Knowledge Quality Assurance in Medical Data Mining

Vassilis S. Moustakis[1,2] and Loukas Tsironis[1]

[1]Department of Production and Management Engineering,

Technical University of Crete, Chania 73100, Greece

[2]Institute of Computer Science, Foundation for Research and Technology - Hellas (FORTH), Science and Technology Park of Crete, P.O. Box 1385, 71110 Heraklion, Greece

E-mail: {moustaki,loukas}@orpheas.dpem.tuc.gr or moustaki@ics.forth.gr

Tel: +30 821 28456, +30 81 391696. FAX: +30 821 69410, +30 81 391601

1.0 Introduction

Knowledge discovery in databases, or KDD for short, represents an active area of research with a handful of fielded industrial applications in place. KDD fits the definition of a process; starting from a mass of raw data we strive to generate knowledge, which should be both useful and interesting according to the objectives of the application. Matheus, Chan, and Piatetsky-Shapiro (1993) defined KDD as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data". Fayyad, Piatetsky-Shapiro and Smyth (1996) encapsulate KDD via a nine step model, part of which is data mining. In other words, Fayyad *et al.* view data mining, or DM for short, as part of KDD. In fact DM emerges as one of the nine KDD steps to capture all methods which are to be used to extract useful patterns from the selected data subset.

Knowledge quality assurance, or KQA for short, in DM affects, and is affected by, KQA in KDD as a whole. KDD steps that precede DM include creation of a target data set, focusing on variables for mining, data reduction and data cleaning. These steps coupled with the fact that most, if not all, DM methods rely on inductive inference to generate useful knowledge certifies that KQA should be approached via systemic application to the whole process rather than via an opportunistic manner to a part of it. Even more, the systematic approach is in tune with total quality management principles and philosophy.

In medicine and medical decision making KDD and DM are critical. Neither medicine nor medical reasoning represent *exact* sciences, thus knowledge, which is hidden in patient records is valuable either to confirm existing theoretical, or textbook based, knowledge or to enlarge formal knowledge. In some sense we may use either knowledge, empirical relates to knowledge hidden in medical records and formal relates to textbook based knowledge, to qualify or to modify or to extend the other. In addition to that medical reasoning, as opposed to reasoning in other branches of sciences, is vertical, which means that the physician strives to integrate scientific based results, such as outcomes of laboratory examinations, with subjective clinical information and patient history. The presence of hedging terms, such as "possibly", "likely", etc. is directly linked with the vertical nature of medical reasoning (Blois, 1988).

This article discusses issues underlying KQA in KDD and DM with medical data. In doing so emphasis is placed upon medical decision support and medical reasoning. We begin by defining quality concepts and metrics relevant to KDD and DM in Section 2.0. In Section 3.0 we discuss the role that machine learning (ML) techniques may play in supporting KQA throughout the process using the seven elements framework for data quality research presented by Wang, Storey and Firth, (1995). Given the use of ML to support KDD/DM, further exploitation of ML to support KQA in KDD/DM contributes to the internalization of QA in the process, both in terms of intermediate results as well as in terms of method(s) used to process sequences of data. Finally, we conclude the article in Section 4.0 pointing to areas for further research on the subject.

2.0 KQA in KDD and medical DM

KQA implies that knowledge learned, or generated, by a KDD application and via the use of DM methods and techniques is both *relevant* and *reliable*. Relevant means that knowledge satisfies the goals of the application or inquiry, which caused the organization to initiate KDD in the first place. Reliable means that knowledge offers a high degree of confidence with respect to its employment in novel decision making tasks or instances. While relevance relates to both validity and comprehensibility, reliability relates to accuracy. Both emerge as conjunctive components of *value* and *utility* of final KDD/DM result. In medical decision support and medical reasoning, accuracy alone is not sufficient; it is however, necessary to achieving relevance. For instance Rule A (see below), learned in the context of a medical KDD/DM inquiry related to patient management in acute abdominal pain in children is a 100% accurate (when tested on previously unseen data); however, from a clinical point of view is irrelevant since it fails to convey the actual site of tenderness in child's abdomen, which is critical in treatment (Gaga, Moustakis, Vlachakis and Charissis, 1996).

Rule A:

IF tenderness at Right Upper Quadrant of the Abdominal area is (i.e., =) none

THEN re-examine the child in six hours

Most, if not all, DM methods rely on inductive inference. Induction, viewed as a heuristic search process, proceeds to generate a theory or model based on a data and some assumptions, often incorporated in initial, or background, knowledge. As such the process is falsity preserving rather than truth preserving (Michalski, 1983) and places a heavy burden on the quality of data, on the DM method and DM representation model used to generate knowledge from data. Thus, for results to be both reliable and relevant, data, DM method and representation model should be relevant and reliable. For example, Rule A (cited above) fails in relevance because the representation model was based on simple attribute value pairs and failed to capture the relational element underlying abdominal pain.

A typical KDD/DM application involves the following:

I. A set **A** of attributes whose members range from hundreds even to few thousands. Each attribute derives values from a domain of reference **R**. So attribute $\mathbf{a}_i$ derives values from the reference set $\mathbf{r}_i$. A $\mathbf{r}_i$ may be numeric, symbolic, or structured; the latter implies that values conform to a generalization hierarchy; for instance the value *right lower quadrant* may evolve as a specialization of the value *right quadrant* or *lower quadrant*.

II. Thousands to millions of data records, which are defined using attributes and values from sets **A** and **R**, defined above. Values may be linked together either via conjunction or disjunction or both. However, most DM methods specialize records (in the ML community the term example refers to a record) in which values are linked with each other using expressions in conjunctive normal form, or CNF.

Typically the user tries to reduce the size of the problem. Reduction is necessary since:

I. With few exceptions most DM methods are good only on medium to large data sets; this means that data should be reduced to few thousands maximum to avoid "getting lost in the search space".

II. Many attributes from set **A** may not be relevant to the goals of the KDD/DM inquiry. Some attributes may be dependent upon others and some other attributes may not convey useful information while other attributes although useful may have either noisy or too many holes (i.e., missing values) in the records. For instance, medical records include a lot of information which is not directly relevant to a clinical inquiry; such information is often recorded and stored following administrative, organizational or legal requirements. Medical protocols often contain many attributes, which are not directly with the medical inquiry, for which the protocol was designed in the first place; however, physicians record this information to satisfy hospital wide requirements or because they are legally bound to do so.

Therefore it is important before we launch the, usually expensive, artillery of DM methods on data to make certain that we are using the right data for the occasion. We bypass the issue of appropriate DM method selection for the time being because it is well documented in previously reported work – see for instance Brodley (1993), Kodratoff, Moustakis and Graner (1994), Moustakis, Lehto and Salvendy (1996), etc.

With respect to medicine and medical reasoning, KDD and DM may play a critical role in:

I. Identifying the most significant clinical variables with respect to a diagnosis or therapy plan (we will be referring to either one as medical decision), or, the most important variables with respect to a category of medical decisions. For instance, a diagnosis may refer to an anemia type while a category may refer to the whole set of the different anemia types. If $\mathbf{w}$ denotes a weight value, $\mathbf{d}_i$ a medical decision (drawn from a domain $\mathbf{D}$ of decisions), and $\mathbf{a}_j$ a clinical variable, using KDD and DM methods we may identify significant $\mathbf{w}_D$, $\mathbf{w}_j$ or $\mathbf{w}_{ij}$ values (Moustakis, Morali, Vassilakis, Patras, 1995).

II. Modeling of subjective and other contextual elements involved in medical decision making. Medical reasoning is socially situated and context dependent. Context and social situation influence subjective elements in physician's medical decision making. For instance, a physician may not implement the best decision but resort to the most appropriate, according to the specifics underlying the situation, decision. Selection of the most appropriate versus the "optimal" decision on the one hand conforms with Herbert Simon's claim about bounded rationality (Simon, 1978) while on the other it enforces subjectivity and contextual bias in decision making. These topics are extensively discussed in previous research, i.e., by Gaga, Moustakis, Orphanoudakis and Charissis (1993), Morik, Potamias, Moustakis and Charissis (1994), and Vassilakis and Moustakis (1996).

III. Modeling of decision making rules. This is what is most often done in studies and published papers exploring the implementation of DM methods on medical cases. Usually, learned knowledge, more often than not represented in production rules, is compared on unseen test medical cases, whose decision, or outcome is known beforehand, to check accuracy. As we have already mentioned accuracy alone is not sufficient in medical reasoning while assessment may be biased by the sample of medical cases used to learn knowledge in the first place. To overcome sample bias advanced experimental techniques, such as boosting (Dietterich, Kearns and Mansour, 1996) or the explicit consideration of the prior probabilities in the sample via the use of information theoretic algebra in classification accuracy assessment (Kononenko and Bratko, 1991). Nonetheless, many papers claim that DM methods beat the expert; however, few evidence exists that such methods have reached acceptability by the expert.

IV. Knowledge revision and refinement. This may be considered as a subset of the previously mentioned activity, modeling of decision making rules. The objective is to revise or to refine existing medical knowledge using empirical knowledge elements hidden in medical records.

3.0 What can ML offer to KQA in medical KDD/DM

Formal data quality framework modeling (Wang *et al.*, 1995) proceeds by examining the process via distinct elements each of which delineates procedural matters and organizational responsibilities. Wang *et al.* (1995) summarize research status using seven elements that capture management responsibilities, operation and assurance costs, research and development, production, distribution, personnel management, and legal function. Then how should these elements be adapted to the context of KQA in medical KDD and DM and what role might ML play in supporting KQA in such a context? This section focuses on this question drawing upon initial results based on on-going research on the subject.

Management responsibilities. This element relates to explicit action which should be adopted by hospital administration and implemented via formal procedural rules to ensure quality of raw data, residing in hospital repositories. This element [should] reside within the hospital information system to ascertain that stored data as well as data transactions meeting predefined quality standards. Management action is necessary to warrant that filled medical forms, such as protocols, do not contain misleading information nor missing important information. ML technology is not directly applicable at this point; however, misleading information may lead to damaged input, which is to be used at later stages, or elements according to the Wang *et al.* (1995) terminology. On the other hand, ML technology may be used to cross validate input to medical forms; this may be achieved by having active agents residing within hospital information system applications, which would act as *critics* of recorded information. In fact critiquing has proved effective as a vehicle of decision support (Miller, 1984; Silverman, 1992). We should, however, stress that a major component of clinical data concerns medical images. Images originate from various modalities and a medium size hospital handles about one million images per year. Quality assurance of image data is directly related with the modality of image origin and with respect to most of these modalities formal procedures do exist to warrant quality.

<u>Operation and assurance costs</u>. This element concerns cost engineering and logistics related to knowledge or data quality assurance. ML may prove helpful in learning about such costs continuously to be able to reengineer and adapt the KQA system according to changing perception, needs and demands regarding quality.

<u>Research and development</u>. ML may significantly contribute to R&D effort to improve KQA. Learning from past data sequences, via the application of inductive learning tools – see for instance Kodratoff *et al.* (1994) may contribute to identifying areas in which quality should be improved in the future. Our work, for instance, in management of children with acute abdominal pain (Gaga et al., 1996; Blazadonakis, Moustakis and Charissis, to appear) had a major impact on the redesign of the departmental information system at the Pediatric Surgery Clinic of the University Hospital of Crete.

<u>Production</u>. ML may contribute significantly to the identification of non-conforming data items and specification of corrective actions. This may be achieved by using learning to generate rules, which should feed a rule based system acting as a critic. A simple analogue of such an operation may be cast from office automation; most modern word processing and text

formatting systems do have critic facilities which provide warnings to the user when something, i.e., a misspelling occurs.

Distribution; Personnel Management; and, Legal Function. These elements, although critical, do not, directly, render themselves to ML technology support. Rather they expect input from ML in adaptation or reengineering phases.

4.0 Concluding remarks

In summary, ML may play a critical role in ensuring quality at all phases of the "from data to information to knowledge" life-cycle. We have used ML to spot inconsistencies, to identify and model subjective or contextual elements affecting objective quality in data and knowledge and to assist information system designers in system redefinition and redesign. Looking ahead to the future we plan to embed learning components, using middleware computational artifacts to:

a) further improve data and knowledge quality;
b) assist expert users in complex human computer interaction task performance; and,
c) improve the customization of system designs according to expert decision needs and requirements.

Likely barriers or challenges to achieving full potential of ML technology in the clinical workplace (and we should stress that these challenges are not limited to medicine, but have a broader appeal with respect to applying and managing ML application development and organizational implementation) are likely to be identified in the following areas:

a) computational cost; learning often involves search over large spaces which limits effectiveness of non polynomial complexity learning systems;
b) absence of enough success stories to convince decision makers about the potential of ML technology in the workplace; decision makers are torn between their need to use modern information processing technology, such as ML, to support organizational activities on the

one hand and on the other by the uncertainty that such technology represents with due regard to robustness, stability, etc.

Fortunately, emerging KDD/DM successes in the marketplace are contributing to changing perception helping decision makers to invest on ML. In addition, health care information system applications are continuously growing and often reach the limit that traditional information processing technology offers. ML represents an advanced technology; its application in many fields (Abu Mostafa, 1995) has already proved effective and holds a prominent record of successful applications in the medical domain. Researchers should realize that the potential of ML is not in proving that a set of rules, learned via the application of a learning algorithm, performs better, in terms of accuracy than the expert. Rather the potential of ML should be tried in embedding the technology in everyday systems to support tasks such as interoperability of processes, transactions and, why not, data and knowledge quality.

Acknowledgment
Work reported in this article represents on-going research, partially supported from the Ministry of Development, General Secretariat for Research and Technology via the basic research grant N° 1248, *Development of a generalized architecture to incorporate machine learning in medical decision making*. The assistance and support of the clinical staff at the Pediatric Surgery Clinic, University Hospital of Crete, Greece, Professor George Charissis, Director, as well as of our colleagues Dr. George Potamias, Mr. Michalis Blazadonakis, and Ms. Lena Gaga are greatly appreciated. Views expressed herein represent the authors and not official Ministry of Development views nor opinions of the clinical staff of the aforementioned Clinic.

References

ABU-MOSTAFA, Y.S. (1995, April). Machines that Learn from Hints. *Scientific American*, pp. 68-73.

BLAZADONAKIS, M., MOUSTAKIS, V. CHARISSIS, G. (to appear). Deep assessment of machine learning techniques using patient treatment in acute abdominal pain in children. *Artificial Intelligence in Medicine*.

BLOIS, M.S. (1988). Medicine and the nature of vertical reasoning. *The New England Journal of Medicine*, 318(13):847-851.

KONONENKO, I. and BRATKO, I. (1991). Information based evaluation criterion for classifiers' performance. *Machine Learning*, 6(1):67-80.

BRODLEY, C. (1993). Addressing the selective superiority problem: Automatic algorithm/model class selection. In *Proceedings of the 10th International Conference on Machine Learning*, pp. 17-24. San Mateo, California: Morgan Kaufmann.

DIETTERICH, T., KEARNS, M. and MANSOUR, Y. (1996). Applying the Weak Learning Framework to Understand and Improve C4.5. In Saitta, L., Editor, *Proceedings of the Thirteenth International Conference*. San Francisco: Morgan Kaufmann.

FAYYAD, U.M., PIATETSKY-SHAPIRO, G. and SMYTH, P. (1996). From Data Mining to Knowledge Discovery. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R., (Eds), *Advances in Knowledge Discovery and Data Mining*, Menlo Park, California: AAAI Press/The MIT Press. (pp. 1-34).

GAGA, L., MOUSTAKIS, V.S., ORPHANOUDAKIS, S. and CHARISSSIS. G. 1993. IDDD: An inductive, domain dependent decision algorithm. In Brazdil, P., Editor, *Proceedings of European Conference on Machine Learning, ECML '93*, pp. 408-413. Berlin: Springer Verlag.

GAGA, L., MOUSTAKIS, V.S., VLACHAKIS, Y. and CHARISSIS, G. (1996). ID+: Enhancing Medical Knowledge Acquisition Using Inductive Machine Learning. *Applied Artificial Intelligence: An International Journal*, 10(2): 101-129.

KODRATOFF, Y., MOUSTAKIS, V.S. and GRANER, N. (1994). Can Machine Learning Solve my Problem? *Applied Artificial Intelligence: An International Journal*, 8(1):1-36.

MATHEUS, C., CHAN, P. and PIATETSKY-SHAPIRO, G. (1993). Systems for Knowledge Discovery in Databases, *IEEE Transactions on Knowledge and Data Engineering*, 5(6): 903-913.

MICHALSKI, R.S. (1983). A theory and methodology of inductive learning. In Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Eds), *Machine Learning: An Artificial Intelligence Approach*, Vol. I. Palo Alto, California: Tioga. (pp. 83-132).

MILLER, P. (1984). *Critiquing Approach to Expert Computer Advice: ATTENDING.* Pitman.

MORIK, K. POTAMIAS, G., MOUSTAKIS, V.S. and CHARISSIS, G. (1994). Knowledgeable learning using MOBAL: a clinical case study. *Applied Artificial Intelligence: An International Journal*, 8(4):579-592.

MOUSTAKIS, V.S., LEHTO, M. and SALVENDY, G. (1996). Which Machine Learning Method Should be Used for which Task?, *International Journal of Human Computer Interaction*, 8(3): 221-236.

MOUSTAKIS, V.S., MORALI, A. VASSILAKIS, P. and PATRAS, Y. (1995). Using typicality to support inference and learning, In J. Janssen, C. Skiadas, and C. Zopounidis, Eds, *Advances in Stochastic Modeling and Data Analysis*, Kluwer Academic Publishers, pp. 357-383.

SILVERMAN, B.G. 1992. Survey of Expert Critiquing Systems: Practical and Theoretic Frontiers. *Communications of the ACM*, 35(4):106-127.

SIMON, H.A. (1978). Rationality as a process and product of thought. *American Economic Review*, 68(2):1-16.

VASSILAKIS, P. and MOUSTAKIS, V.S., (1996). Identifying decision maker's preferences through a prototype based inductive learning method: a medical case study. *Behaviour and Information Technology*, 15(2): 113-122.

WANG, R.Y., STOREY, V.C., and FIRTH, C.P. (1995). A Framework for Analysis of Data Quality Research, *IEEE Transactions on Knowledge and Data Engineering*, 7(4): 623-639.