Expert Systems with Applications 39 (2012) 7524–7535

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Cooperation between expert knowledge and data mining discovered knowledge: Lessons learned

Fernando Alonso *, Loïc Martínez [1], Aurora Pérez [2], Juan P. Valente [2]

Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28600 Boadilla del Monte, Madrid, Spain

ARTICLE INFO

Keywords:
Data mining
Discovered knowledge
Expert systems
Expert knowledge
Mined knowledge
Lessons learned

ABSTRACT

Expert systems are built from knowledge traditionally elicited from the human expert. It is precisely knowledge elicitation from the expert that is the bottleneck in expert system construction. On the other hand, a data mining system, which automatically extracts knowledge, needs expert guidance on the successive decisions to be made in each of the system phases. In this context, expert knowledge and data mining discovered knowledge can cooperate, maximizing their individual capabilities: data mining discovered knowledge can be used as a complementary source of knowledge for the expert system, whereas expert knowledge can be used to guide the data mining process. This article summarizes different examples of systems where there is cooperation between expert knowledge and data mining discovered knowledge and reports our experience of such cooperation gathered from a medical diagnosis project called Intelligent Interpretation of Isokinetics Data, which we developed. From that experience, a series of lessons were learned throughout project development. Some of these lessons are generally applicable and others pertain exclusively to certain project types.

1. Introduction

Expert knowledge and discovered knowledge are two powerful tools that can be combined. Used together they maximize the qualities that they have separately.

An expert system operates on a knowledge base that contains the knowledge elicited from the expert (EK). This knowledge base is represented by some formalism (rules, frames, Bayesian networks, etc.) and is built by the knowledge engineer from elicited expert knowledge and, later, validated by the expert. Evidently, the system is subject to and limited by the amount of knowledge entered, that is, represented in its knowledge base. And, precisely, the bottleneck in expert system construction is knowledge elicitation, a phase conditioned by countless constraints ranging from the number of available experts, or how much expertise the experts have, to the complexity of the actual knowledge elicitation process.

Recently, automatic knowledge acquisition techniques have attracted a lot of interest as they are potentially a big help for remedying this bottleneck. The knowledge discovery in databases (KDD) process, especially data mining techniques, is used to automatically discover knowledge from data. The knowledge discovered by data mining (DMK) is implicit in the data and can take the shape of patterns or models that fit the data, trends in temporal data, associations among different data features, rules, etc.

The key point is that these two approaches, knowledge elicitation from experts and knowledge discovery from data, complement each other (da Silva, Amorim, Campos, & Brasil, 2002; Daniels & van Dissel, 2002; de la Vega et al., 2010; Weiss, Buckley, Kapoor, & Damgaard, 2003). Applied together, they can be used to build better systems: data mining techniques can be used to support the different tasks involved in expert system (ES) or knowledge-based system (KBS) development (Flior et al., 2010; Mejía-Lavalle & Rodríguez-Ortiz, 1998; Phuong, Phong, Santiprabhob, & Baets, 2001; Wang, Liu, & Cheng, 2004), and expert knowledge can be used to facilitate and improve the results of the different stages of the KDD process (Kusiak & Shah, 2006; Zhang & Figueiredo, 2006).

The aim of this article is to describe the key results of this interaction between EK and DMK, while highlighting the lessons learned over the years from our own experience of these issues in the medical field, presenting a long-term project called I4 (Intelligent Interpretation of Isokinetics Data). This project integrates expert systems and data mining techniques to process isokinetics data. We believe that the results of and the lessons learned from this project are potentially useful for developing systems incorporating EK and DMK.

The remainder of the article is organized as follows. Section 2 describes related work analyzing other applications that present some facet of this type of cooperation. In Section 3 we outline our I4 project. Sections 4–6 describe the three I4 project phases:

* Corresponding author. Tel.: +34 913522546; fax: +34 913367411.
E-mail addresses: falonso@fi.upm.es (F. Alonso), loic@fi.upm.es (L. Martínez), aurora@fi.upm.es (A. Pérez), jpvalente@fi.upm.es (J.P. Valente).
[1] Tel.: +34 913522546; fax: +34 913367411.
[2] Tel.: +34 913367389; fax: +34 913367411.

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.133

expert system development, data mining and symbolic data mining. In Section 7 we summarize the lessons learned. And, finally, Section 8 outlines some conclusions.

2. Cooperation between expert knowledge and discovered knowledge in different fields

Cooperation between EK and DMK has emerged in different fields, like medicine, engineering, finance, etc., and has diverse features. We have grouped different examples of this cooperation into two separate sections. In Section 2.1, we present cases where DM techniques have been used to aid the development of an ES or KBS. In Section 2.2, we present cooperation in the opposite direction, that is, expert system techniques used to improve a knowledge discovery process or system.

2.1. Contribution of discovered knowledge to expert knowledge

Cooperation by applying data mining techniques to an expert system built from heuristic knowledge elicited beforehand from the expert has the goal of optimizing and maximizing the performance of the resulting KBS. This contribution of discovered knowledge to expert knowledge has been designed for different purposes, as we describe below.

- When an expert's heuristic knowledge is the basis for developing an ES and there is a consistent database of solved cases, it is usual practice to check ES robustness by applying DM techniques to the generated database. An example of this type of cooperation is to be found in (Cooke et al., 2000). PERFEX is a rule-based expert system for the automatic interpretation of cardiac SPECT (single photon emission computed tomography) data. This system infers the extent and severity of CAD (coronary artery disease) from perfusion distributions, and outputs a patient report summarizing the condition of the three main arteries and other pertinent information. The overall goal is to assist in the diagnosis of CAD. The expert system presents the resulting diagnostic recommendations in both visual and textual forms in an interactive framework, thereby enhancing overall utility. DM techniques were applied to the patient database containing images and text to validate the system and the confidence (certainty factors) in the heuristic rules of the ES.
- Frequently, KBS users have to answer a great many questions formulated by the system to gather information about the problem. In consultation systems like these, the application of DM techniques to the stored results of previous system executions can optimize future question/answer sequences by reducing the number of questions. This is what the research presented in Bethel, Hall, and Goldgof (2006) did. The developed system is a web-based expert system to match eligible breast cancer patients with open clinical trials, or categorize the reasons why an eligible patient was not put onto the trial. Through interviews with clinicians, implications were discovered that reduced the number of questions/answers required to determine eligibility. The idea is that a physician will immediately know the answers to some questions based on answers to others, that is, there is a clear implication that always holds.
- When the knowledge base is built with generic global knowledge and ES behavior on a specific local problem is unknown (i.e., we do not know whether it will correctly and completely account for all the cases that can occur in that local situation), discovered knowledge can be used to verify whether the rules obtained are representative enough for the local data and if the local data have any new correlation that the knowledge base does not contain. This cooperation between discovered knowledge and expert knowledge was applied in Lama et al. (2006) to treat nosocomial infection. The MERCURIO system was designed to support medical practitioners in the complex task of controlling nosocomial infections. Its knowledge base was built from both the NCCLS (the National Committee for Clinical Laboratory Standards) guidelines and the expert's suggestions. These guidelines are quite general, since they were built considering data regarding many laboratories around the world. However, it is not clear that they can completely and correctly interpret the infections developed inside a particular hospital environment. For this reason, it was necessary to verify if the rules obtained from the NCCLS document were representative of the local hospital infections, and if there were other correlations in the local hospital infection data that are either not considered in the NCCLS document or unknown to the expert microbiologists. To address these problems, data mining was applied to local hospital infection data to generate association rules that show the susceptibility or resistance of a bacterium to different antibiotics. The discovered association rules were transformed into alarm rules that were confirmed by experts and then used for data validation in the system.
- Another type of cooperation of discovered knowledge with expert knowledge takes place when the application domain changes too quickly, and it is very hard or even impossible to timely update the KBS developed by the expert. Discovered knowledge is applied for this purpose, as it can be used to update the system. An example of this type of cooperation is described in WuJing (2001) in a project concerned with optimizing a mobile network. The problem with using a mobile network is that it changes very often, and its management software also upgrades very often. So the knowledge in the expert system must be upgraded accordingly. Traditional knowledge extraction methods cannot conform to the fast changing environment, and DM techniques have been applied to update the knowledge, using the Operation and Maintenance Center repository as a source.

2.2. Contribution of expert knowledge to discovered knowledge

Expert knowledge supports the extraction of knowledge applying DM techniques and the later validation of this knowledge in a wide variety of manners, as described in the following.

- One of the problems of automatically extracting knowledge is how to validate and check its fitness. In this respect, Holmes and Cunningham describe a typical case of KBS construction where expert knowledge can debug and validate the set of rules gathered from DM (Holmes & Cunninghan, 1993). The "Explora" data mining tool is used to build and maintain the ES. "Explora" is a statistics-based mining program that applies domain knowledge and statistical measurements of the database to identify patterns of attribute values and value combinations that occur more or less frequently in the database. The discovered knowledge is debugged and validated by the expert and built into the ES.
- DM can be applied in personalized applications, typical in e-commerce (dynamic content presentation, personalized ad targeting, individual customer recommendations, etc.). In these types of applications rule discovery methods are applied individually to the transactional data of every user to capture each user's truly personal behavior. EK can be used to globally validate all this locally mined knowledge and discard irrelevant rules. Adomavicius and Tuzhilin deal with the problem of a human expert validating large numbers of locally mined rules with relatively little input from the expert (Adomavicius & Tuzhilin, 2001). This is done by applying different rule