Knowledge Management Challenges in Knowledge Discovery Systems

Mykola Pechenizkiy
Dept. of CS and ISs
University of Jyväskylä
Finland
mpechen@cs.jyu.fi

Alexey Tsymbal
Dept. of CS
Trinity College Dublin
Ireland
tsymbalo@tcd.ie

Seppo Puuronen
Dept. of CS and ISs
University of Jyväskylä
Finland
sepi@cs.jyu.fi

Abstract

Current knowledge discovery systems are armed with many data mining techniques that can be potentially applied to a new problem. However, a system faces a challenge of selecting the most appropriate technique(s) for a problem at hand, since in the real domain area it is infeasible to perform a comparison of all applicable techniques. The main goal of this paper is to consider the limitations of data-driven approaches and propose a knowledge-driven approach to enhance the use of multiple data-mining strategies in a knowledge discovery system. We introduce the concept of (meta-) knowledge management, which is aimed to organize a systematic process of (meta-) knowledge capture and refinement over time.

1. Introduction

Current electronic data repositories are growing quickly and contain big amount of data from commercial, scientific, and other domain areas. The capabilities for collecting and storing all kinds of data exceed the abilities to analyze, summarize, and extract knowledge from this data. Knowledge discovery systems (KDSs) use achievements from many technical areas, including databases, Data Mining (DM), statistics, AI, machine learning, pattern recognition, decision support systems, and knowledge-based systems. Knowledge discovery from databases is an innovative approach to information management and is associated commonly with the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns and relations in large databases [2].

Present-day KDSs are armed with a number of available techniques to process data; and, potentially, there are many possible combinations of these techniques to construct a DM strategy for mining a current problem. Many empirical studies are aimed to show that one learning strategy can perform significantly better than another on a group of problems that are characterised by some properties [6]. Selection of the most appropriate DM technique or a group of the most appropriate techniques is usually not straightforward. In a real problem-solving situation it is not computationally feasible to apply every DM strategy. Therefore, dynamic selection of DM methods in KDSs has been under active study (see, e.g., [15]).

However, at least two contexts of dynamic selection can be distinguished. First, the so-called multi-classifier systems that apply different ensemble techniques [4]. Their general idea is usually to select one classifier on dynamic basis taking into account the local performance (e.g. generalisation accuracy) in the instance space. Second, multistrategy learning that applies a strategy selection approach which takes into account the classification problem related characteristics (meta-data). We are interested in the second context in this study.

Meta-learning is the effort to automatically induce dependencies between learning tasks and appropriate learning strategies. In the context of classifier ensembles, where only the data itself is used to make decisions about method selection, rather good practical results are shown in experiments supported by theoretical studies as well [15]. Unfortunately, this is not the case with meta-learning for dynamic integration of DM strategies for a data set at hand. This area is less studied and hardly ever applied in practice.

There are works on the multistrategy approach based on the ideas of constructive induction and conceptual clustering for conceptual data exploration, i.e. the derivation of high-level concepts and descriptions from data through symbolic reasoning involving both data and background knowledge [10]. Recently, several studies on automatic classifier selection via meta-learning have been reported, see e.g. [5] for an overview. However, the experimental results

of presented approaches are not so promising.

In this paper we propose a knowledge-driven approach to enhance the dynamic integration of DM strategies in KDSs. Our focus here is on knowledge management (KM) aimed to organise a systematic process of knowledge capture and refinement over time. In order to distinguish between the knowledge extracted from data and the higher-level knowledge (from the KDS perspective) required for managing techniques' selection, combination and application we will refer to the latter as *meta-knowledge*. We consider the basic KM processes of knowledge creation and identification, representation, collection and organization, sharing, adaptation, and application with respect to the introduced concept of meta-knowledge.

The rest of the paper is organized as follows. In Section 2 we consider the basics of multistrategy knowledge discovery. In Section 3 we consider meta-learning approaches for automatic technique selection. A knowledge-driven approach for dynamic integration of DM techniques is introduced in Section 4. We conclude with a brief summary and assessment of further research directions in Section 5.

2. Multistrategy Knowledge Discovery

Numerous process-oriented KDSs have recently been developed. At the beginning of the millennium there exist about 200 tools that could perform a few tasks each (such as clustering, classification, regression, and visualization) for specialized applications [12]. This growing trend towards integrating DM tools with specialized applications has been associated with the development of KDSs that are often called "vertical solutions" [3].

However, adaptability to variations in data characteristics and dynamics of business scenarios becomes increasingly important for data processing systems, as they become an integral part of an organizational decision support system [6]. KDSs should be able to discover knowledge by combining several available techniques, and provide a more automatic environment, or an application envelope, surrounding a highly sophisticated DM system [3]. A similar trend was reported in [10] with respect to the orientation of machine-learning systems from single-strategy to multistrategy task-adaptive learning.

Let us consider briefly the basics of the knowledge discovery process according to [13] presented in Figure 1. The life cycle of an idealized knowledge discovery project consists of seven sequential phases from business understanding to deployment. Generally, these phases are not so strictly sequential, and moving back and forth between different phases,

caused by the outcome of each phase, is rather natural.

The business-understanding phase is aimed to formulate business questions and translate them into DM goals. The data-understanding phase aims to analyse and document available data and knowledge sources in the business according to the formulated DM goals and provide initial characterization of data. The data preparation phase starts from target data selection that is often related to the problem of building and maintaining useful data warehouses. After selection, the target data is preprocessed in order to reduce the level of noise, handle missing information, reduce data, and remove obviously redundant features. The data exploration phase aims to provide the first insight into the data, evaluate the initial hypotheses, usually, by means of descriptive statistics and visualization techniques. The DM phase covers selection and application of DM techniques, initialization and further calibration of their parameters to optimal values. The discovered patterns that may include a summary of a subset of the data, statistical or predictive models of the data, and relationships among parts of the data are locally evaluated. The evaluation and interpretation phase aims to analyse the discovered patterns, to determine the patterns that can be considered as new knowledge, and to draw conclusions about the whole discovery process as well. The deployment phase aims to transfer DM results that meet the success criteria into business [13].
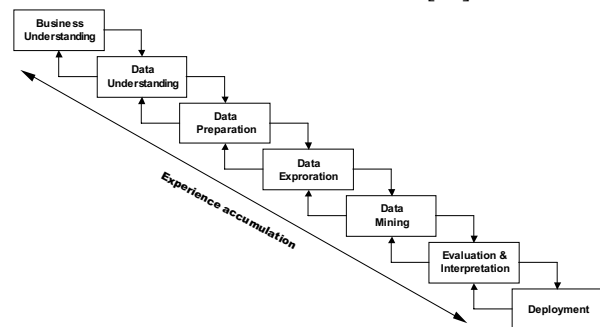

Figure 1. Knowledge discovery process: from problem understanding to deployment (adopted from [13])

As one can see from Figure 1, the knowledge discovery process is iterative and consists of various phases and tasks, although the core of the process is DM. The important issue here is that during the KDD process positive or negative (but still valuable) experience is achieved that can be and should be accumulated for better understanding of KDD and further use. We have to admit that current research in knowledge discovery concentrates mostly on the technical details of DM algorithms, whereas the relations between techniques from different fields and how they would fit into the overall knowledge

discovery process is often not so clear. It seems that today still there does not exist, unfortunately, any unified methodology (beside general guidelines like CRISP-DM [1]) that would help practitioners to manage the knowledge discovery process.

An end user of a present-day KDS needs to be a DM expert, or he or she should work in close collaboration with professional data miners, since clear and complete understanding of the knowledge discovery process is essential for successful discovery. And even for business analysts and experienced DM engineers it often remains difficult to find the best-suited techniques to solve a business problem under consideration.

3. Automatic DM Technique Selection

Meta-learning (also known as bias learning) can be seen as an effort to automatically induce correlations between tasks and inductive learning strategies [5]. Successful meta-learning in the context of (automatic) DM technique(s) selection would be really very important and beneficial in the DM practice. It is obvious that in a real-world situation it is very unlikely to accomplish the brute-force search comparing all the applicable approaches.

Several meta-learning approaches for automatic DM technique selection have been introduced in the literature. A comprehensive overview can be found in [5]. The most popular strategies for meta-learning are the characterization of a data set in terms of its statistical/information properties and the characterization of algorithms.

The general idea of meta-learning with respect to the selection of a DM technique for a data set at hand is rather straightforward. Having a collection of data sets and a collection of classifiers, we can produce their meta-data as a result of their (algorithms and data sets) characterization. When meta-data is available, a machine-learning algorithm can be applied to it. As a result, a meta-learning model that maps data set characteristics onto classifiers' characteristics with respect to the introduced performance criteria is built. When a new data set is introduced to the system, necessary data set's characteristics are estimated so that the meta-model is able to suggest an algorithm or a combination of algorithms according to a performance criterion.

The characteristics of a data set that can be used for meta-learning are commonly divided into those that describe the nature of attributes, attributes themselves, associations between attributes, and associations between attributes and a target variable. The

exhaustive list of currently known statistical and information measures of a data set and the data characterization tool is proposed in [9].

Beside characterization of a data set, some application restrictions or priorities can be introduced. Thus, a user may like to define the most (un)desirable or the most crucial characteristic(s) of an algorithm to be selected for a certain application. The most common characteristics that are taken into account are: algorithm taxonomy, interpretability of the model, importance of results interpretability and algorithm transparency, explanation of decision, training and testing time, accuracy etc. A good overview of potential algorithms' characteristics is given in [5].

Although having some degree of success, the meta-learning approach as such has several shortcomings. In [9] two general problems with meta-models that associate a data set and algorithm characteristics are reported. The first problem is the representativeness of meta-data examples. The possible space of learning problems and thus a meta-learning-space is vast and is getting larger with the invention of new algorithms, consideration of new characteristics and parameters. But the size of meta-data sets used in the studies is naturally rather small because of the computational complexity of producing a single meta-example – usually a time-consuming cross-validation process is used to estimate the performance of every algorithm used in the study. The other problem is the computational complexity of some sophisticated statistical measures.

We believe that a deeper analysis of a restricted set of DM techniques both at the theoretical and experimental level is a more beneficial approach rather than application of the meta-learning approach only to the whole range of machine learning techniques at once.

In the next section we propose to treat a KDS as a complex adaptive system that creates, receives, stores, retrieves, transforms, and transmits (meta-) knowledge to improve its own ability to adapt to the environment and to utilize available DM techniques more efficiently and effectively. We emphasize the necessity to integrate meta-knowledge produced by DM experts, DM practitioners and meta-learning approaches.

4. A Knowledge-Driven Approach for Efficient Use of DM Techniques

According to the presented in [7] analysis of the most important issues in KM, there are 4 groups of such issues: (1) executive/strategic management, (2) operational management, (3) costs, benefits, and risks

management, and (4) standards in the KM technology and communication. In this paper we address the issues of the second group that include the identification of available knowledge, seeking ways to capture it in the KM process, and analyzing the ability to design a (meta-)knowledge management system including its tools and applications. According to [16] the most practical way to define KM is to show on the existing IT infrastructure the involvement of: (1) knowledge repositories, (2) best-practices and lessons-learned systems, (3) expert networks [DM experts], and (4) communities of practice [these are end-users].

The main idea of the continuous KM process is presented in Figure 2. We separate five key phases of this process. The first phase deals with knowledge acquisition or creation. Having a collection of data sets and a collection of classifiers, we can characterize them and collect the results to produce their meta-data (both for algorithms and data sets). A functional approach to KM with regard to both symbolic and numerical data exploration tasks and at both the data and knowledge bases levels was proposed in [10]. Knowledge generating operators (KGOs) that can be used for creating, modifying, combining, deleting, and selecting rules and other structures in the knowledge base were introduced. KGOs can be applied (1) to generate and test meta-rules (hypotheses), which are inferred from summarizing the facts or discriminating between the groups of facts about DM strategy performance on problems with certain characteristics; (2) to construct a decision structure from a set of decision rules; (3) to generate new features/concepts.
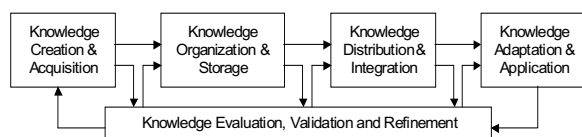

Figure 2. The knowledge management process

The second phase deals with knowledge organization and storage. In our context these processes are related mainly to knowledge representation issues. Minsky [11] discusses pros and cons of connectivist and structural approaches to knowledge representation, concluding that their combination would be natural, since usually at the lower levels of abstraction it tends to have a net architecture and tends to organize clusters and hierarchical structures at the higher levels of abstraction.

The third phase is related to knowledge distribution and knowledge integration processes. Generally, we distinguish four potential sources of knowledge that has to be integrated in the repository of KDS system:

(1) meta-knowledge from an expert in DM and related fields; (2) meta-knowledge from a DM practitioner; (3) meta-knowledge from laboratory experiments on synthetic data sets; and, finally, (4) meta-knowledge from field experiments on real-world problems. Beside this, research and business communities, or similar KDSs themselves can organize different so-called trusted networks, where participants are motivated to share their knowledge.

The fourth phase deals with the knowledge adaptation and application processes. It is often impossible to apply certain elements of knowledge directly. Therefore the knowledge adaptation process needs to be undertaken. Case-based reasoning with multiple case bases [8] might be one approach to perform such adaptation.

The fifth phase deals with the knowledge evaluation, validation and refinement processes. In order to keep the meta-knowledge updated there is a need to have a monitoring process to control whether the discovered (meta-)knowledge remains valid and a technique for continuous enhancement of knowledge. Since the repository is created it tends to grow and at some point it naturally begins to collapse under its own weight, requiring major reorganization [16]. Therefore, the repository needs to be continuously updated, and some content needs to be deleted (if misleading), deactivated or archived (if it is potentially useful). Content may become less fragmented and redundant if similar contributions are combined, generalized and restructured. The process of filtering knowledge claims (especially when produced automatically) into the accepted and rejected ones is often applied in KM.

It is highly desirable to make the knowledge repository adaptive, i.e. some knowledge should exist that would guide an organization to change the repository when the environment calls for it. The basic idea here is that when the environment changes (that in general may happen all the time), all of the general rules without specifying the context (so-called "knowing when" and "knowing where" contexts) could become invalid.

Some knowledge claims are naturally in constant competition with the other claims. Disagreements within the knowledge repository need to be resolved by means of generalization of some parts and contextualization of the others. In order to increase the quality and validity of knowledge, it needs to be continually tested, and enhanced including, improving, and removing (deactivating) knowledge claims. After changes a new process of testing and validation is needed.

We would like to clarify the notions of *knowledge validity* and *knowledge quality* with respect to the

knowledge refinement process. The contexts "knowing when" and "knowing where" can be discovered before it appears in a real situation. So-called zooming in and zooming out procedures can be used to find a context where a theory can be falsified or supported. The goal of such procedures is in search for balance between generality, compactness, interpretability, and understandability and sensitiveness to the context, exactness, precision, and adequacy of meta-knowledge.

The quality of knowledge can be estimated by its ability to help a KDS produce solutions faster and more effectively. To determine the relative quality of a validated knowledge claim, its value needs to be compared to the values of the other claims according to the existing criteria. In any case knowledge claims have both a degree of utility and a degree of satisfaction. However, the quality of knowledge is often context-dependent. Therefore *where* and *when* context conditions may be important in many situations not only for knowledge validation but also for quality estimation.

The quality of a knowledge claim is further dependent on the accuracy of the criteria used to evaluate it. Such criteria as complexity, usefulness, and predictive power are well formalized and easy to estimate. On the contrary, such criteria as understandability, reliability of source, explanatory power are rather subjective and therefore inaccurate.

5. Conclusions

Although there is a huge number of DM techniques – one can hardly find a technique that would be best for all data sets. Unfortunately, there does not exist canonical knowledge, a perfect mathematical model, or any relevant tool to select the best technique for certain problem at consideration. Instead, a volume of accumulated empirical findings, some trends, and some dependencies have been discovered in a number of various studies. On the other hand there are certain assumptions on performance of DM techniques under certain conditions.

In this paper we proposed a knowledge-driven approach to enhance the dynamic integration of DM strategies in KDSs. Our focus was on KM aimed to organize a systematic process of meta-knowledge capture and refinement over time. We considered the basic KM processes of knowledge creation and identification, representation, collection and organization, sharing, adaptation and application with respect to the introduced concept of meta-knowledge.

We see our further research efforts in the implementation of presented knowledge-driven framework for a KDS that contains a limited number of DM techniques of a certain type (presumably different feature transformation methods as we have experience with them and tools with their implementation) and evaluation of this framework in practice for real-world problems in a distributed environment.

References

[1] CRISP-DM: 1.0 Step-by-step DM guide, SPSS Inc.
[2] Fayyad, U.M. "Data Mining and Knowledge Discovery: Making Sense Out of Data", *IEEE Expert 11(5)*, 1996, pp. 20-25.
[3] Fayyad, U.M., Uthurusamy, R. "Evolving data into mining solutions for insights", *Communications of the ACM 45(8)*, 2002, pp. 28-31.
[4] Dietterich, T.G. "Machine Learning Research: Four Current Directions", *AI Magazine 18 (4)*, 1997, pp. 97-136.
[5] Kalousis, A. "Algorithm Selection via Meta-Learning", University of Geneve, PhD Thesis, 2002.
[6] Kiang, M. "A comparative assessment of classification methods", *Decision Support Systems 35*, 2003, pp. 441-454.
[7] King, W., Marks, P., McCoy, S. "The most important issues in knowledge management", *Communications of the ACM 45(9)*, 2002, pp. 93-97.
[8] Leake, D., Sooriamurthi, R., "Automatically Selecting Strategies for Multi-Case-Base Reasoning". In: *Proc. ECCBR 2002*, LNCS 2416, 2002, pp. 204-233.
[9] Lindner, G., Studer, R. "AST: Support for algorithm selection with a CBR approach", In *Proc. of the 3rd PKDD Conference*, 1999, pp. 418-423.
[10] Michalski, R.S. "Seeking Knowledge in the Deluge of Facts", *Fundamenta Informaticae 30*, 1997, pp. 283-297.
[11] Minsky, M. "Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy", *AI Magazine 12(2)*, 1991, pp. 34-51.
[12] Piatetsky-Shapiro, G. "Knowledge Discovery in Data-bases: 10 years after", *SIGKDD Explorations 1(2)*, 2000, pp. 59-61.
[13] Reinartz, T. *Focusing Solutions for Data Mining*. LNAI 1623, Berlin Heidelberg, 1999.
[14] Schmidhuber, J., Zhao, J., Wiering, M. "Simple principles of metalearning", Technical Report IDSIA-69-96, 1996.
[15] Tsymbal, A. "*Dynamic Integration of Data Mining Methods in Knowledge Discovery Systems*", PhD Thesis. University of Jyväskylä, Finland, 2002.
[16] Zack, M. "Managing codified knowledge", *Sloan Management Review 40(4)*, 1999, pp. 45-58.