In its simplest case, it describes a hierarchy of concepts related by is-a relations, a *taxonomy*. In more complex cases, other relationships can be added, as well as a set of axioms to help and constrain the interpretation of concepts. One of the most important features of ontologies is that they are valid, independently of the individuals or instances belonging to the domain.

Formally, an ontology is a tuple $O:=(C,\leq_C,R,A)$, where C corresponds to the set of concept identifiers (or just *concepts*) in the ontology, $\leq_C$ to the *hierarchy of concepts*, i.e. the *is-a* relations between concepts, R is the set of *relation* identifiers, and A a set of *axioms*. Relations referring to just one concept are called the *attributes*.

Among the formalisms proposed by Ontological Engineering for the construction of ontologies, the most currently used are Description Logics (DL) (Baader et al. (eds.) 2003). DLs are a family of First Order Logic fragments that allow for the specification of knowledge in terms of classes (concepts), relations (roles) and instances (individuals). A DL *knowledge base* (KB) consists in a terminological (schema) part, called T-Box, and an assertional (data) part, called A-Box. The T-Box part is where an ontology can be defined, and the A-Box corresponds to the database, with all the instances. The T-Box is usually referred to simply as the *ontology,* and the A-Box as one of the possible *knowledge bases* associated with that ontology. Moreover, ontologies describe the context in which the instances should be understood.

In a pragmatically view, an ontology just defines a directed graph, with concepts represented by nodes and relations by edges, which can be efficiently traversed by search domain-independent algorithms. In addition they incorporate axioms, which can be useful for describing additional constraints. In this sense, ontologies are a perfect tool to incorporate the represented knowledge deep into the mining process.

3.2 Knowledge Exploration

As discussed before, the idea of using knowledge to enrich the mining process is not new, and has been explored by data mining methods, but only seldom, and without having any significant results.

Examples of such approaches are the ones that used taxonomies to bias the discovery process (Srikant and Agrawal 1995) for pattern mining, but also (Zhang, Silvescu and Honavar 2002) in the classification context). In those methods, the main idea was to use the taxonomy to decide the best level of abstraction to consider, given the data under analysis. Indeed, items have different levels of support, when represented at different levels of abstraction.

As a more expressive form of knowledge representation than taxonomies, it is expected that ontologies can help guiding data mining algorithms finding more interesting results. The use of ontologies in data mining with this purpose is recent, and great parts of existing works are ad-hoc applications to specific problems.

Our claim is based on three points. First, through the taxonomy present in the domain ontology, data may be mined at the most interesting granularity, which is chosen on the fly, and allowing for spanning patterns at different granularities. Along with this use, it is then possible to reduce the time spent on data understanding and preparation steps in the mining process (Wirth and Hipp 2000).

Second, relations other than is-a relations, described in the ontology, may be used to filter the patterns that should be considered useful, reducing the complexity of the models, either the number of patterns or the depth / number of rules in classification models.

Third, axioms may be used to automatize the annotation of the data, by allowing for the automatic inference of information from the original data, and the posterior use of this information for being mined along with the data.

Recent work (Antunes and Bebiano 2012) has demonstrated that constraints defined over ontologies may be used by constrained adaptations of the most-well known algorithms for pattern mining, namely *apriori* (Agrawal and Srikant 1994) and *FP-growth* (Han, Pei and Yin 2000), without impairing their efficiency, but enabling the reduction of the number of patterns discovered, focusing the discovery according to user expectations.

In this new context, constraints have to perform three complementary roles – being the *mapper*, the *matcher* and the *filter*.

The *mapper* is the responsible for linking the data to be mined to the knowledge represented in the domain ontology. It should be a function from the set of items in the dataset to the concepts in the ontology. By making this connection, it is then possible to mine the concepts instead of the individuals, reducing the number of distinct elements during the mining process, which results in a smaller number of patterns and a simplification of the information discovered.

As a *matcher*, the constraint should be able to apply equivalences among entities. It should be a predicate among entities, either individuals or concepts. The goal is to allow for considering known and represented knowledge about the items, for example equivalences, which may be useful for choosing the right granularity and for counting the frequency of each item. Like the mapper, the matcher would allow for the simplification of the models, and the simplification of the pre-processing step.

At last, being a *filter*, the constraint avoids exploring search branches that are not interesting in accordance to user expectations. This filter would be a predicate over sets of entities (patterns in pattern mining and rules in classification).

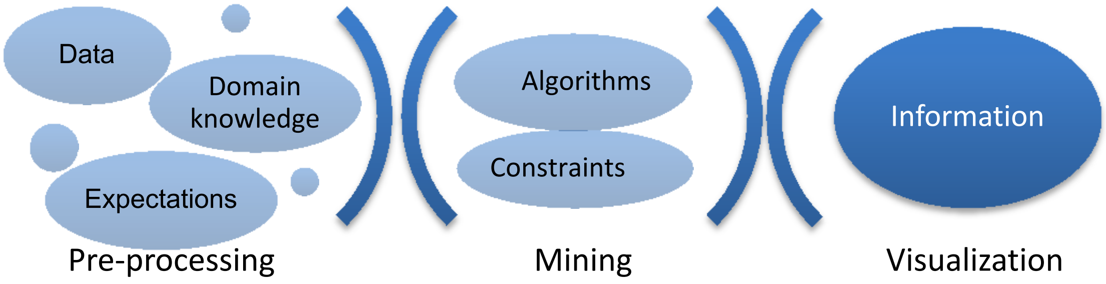

Figure 1: Process for knowledge exploration through the use of ontology-based constraints.

In all cases, and despite the different roles played, constraints should be used as a decision - maker by the mining algorithm. They should be the responsible for all decisions in the mining process, in particular for counting the frequency of each entity and for deciding which entities should be considered.

In this manner, the mining process becomes to be controlled by constraints, which at the end are just the mean to incorporate the domain knowledge deep into the mining process Figure 1.

4 VALIDATION

The simplification of the models discovered during the mining process is mandatory, either for reducing the bias of the models, or just for simplifying user understanding. However, how can we measure the improvements made by knowledge-driven approaches?

This measure is not straightforward since a simple count of the reduction on the number of patterns or rules discovered, against the counterpart in non-constrained algorithms, is not enough to assess the gains. Indeed, through the use of taxonomies is always possible to find just one rule with 100% of support, however this rule does not envisage any information.

In terms of efficiency, it is expected that constrained algorithms would follow a more complex process, more time consuming, since it is necessary to assure that the results are accordingly to the constraints. However, and as previous experiments show (Silva and Antunes 2013), the more expensive tasks in the mining step are the ones that scan the database for counting frequency of entities. So, and since constrained algorithms may reduce considerably these scans, it is expected that time efficiency of original algorithms remain unaffected. In order to measure the increase of time spent due to the complexity of the constrained task, we may measure the average time spent for finding each pattern, against the time spent in non-constrained approaches.

In terms of efficacy, the important issues are the quality of the discovered models and their simplicity. Naturally, the first issue may be measured through available measures, like support, confidence and lift for patterns, and accuracy, sensibility and specificity for classification.

The second assessment is much harder, and has to pass through the analysis of the compactness of the models discovered. One of the possible approaches is to compute the average number of non-constrained patterns / rules covered by a single constrained one, and the lost of interest of that condensed pattern / rule.

Note that this coverage identification is only possible, recurring to the ontology, which establishes the relation among the discovered entities.

5 CONCLUSIONS

The incorporation of domain knowledge, deep into the mining process, is expected to focus the search and modeling process, allowing for finding more interesting results. Despite the advances, most of the existing work is designed for some specific domain, and therefore cannot be reused.

In this position paper, we claim for a change in the data mining research, from a data-driven to a knowledge-driven process. We argue that knowledge represented through mature formalisms for knowledge representation, such as ontologies, can be used to define constraints, which may guide the mining process, by being responsible for the most important decisions in the mining step.

Beside the definition of the main roles to be played by constraints, we discuss different ways to assess the improvements on the discovered information, in terms of both efficiency and efficacy.

ACKNOWLEDGEMENTS

This work is partially supported by FCT Fundação para a Ciência e a Tecnologia, under research project D2PM (PTDC/EIA-EIA/110074/2009).

REFERENCES

Agrawal, R and Srikant, R 1994, 'Fast Algorithms for Mining Association Rules', *Int'l Conf on Very Large Data Bases*, Morgan Kaufmann, Chile.

Antunes, C 2009, 'Mining Patterns in the Presence of Domain Knowledge', *Int'l Conf on Enterprise Information Systems*, INSTICC, Italy.

Antunes, C and Bebiano, T 2012, 'Mining Patterns with Domain Knowledge: a case study on multi-language data. In', *Int'l Conf on Information Systems (ACIS ICIS 2012)*, IEEE Press.

Baader, Franz, Calvanese, D, McGuinness, DL, Nardi, D, Patel-Schneider, PF (eds.) 2003, *The Description*

Logic Handbook: Theory, Implementation, and Applications, Cambridge University Press.

Cao, L 2008, 'Domain driven data mining (d3m)', *2008 IEEE Int. Conf. on Data Mining Workshops (DDDM 08)*, IEEE Computer Society.

Cao, L 2010, 'Domain-Driven Data Mining: Challenges and Prospects', *IEEE Transactions on Knowledge and Data Engineering*, vol 22, no. 6, pp. 755-769.

Cao, L, Luo, D and Zhang, C 2007, 'Knowledge actionability: satisfying technical and business interestingness', *Int'l Journal of Business Intelligence and Data Mining*, vol 2, no. 4, p. 496–514.

Cao, L, Yu, P, Zhang, C and Zhang, H 2010, *Data Mining for Business Applications*, Springer.

Cao, L and Zhang, C 2006, 'Domain-driven data mining: A practical methodology', *Int. Journal Data Warehousing and Mining*, vol 2, no. 4, p. 49–65.

Chandrasekaran, B, Josephson, JR and Benjamins, VR 1999, 'What Are Ontologies, and Why Do We Need Them?', *IEEE Intelligent Systems*, vol 14, no. 1, p. 20–26.

Diamantini, C and Potena, D 2008, 'Semantic annotation and services for kdd tools sharing and reuse ', *2008 IEEE Int. Conf. on Data Mining Workshops (ICDMW 08)*, IEEE, Pisa, Italy.

Druck, G and McCallum, A 2011, 'Toward interactive training and evaluation', *ACM Int'l Conf on Information and Knowledge Management*, ACM.

Dzeroski, S 1996, 'Inductive logic programming and knowledge discovery in databases', in *Advances in Knowledge Discovery and Data Mining*, MIT Press.

Goethals, B and Bussche, J 2000, 'On Supporting Interactive Association Rule Mining', *Int'l Conf Data Warehousing and Knowledge Discovery*, Springer.

Goethals, B, Moens, S and Vreeken, J 2011, 'MIME: a framework for interactive visual pattern mining', *ACM SIGKDD Int'l Conf on Knowledge discovery and data mining (KDD 11)*, ACM.

Han, J, Pei, J and Yin, Y 2000, 'Mining Frequent Patterns without Candidate Generation', *Int'l Conf. on Management of Data*, ACM Press, TX.

Heckerman, D, Geiger, D and Chickering, DM 1995, 'Learning Bayesian Networks: The Combination of Knowledge and Statistical Data', *Machine Learning*, vol 20, pp. 197-243.

Jozefowska, J, Lawrynowicz, A and Lukaszewski, T 2010, 'The role of semantics in mining frequent patterns from knowledge bases in description logics with rules', *Theory Practical Logical Programming*, vol 10, no. 3, p. 251–289.

Kononenko, I 1997, 'Machine learning for medical diagnosis: history, state of the art and perspective', in RS Michalski, I Bratko, M Kubat (eds.), *Machine Learning and Data Mining: Methods and Applications*, Wiley.

Lavrac, N, Vavpetic, A, Soldatova, LN, Trajkovski, I and Novak, PK 2011, 'Using ontologies in semantic data mining with segs and g-segs', *Int'l Conf on Discovery Science (DS 11)*, Finland.

Levy, A and Rousset, M-C 1998, 'Combining horn rules and description logics in carin', *Artificial Intelligence*, vol 104, no. 1, p. 165–209.

Lisi, F and Esposito, F 2009, 'On ontologies as prior conceptual knowledge in inductive logic programming', *Studies in Computational Intelligence*, vol 220, p. 3–17.

Lisi, F and Malerba, D 2004, 'Inducing multi-level association rules from multiple relations', *Machine Learning*, vol 55, no. 2, p. 175–210.

Liu, H 2010, 'Towards semantic data mining', *Int'l Semantic Web Conf. (ISWC 10)*.

Malerba, D and Lisi, F 2001, 'Discovering associations between spatial objects: An ilp application', *Int'l Conf. on Inductive Logic Programming (ILP 01)*, Springer-Verlag, London, UK.

Nag, B, Deshpande, PM and DeWitt, DJ 1999, 'Using a knowledge cache for interactive discovery of association rules', *ACM SIGKDD Int'l Conf Knowledge Discovery and Data Mining*, ACM.

Nienhuys-Cheng, S-H and Wolf, RD 1997, *Foundations of Inductive Logic Programming*, Springer-Verlag.

Novak, P, Vavpetic, A, Trajkovski, I and Lavraˇc, N 2009, 'Towards semantic data mining with g-segs', *Int'l Multiconference Information Society (IS 09)*.

Pearl, J 1988, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann.

Raedt, LD and Ramon, J 2004, 'Condensed representations for inductive logic programming', *Int'l Conf. on Principles of Knowledge Representation and Reasoning*, AAAI Press.

Rouveirol, C and Ventos, V 2000, 'Towards learning in carin-aln', *Int'l Conf. on Inductive Logic Programming (ILP 00)*, Springer-Verlag.

Silva, A and Antunes, C 2013, 'Pushing Constraints into a Pattern-Tree', *Int'l Conf. on Modeling Decisions for Artificial Intelligence (MDAI 2013)*, Springer.

Srikant, R and Agrawal, R 1995, 'Mining Generalized Association Rules', *Int'l Conf on Very Large Databases*, Morgan Kaufmann, Switzerland.

Wirth, R and Hipp, J 2000, 'CRISP-DM: Towards a Standard Process Model for Data Mining', *Int'l Conf. on the Practical Application of Knowledge Discovery and Data Mining*.

Zhang, J, Silvescu, A and Honavar, V 2002, 'Ontology-driven induction of decision trees at multiple levels of abstraction', in *Abstraction, reformulation, and approximation*, Springer.