Design of a Method to Integrate Knowledge Discovery Techniques with Prior Domain Knowledge for Better Decision Support

K. Niki Kunene
Virginia Commonwealth University
Tel: 1-804-828-0070; Fax: 1-804-828-3199
Email: s2knkune@vcu.edu

H. Roland Weistroffer
Virginia Commonwealth University
Tel: 1-804-828-7118; Fax: 1-804-828-3199
Email: hrweistr@vcu.edu

ABSTRACT

The current data mining methodology uses knowledge discovery techniques to predict the likelihood of specific events using information residing in databases, or determines specific clusters or classes from some data or text. The knowledge discovered is assumed to be interpretable and useful for decision making purposes; however data mining experts are required to make the interpretation in the context of limited understanding of the business environment, its objectives and constraints, managerial decision preferences and options. This paper describes the design of a method for integrating domain specific decision preferences, options, objectives and constraints into the knowledge discovery process, extending the use of data mining technology to include direct support for decision making by managers. This is accomplished by integrating data mining and knowledge discovery techniques iteratively with rule sets, decision maker specified objective functions, and user qualitative preferences.

Keywords: Decision support systems, knowledge discovery, data mining, knowledge management.

INTRODUCTION

Knowledge discovery in databases (KDD), the "non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data," was first articulated as a process-framework by Fayyad et al. (1996); and *data mining* — the application of specific algorithms for the extraction of patterns (or models) from data — is only a single step in this broader process, even though much of the published research has focused on this data mining step. Fayyad et al. (1996), in what they loosely call a *unifying process-centric framework* for KDD, formulate a nine-step process that includes *data preparation*, *data selection*, *data cleaning*, the *incorporation of appropriate prior knowledge*, and a *proper interpretation of the results***.** Han & Kamber (2001) recognized that there is not a broadly accepted methodology for KDD and data mining, but that, notwithstanding, any such methodology should contain the following steps: (1) *problem analysis,* (2) *data preparation,* (3) *data exploration,* (4) *pattern generation* (data mining), (5) *pattern monitoring,* and (6*) pattern deployment*. At the same time, many practitioners use *CRISP-DM* (The *CRoss Industry Standard Process for Data Mining)* methodology, making it a de facto standard. The design-logic of CRISP-DM methodology is not dissimilar to the framework of Fayyad et al. (1996), or that of Han & Kamber (2001); CRISP-DM stipulates the following phases: (1) *business understanding*, (2) *data understanding,* (3) *data preparation*, (4) *modeling,* (5) *evaluation, and* (6) *deployment.*

DESRIST 2006. February 24-25, 2006, Claremont, CA.

None of the above-mentioned KDD process-frameworks (methodologies) offer any specific formulation for the integration of domain knowledge and discovery techniques. There is no systematic, easy to generalize, and repeatable method for framing domain knowledge and incorporating it into knowledge discovery. Using design science as our research methodology (March and Smith, 1995; Hevner, March, Park and Ram, 2004; Vaishnavi and Kuechler, 2005), we here describe the design of a method for integrating domain knowledge into the knowledge discovery process, thereby extending the use of data mining technology to include direct support for decision making by managers.

The remainder of this paper is organized around the design cycle as discussed by Vaishnavi and Kuechler (2005): (1) Awareness of problem, resulting in a proposal; (2) suggestion, resulting in a tentative design; (3) development, resulting in the design artifact; (4) evaluation, resulting in performance measures; and (5) conclusion. Reference is also made to the general design science guidelines, as far as applicable, provided by Hevner et al. (2004):

1. Design as an Artifact – design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
2. Problem Relevance – The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
3. Design Evaluation – The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
4. Research Contributions – Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
5. Research Rigor – Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
6. Design as a Search Process – The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
7. Communication of Research – Design-science research must be presented effectively to technology-oriented as well as management-oriented audiences.

AWARENESS OF PROBLEM

Domain knowledge is background or prior knowledge about a decision context. Within the data mining process, it is the knowledge used above and beyond the data itself. This includes semantic meta-data, prior expectations and intuitions, as well as any formalized or implicit knowledge about the application domain as employed by the analyst or decision maker. Domain knowledge includes the knowledge used to analyze, select the required data, interpret, and evaluate the results of the knowledge discovery process (Pohle, 2003; Fayyad et al., 1996).

The challenge for research and application is incorporating domain knowledge into data discovery so that it either helps improve the discovery process computationally, or more significantly, improves the interestingness of the patterns (or models) discovered. Evidence from the literature shows that the incorporation of domain knowledge into the KDD process has been accomplished predominantly by the introduction of the statistical deviations of attributes, other descriptive statistics, and interfield correlations into the data mining phase (de Abajo et al., 2004;

Hotz et al., 2001). For example, Yoon et al. (1999) look at interfield correlations between variables, which are then used to restrict, or eliminate irrelevant data mining queries. Kopanas et al. (2002) use the explicit domain knowledge about "the structure of available information" and its semantic value, where this knowledge is obtained from the data processing department and particularly the data entry employees. This enabled the authors to eliminate irrelevant and redundant attributes, determine missing values**,** aggregate data, and reduce data by sampling and transaction elimination**.** In the Kopanas study, the identified domain knowledge contributes to the reduction of the data mining search space. Shen et al. (1996) develop second-order predicate (or template) meta-queries that are run on the database, with the results used as input into the data mining process. Then there are Bayesian approaches, which in general use prior probabilities over data distributions as one way of encoding knowledge. In addition, some web mining research also uses domain knowledge in data mining by incorporating site semantics for more "intelligent mining" (Pohle, 2003; Kosala & Blockeel, 2000; Pal et al., 2002; Zaki et al., 2001).

Notwithstanding the fact that the potential usefulness of domain knowledge in KDD has been recognized for a number of years, it has not been adequately researched nor understood, and most data mining tools do not support it. Kopanas et al. (2002) examined the use of domain knowledge and the associated tools employed in the various stages of knowledge discovery (Table 1 below). They concluded that although domain knowledge plays a crucial role mainly in the initial and final stages of the knowledge discovery process, it nevertheless contributes to all stages of a KDD project, and that "one should consider a data mining project as a knowledge-driven process." Furthermore, and of significance to this study, the authors point out that "more support and adequate tools are needed which model the domain knowledge and track the contribution of domain experts." We further note that in their assessment there are currently no formalized tools used during the *problem definition* phase (Table 1).

Table 1: Overview of Use of Domain Knowledge in a Data Mining Project (adapted from Kopanas et al., 2002)

Stage	***Use of Domain Knowledge (DK)***	***Type of DK***	***Tools used***
(a) Problem definition	HIGH	Business and domain knowledge, requirements Implicit, tacit knowledge	
(b) Creating target data set	MEDIUM	Attribute relations, semantics of corporate DB	Data warehouse
(c) Data prepossessing and transformation	HIGH	Tacit and implicit knowledge for inferences	Database tools, statistical analysis
(d) Feature and algorithm selection	MEDIUM	Interpretation of the selected features	Statistical analysis
(e) Data Mining	LOW	Inspection of discovered knowledge	Data mining tools
(f) Evaluation of learned knowledge	MEDIUM	Definition of criteria related to business objectives	Data mining tools
(g) Fielding the knowledge base	HIGH	Supplementary domain knowledge necessary for implementing the system	Knowledge-based system shells and development tools

In this paper, a KDD process-centric method is developed that integrates user tacit domain knowledge into the pattern discovery process, as well as other domain knowledge that is formally known and applied to the decision problem at hand, but not included in the databases or warehouses. This method employs multi-criteria decision techniques and tools as a way of systematically and formally eliciting domain knowledge starting in the *problem definition* phase of the KDD process.

As indicated above, there seems to be a domain knowledge utilization gap in conventional data mining activities, which is evidence for the relevance of our research problem (relevance is guideline 2 of Hevner et al. (2004)). Our proposed method is the artifact resulting from the design process described in this paper.

SUGGESTION

Figure 1 shows the traditional approach to knowledge discovery as a process. Data mining is performed on data captured in a database to generate patterns. Decision makers use their experience and judgment together with the discovered patterns to make decisions.

Figure 1: Traditional Approach

The proposed method is illustrated in Figure 2. Whereas with traditional practice, the data passed to the data mining process consists only of knowledge captured in databases, the proposed method combines the database knowledge with extra domain knowledge before applying data mining. In other words, whereas in the traditional approach the pattern discovery is bounded by the depth and limitations of the database, the proposed method seeks out and incorporates additional knowledge that is available to the decision makers, but is not stored in the database.

To summarize the motivation for the design intervention before next describing the development: it is a well established fact that the absence of domain knowledge from data mining activities is a constraint on the quality of discovered patterns, and therefore that mechanisms seeking to incorporate domain knowledge into data mining activities may well serve the performance of data mining activities and relevance of resulting data mining patterns. Our method is one way by which to incorporate domain knowledge into data mining activities using formalized tools and techniques with an established research and practical track records. Our method relies on multi-criteria decision making analysis (MCDA) tools and the associated methodology to systematically inquire into the domain. The knowledge thus extracted is (machine) represented using MCDA in one instantiation of our method, and in the second instantiation using decision rules and their concomitant object-attribute-value (OAV) triplets. We note that the second instantiation is partially dependent on the first.

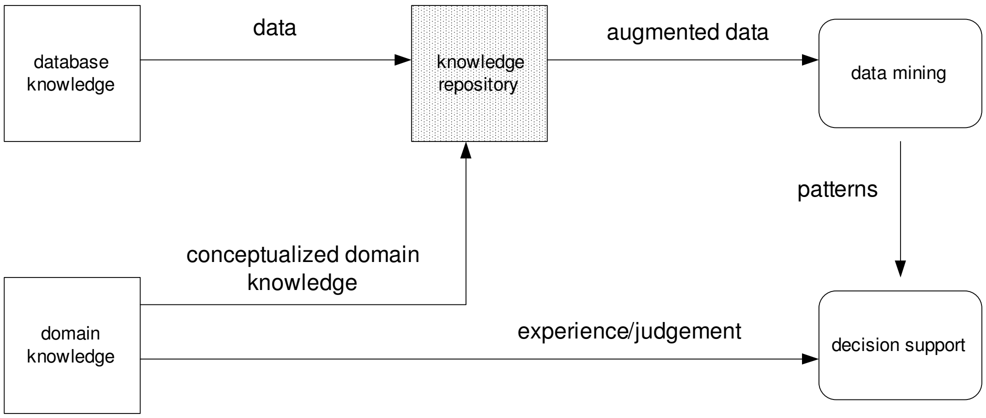

Figure 2: Proposed Method

Using the analytic hierarchy process (AHP), an MCDA tool and techniques, we set out to inquire into the decision structure of the domain. In short, the AHP achieves two distinct objectives: problem structuring, and then preference evaluation. The AHP uses hierarchical structures of the *goal – criteria – sub-criteria – alternatives* form to represent a decision problem, and then assigns priorities for the decision alternatives based on user judgments. The latter are captured progressively in the form of pairwise comparisons of criteria with respect to the goal, sub-criteria with respect to each criterion, and the alternatives with respect to the sub-criteria. Thus, the elements of one level of a hierarchy can be compared in a pairwise fashion with respect to one element of the next higher level. The AHP computes relative weights for the criteria, sub-criteria and alternatives based on the pairwise comparisons, which are in turn synthesized to yield the composite priorities of all criteria and ultimately the relative, global weights of the alternatives.

The decision/problem structure (the hierarchy) yielded using the AHP can also be articulated using decision rules (of the IF-THEN-ELSE type). We do just this in the second instantiation of our method, however here the preference assessments (of the AHP) that yield global weights for the decision alternatives are excluded.

DEVELOPMENT

So far we have discussed the initial conceptualization of how we incorporate domain knowledge into data mining. In reality, the design process is an open system that imports and exports materials from its environment resulting in a change of components (Katz & Kahn, 1966/1978). The following is a discussion of that process. The building blocks or components of our artifact are drawn from established knowledge bases, namely: the analytic hierarchy process (AHP), which is a multi-criteria decision analysis (MCDA) method and technique; and knowledge representational methods from expert systems, and decision tree classification/prediction from data mining. Our artifact, which is related to unpacking the decision structure of a domain, is closely coupled to the domain of practice. Our method is developed within the domain of the treatment and management of traumatic brain injury (TBI) patients in a neuroscience intensive care unit (ICU) of a large university hospital in the United States. The data mining objective for our method is used to classify patients by patient outcome, the target variable. The following description of the development of the method illustrates the design of the artifact within a search process, as specified in guideline 6 of Hevner et al. (2004).

1. Unpacking the Decision Structure using the Analytic Hierarchy Process

The development and construction of a decision (AHP) hierarchy is an iterative process shaped by the interactions between the analyst and the domain expert. In the early stages, the analyst can only ask basic questions about the domain, e.g. "what factors are important to patient outcome?" The expert provides information that he/she believes is primarily important. As the analyst's understanding of the domain grows, (s)he can inquire more intelligently into the domain, and as such the constructed decision structure is progressively modified reflecting the maturity of this interaction. Furthermore in the case of database management systems and data mining, we may eventually discover that factors that are important to the patient are for whatever reason not rendered in the data collection.

Knowledge about the domain was gathered, through a series of unstructured interviews with two clinicians (one clinician was a neurosurgeon, the other an anesthesiologist who also served as a research fellow for the ICU; regular interaction and validation was with the anesthesiologist with the neurosurgeon consulted when the former could not sufficiently address inquiries; and other times the two of them would resolve whatever inquiry on hand and the anesthesiologist would relay the consensus outcome), by plowing through published research materials in the field (identified by clinicians; particularly the Brain Trauma Foundation's guidelines to the management of Traumatic Brain Injury) and providing statistical support for their research efforts. This knowledge was used to construct the decision hierarchy.

Figures 3 through 6 below illustrate some of the progressive changes in conceptualization and implementation of the decision hierarchy over time. Figure 3 was the first stab at representing the decision structure. Figure 4 is a subsequent iteration which reflects a better understanding of the domain and organization, and of relationship among its components.

It is not the intension of this paper to dwell on the knowledge specifics of the TBI domain per se, but rather to discuss the iterative design process. Briefly, Figure 3 and subsequent figures depict the decision structure of the domain. Level 1 is always the focus of the hierarchy, i.e. the prognosis of TBI patient-outcome (also called survivability). Lower levels represent the factors (sub-factors, and in some cases their related intensities) bearing on survivability. At the bottom level are the alternatives, which are too numerous to show. In some cases, the identified factors are proxy measures derived using domain knowledge, e.g. the presence of hypotension.

Figures 5 and 6 reveal two important steps. First, they demonstrate the introduction of the importance of clinical certainty which hitherto had been absent in the model. Second, the two figures individually model the so-called *epidemiology* factors, and *treatment and monitoring* factors. The first consideration reflects the continuous learning that inheres in the structuring process. The second consideration reflects a design decision made with both the domain and technology in mind.

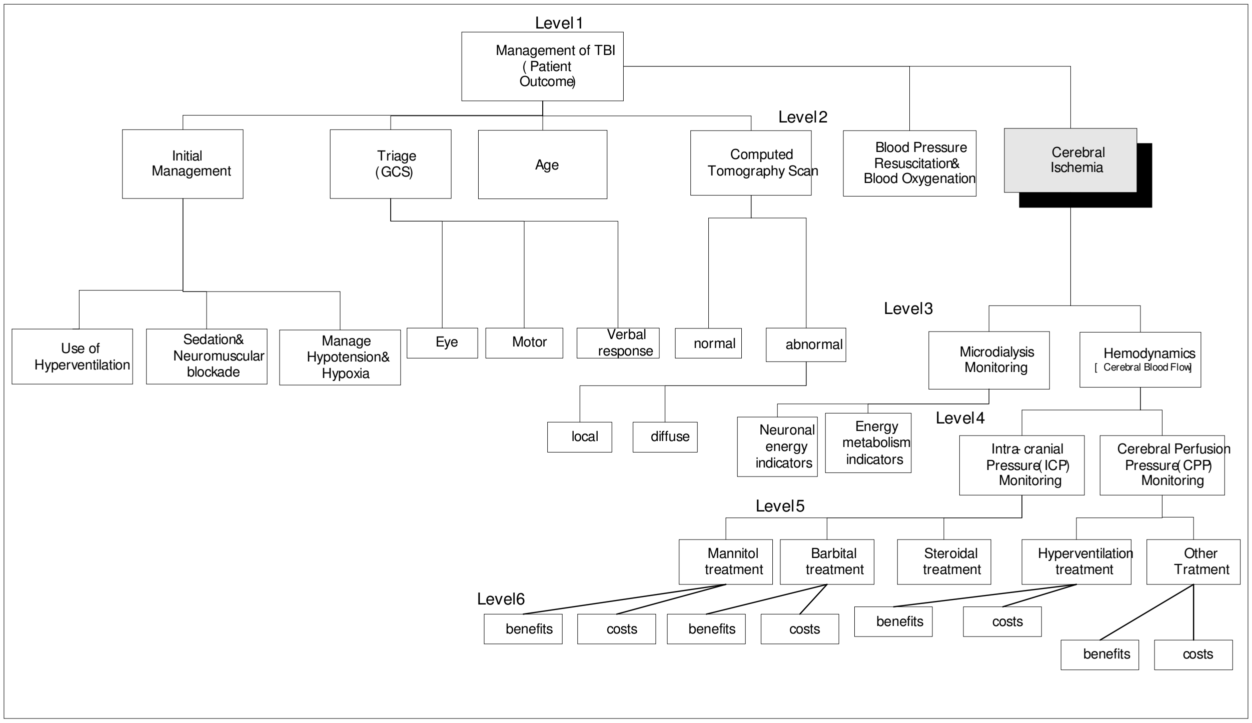

Figure 3: An Initial Understanding of the Decision Structures of the Domain

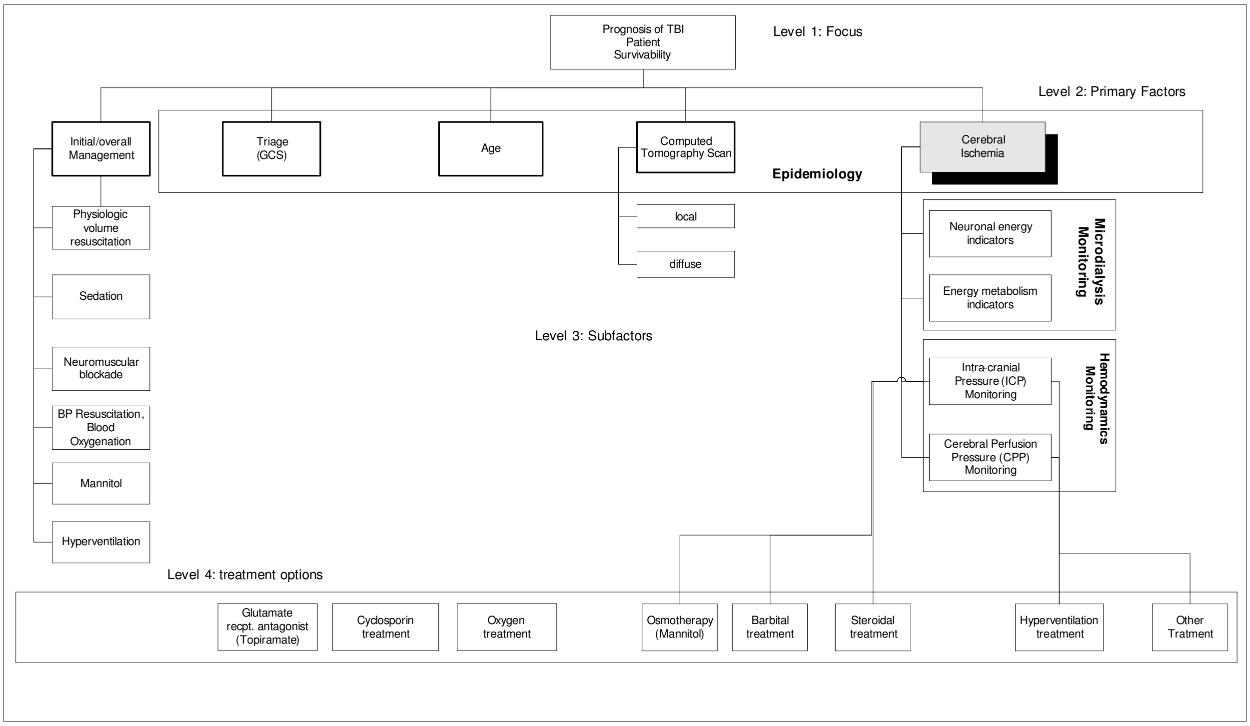

Figure 4: An Intermediate Understanding of the Decision Structures of the Domain

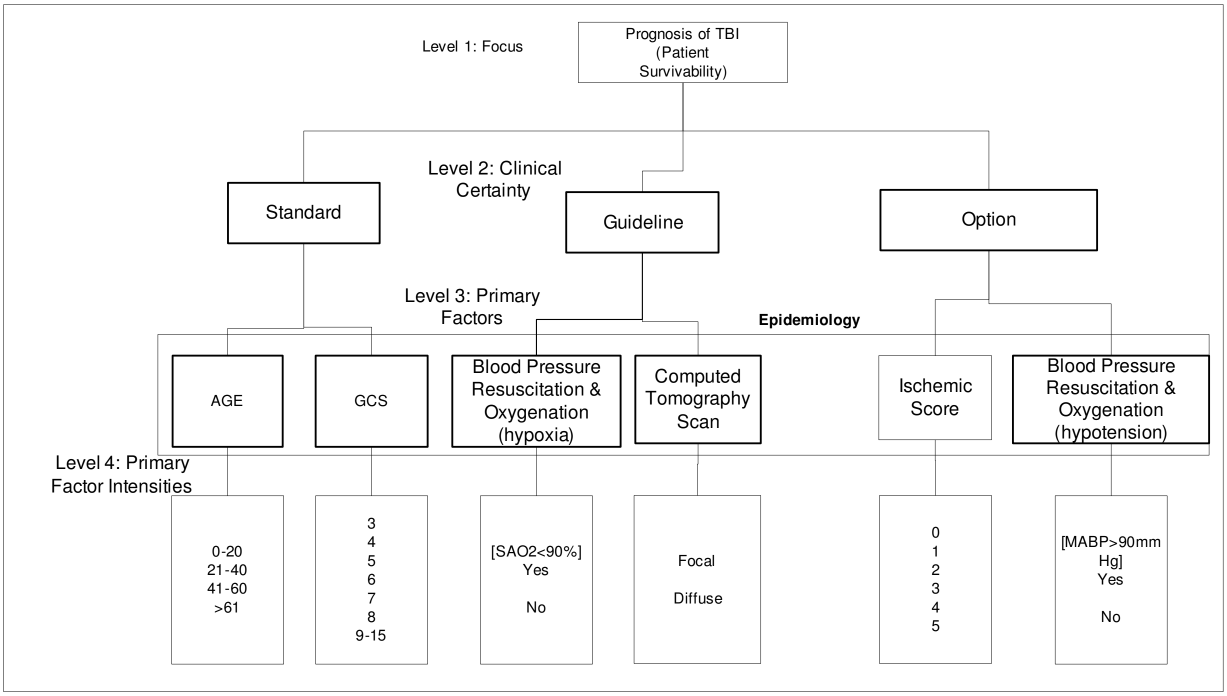

Figure 5: A Post-Intermediate Understanding of the Decision Structures of the Domain – Epidemiology

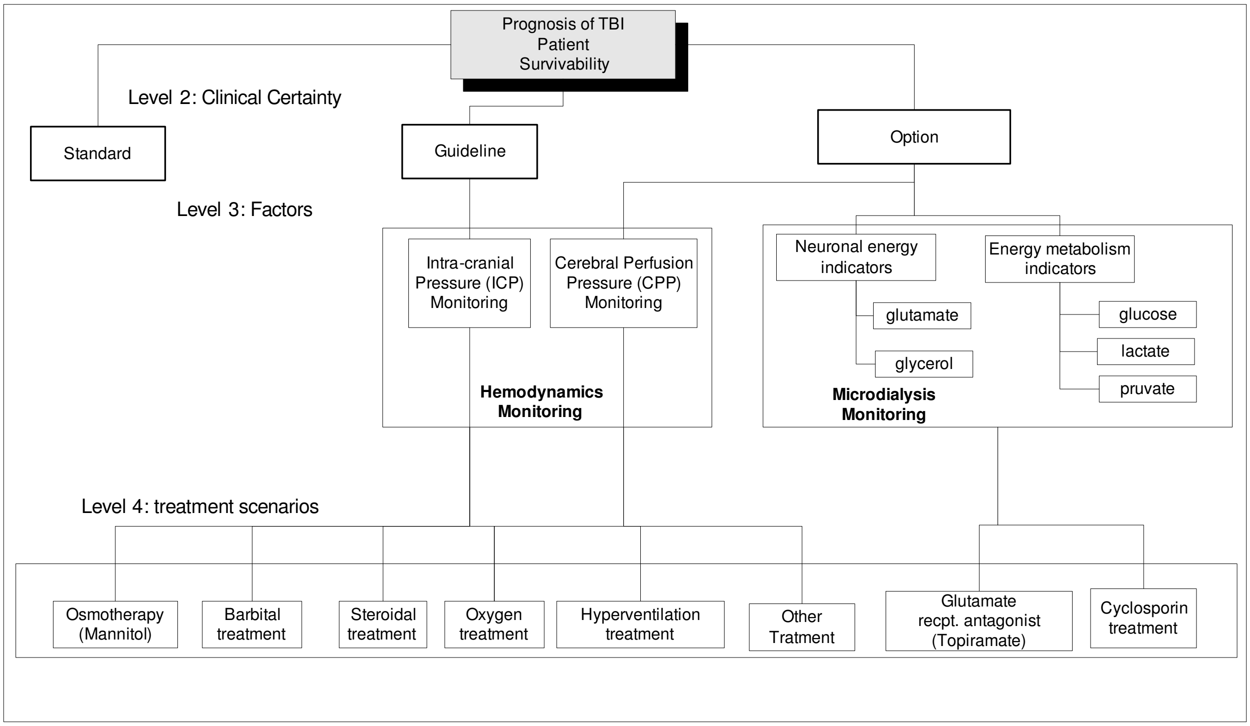

Figure 6: A Post-Intermediate Understanding of the Decision Structures of the Domain – Treatment and Monitoring

Note that in Figure 5, Level 2 of the hierarchy reflects the degree of clinical certainty associated with the known factors/variables bearing on survivability: where standard, guideline, and option represent varying degrees of clinical certainty, with respect to patient *age*; the level of patient of consciousness at admission measured using the GCS; the CT scan result using Marshall Classification; and the ischemic score.

Similarly Figure 6 also incorporates clinical certainty, but is related to patient treatment and monitoring factors effect on the prognosis of patient outcome. We must note that the items displayed in the decision hierarchy do not always have data about them residing in the operational database.

2. Generating the Production Rules and OAV Triplets

The extracted domain knowledge needs to be represented in machine-readable format to allow for this knowledge to be incorporated into existing data repositories. We used production rules and their related object-attribute-value (OAV) triplets for knowledge representation; any IF-THEN ELSE statement can be represented as a series of OAV triplets. The derivation of decision rules from the decision hierarchy is fairly straight-forward. For the database, OAV triplets are clearly just column/attribute definitions plus their values for each alternative.

3. Mining the Extended Repository

Having represented the extracted knowledge in a machine-readable format (OAVs), we populate the data repository and mine the domain knowledge extended repository.

4. The Artifact

Figure 7 conceptualizes our method, the design artifact. It consists of the *inquiry mechanism*, where first, domain knowledge bearing on the decision problem not contained in the database is obtained, possibly from domain experts and from available documentation. The *conceptualization and articulation mechanism* is where the domain knowledge is articulated into, for example, AHP hierarchies or decision rules. The *machine representation mechanism* represents the articulated knowledge in machine readable form, such as object-attribute-value triplets or the scores resulting from the pair-wise comparisons of the AHP. This machine represented domain knowledge is included into the knowledge repository (together with the database knowledge), where the data mining then takes place.

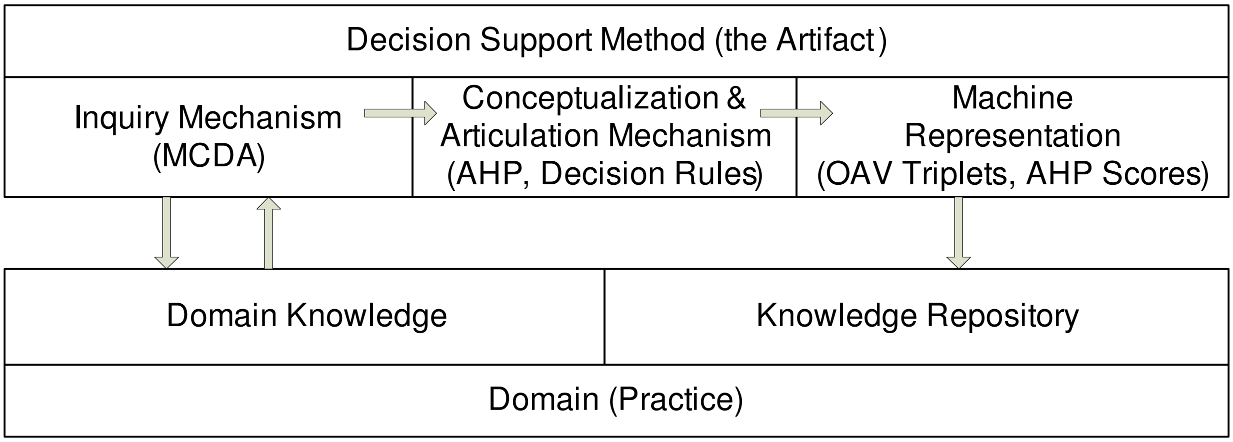

Figure 7: The Design Artifact

EVALUATION

For data preparation, data cleansing, and data exploration (including statistical analysis) SPSS 11.5 for Windows as well MS Excel 2003 were used. MS Excel macros were used to transform knowledge domain decision-rules into object-attribute-value (OAV) triplets. The multi-criteria decision structure and related pairwise comparison were captured using Expert Choice 11.1™, which is a decision support tool for the AHP. For data mining, we performed only decision tree classifications using SAS Enterprise Miner. Enterprise Miner decision tree methodology implements CHAID, a mixed algorithmic strategy.

For the purpose of evaluation, the approach of our method is also compared to the traditional approach to data mining, which does not systematically elicit and extract domain knowledge using MCDA tools and techniques. It is from such comparisons that we use KDD performance metrics to compare the interventions of our method with the traditional approach. Guideline 3 of Hevner et al. (2004) recommends that the designed artifact, i.e. our proposed method, must be evaluated to show its utility, quality, and efficacy. To that effect, we implemented competing data mining models, some using our method, and some using the traditional approach to data mining. In addition because we are using CHAID, for each approach we implemented with different information gain measures, n-ary splits, and so on to find the best tree for each approach.

There are two types of measures for assessing the interestingness of data mining results: objective and subjective measures (Hildeman and Hamilton 2000; 2001). For objective measures of performance we use:

1. **Accuracy**, usually measured as the misclassification rate.
2. **Stability** – a predictive or classification model is stable when it performs just as well on both seen (training) and unseen (validation) data sets. We will measure stability as a number between 0 and 1, where 0 means completely stable, and 1 means completely unstable. We calculate this value as the arithmetic difference divided by arithmetic sum of the training and validation classification rates: $(CR_T\text{-}CR_V)/(CR_T\text{+}CR_V)$.
3. **Lift Charts** to plot the relative predictive/classification improvement from the baseline (chance). A lift chart consists of a set of line segments for a decision tree, with each line segment corresponding to one of the leaves of the decision tree.
4. **Complexity**, measured by a combination of the depth of the tree, and the number of leaves, following both pre-pruning and post-pruning activities. If tree depth is fixed across all models, then the number of leaf-nodes (i.e. distinct classes) is the main determinant. Note, tree complexity is partly a subjective notion, as human input is required to set a complexity threshold. There are no objective stipulations here except that less is more. However, a classification model can be too simple. The simplest model has a single leaf-node or class, which is equal to no classification at all. The threshold for complexity should also be contingent on the nature of the problem.

For subjective assessment, we examine the generated decision tree rules themselves for relative surprisingness and for clinical actionability (Freitas 1999). Note, we ask the clinicians to make these assessments using their expert knowledge and experience:

1. **Surprisingness** refers to how surprising the resultant patterns are to the clinician.

2. **Actionability** refers to how practical it is to take action based on the results.

We treat surprisingness as more important than actionability. In fact, a pattern is either surprising or not, and where it is not surprising we should not be interested in further ascertaining its actionability. Results that are not surprising (i.e. reveal only information that is already known) are not really useful in KDD.

Using the above described measures, we compared various implementations of our method to implementations of traditional methods. We implemented two variations of our method, one incorporating the results of AHP, and the other using only decision rules. To summarize the results, we found the following: With respect to classification *accuracy,* the most successful implementations of both instantiations of our method performed almost equivalently, however the approach using AHP results was marginally better. Both instantiations performed better than the traditional data mining approach, with classification accuracy improved by between 33% and 40% for training and validation datasets. *Stability* also improved using our method by a margin of 27% over the traditional approach. The traditional approach produced models that were not sufficiently complex and therefore not explanatory or descriptive enough, whereas the approaches using our model generated more complex descriptions that utilized more of the pertinent variables. Finally, with respect to lift, the approach using our method generated better lift up to the 70th percentile. On the other hand, with respect to subjective interestingness measures, our users found the patterns generated using our method relatively more surprising without being unactionable, whereas the traditional approach generated responses like, “yes, we know that.”

CONCLUSION AND LIMITATIONS

In this paper we used design science as a research methodology to develop and evaluate a new *method* for knowledge discovery in which we seek to apply data mining decision tree classification to repositories incorporating prior domain knowledge that hitherto has been excluded. Given the domain task, we explicated the related domain decision hierarchy using the AHP, elicited preference valuations from the user, and in a second instantiation, transformed the hierarchy into decision rules and concomitant OAV-triplets to incorporate into a data mining repository.

The contribution of our study (Guideline 4 of Hevner et al. (2004)) is primarily in the design artifact, i.e. our proposed method. Our method proposes a more systematic manner in which to inquire into a domain, formally articulate the subsequent domain knowledge and subsequently incorporate it into the repositories we explore using data mining using tools and techniques with a proven track record in research and practice. Multi-criteria decision analysis is a vibrant field of study and inquiry with several schools of thought methods, notwithstanding any weaknesses within any school of thought; these are systematic and rigorous approaches to exploring decision problems. Preliminary evaluation in our study suggests that our method is more effective than the traditional approach: both the objective and subjective quality of discovered data mining patterns are better for our method. Although, we instantiated in a medical domain, there is no reason why our approach could not be used in a business domain to incorporate strategic objectives that are not typically contained in operational databases. Furthermore, if database schemas were designed with the strategic intent or decision making in mind, then perhaps there would be no need for our method. Alternatively, this could contribute to thinking about databases that directly affect decision making goals and needs.

There are of-course limitations to our results. The implementation of our method is a time consuming affair that calls on analysts to deeply immerse themselves in the problem domain. It may be quicker and easier to follow the traditional approach in many cases. Our method also requires the data mining analysts to learn MCDA techniques. Furthermore, for the purposes of validating the method, we have not studied the use of the method under a different domain of practice as yet.

REFERENCES

de Abajo, N., Diez, A.B., Lobato, V. and Cuesta, S.R. (2004) ANN Quality Diagnostic Models for Packaging Manufacturing: An Industrial Data Mining Case Study, in *Proceedings of The 2004 ACM SIGKDD international conference on knowledge discovery and data mining,* ACM Press,

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) The KDD Process for extracting useful knowledge from volumes of data, *Communications of the ACM,* **39,** 27-34.

Freitas, A.A. (1999) On Rule Interestingness Measures, *Knowledge-Based Systems,* **12,** 309-315.

Han, J. and Kamber, M. (2001) *Data Mining: Concepts and Techniques,* Morgan Kaufmann, San Francisco (CA).

Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004) Design Science in Information Systems Research, *MIS Quarterly,* **28,** 75-105.

Hotz, E., Grimmer, U., Heuser, W., Nakhaeizadeh, G. and Wieczorek, M. (2001) REVI-MINER, a KDD-environment for deviation detection and analysis of warranty and goodwill cost statements in automotive industry, in *Proceedings of The seventh ACM SIGKDD international conference on Knowledge discovery and data mining,* ACM Press, pp.432-427

Kopanas, I., Avouris, N.M. and Daskalaki, S. (2002) In *Methods and Applications of Artificial Intelligence, Lecture Notes in AI -LNAI*, Vol. 2308 (Eds, Vlahavas, I. P. and Spyropoulos, C.D.) Springer-Verlag, Berlin, pp. 288-299.

Kosala, R. and Blockeel, H. (2000) Web Mining Research: A Survey in *Proceedings of SIGKDD Explorations 2000,* ACM Press, pp.1-15

March, S.T. and Smith, G.F. (1995) Design and Natural Science Research on Information Technology, *Decision Support Systems,* **15,** 251-266.

Mitra, S., Pal, S.K. and Mitra, P. (2002) Data Mining in Soft Computing Framework: A Survey, *IEEE Transactions on Neural Networks,* **13,** 3-14.

Pal, S.K., Talwar, V. and Mitra, P. (2002) Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, *IEEE Transactions on Neural Networks,* **13,** 1163-1177.

Piatetsky-Shapiro, G. and Matheus, C.J. (1994) The interestingness of deviations, in *Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases,* pp.25-36.

Pohle, C. (2003) Integrating and Updating Domain Knowledge with Data Mining, Citeseer.IST, Last accessed 2004, Jan, http://citeseer.ist.psu.edu/668556.

Saaty, T.L. (1992) *MultiCriteria Decision Making – The Analytic Hierarchy Process,* RWS Publications,

Saaty, T.L. (1994) *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process,* RWS Publications, Pittsburgh, PA.

Shen, W.-M., Ong, K., Mitbander, B.G. and Zaniolo, C. (1996) In *Advances in Knowledge Discovery and Data Mining*, Vol. (Eds, Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.) AAAI/MIT Press, 375-398.

Vaishnavi, V., and Kuechler, B. (2005) Design Research in Information Systems, http://www.isworld.org/Researchdesign/drisISworld.htm.

Yoon, S.-C., Henschen, L.J., Park, E.K. and Makki, S. (1999) Using Domain Knowledge in Knowledge Discovery, in *Proceedings of The Eighth International Conference on Information and Knowledge Management,* ACM Press, pp. 243-250.

Zaki, M.J., Punin, J. and Krishnamoorthy, M. (2001) LOGML for Web Usage Mining in *Proceedings of INFORMS Annual Meeting.*