A Domain-Expert Centered Process Model for Knowledge Discovery in Medical Research: Putting the Expert-in-the-Loop

Dominic Girardi[1(✉)], Josef Kueng[2], and Andreas Holzinger[3]

[1] Research Unit Medical Informatics, RISC Software GmbH, Johannes Kepler University Linz, Linz, Austria
dominic.girardi@risc.uni-linz.ac.at
[2] Institute for Application Oriented Knowledge Processing, Johannes Kepler University Linz, Linz, Austria
[3] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria

Abstract. Established process models for knowledge discovery see the domain expert in a customer-like, supervising role. In the field of biomedical research, it is necessary for the domain experts to move into the center of this process with far-reaching consequences for their research work but also for the process itself. We revise the established process models for knowledge discovery and propose a new process model for domain-expert driven knowledge discovery. Furthermore, we present a research infrastructure which is adapted to this new process model and show how the domain expert can be deeply integrated even into the highly complex data mining and machine learning tasks.

Keywords: Expert-in-the-loop · Interactive machine learning · Process model · Knowledge discovery · Medical research

1 Introduction

Scientists in the life sciences are confronted with increasingly large, complex and high-dimensional data sets [17]. Consequently the application of machine learning techniques for knowledge discovery is indispensable. However, automated machine learning algorithms work well in lower dimensional spaces and well-defined environments, but in the biomedical domain we are confronted with probability, uncertainty, incompleteness, vagueness, noise, etc., which makes the application of automated approaches difficult and the complexity of machine learning algorithms have kept away non-experts from the application of such solutions. However, a smooth interaction of the domain expert with the data would greatly enhance the whole knowledge discovery process chain [18]. In everyday clinical research the actual process differs significantly from established process descriptions. In the commonly known definitions (see [21] for a

Y. Guo et al. (Eds.): BIH 2015, LNAI 9250, pp. 389–398, 2015.
DOI: 10.1007/978-3-319-23344-4_38

good overview) the domain expert is seen in a super-vising, consulting and customer role. A person that is outside the process and assists in crucial aspects with domain knowledge and receives the results. All the other steps of the process are performed by so called data analysts, who are supported by the domain experts in understanding the domain and interpreting the results. However, for the analysis of medical data, detailed and explicit medical expert knowledge and knowledge of clinical processes is urgently required. Roddick et al. [29] point out that data mining in medical domain requires significant domain expertise and can not be performed without the intense cooperation of medical domain experts. This clearly distinguishes data mining in the medical domain from data mining in market basket or financial trading data. Furthermore, Roddick et al. suggest the findings of data mining in medical research should only be interpreted as suggestions for further research. Cois and Moore [7] stress the uniqueness of medical data mining, caused by the nature of its data and other aspects. This is also supported by Bellazi and Zupan [3], who stress the safety aspect of medical knowledge discovery and is an often neglected part, as the expert-in-the-loop (in the biomedical sciences we speak of a "doctor-in-the-loop") is a new paradigm in information driven medicine, seating the expert as authority inside a loop supplying him/her with information on the actual patient data [20].

To integrate the domain expert more deeply into the machine learning and data mining tasks is a very recent approach, however, data mining is only one of many steps of the knowledge discovery process chain (see Figure 2 in [18]). Consequently, it is mandatory to investigate which tasks arise for the domain experts as central actors of the whole knowledge discovery process and what consequences this paradigm shift has for the process itself. In this paper we focus on aspects of a novel process model.

2 Related Research

There is not much research yet on this hot topic. A recent work by Mirchevska et al (2014) [25], presents a method for combining domain knowledge and machine learning for classifier generation and online adaptation, which exploits advantages in domain knowledge and machine learning as complementary information sources. The authors state that whilst machine learning methods may discover patterns in domains that are too subtle for humans to detect, domain knowledge of an expert may contain information on a domain even not present in the available domain data! This has essential influences for medical research.

The essence is that knowledge elicitation from domain experts and empirical machine learning are two completely distinct approaches for knowledge discovery with different and mutually complementary capabilities [33].

3 A New Process Model

3.1 Established Process Models

In 1996 Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth published a number of articles [9], [10], [8] which build the base for what we call now

the process of knowledge discovery in databases. Soon, further process models were published with different focus and degree of detail [30], [6], and many more. Generally, there is a big consensus among these process models. In their review paper from 2006 Kurgan et al. [21, Table1onpage6] even managed to extract a generic process model out of the most established process model.

Aside from the significant consensus concerning the steps of these process models, there is also a huge agreement about the roles within these processes. The process is executed by a so-called data analyst, a person who's profile varies from computer scientist, to statistician or data mining expert. The domain expert is always seen in an external position, as a customer and/or supervisor. This fact is clearly reflected by the first steps of the generic process model (and hereby of most other process models): *1 - Understanding the Domain* and *2 - Understanding the data*. Both steps would be unnecessary for domain-experts within the process loop.

3.2 A New Process Model

When kept in mind, that medical domain experts are required to be deeply involved into the process of medical knowledge discovery [7], the known process models are hardly suitable. A new process model is needed, which regards the central role of the domain-experts.

We present a new process model for domain-expert centered knowledge discovery in (bio-)medical research — see Figure 1. It is, of course, closely related to and derived from existing models, but differs in crucial aspects. The major different to established definitions can not be seen in this process description, it takes place at another level. It is the switched role of the medical domain experts from the edge of the process to the center. Subsequently, the first significant difference is the absence of the step 'Understanding of the Problem', which is of course caused by the new major player of the process, who does no longer need invest time in getting into the research matter. So, the steps of the new process are defined as follows:

1. **Data Modeling.** This step is closely related to the step 'Understanding of the data' in the definitions of [27]. It is necessary for the researcher(s) to be aware what kind of data is needed to be able to answer the research questions. What data entities from my research domain are relevant for the current research projects, which of their attributes are needed and in what kind of relation are they in. This data definition, which will be called the domain ontology from now on, builds the base for all further data-based operations and differs from research project to research project and from domain to domain. This distinguishes this process definition from many conventional definitions, where only available data — data which is produced in every day routines — is analyzed. For being able to answer medical research questions it is necessary to overcome the bias of using only what is easily available.

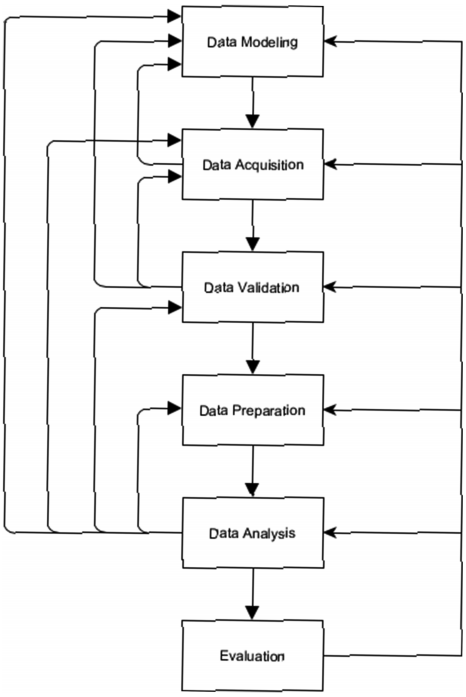

Fig. 1. A new process model for domain-expert centered knowledge discovery in (bio-)medical research

2. **Data Acquisition.** Especially in medical, scientific research it is often necessary to acquire the data of need. Data which is stored in electronic hospital information systems (HIS) is hardy suitable for scientific research because it is often semi-structured, textual data [5] or contains data mostly for billing and documentation purposes [22]. Especially medical diagnoses and interpretations of medical test are often stored as free text. Furthermore, redundant and contradictory data also occurs. Although data mining has already been performed directly on HIS, its results are less scientifically applicable than for management purposes [31], [32]. The missing or insufficient re-usability of data stored in clinical information system has already been identified as a major challenge to medical informatics [28].
3. **Data Validation.** The quality of the outcome of a research projects strongly depends on the quality of the underlying data. As already mentioned above, data quality is a widely underestimated issue in medical data sets, and even data from electronic sources (hospital information systems, etc.) are erroneous and inconsistent. Considering the complexity and amount of medical data needed for medical research the need for an automatic data validation becomes obvious. Data quality is known to be a generally underrepresented topic in medical publications [4].

4. **Data Preparation.** Data analysis rarely performed directly on the whole data set directly. Usually, data set of interest are created for certain hypothesis and erroneous or implausible data is removed from these sets. Furthermore, in medicine, very often changes or differences (functions on data values) rather than raw data contains valuable information [29]. Consequently, it is necessary to define these desired functions on the data and make their result accessible as new calculated variables.
5. **Data Analysis.** In this phase the actual step of knowledge discovery is performed, using either conventional statistics or methods of data mining, machine learning or means of visual analytics.
6. **Evaluation.** In a final step, the gain knowledge must be clinically evaluated and verified.

The steps of this process are not aligned in a strict sequence. On the one hand, steps happen in parallel or are strongly interwoven with each other. So, it is possible to see the steps data acquisition and data validation in a sequential order, where validation is performed as soon as all the data is acquired. Alternatively, data acquisition and data validation can occur in parallel where each newly entered piece of data is immediately checked. Furthermore, it is of course possible to already perform data preparation and subsequent analysis while the data acquisition is still in progress. On the other hand there exist a number of feedback loops, such as from almost any step of the process to data modeling. This means at any of these steps it may become necessary to adapt the actual domain ontology. Furthermore, insights gained from data validation and data analysis may cause re-acquisition or revision of existing data. And results from first data analysis may reveal systematic data errors which results in a revision of the data validation rules or the data preparation algorithms.

3.3 Consequences and Challenges

The researching medical domain experts face a number of challenges and obstacles when they try to perform medical research and knowledge discovery. The situation is worsened by the fact that research projects with limited funding often complete lack an explicit IT support. So the researchers find themselves in a situation where they have to deal with both, the complexity of their research domain and the complexity of their own data and data structures with all its consequences.

The selection, setup and maintenance of a research data infrastructure has already been identified as a major obstacle in biomedical research [12]. In 2007, a survey among biomedical researchers [1] found out that data handling in general had become a major barrier in a number of bio-medical research projects. Furthermore, biomedical researchers are often hardly able to cope with the complexity of their own data. The fact that many researchers use general-purpose office applications, which do not provide any support in data handling, worsens the situation.

Although highly sophisticated data mining (DM) and machine learning (ML) algorithms have been used in other domains for decades, their usage in the field of medical research is still limited. A survey from 2012 among hospitals from Germany, Switzerland, South Africa, Lithuania, and Albania [26] showed that only 29% of the medical personnel of responders were familiar with a practical application of DM. Although the survey is sure not globally representative, it clearly shows the trend that medical research is still widely based on basic statistical methods. One reason for this rather low acceptance rate is the relatively high technical obstacle that needs to be taken in order to apply these algorithms combined with the limited knowledge about the algorithms themselves and their output. A view that is shared by [16] who states that '*the grand challenge is to combine these diverse fields to support the expert end users in learning to interactively analyze information properties thus enabling them to visualize the relevant parts of their data*'.

Since the medical domain itself is a very complex one and data acquisition is usually done by multiple persons over a certain period of time, it is crucial for subsequent data analysis to check the plausibility and validity of the collected data. Simple recording errors can usually be detected by simple rules, but systematic and procedural errors, which are known to cause severe bias to the study outcome [2], can rather be detected by high complex rules. In general, data quality in medical research project is not a well researched topic [4].

4 Application and Implementation

In order to address all these challenges we developed a generic, ontology-centered research infrastructure. The main principle is the following: By modeling the actual research domain in form of a domain-ontology (Step 1 of the process) the domain-experts builds the base for all subsequent steps. The whole research infrastructure derives its structure and behavior from the central domain ontology — at run-time. Changes to the ontology have immediate effects on the whole system, which consists of three main modules. Firstly, a management tool, which allows the user to model and maintain the domain ontology, but also process and analyze the research data. The other two components are an ontology-derived electronic data interface based upon and open-source ETL (Extract-Transform-Load) suite, and an ontology-derived web interface for manual data input and processing. Wherever possible the elaborate structural meta-information is used to actively support the user in data handling, processing and analyzing. The system always appears to the user as if it was especially tailored for his domain. For a more detailed information on the infrastructure itself, the reader is kindly referred to [13], [14], [15].

Based on one particular example we now want to show how this process in combination with an appropriate software system can enable the domain expert to utilize advanced machine learning algorithms. Given the following situation: The researcher used the above-mentioned research infrastructure for collecting his research data and now wants to investigate a (possibly non-linear) influence

of a number of features on a target class. Experts in the field of computer science will recognize this problem as a binary classification problem. In order to answer this question to the researcher, the following approach was made: After the user selects the potential features and the desired target class for a given data set, a number of classification algorithms in numerous configurations are launched parallelly in background. The whole data transformation and pre-processing is performed automatically by using the extensive structural meta-information from the current domain ontology. For all resulting classification models a 10-fold cross validation is performed and the area under the RoC curve of each classification algorithm and configuration is calculated. As a result, the best area under RoC of each algorithm are consolidated and presented in a user-friendly way. In this way the research gets an indication whether the assumed influence is measurable or not. This approach is based upon the following assumptions:

1. The quality of the classification model that is developed by a classification algorithm in a reasonable (default) configuration or an automatically optimized configuration provides an indication whether a reliable classification is possible at all or not. E.g. if such a classification model shows an area under the ROC curve of something close to 0.5 then it is rather unlikely to increase the quality of the classification model to a satisfying level just by adjusting and tuning the algorithms parameters. It is way more promising to adjust the input set of input variables.
2. If none of the applied classification algorithms in any of the used configuration is able to yield a satisfying classification model then it is assumed that there is no measurable influence of the input features on the target class within the available data set.

It has to be kept in mind, that this approach shows a number of limitations and restrictions: The yielded result is an indication whether an influence can be assumed, not a classification model. The models themselves are only a mean to get a result. The result doesn't provide any information an statistical significance of the discovered phenomena. The result doesn't provide any information on causalities and reasons for the discovered phenomena. This approach does not (yet) take into account the correlations among the input features. This approach does not (yet) provide any information whether a subset of the chosen features would have been sufficient to predict the class label. This approach does not provide any explanation component on how strong or in which way the features influence the target class. Nonetheless, it does yield an easy to use and easy to interpret indication whether the assumed (even non-linear) influence can be measured in the data.

For a first test set up the following algorithms were used: A Naive Bayes classifier, a Random Forest, a Logistic Regression, a Support Vector Machine with Grid Search optimization [19], and a Multi-Layer Perceptron. The ontology-guided meta-classifier was tested with the following publically available test data sets: The Iris flower data set [11], a randomly generated numeric data set, a heart disease data set, and a diabetes data set from lima indians. The real-word data set were provided by the UCI Machine Learning Repository[23].

Table 1. Area under ROC curve for the test data sets for all selected algorithms. NB = Naive Bayes, RF = Random Forest, MLP = Multi Layer Perceptron, LR = Logistic Regression, SVM = Support Vector Machine

Data Set	Target	NB	RF	MLP	LR	SVM	Median	Variance
Iris	Setosa	1.00	1.00	1.00	1.00	1.00	1.00	0.000
Iris	VersiColor	0.98	0.99	0.99	0.81	0.97	0.98	0.005
Iris	Virginica	0.98	0.98	0.99	0.99	0.97	0.98	0.00003
Random	Target	0.64	0.70	0.55	0.55	0.50	0.55	0.004
Heart	Diagnose	0.88	0.93	0.90	0.91	0.84	0.90	0.0003
Diabetes	Class variable	0.82	0.88	0.82	0.83	0.80	0.82	0.0006

First test results clearly showed, that the classification algorithms show a big degree of consensus in their results and were able to identify the already known influences of the features on the target class variables and were able to reject a possible influence in the random number data set. The most remarkable fact is that all these results were yielded without any IT-expert driven parametrization of the algorithms. All the user did, was selecting the target class, features and data sets of interest. The whole pre-processing was done automatically by using the meta-information from the domain ontology, and the algorithm parametrization was either not necessary or done by automatic optimization.

5 Results and Discussion

All known and relevant process models for knowledge discovery see the (medical) domain expert in a customer-like, supervising role [21], [24]. While the scientific community is slowly realizing what benefits can be gained when the domain expert is deeply integrated into the data mining and machine learning loop, no accordingly research on the knowledge discovery process could be found.

We propose a new process model for expert driven knowledge discovery in medical research. It eliminates the frequent task *Understanding the Domain* and *Understanding the Data* from known models and replaces this tasks by the tasks: *Data Modeling*, *Data Acquisition*, and *Data Validation*. For the software support of this new process model, an ontology-centered approach was chosen. In the first step of the new process (Data Modeling) the domain experts defines what data (structures) are necessary for the current research questions to be answered. This definition is stored in the form of a domain-ontology, which is subsequently used to actively support the user in all the tasks of the process.

In this paper we demonstrated how the extensive use of ontology-originated, structural meta-information can help to allow medical domain-expert using advanced machine learning algorithms — algorithms that are usually preserved for IT and machine-learning experts. By automatizing the data pre-processing and algorithm parametrization to a very high degree, it is possible for a Non-IT user to apply these algorithms and answer research questions.

References

1. Anderson, N.R., Lee, E.S., Brockenbrough, J.S., Minie, M.E., Fuller, S., Brinkley, J., Tarczy-Hornoch, P.: Issues in biomedical research data management and analysis: Needs and barriers. Journal of the American Medical Informatics Association **14**(4), 478–488 (2007). http://jamia.bmj.com/content/14/4/478.abstract
2. Baigent, C., Harrell, F.E., Buyse, M., Emberson, J.R., Altman, D.G.: Ensuring trial validity by data quality assurance and diversification of monitoring methods. Clinical Trials **5**(1), 49–55 (2008). http://ctj.sagepub.com/content/5/1/49.abstract
3. Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics **77**(2), 81–97 (2008)
4. Van den Broeck, J., Cunningham, S.A., Eeckels, R., Herbst, K.: Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Medicine **2**(10), e267 (2005)
5. Bursa, M., Lhotska, L., Chudacek, V., Spilka, J., Janku, P., Huser, M.: Practical Problems and Solutions in Hospital Information System Data Mining. In: Böhm, C., Khuri, S., Lhotská, L., Renda, M.E. (eds.) ITBAM 2012. LNCS, vol. 7451, pp. 31–39. Springer, Heidelberg (2012)
6. Cios, K.J., Teresinska, A., Konieczna, S., Potocka, J., Sharma, S.: Diagnosing myocardial perfusion from pect bull-eye maps-a knowledge discovery approach. IEEE Engineering in Medicine and Biology Magazine **19**(4), 17–25 (2000)
7. Cios, K.J., William Moore, G.: Uniqueness of medical data mining. Artificial Intelligence in Medicine **26**(1), 1–24 (2002)
8. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Communications of the ACM **39**(11), 27–34 (1996)
9. Fayyad, U., Piatetsky-shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine **17**, 37–54 (1996)
10. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.: Advances in knowledge discovery and data mining (1996)
11. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics **7**(2), 179–188 (1936)
12. Franklin, J.D., Guidry, A., Brinkley, J.F.: A partnership approach for electronic data capture in small-scale clinical trials. Journal of Biomedical Informatics 44(suppl. 1), S103–S108 (2011)
13. Girardi, D., Arthofer, K.: An ontology-based data acquisition infrastructure - using ontologies to create domain-independent software systems. In: KEOD 2012, Proceedings of the International Conference on Knowledge Engineering and Ontology Development, Barcelona, Spain, October, 4-7, pp. 155–160. SciTePress, Barcelona (2012)
14. Girardi, D., Dirnberger, J., Trenkler, J.: A meta model-based web framework for domain independent data acquisition. In: ICCGI 2013, The Eighth International Multi-Conference on Computing in the Global Information Technology, pp. 133–138. International Academy, Research, and Industry Association, Nice, France (2013)
15. Girardi, D., Küng, J., Giretzlehner, M.: A Meta-model Guided Expression Engine. In: Nguyen, N.T., Attachoo, B., Trawiński, B., Somboonviwat, K. (eds.) ACIIDS 2014, Part I. LNCS, vol. 8397, pp. 1–10. Springer, Heidelberg (2014)

16. Holzinger, A.: On knowledge discovery and interactive intelligent visualization of biomedical data-challenges in human-computer interaction & biomedical informatics. In: DATA (2012)
17. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinformatics 15(S6), I1 (2014). http://www.biomedcentral.com/1471-2105/15/S6/I1
18. Holzinger, A., Jurisica, I.: Knowledge Discovery and Data Mining in Biomedical Informatics: The Future Is in Integrative, Interactive Machine Learning Solutions. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 1–18. Springer, Heidelberg (2014)
19. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003)
20. Kieseberg, P., Schantl, J., Frhwirt, P., Weippl, E., Holzinger, A.: Witnesses for the doctor in the loop. In: Brain and Health Informatics BIH 2015, Lecture Notes in Artificial Intelligence LNAI. Springer, Heidelberg (in print, 2015)
21. Kurgan, L.A., Musilek, P.: A survey of knowledge discovery and data mining process models. The Knowledge Engineering Review **21**(01), 1–24 (2006)
22. Leiner, F., Gaus, W., Haux, R., Knaup-Gregori, P.: Medical Data Management - A Practical Guide. Springer (2003)
23. Lichman, M.: UCI machine learning repository (2013). http://archive.ics.uci.edu/ml
24. Mariscal, G., Marbán, Ó., Fernández, C.: A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review **25**(2), 137–166 (2010)
25. Mirchevska, V., Lustrek, M., Gams, M.: Combining domain knowledge and machine learning for robust fall detection. Expert Systems **31**(2), 163–175 (2014)
26. Niakšu, O., Kurasova, O.: Data mining applications in healthcare: Research vs practice. Databases and Information Systems Baltic DB&IS 2012, p. 58 (2012)
27. Pal, N.R., Jain, L.: Advanced techniques in knowledge discovery and data mining. Springer, New York (2004)
28. Prokosch, H.U., Ganslandt, T.: Perspectives for medical informatics. Methods Inf. Med. **48**(1), 38–44 (2009)
29. Roddick, J.F., Fule, P., Graco, W.J.: Exploratory medical knowledge discovery: experiences and issues. SIGKDD Explor. Newsl. 5(1), 94–99 (2003). http://doi.acm.org/10.1145/959242.959243
30. Shearer, C.: The crisp-dm model: the new blueprint for data mining. Journal of Data Warehousing **5**(4), 13–22 (2000)
31. Tsumoto, S., Hirano, S.: Data mining in hospital information system for hospital management. In: ICME International Conference on Complex Medical Engineering, CME 2009, pp. 1–5 (April 2009)
32. Tsumoto, S., Hirano, S., Tsumoto, Y.: Information reuse in hospital information systems: A data mining approach. In: 2011 IEEE International Conference on Information Reuse and Integration (IRI), pp. 172–176 (August 2011)
33. Webb, G.I.: Integrating machine learning with knowledge acquisition through direct interaction with domain experts. Knowledge-Based Systems **9**(4), 253–266 (1996)