Information Sciences 288 (2014) 254–278

Contents lists available at ScienceDirect

Information Sciences

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

Combining human analysis and machine data mining to obtain credible data relations

Vedrana Vidulin *, Marko Bohanec, Matjaž Gams

Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 19 November 2012
Received in revised form 25 July 2014
Accepted 2 August 2014
Available online 12 August 2014

Keywords:
Interactive data mining
Interactive machine learning
Interactive explanation structure
Relation-extraction scheme
Domain analysis
Human–computer interaction

ABSTRACT

Can a model constructed using data mining (DM) programs be trusted? It is known that a decision-tree model can contain relations that are statistically significant, but, in reality, meaningless to a human. When the task is domain analysis, meaningless relations are problematic, since they can lead to wrong conclusions and can consequently undermine a human's trust in DM programs. To eliminate problematic relations from the conclusions of analysis, we propose an interactive method called Human–Machine Data Mining (HMDM). The method constructs multiple models in a specific way so that a human can reexamine the relations in different contexts and, based on observed evidence, conclude which relations and models are credible—that is, both meaningful and of high quality. Based on the extracted credible relations and models, the human can construct correct overall conclusions about the domain. The method is demonstrated in two complex domains, extracting credible relations and models that indicate the segments of the higher education sector and the research and development sector that influence the economic welfare of a country. An experimental evaluation shows that the method is capable of finding important relations and models that are better in both meaning and quality than those constructed solely by the DM programs.

1. Introduction

In data mining (DM) and machine learning (ML), a human supplies the data and tunes the parameters of used methods. The obtained model is typically the result of several iterative, parameter-tuning steps. This paper aims to improve the interaction between humans and DM and ML programs and, therefore, belongs to the field of interactive DM (IDM) or interactive ML (IML) (the terms are used interchangeably in the literature) [71,73].

The goal of IML is to "help scientists and engineers exploit more of their specialized data" [53]. IML "focuses on methods that empower domain experts to control and direct machine learning tools from within the deployed environment, whereas traditional machine learning does this in the development environment" [53].

The field of IML has recently received a great deal of attention. The preface of the IUI 2013 Workshop on Interactive Machine Learning stated, "Many applications of Machine Learning (ML) involve interactions with humans... a growing community of researchers at the intersection of ML and human–computer interaction are making interaction with humans a central part of developing ML systems. These efforts include applying interaction design principles to ML systems, using

* Corresponding author. Tel.: +386 1 477 3147; fax: +386 1 477 3131.
E-mail addresses: vedrana.vidulin@ijs.si (V. Vidulin), marko.bohanec@ijs.si (M. Bohanec), matjaz.gams@ijs.si (M. Gams).

http://dx.doi.org/10.1016/j.ins.2014.08.014
0020-0255/

human-subject testing to evaluate ML systems and inspire new methods, and changing the input and output channels of ML systems to better leverage human capabilities" [54]. The mission statement of one of the Microsoft research groups dealing with IML notes: "... with the advancement of computational techniques such as machine learning, we now have the unprecedented ability to embed 'smarts' that allow machines to assist users in completing their tasks. We believe that trying to fully automate tasks is extremely difficult and even undesirable, but instead there exists a computational design methodology which allows us to gracefully combine automated services with direct user manipulation" [12].

When supervised DM and ML methods construct models in complex domains, such as economic and social domains, the models often contain *less-credible relations* [1,27,57]. Here, the *relation* is a pattern that connects a set of attributes describing the properties of a concept underlying the data with a class/target attribute, which represents the concept. The term *less-credible* means that the relation is of either low quality or high quality, but is meaningless to a human analyst. *Meaningless* means that a relation's semantic is contradictory to the human's common sense or domain knowledge, and a meaningless state can only be determined by including the human in the DM process. When the task is domain analysis, less-credible relations must be eliminated from the constructed models, since they lead to incorrect conclusions about the most important relations in the domain and can, consequently, undermine the human's trust in the DM system [59].

The problem is illustrated by the example in Fig. 1. The decision-tree model on the right side of Fig. 1 represents a domain model. The tree is constructed from the data (the table on the left side of Fig. 1) using the J48 algorithm in Weka [69] with default parameters. The first three columns, or attributes, of this table represent properties of a person, while the final column, or class, indicates the person's gender. Each row, or example, represents a person. In the tree, the node represents an attribute, and the leaves represent the class. In each leaf, the number in brackets represents the number of examples that reach that leaf. The tree contains a single relation, indicating that a person is a woman if the person has long hair and that a person is a man if the person has short hair. The relation is of high quality, since the tree's accuracy (ACC) is 100%. ACC denotes the overall performance of the tree, expressed as the percentage of correctly classified examples out of all the examples classified by the tree. The relation is meaningless, however, since several men have long hair but are not women (as the left branch of the tree suggests).

The problem that this paper examines most commonly stems from an incompleteness of data [52]. For example, adding more rows and columns to the table in Fig. 1 would likely result in a different relation, but adding the right additional data might be a demanding task. Humans, however, can detect weak relations in domain models using domain knowledge and common sense.

The knowledge that men and women have long and short hair is objective in terms of common sense, as is the case in Fig. 1, but it is hard to take a purely objective position when humans are involved. Humans can also be subjective in terms of fairness; however this discussion is beyond the scope of this paper.

Although the relation in Fig. 1 is of high quality, its meaninglessness makes it less credible.

Another example was obtained through DM in a real-life domain. The decision-tree model presented in Fig. 2 is constructed with the J48 algorithm in Weka using the default parameters and a minimum number of instances per leaf (MNIL) of 5. The tree is constructed from a data set composed of 37 attributes describing the research and development (R&D) sector of a country, 167 examples representing countries and the class that differentiates countries according to their economic welfare into "low", "middle" and "high" (see Section 4.1 for more information on this data set). In the tree, the subtrees form the relations. In each leaf, the first number in brackets represents the number of examples that reach that leaf. The second number represents the number of the examples of the class value other than the one represented by the leaf. The quantities are expressed in decimals to account for the weights of the examples with missing values.

The tree contains three interesting relations. The first is that countries with better welfare invest extensively in R&D. The relation contains attribute "GERD per capita (PPP$)" (GERD stands for Gross Domestic Expenditure on R&D and PPP$ for purchasing power parity in American dollars), which represents the level of investment in R&D. This relation appears twice in the tree. Both times, the "higher than" side of the subtree (>10.8 and >105.5) leads to leaves representing welfare better than that on the "less than" side. One could conclude that the first relation is a valid candidate for a *credible relation* in the tree because it is *meaningful*; that is, it is in accordance with domain knowledge [63] and common sense, it appears twice in the tree and, both times, it makes a clear distinction between countries with different levels of welfare. This relation is marked in bold in Fig. 2. The second relation—"Sector investing the most in R&D" (the right subtree)—seems to be meaningless, since all but one of the leaves represent the class "high", and the single "middle" leaf represents the countries for which the sector is unknown ("N/A" value). Therefore, the entire subtree can be replaced with a single node: "high". A detailed analysis shows

Wears a skirt?	Has long hair?	Has a woman-like body shape?	Sex of a person?
yes	yes	yes	woman
no	yes	no	woman
yes	yes	yes	woman
yes	yes	no	woman
no	no	no	man
no	no	no	man

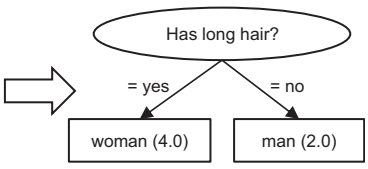

Fig. 1. An example of a domain model ("women have long hair and men do not"), correctly constructed from incomplete data.