Applying Trusted Knowledge in Evaluation Phase of Data Mining

Viktor Nekvapil

Department of Information and Knowledge Engineering, FIS VSE Praha
nam. W. Churchilla 4
130 67 Praha 3

viktor.nekvapil@vse.cz

Abstract. New concept of Trusted Knowledge (TK) is introduced. Trusted Knowledge are data from trusted organizations such as ministries, statistical offices and so on which can replace domain expert in the evaluation phase of the data mining task. The approach called "A/TK-formulas" enables to filter out resulting patterns which are consequences of Trusted Knowledge and thus enables user to concentrate on interesting ones. Conversely, user can request to show only resulting patterns which are consequences of TK to see which of them are in line with TK. The third option enables to request patterns which are in contradiction to the TK. Further new features of Trusted Knowledge framework are introduced in this paper – Trusted Knowledge for mining histograms and Trusted Knowledge hints.

Key words: Trusted Knowledge, evaluation of data mining.

1 Introduction

The approach presented in this paper incorporates additional knowledge in the evaluation phase of data mining but avoids lengthy and complex task of building a belief system of the user (see e.g. [4], [7], more recently in [2]). The idea is to enhance user's domain knowledge using available trusted sources of data – that is data from trusted organisations such as statistical offices, ministries and so on. I refer to this knowledge as ***Trusted Knowledge***. The Trusted Knowledge Framework has been introduced in [3]. In this paper, new features are presented.

The concept of Trusted Knowledge is inspired by FOFRADAR framework [5]. FOFRADAR is based on a logical calculus of association rules. The interpretation is based on mapping important items of knowledge to the sets of association rules which can be considered as their consequences. Important items of knowledge are expressed using a simple mutual influence among attributes. These are predefined relationships of attributes which are used to determine whether the association rule can be seen as a consequence of the item of knowledge or not. For example, the simple mutual influence (SI-formula) *Income* ↑↑ *Loan* means: "*if Income increases, then Loan increases as well*". The set of atomic consequences of this SI-formula can be expressed by the

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 198-203.

following union: $Low_{Income} \times Low_{Loan} \cup Medium_{Income} \times Medium_{Loan} \cup High_{Income} \times High_{Loan}$, saying that "*if Income is high, then Loan is high or if Income is medium then Loan is medium or if Income is high then Loan is high*". Based on the *levels* in the union, it is possible to say whether the resulting rule is a consequence of the defined *SI-formula* or not. This feature is used in the proposed framework and further developed, as obvious in the following sections.

2 Trusted Knowledge

I define Trusted Knowledge as follows: ***Trusted Knowledge*** (TK) is the data from trusted sources which can be connected to the results of a data mining task and are used in the evaluation phase of the data mining task to help with the understanding of the results. Trusted Knowledge can be seen as special case of domain knowledge.

Trusted Knowledge is obtained from a trusted organisation. An example of such knowledge is average and median income per district in the Czech Republic obtained from Czech Statistical Office [1].

Measure of Trusted Knowledge (measure of TK) is a formalised piece of Trusted Knowledge. An example of the measure of TK is depicted in **Table II**. Basic feature of measure of TK is its close connection to the results of a data mining task (resulting patterns). I use association rules as an example. Geographical dimension (locality) is used as a *connecting element* between measure of TK and resulting patterns. *An average income in District X* as a measure of TK and *The loan amount taken by a client in District X* as an attribute from analysed data can be examples of such a connection.

To distinguish between data and *Trusted Knowledge*, I use term *attribute* for variables derived from analysed data and *measure of TK* for variables used as *Trusted Knowledge*. Note that both measure of TK and the attribute connected via *connecting element* are ordinal.

Levels of measures of TK enables us to easily compare attributes and measures of TK. The way how domain experts evaluate the found patterns is commonly expressed by easily interpretable phrases saying for example *"Income is low"*, *"Amount is high"* and so on. Recall the set of atomic consequences of SI-formula *Income* $\uparrow\uparrow$ *Loan*: $Low_{Income} \times Low_{Loan} \cup Medium_{Income} \times Medium_{Loan} \cup High_{Income} \times High_{Loan.}$ Now we have to define, what means for example *"Income is low"* (that is to define the level Low_{Income}).

Expert-based approach means that domain expert decides which category is assigned to each *level*. ***Rank-based approach*** is the newly proposed way of automatic definition of *levels*. Categories of a particular attribute or measure of TK are sorted from the lowest to the highest. Then, we assign rank to each of the category according to the value of attribute or measure of TK. Last step comprises of assigning *Level(l)* to each rank. For example, consider the categories of attribute *Loan_amount* depicted in **Table I**. Based on the rankings of the categories, we can assign respective categories to levels. Having the levels of attributes and measures of TK defined, we can compare levels and draw consequences based on values of the levels. This is further elaborated upon in section 3.

Table I: Levels for attribute Loan amount

Loan_amount	Rank	Level
<0; 100000)	1	Very low
<100000; 150000)	2	Very low
<150000 ;200000)	3	Low
…	…	…
<500000; 550000)	8	High
<550000; 650000)	9	Very high
<650000; 2600000>	0	Very high

Table II: Levels for measure of TK Income

District	Income	Rank	Level
Hlavni mesto Praha	35 115	1	Very high
Stredocesky kraj	27 345	2	Very high
Jihomoravsky kraj	26 116	3	Very high/High
Plzensky kraj	26 026	4	High
…	…	…	…
Pardubicky kraj	24 067	12	Low/Very low
Zlinsky kraj	23 873	13	Very low
Karlovarsky kraj	22 707	14	Very low

2.1 Applying Trusted Knowledge

One of the possible solutions of the automatic formulation of conclusions using domain knowledge is presented in the FOFRADAR framework, as described above.

Using the measures of TK, we can define mutual influence between an attribute and measure of TK. I call this mutual influence ***Attribute / Trusted Knowledge-formula* (*A/TK-formula*)**. The principle of A/TK-formula is the same as for SI-formulas in FOFRADAR, but instead of one of the attributes, measure of TK is used in the mutual influence.

The proposed framework works as follows: After the results are obtained, Trusted Knowledge repository is queried for A/TK-formulas which are available and are relevant for the resulting patterns. Afterwards, A/TK-formulas can be applied. In [3], two ways how the consequences of A/TK-formulas can be applied are presented:

1. to obtain patterns which are consequences of A/TK-formula – this way is useful when the user wants to know which resulting patterns are in line with the overall knowledge (trusted knowledge);
2. to filter out patterns which are consequences of A/TK-formula – this way the user can filter out resulting patterns which are in line with Trusted Knowledge and concentrate on patterns which are not consequences of TK;

In this paper, I introduce the third possible way how the consequences of A/TK-formulas can be applied:

3. to obtain patterns which are in contradiction to the A/TK-formula – this way the user can obtain only rules which are in contradiction to the A/TK formula and concentrate on this resulting patterns (exceptions).

As an example, let us discuss the A/TK-formula *Income* ↑↑ *Loan amount*. *Income* is a measure of TK. Using *rank-based approach*, it is possible to assign values to respective levels, as shown in **Table II**. The categories of the attribute *Loan amount* can be

assigned to the levels, as depicted in **Table I**. Then the set of consequences of the A/TK-formula *Income* ↑↑ *Loan amount* is defined by the following union:

Very low$_{INCOME}$ × *Very low*$_{LOAN}$ ∪ *Low*$_{INCOME}$ × *Low*$_{LOAN}$ ∪ *Medium*$_{INCOME}$ × *Medium*$_{LOAN}$ ∪ *High*$_{INCOME}$ × *High*$_{LOAN}$ ∪ *Very high*$_{INCOME}$ × *Very high*$_{LOAN}$.

To obtain patterns which are in contradiction to the A/TK-formula *Loan amount* ↑↑ *Income* (way 3 above), I modify the union in the following way: *Very low*$_{INCOME}$ × *Very high*$_{LOAN}$ ∪ *Low*$_{INCOME}$ × *High*$_{LOAN}$ ∪ *High*$_{INCOME}$ × *Low*$_{LOAN}$ ∪ *Very high*$_{INCOME}$ × *Very Low*$_{LOAN}$

Note that medium Levels of *Income* and *Loan amount (Medium*$_{INCOME,}$ *Medium*$_{LOAN}$*)* are not present in the union, because they appear together in consequences of A/TK-formula *Loan amount* ↑↑ *Income* and thus they cannot be part of contradictions of A/TK-formula *Loan* amount ↑↑ *Income*.

Main difference between way 2 and 3 is the fact that in 3, we explicitly obtain rules which are in contradiction with the A/TK-formula while in way 2, we obtain rules which are not consequences of A/TK-formula (meaning that also rules with no relation to the A/TK-formula are present).

As an example, let us use the resulting association rule: `District (Zlinsky kraj) -> Loan amount (<100000; 150000)`. Level of *Loan amount* is '*very low'*, connecting element *District* with value *Zlinsky kraj* is used to link the rule to the measures of TK *Income*. If one looks at the level of the measure *Income*, it is *very low* according to **Table II**. So we can conclude that this rule is consequence of the A/TK-formula *Income* ↑↑ *Loan amount* and is not a contradiction of A/TK-formula *Income* ↑↑ *Loan amount.* We can determine the relationship of each rule to the three ways mentioned above.

Further newly defined features of Trusted Knowledge framework include Trusted Knowledge for mining histograms and Trusted Knowledge hints.

Trusted Knowledge for mining histograms

Data mining with histograms has been introduced in [6] using the CF-Miner procedure of the LISp-Miner system. In a simplified manner, the task is to find 'interesting' histograms. Each histogram *Hsg* is in a form *Hsg*(*Attribute*, *Condition*, *Data Matrix*, *Abs/Rel*), where *Condition* is Boolean attribute which each row of the *Data Matrix* must satisfy and *Abs/Rel* states whether the frequencies for *Attribute* are absolute or relative (relative to the overall data matrix without the *Condition*). Furthermore, interestingness measure ≈ called *CF-quantifier* is used to find interesting histograms. For example, a *CF-quantifier* $\approx^{U}_{100,6}$ defines that histogram is interesting if it has at least 100 of objects satisfying *Condition* and there are 6 steps up. That means that 6 consecutive categories of *Attribute* has higher frequency than the previous category.

In [6], domain knowledge is used to filter out resulting histograms which are consequence of defined SI-formula for the *CF-quantifier* ≈ in a similar manner as mentioned above in the FOFRADAR framework. Firstly, a set of atomic consequences of SI-formula for the *CF-quantifier* ≈ needs to be defined. For example, we can define atomic consequences of SI-formula *Price of flat* ↑↑ *Loan amount* for a CF-quantifier $\approx^{U}_{100,6}$ as a set of histograms $\approx^{U}_{100,6}$ *Price of flat / Loan amount(α)* satisfying that level α of *Condition Loan amount* is HIGH or VERY HIGH and the *Attribute Price of*

flat has 7 categories (to ensure that there are only steps up). Levels of *Loan amount* are defined as stated in **Table I**. For example, resulting histogram $\approx^{U}_{100,6}$ *Price of flat / Loan amount (<550000; 650000)* is an atomic consequence of SI-formula *Price of flat* ↑↑ *Loan amount* for *CF-quantifier* $\approx^{U}_{100,6}$. The consequences can be then used for evaluation of the found histograms (for example, filter out histograms which are consequences of SI-formula). The business interpretation of the SI-formula *Price of flat* ↑↑ *Loan* is that if level α of *Loan amount* is HIGH or VERY HIGH, then the level of *Price of flat* will probably also be HIGH or VERY HIGH. Furthermore, agreed consequences are defined in [6], which I do not further discuss here.

Using A/TK-formulas of the Trusted Knowledge framework defined above, we can proceed analogously as in [6] as follows. Let us define atomic consequences of the A/TK-formula *Loan amount* ↑↑ *Income* for *CF-quantifier* $\approx^{U}_{100,9}$ as a set of histograms $\approx^{U}_{100,9}$ *Loan amount / District(α)*, level α being VERY HIGH or HIGH. Then, the resulting histogram $\approx^{U}_{100,9}$ *Loan amount / District(Hlavni mesto Praha)* of relative frequencies is an atomic consequence because the *District Hlavni mesto Praha* in Trusted Knowledge Repository has the level of measure of TK *Income* '*very high*' (see **Table II**). Ways how to apply the consequences of A/TK-formulas for histograms are the same as for association rules mentioned above and are now studied in detail.

Moreover, there could be more flexible ways of applying *CF-quantifiers*. One issue of the *CF-quantifiers* is the fact that they are 'strict' in a sense that if there is one category in a histogram that breaks the overall trend in histogram, the histogram will not be considered as interesting and will not be in resulting histograms. For example, if an attribute in histogram has 6 categories, all but one satisfying the steps up quantifier but the 3rd and 4th category does not satisfy the *CF-quantifier* steps up (but only slightly), the *CF-quantifier* will not be satisfied while from the business perspective, the histogram has the upwards tendency and thus is interesting. Ways how to overcome this issue will be further studied.

Trusted Knowledge hints

Another way how to exploit Trusted Knowledge is to compare a measure of TK and attribute in the analysed data in case they have similar content. For example, it is possible to compare average *Income* presented in the analysed data to the average *Income* as a Trusted Knowledge aggregated to districts. In case the analysed data has sufficient number of objects in each of the groups (groups according to geographical dimension), we can get additional knowledge. For example, we can get following information: The average income of clients in data in district Praha is lower than the average income of people in district Praha of the whole population (as Trusted Knowledge). This can bring us to the deeper investigation of the origin of the data. For example, we can say that in general, affluent clients do not take a consumer loan. If we have a data of clients who took a consumer loan, we can derive that their income will be below the average income in district Praha.

3 Conclusions and future work

First experiments has shown that using all three ways how the consequences of A/TK-formulas can be applied significantly reduce the amount of work user needs to evaluate the resulting rules. This helps the user to concentrate on rules which are interesting from the user's perspective. New features of Trusted Knowledge framework were introduced: Trusted Knowledge for mining histograms and Trusted Knowledge hints. Both features brings new ways of applying Trusted Knowledge.

Another way how to elaborate upon the Trusted Knowledge framework is to study the situation when one pattern is supported by more than one A/TK-formula. Furthermore, different sources of Trusted Knowledge could be evaluated according to their trustworthiness. For example, one source is better than another one from the user's perspective and this information could be further incorporated into the Trusted Knowledge framework. Both features will be further studied.

References

1. Czech Statistical Office (CSO), 2015. Výsledky sčítání lidu, domů a bytů 2011 (Census 2011 – in Czech) [online]. https://www.czso.cz/csu/czso/otevrena_data_pro_vysledky_scitani_lidu_domu_a_bytu_2011_-sldb_2011- Last modified on 14 th April 2015.
2. De Bie, T., 2013. Subjective interestingness in exploratory data mining. In Advances in Intelligent Data Analysis XII: 12th International Symposium, IDA 2013, London, UK, October 17-19, 2013.
3. Nekvapil, V. 2017. Data Mining with Trusted Knowledge. FedCSIS Conference, Prague. 3-6 September 2017. Accepted for publication.
4. Padmanabhan, B., Tuzhilin, A., 1998. A belief-driven method for discovering unexpected patterns. In Proc. of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 94-100, 1998.
5. Rauch, J., 2015. Formal Framework for Data Mining with Association Rules and Domain Knowledge – Overview of an Approach. Fundamenta Informaticae, 137 No 2, pp. 1–47
6. Rauch, Jan, Šimůnek, Milan. Data Mining with Histograms – A Case Study. In: *Foundations of Intelligent Systems* [online]. Lyon, 21.10.2015 – 23.10.2015. Cham : Springer International Publishing, 2015, s. 3–8. ISBN 978-3-319-25251-3. DOI: 10.1007/978-3-319-25252-0.
7. Silberschatz, A., Tuzhilin, A., 1995. On subjective measures of interestingness in knowledge discovery. In Proc. of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 275-281, 1995.

Acknowledgment: The work described here has been supported by the internal grant agency of the University of Economics, Prague under project IGA 29/2016.