Roles of Medical Ontology in Association Mining CRISP-DM Cycle

Hana Češpivová[1], Jan Rauch[1,2], Vojtěch Svátek[1,2], Martin Kejkula[1], Marie Tomečková[3]

[1] Department of Information and Knowledge Engineering, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
{rauch,svatek,kejkula}@vse.cz
[2] European Centre for Medical Informatics, Statistics and Epidemiology – Cardio University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
[3] European Centre for Medical Informatics, Statistics and Epidemiology - Cardio, Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic
tomeckova@euromise.cz

Abstract. We experimented with introduction of medical ontology and other background knowledge into the process of association mining. The inventory used consisted of the *LISp-Miner* tool, the UMLS ontology, the STULONG dataset on cardiovascular risk, and a set of simple qualitative rules. The experiment suggested that an ontology may bring benefits to all phases of the KDD cycle as described in CRISP-DM.

1 Introduction

With growing amounts of medical knowledge available in computer-processable form, its exploitation in the KDD process becomes highly desirable. Domain *ontologies*, being hot topic in today's knowledge engineering research, are the most promising candidates: they express domain concepts and relationships in a way that is consensual and comprehensible to the given professional community. The role to be played by ontologies in KDD (and even their mere usability) depends on the given mining *task and method*, on the *stage of the KDD process*, and also on some characteristics of the *dataset*. The experiment described in this paper is connected with *association mining*, namely, with the *LISp-Miner* tool [20, 23]. As a generally-acceptable guideline for distinguishing various uses of ontologies, we opted for *CRISP-DM* [1], the most widespread methodology describing the steps of the KDD process. Finally, we adopted the medical *dataset STULONG*, previously used as benchmark data for various KDD tools.

Section 2 attempts to characterise the potential roles of ontology in the CRISP-DM cycle. Section 3 introduces the *LISp-Miner* tool and explains the main principles of its *4ft-Miner* procedure. Section 4 describes an experiment in the domain of cardio-vascular risk, focused on background knowledge (general UMLS ontology and a qualitative rule base) exploitation. Section 5 evaluates the experiment and attempts to draw lessons from it. Section 6 surveys related work. Finally, section 7 wraps up the paper and outlines future perspectives.

2 Domain Ontology in the CRISP-DM Lifecycle

The CRISP-DM model [1] distinguishes six main phases of a KDD process: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. In this section, we suggest some general roles for ontologies in every phase. Each phase will later (section 4) be instantiated with our experiments with association mining in the cardiovascular risk domain.

1. The role of ontologies in *Business Understanding* is not peculiar to KDD. Domain ontologies are an important vehicle for inspecting a domain prior to committing to a particular task. Semi-formal ontologies can help a newcomer get familiar with most important concepts and relationships, while formal ontologies allow to identify conflicting assumptions that might not be obvious at the first sight.
2. For improved *Data Understanding*, elements of an ontology have to be (presumably, manually) mapped on elements of the data scheme and vice versa. This will typically lead to selecting a relevant part of an ontology (or, multiple ontologies) only. The benefits of this effort might be e.g.:
   - identification of missing attributes that should be added to the data set
   - identification of redundant attributes (e.g. measuring the same quantity in different units) that could be deleted from the dataset.
3. The *Data Preparation* phase is already connected with the subsequent Modelling phase. Concrete use of domain ontology thus partially depends on the chosen mining tool/s. An ontology may typically help by identifying multiple groupings for attributes and/or values according to semantic criteria.
4. In the *Modelling* phase, ontologies might help design the individual mining sessions. In particular for large datasets, it might be worthwhile to introduce some ontological bias, e.g. to skip the quantitative examination of hypotheses that would not make sense from the ontological point of view, or, on the other hand, of too obvious ones.
5. In the *Evaluation* phase, the discovered model/s have the character of structured knowledge built around the concepts (previously mapped on data attributes), and can thus be interpreted in terms of ontology and associated background knowledge.
6. In the *Deployment* phase, extracted knowledge is fed back to the business environment. Provided we previously modelled the business using ontological means, the integration of new knowledge can again be mediated by the business ontology. Furthermore, if the mining results are to be distributed across multiple organisations (say, using the semantic web infrastructure), mapping to a shared ontology is inevitable.

Presumably, in particular in the middle part of the KDD process, ontologies may be used with different scenarios, going beyond the presented framework and conforming to the mining methods used[4]. In section 4, we will instantiate the general framework with concrete experience connected to association mining.

[4] The framework is at least restricted to 'tabular' data mining, i.e. mining in tables containing values for a fixed set of attributes.

3 Association Mining with *4ft-Miner*

3.1 *LISp-Miner* Procedures

The core of *LISp-Miner* are *procedures* for analysis of data represented as strings of bits. In this way, it is possible to efficiently generate and verify various patterns. The most widely used *4ft-Miner* procedure relies on analysis of *four-fold contingency table*; it mines for 16 types of *association rules* (ARs) including ARs corresponding to statistical hypothesis test and conditional ARs [19]. Let us also mention the procedure *KL-Miner* for patterns based on evaluation of two-dimensional contingency tables of two categorical attributes [21], and the procedure *SDS-Miner*, which mines for couples of disjoint subsets of observed objects that differ in some attribute property [14]. For the sake of this paper, we will only describe *4ft-Miner*, since we used this procedure in the 'ontological' experiment.

3.2 Structure of *4ft-Miner* Relevant Questions

4ft-Miner mines for ARs of the form $\varphi \approx \psi$ where φ and ψ are called *antecedent* and *succedent*, respectively[5]. The symbol $\approx$ stands for *4ft-quantifier*, i.e. a condition over the four-fold contingency table of φ and ψ. The *four-fold contingency table* of φ and ψ in data matrix $\mathcal{M}$ is a quadruple $\langle a, b, c, d\rangle$ of natural numbers such that a is the number of data objects from $\mathcal{M}$ satisfying both φ and ψ, b is the number of data objects from $\mathcal{M}$ satisfying φ and not satisfying ψ, c is the number of data objects from $\mathcal{M}$ not satisfying φ and satisfying ψ, and d is the number of from $\mathcal{M}$ from $\mathcal{M}$ satisfying neither φ nor ψ.

There are 16 4ft-quantifiers in the *4ft-Miner*. An example is *above-average dependence* $\sim^{+}_{p,Base}$, defined for $p > 0$ and $Base > 0$ by the condition

$$\frac{a}{a+b} \;\geq\; (1+p)\,.\,\frac{a+c}{a+b+c+d} \;\wedge\; a \geq Base\;.$$

$\varphi \sim^{+}_{p,Base} \psi$ hence means that among the objects satisfying φ there are at least $100.p$ per cent more objects satisfying ψ than among all observed objects, and that there are at least $Base$ observed objects satisfying both φ and ψ.

The input of *4ft-Miner* consists of the data matrix and of several parameters defining the set of ARs to be automatically generated and tested. All properties of such AR are incorporated into the so-called *relevant questions* [12]. The output of *4ft-Miner* consists of all *prime* ARs, i.e. all ARs that are true in the data matrix and do not immediately follow from simpler ARs also output. Defining the set of relevant questions amounts to defining the set of relevant antecedents and succedents and to choosing the 4ft-quantifier. Antecedent and succedent are conjunctions of *literals*. Literal is a Boolean variable $A(\alpha)$ or its negation $\neg A(\alpha)$, where α (a set) is *coefficient* of the literal $A(\alpha)$.

[5] *4ft-Miner* also mines for *conditional* hypotheses (i.e. with a third symbol representing a restrictive condition). We will not discuss them here, for brevity.

As an example of AR, let us present the expression

$$A(a_1, a_7) \wedge B(b_2, b_5, b_9) \sim^{+}_{p,Base} C(c_4) \wedge \neg D(d_3) \ .$$

Here, $A(a_1, a_7)$, $B(b_2, b_5, b_9)$, $C(c_4)$ and $\neg D(d_3)$ are literals, a_1 and a_7 are categories of A, and $\{a_1, a_7\}$ is the coefficient of $A(a_1, a_7)$[6], and analogously for the remaining literals.

In order to determine the set of relevant questions more easily, we can define *cedents* (i.e. antecedent and/or succedent) φ as a conjunction

$$\varphi = \varphi_1 \wedge \varphi_2 \wedge \ldots \wedge \varphi_k$$

where $\varphi_1, \varphi_2, \ldots, \varphi_k$ are *partial cedents*. Each φ_i is chosen from one *set of relevant partial cedents*.

The set of partial cedents is given in the following manner:

- the minimum and maximum *length* of the partial cedent is defined
- a set of *attributes* from which literals will be generated is given
- some attributes can be marked as *basic*, each partial cedent then must contain at least one basic attribute
- a simple definition of the set of all *literals* to be generated is given for each attribute
- *classes of equivalence* can be defined, such that each attribute belongs to at most one class of equivalence; no partial cedent then can contain two or more attributes from the same class of equivalence.

The *length of the literal* is the number of categories in its coefficient. The set of all literals to be generated for a particular attribute is given by:

- the type of coefficient; there are six types of coefficients: subsets, intervals, left cuts, right cuts, cuts, one particular category
- the minimum and the maximum length of the literal
- positive/negative literal option:
  - only generate positive literals
  - only generate negative literals
  - generate both positive and negative literals.

It is clear that the design of *4ft-Miner* relevant questions is a demanding task. This was also one of motivations for bringing ontologies into the game.

4 Experiment in the Domain of Cardio-Vascular Risk

Our focus on the medical domain, in particular, on the sub-domain of cardiovascular diseases, stems from our involvement in the EuroMISE Centre – Cardio, a government-funded co-operation of multiple academic institutes and hospitals. Within this centre, many medical datasets as well as knowledge bases are collected, shared and analysed.

In this section, we will trace our small experiment along all CRISP-DM phases, and thus instantiate the general framework from section 2.

[6] For convenience, we can write $A(a_1, a_7)$ instead of $A(\{a_1, a_7\})$.

4.1 'Business Understanding' via Medical Ontologies

Ontological resources of medical knowledge typically have the form of *terminological taxonomies* such as the *International Classification of Diseases* (ICD) [3], *MeSH* [4] or *Snomed* [5]. Most of them are subsumed by *UMLS* (Unified Medical Language System) [7], which consists of a high-level *semantic network* and a *meta-thesaurus* mapping the concepts picked from third-party resources onto each other. Although the central construct of UMLS is the concept-subconcept relation, the semantic network also features lots of other binary relations such as 'location of' or 'produces'. However, since the network only covers 134 high-level 'semantic types' (such as 'Body Part' or 'Disease'), the relations are only 'potentially holding' (it is by far not true that every Body Part can be 'location of' every Disease...). The meta-thesaurus, in turn, covers (a large number of) more specific concepts but relations are only scarcely instantiated, and nearly all relation instances belong to the 'location of' relation. Few medical ontologies go significantly beyond the terminological level. A slightly richer in structure is the *Foundational Model of Anatomy* [2], which is limited to anatomical concepts but covers concrete instances of e.g. 'Body Part' or 'Body Liquid' and adequate relations among them. There are also a few deeper medical ontologies containing formal axioms, such as the medical ontology developed in the *ONIONS* project [11], they however only cover a fraction of important domain concepts.

Aside, there are domain knowledge resources that are not centred around concepts but still in some sense *consensual*. This is the case of knowledge accumulated in the Czech medical community with respect to risk factors of cardiovascular diseases, in connection with the STULONG project described below. The knowledge base consists of 36 qualitative rules, most of which can be characterised as medical background knowledge or common-sense knowledge, e.g. "increase of cholesterol level leads to increase of triglycerides level", "increase of age leads to increase of coffee consumption", "increase of education leads to increase of responsibility in the job" or the like. Only a few of them can be interpreted as causalities. Yet, in the absence of 'truly ontological' inter-concept relationships, we can use them, provisionally, in the role of ontological relations.

Procedure and benefits: Since most authors of this paper were not medical experts, reference to the mentioned resources (in particular, to UMLS and to the knowledge base on cardio-vascular risk factors) was important throughout the whole experiment, as vehicle for 'business understanding'.

4.2 Data Understanding: Mapping the STULONG Dataset on Medical Concepts

The *STULONG dataset*, which we chose for our experiments, concerns a twenty-years-lasting longitudinal study of risk factors for atherosclerosis in the population of middle-aged men (see `http://euromise.vse.cz/stulong-en/`). The dataset is popular in the KDD community thanks to its role of benchmark data in the Discovery Challenge (2002-2004, see `http://lisp.vse.cz/challenge/`) held within ECML/PKDD conferences. It consists of four data matrices:

Entrance. Each of 1 417 men has been subject to entrance examination. Values of 244 attributes have been surveyed with each patient. These attributes are divided into 11 groups e. g. *social characteristics*, *physical activity* etc.

Control. Risk factors and clinical demonstration of atherosclerosis have been followed during the control examination for the duration of 20 years. Values of 66 attributes have been recorded for each one. There are 6 groups of attributes, e.g. *physical examination*, *biochemical examination* etc.

Letter. Additional information about health status of 403 men was collected by postal questionnaire. There are 62 attributes divided into 8 groups such as *diet* or *smoking*.

Death. There are 5 attributes concerning the death of 389 patients.

Procedure: Mapping the STULONG data on UMLS concepts required to bridge the semantic gap between a data schema and a terminological ontology. We succeeded in mapping 53 of STULONG attributes (from the Entrance dataset) on 19 UMLS semantic types and 25 metathesaurus concepts. Six attributes for which a concept could not be found were only assigned semantic type, for example, 'responsibility in job' was assigned to semantic type Occupational Group. For subsequent processing, we only kept a light-weight fragment of UMLS, containing, for each data attribute, the most adequate metathesaurus concept and the least-general semantic type subsuming this concept. We obtained a structure, to be called STULONG_UMLS, with five taxonomy roots: *Finding*, *Activity*, *Group*, *Food*, and *Disease or Syndrome*.

The mapping between STULONG data and the background knowledge base on cardio-vascular risk factors was straightforward, since the data were collected (more-or-less) by the same community of physicians who also formulated the knowledge base, within the same project.

Benefits: The mapping of data to ontology helped us to identify *redundant attributes*, which, although necessary for data management purposes, were not useful as input to data mining. For example, since the dataset contained the attribute 'age on entrance to STULONG study', the attributes 'birth year' and 'year of entrance to STULONG study' (all of these being mapped to the Age Group semantic type) were of little use.

4.3 Data Preparation: Building the Partial Cedents

Using the mapping on ontology concepts from the previous phase, we can identify attributes that should be semantically grouped into partial cedents (Fig. 1).

Procedure: We created partial cedents covering the attributes mapped on the five upmost classes of UMLS_STULONG. Although we carried out this part of the task manually, it could easily be automated.

Benefits: We could see in section 3 that the apparatus of *4ft-Miner* analytic questions is rather complex. The effort (once) invested to mapping data attributes to ontology concepts in the Data Understanding phase could (repeatedly) pay off in the Data Preparation phase, by reducing the effort of building relevant analytic questions.

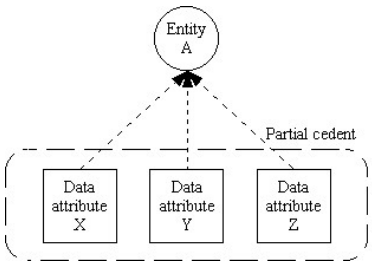

Fig. 1. Construction of a partial cedent corresponding to an ontology concept

4.4 Modelling: Introducing Ontological Bias to Analytic Questions

The mining process in the most narrow sense—*running* an individual mining session—is probably not amenable to ontologies in the case of *LISp-Miner*. Fast analysis of large data tables using the bitstring approach relies on optimised database-oriented algorithms, which could hardly accommodate the heterogeneity of ontological information. On the other hand, there is room for ontologies in the process of *designing* the individual sessions.

Procedure: A very general mining task setting can be decomposed into more specific tasks, which can be run faster, their results will be conceptually more homogeneous, and thus can be interpreted more easily (see below). As an example of task decomposition (for associations between activities of the patient and diseases/syndromes) is at Fig. 2. The base task (left branch) might lead to a high number of hypotheses that would be hard to interpret. We can thus e.g. separately refine the antecedent (middle branch) or succedent (right branch) of the base task, to obtain a reasonable amount of results per session.

Benefits: Presumably, the hierarchical decomposition of mining tasks will not pay off in the Modelling phase itself, but rather in the subsequent Result Evaluation phase. Namely, the collections of hypotheses entering the next phase will be smaller and homogeneous, hence easier to examine for a human evaluator.

4.5 Result Evaluation: Comparison with Prior Knowledge

Given the data-to-ontology mapping, concrete associations discovered with the help of *4ft-Miner* can be matched to corresponding semantic relations from the ontology, see Fig. 3. The semantic relation represents a potential context (e.g. explanation) for the discovered association.

Procedure and benefits: Since the data were prepared with respect to ontological concepts, each mining task corresponds to a meaningful 'framing' question, such as "Searching for relationships between Activity of the patient and Diseases/Syndromes". Concrete associations discovered with the help of *4ft-Miner* can be then compared with the more specific layer of the ontology; in our case, with additional background knowledge. The relationship of an association with prior knowledge will typically be one of the following:

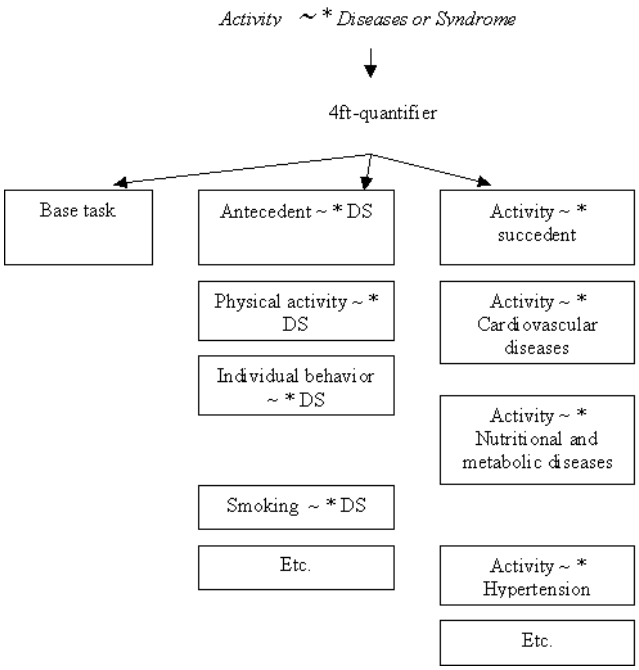

Fig. 2. Decomposition of 4ft tasks with respect to ontology

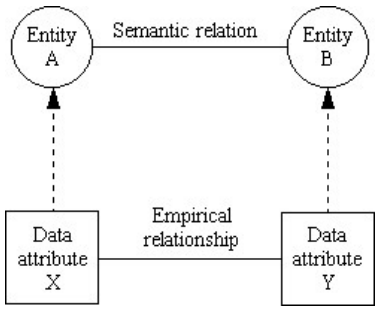

Fig. 3. Semantic relation as context to empirical relationship between attributes

Table 1. Four-fold table for a 'confirmation' association

	Succedent	NOT Succedent	
Antecedent	10	22	32
NOT Antecedent	66	291	357
	76	313	389

Table 2. Four-fold table for a 'conflicting' association

	Succedent	NOT Succedent	
Antecedent	216	14	230
NOT Antecedent	145	14	159
	361	28	389

- Confirmation of prior knowledge, without additional information
- New knowledge compatible with prior knowledge
- Exception to or conflict with prior knowledge.

We have not yet implemented an automated tool for comparison of new associations to prior knowledge, and only made some sample comparisons by hand. Let us show two examples, with their four-fold tables:

- The discovered association "Patients who are not physically active within the job nor after the job (Antecedent) will more often have higher blood pressure (Succedent)" was derivable from the background knowledge rule "Patients who are physically active after the job will more often have lower blood pressure" (Table 1).
- The discovered association "94% of patients smoking 5 or more cigarettes a day for more than 21 years (Antecedent) have neither myocardial infarction nor ictus nor diabetes (Succedent)" was in conflict with prior knowledge "Increase of smoking leads to increase of cardio-vascular diseases" (Table 2).

The examples are merely illustrative. In order to draw medically valid conclusions from them, we would at least need to examine the statistical validity of the hypotheses in question. As the STULONG dataset is relatively small, few such hypotheses actually pass conventional statistical tests.

4.6 Deployment: Shareable Analytic Reports

The role of ontology in the deployment phase is most crucial if the mining results are to be supplied to a wider (possibly unrestricted) range of consumer applications. A promising approach would be to incorporate the mining results into *semantic web* documents. In 2000-2002[7], we (semi-manually) created a collection of analytic reports based upon results of *4ft-Miner* tasks on STULONG

[7] Solely for this phase, we refer to earlier work rather than to the current experiment.

data. Furthermore, we suggested a way to embed the formal representation of *4ft-Miner* results themselves, i.e. ARs, into the text of the reports [15]. For this purpose, we used an original language conforming to RuleML specifications [6], which act as de facto standard for rules on the semantic web. A natural next step would be to combine such rules with an ontology (mapped on data attributes, cf. section 4.2), as proposed by the new Semantic Web Rule Language [13].

5 Evaluation of Resources Used

In this section, we try to assess to what extent the dataset and ontology were suitable for this kind of experiment. Such assessment is important for generalising the outcomes of the experiment as well as for setting directions for future work.

1. The STULONG dataset is only medium-sized. This makes it suitable for experiments with knowledge-engineering aspects, requiring human intervention. On the other, benefits of background knowledge in the *modelling phase* thus diminish, since *LISp-Miner* is able to explore the entire space of hypotheses by brute force. A real disadvantage of the STULONG dataset however is the fact that it covers areas that are *conceptually disparate*, such as laboratory tests on blood and urine, physical measurements (such as skinfold thickness) as well as aspects of patient's behaviour (ways of transport) and even social status. All of these are to some extent covered by UMLS; however, picking distant 'knowledge islands' from a large resource may result in poor quality and usability of the resulting model.
2. The deficiency of UMLS, namely, the lack of concrete *relation instances*, seems to be more serious. We only partially solved this problem using external background knowledge. Another problem we experienced with UMLS, similar to those observed e.g. in [18], is insufficient commitment to uniform semantic modelling patterns. For example: concepts belonging to semantic type Group may 'exhibit' Alcohol Consumption as concept belonging to semantic type Individual Behaviour. However, a Group may also be related by relation 'consume' to Alcoholic Beverages (or e.g. Beer) as concept belonging to semantic type Food. There is no logical connection between these two constructions in UMLS, except that Alcohol Consumption is listed among the 'Other related concepts' for Alcoholic Beverages and vice versa.

We can conclude that both resources were usable and compatible in principle. Experiments run on them however only provide preliminary hypotheses that should be verified and possibly revised based on more comprehensive resources.

6 Related Work

Although domain ontologies are a popular instrument in many diverse applications, they only scarcely appeared in 'tabular' KDD. An exception is [17], where

'common-sense' ontologies of time and processes were exploited to derive constraints on attributes, which were in turn used to construct new attributes. Not explicitly talking about ontologies, [10] used qualitative models as bias for inductive learning. Finally, [24] and [25] used problem-solving method descriptions (a kind of 'method ontologies') for the same purpose. There have also been efforts to employ taxonomies over domains of individual attributes [8, 9, 16, 22] to guide inductive learning. None of these projects however attempted to systematically map the possible roles of domain ontology throughout the KDD process.

7 Conclusions and Future Work

We outlined the possible roles of a domain ontology in the phases of the CRISP-DM process, and illustrated them on an experiment in the domain of cardiovascular risk, in the context of association mining and with UMLS as start-up ontology. Although the experiment had limited extent, it revealed many opportunities for ontologies in (at least, association mining) KDD process, as well as many problems tied to the characteristics of medical ontologies and/or datasets.

In contrast to the highly-optimised mining processes of *LISp-Miner*, the incorporation of ontology information has so far been done mostly manually. Further effort is thus needed to complement the *LISp-Miner* architecture with *automated tools* for ontology handling. We also want to explore *different types of medical ontologies*, possibly with higher ontological accuracy than UMLS. We would also like to experiment, in a similar spirit, with *other LISp-Miner procedures* such as KL-Miner or SDS-Miner mentioned in section 3.1. Finally, our long-term goal is to develop an extended architecture (nicknamed EverMiner) that would be capable to design (to some extent) automatically *new relevant questions* based on results of previous mining sessions.

The research is partially supported by project no.LN00B107 of the Ministry of Education of the Czech Republic and by grant no.2003/23 of the Internal Grant Agency of the University of Economics, Prague. The STULONG study was carried out at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital (head Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudík, MD, ScD, with collaboration of M. Tomečková, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. The data were transferred to electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head Prof. RNDr. J. Zvárová, DrSc).

References

1. CRoss Industry Standard Process for Data Mining, `http://www.crisp-dm.org/`.
2. Foundational Model of Anatomy, `http://sig.biostr.washington.edu/projects/fm/`.
3. ICD-10, The International Statistical Classification of Diseases and Related Health Problems, tenth revision, `http://www.who.int/whosis/icd10/`.
4. Medical Subject Headings, `http://www.nlm.nih.gov/mesh/meshhome.html`.

5. SNOMED Clinical Terms, http://www.nhsia.nhs.uk/snomed.
6. The Rule Markup Initiative, http://www.ruleml.org/.
7. Unified Medical Language System, http://www.nlm.nih.gov/research/umls/.
8. Almuallim, H., Akiba, Y. A., Kaneda, S.: On Handling Tree-Structured Attributes in Decision Tree Learning. In: Proceedings of the Twelfth International Conference on Machine Learning (ML-95). Morgan Kaufmann, 12-20.
9. Aronis, J.M., Provost, F.J., Buchanan, B.G.: Exploiting Background Knowledge in Automated Discovery. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996 (KDD-96).
10. Clark, P. Matwin, S.: Using Qualitative Models to Guide Inductive Learning. In: Machine Learning - ECML'94, European Conference on Machine Learning, Catania 1994. Lecture Notes on Artificial Intelligence, Springer Verlag 1994, 360-365.
11. Gangemi, A., Pisanelli, D.M., Steve, G.: An Overview of the ONIONS project: Applying Ontologies to the Integration of Medical Terminologies. *Data and Knowledge Engineering*, vol.31, 1999.
12. Hájek, P., Havránek, T.: Mechanizing Hypothesis Formation (Mathematical Foundations for a General Theory), Springer–Verlag 1978.
13. Horrocks I. et al. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. http://www.daml.org/2004/04/swrl/.
14. Karban T., Rauch J., Šimůnek, M.: SDS-Rules and Association Rules, accepted for ACM Symposium on Applied Computing (SAC 2004).
15. Lín, V., Rauch, J., Svátek, V.: Content-based Retrieval of Analytic Reports. In: Schroeder, M., Wagner, G. (eds.). Rule Markup Languages for Business Rules on the Semantic Web, Sardinia 2002, 219-224.
16. Núñez, M.: The Use of Background Knowledge in Decision Tree Induction. *Machine Learning*, 6, 231-250 (1991).
17. Phillips, J., Buchanan, B.G.: Ontology-guided knowledge discovery in databases. In: International Conf. Knowledge Capture, Victoria, Canada, 2001.
18. Pisanelli, D.M., Gangemi, A., Steve, G.: An Ontological Analysis of the UMLS Metathesaurus. Proc. AMIA 98, 1998.
19. Rauch, J.: Interesting Association Rules and Multi-relational Association Rules. In: Lee, H. C., Lai, F. (eds), Communications of Institute of Information and Computing Machinery, IICM, Taiwan, 2002.
20. Rauch, J., Šimůnek, M.: Alternative Approach to Mining Association Rules. In: Lin T Y, Ohsuga S (eds) The Foundation of Data Mining and Knowledge Discovery (FDM02), IEEE 2002.
21. Rauch, J., Šimůnek, M., Lín, V.: Mining for Patterns Based on Contingency Tables by KL–Miner, First Experience. In: Lin T Y, Ohsuga S, Liau C J (eds) Foundations and New Directions of Data Mining (FDM03), IEEE 2003.
22. Svátek, V.: Exploiting Value Hierarchies in Rule Learning. In: ECML'97 Poster Papers. Prague 1997, 108-117.
23. Šimůnek, M: Academic KDD Project LISp-Miner. In: Abraham, A. et al (eds), Advances in Soft Computing - Intelligent Systems Design and Applications, Springer 2003.
24. Thomas J., Laublet, P., Ganascia, J. G.: A Machine Learning Tool Designed for a Model-Based Knowledge Acquisition Approach. In: EKAW-93, European Knowledge Acquisition Workshop, Lecture Notes in Artificial Intelligence No.723, N.Aussenac et al. (eds.), Springer-Verlag, 1993, 123-138.
25. van Dompseler, H. J. H., van Someren, M. W.: Using Models of Problem Solving as Bias in Automated Knowledge Acquisition. In: ECAI'94 - European Conference on Artificial Intelligence, Amsterdam 1994, 503-507.