International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol. 3, Issue. 3, May - June 2013 pp-1610-1616 ISSN: 2249-6645

A Novel Approach to Ontology Based Hybrid Intelligent Data Mining Assistance (HIDMA)

Syed Abusali.V[1], Syed Ali.A[2]
[1]*Lecturer, Department of Computer Science & Engineering, TJ Institute of Technology, Chennai, India*
[2]*Assistant Professor, Department of Computer Science & Engineering, TJ Institute of Technology, Chennai India*

Abstract : *An efficient application of a data mining process is littered with many difficult and technical decisions such as data refining, feature transformations, algorithms, parameters, evaluation. Subsequently most data mining products provide a large number of models and tools but few provide intelligent assistance for addressing the above mentioned challenges that face the non specialist data miner. In this paper, we propose the realization of a hybrid intelligent data mining assistant(HIDMA) based on the synergistic combination of both declarative and procedural ontology knowledge in order to empower the non specialist data miner throughout the key phases of the Cross Industry Standard Process for Data Mining (CRISP-DM) process.*

Keywords: *Knowledge Acquisition, Knowledge Representation, Ontologies, Case based Reasoning, Data Mining, Declarative and Procedural knowledge.*

I. INTRODUCTION

In order to remain competitive in the business world, decision makers have begun to turn to data mining (DM) technology to cope with the information deluge and meet their informational needs. Although data mining does promise to uncover potentially valuable, useful and implicit knowledge from one's abundant data repositories, the effective application of data mining still faces some very serious challenges:

- Data mining research seems to be based on specialized techniques (statistics, machine learning, information theory, database technology etc.) whereas research on and even epistemological aspects of DM are rare [1].
- Current DM processes make very little use of already existing corporate knowledge. Consequently, DM is more tedious than is necessary and can tend to produce already known information.
- Existing DM methodologies only provide general directives, however what a non specialist really needs are explanations, heuristics and recommendations on how to effectively carry out the particular steps of the methodology.

In this paper, Section 2 presents some key challenges with providing intelligent DM assistance. Section 3 summarizes related work both in the fields of data mining and ontologies, and the use of CBR (Case Based Reasoning) and ontologies. In Sections 4 and 5 respectively, we provide a system overview and design details of our proposed intelligent DM assistant. Section 6 provides a brief discussion and Section 7 presents the conclusions and future work.

II. THE REAL CHALLENGES OF DATA MINING PROCESS IN HIDMA

2.1 Supports for the Non Expert Data Miner

Most Commercial data mining products either do not offer any intelligent assistance (decision support) or tend do so in the form of rudimentary "wizards like" interfaces. These wizards like interfaces make hard assumptions about the level of background knowledge required by a user in order to effectively use the system (i.e. Oracle Data Miner, SAS Enterprise Miner, etc.). This fact has been further supported by [2]. For instance, the following is a concise list of important decisions that must be considered during a DM process:

- How to effectively perform data quality verification?
- How to efficiently perform the data preparation phase (i.e. sampling, missing values, discretization)?
- Which statistical or machine learning algorithm is most appropriate?
- Which training parameters are most appropriate?
- How to deal with a potential class imbalance problem?
- How to avoid model over fitting?
- How to improve the accuracy rate (i.e. error rate)?
- Which evaluation method is most appropriate?

Over the past several decades the fields of statistics and machine learning have produced a myriad of models/algorithms that can be readily exploited by data miners. Consequently this profusion of algorithms has dramatically burdened the data miner with difficult decisions that must be addressed in order to effectively apply DM to produce useful and meaningful results.

2.2 Fostering Knowledge Reuse

With respect to the overall data mining process, most enterprises do not directly manage tacit knowledge (i.e. useful generalizations for answering above questions) in a form that can be effectively stored, refined and reused. Most products simply archive DM activities, but leave it up to the user to intelligently manage this knowledge. An intelligent DM assistant should possess the necessary characteristics that allow it to learn from past experience and empower the user of the system to avoid the repetition of mistakes.

2.3 Beyond Model Selection Support

Previous research efforts into intelligent DM assistants have focused on providing a user with model selection support ([3], [4], [5], [6]). The selection of an appropriate algorithm for a given data mining task may be considered necessary, but is definitely not sufficient for ensuring the successful outcome of a DM project. Our appeal to an intelligent DM assistant implies the realization of a system that is capable of aiding a user throughout the key phases of the data mining process. To complicate matters, and add to the requirements for an intelligent assistant, it must be emphasized that these phases tend to be strongly inter dependent. For instance the choice of a given DM algorithm (data modeling phase) is dependent on the inherent characteristics of the data being mined (data understanding phase), while the activities carried out during the data preparation phase depend both on data quality (data understanding phase) and the chosen DM algorithm (data modeling phase).

2.4 A Need for Detailed Knowledge

DM processes adequately specify the phases, tasks and activities that need to be carried out during a DM project but provide very little detailed knowledge for the novice miner on how to actually carry out a given step. For example, the proper application of a simple linear regression model requires that the user possess some detailed knowledge (or basic heuristics) for effectively carrying out the model evaluation phase for a given DM task (i.e. significance testing, residue normality and model variance).

III. RELATED WORK

Several previous research efforts have demonstrated the effectiveness of using ontologies for supporting the knowledge discovery process. Bernstein et al. have proposed an intelligent data mining assistant based on the use of ontology [8]. Their ontology contains constraints and performance knowledge that is eventually searched for in order to find a ranking of possible satisfactory DM processes (based on user input). As shall be elaborated upon our aim is to provide a hybrid data mining assistant that leverages the synergy between the CBR paradigm and ontology based DM knowledge. Phillips and Buchanan have used ontologies to guide the feature selection step of the knowledge discovery process [7]. Bauer and Baldes have used an ontology based interface to aid non expert users of machine learning (ML) better understand and influence an ML system from a semantic perspective [9]. Canataro and Camito have demonstrated the use of DM ontology to simplify the developmentof distributed knowledge discovery applications in the area of grid computing [10]. Although our DM ontology has similar high level concepts, we have significantly extended our ontology to provide detailed knowledge (i.e. data quality verification, data preparation and model evaluation) using a rule base and associated rule based reasoner. In addition, we have also based our DM ontology model explicitly on the structured approach used by the CRISP DM methodology (industry recognized and virtually the de facto DM process) [11]. Moreover, a number of previous research efforts have demonstrated the effectiveness of combining ontologies with the CBR paradigm. Aamodt et al. have developed a KI CBR framework (CREEK) based on the use of ontologies [12]. Bello Thomas et al. have developed a framework for building CBR systems that use task/method ontology for promoting problem solving methods reuse [13]. Bichindaritz has demonstrated the use of ontologies for facilitating case structuring and acquisition [14].

IV. SYSTEM OVERVIEW

The following is a continuation from previous work where a strong case has been established for exploiting the synergistic combination of DM ontology and the case based reasoning paradigm [15]. This section briefly introduces the key features of both our CBR system and DM ontology implementations. As illustrated in Figure 1, our hybrid DM assistant consists primarily of six major components: a DM Case Base, a DM Ontology, a Case Reasoner, Rule Reasoner, a DL Reasoner (Description Logic), and a DM Assistant Interface. For the DM Case Base, we chose to use the CRISP DM data mining process as a basis for eliciting a set of representative features for our case representation. CRISP DM efficiently captures “knowledge” (in the form of a series of well defined phases, tasks and activities) of the entire data mining effort. From this, we were able to define a DM case representation consisting of 53 features. The majority of our indexes were derived from measures used in the area of data characterization (i.e. general, statistical and information theoretic) [16]. For the reasoning component of our CBR system, we implemented a feature weighted, instance based learning algorithm (IBL) [17].

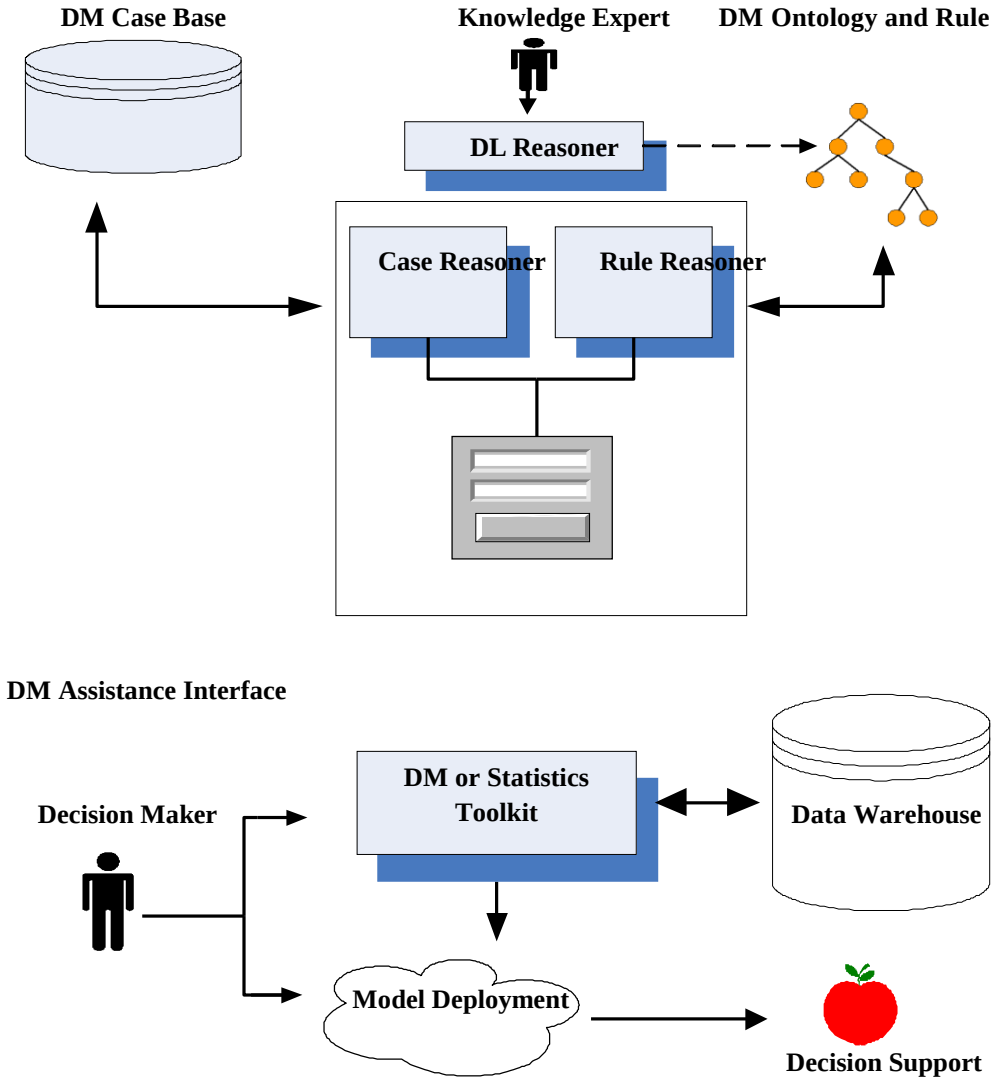

Figure1. System Architecture of Intelligent DM Assistant

Our preliminary CBR evaluation has yielded some promising DM assistant results. Although through its use, the CBR system is capable of learning useful "business problem to DM case" knowledge, it fails to provide the "deeper" knowledge that is essential for supporting a non specialist data miner (i.e. for case adaptation). As such our DM ontology (by formally capturing concepts, relationships, constraints and rules) is capable of complementing the CBR system and addressing this need by assisting the non expert data miner by means of recommendations and heuristics during the course of a DM task. The implementation of our ontology has consisted of two separate phases: (1) the high level knowledge representation of the CRISP DM methodology and (2) the representation of detailed DM knowledge in the form of concepts and rules.

During the early stages of our investigations on the nature of detailed DM knowledge (i.e. how to deal with the class imbalance problem), we concluded that such knowledge tends to most appropriately take a procedural or rule like form. Hence, we have elicited a preliminary set of rules for providing intelligent DM assistance (i.e. heuristics, recommendations, automatic responses) during the key phases of the DM process (Data Understanding, Data Preparation, Data Modeling, Evaluation). These rules are implemented using a proposed rule language standard for the semantic web (SWRL [18]). More specifically, these were implemented using the Protégé SWRL Tab plug in [19]. Subsequently during operation of our DM assistant, a rule based reasoner (JESS [20]) operates on a set of "facts". The facts consist of both automatically supplied and user supplied case attributes. When appropriate the rule based reasoner "fires" rules which then provide heuristics and recommendations to a user or automatically.

We shall be using the general term "advice" to represent any assistance provided by the system (i.e. text message or automated fact response), we do make a clear distinction between a recommendation and a heuristic (both are sub types of the term advice). A recommendation is a more formal type of advice (assertion) while a heuristic should be interpreted less formally by a user (i.e. rule of thumb). The CBR and DM ontology subsystems have well defined knowledge representation roles. The DM ontology defines and manages high level concepts (i.e. tasks, activity types, algorithms, etc.) while the CBR holds detailed case information (i.e. data preparation steps, model parameters, etc.). From another perspective the CBR learns problem (i.e. business and data characteristics) to solution (i.e. data preparation, modeling and evaluation) mappings while DM ontology provides additional assistance (complements where the CBR lacks knowledge) to a user during the

various phases of the DM process. Although not the focus of this paper, the DM ontology also provides the user with basic definitions for all the vocabulary terms used within the DM Assistant Interface. For the moment we are mostly making use of a DL reasoner for the purposes of assuring the consistent evolution of our ontology (consistency checking).

V. ONTOLOGY BASED HYBRID INTELLIGENCE DM ASSISTANCE

This section shall be addressing how the above mentioned system components are synergistically combined to provide a novice data miner with ontology guided DM assistance. In order to facilitate the discussion that follows Figure 2 essentially provides a abstract view of the principle components, some important DM case attributes are represented by the DM Assistance Information Grid, an ontology segment represents some detailed knowledge and several SWRL rules are given. In addition, for the moment ontology guided DM assistance is primarily restricted to the three most important phases of the CRISP DM methodology (i.e. Data Understanding, Data Preparation and Data Modeling). Although the CBR paradigm provides the benefit of retrieving similar cases, the required solution part is rarely an exact match to the current DM problem being attempted.

Hence after the retrieval and reuse CBR phases are completed, the user is faced with the grand challenge of examining the chosen cases' contents and revising certain attributes (i.e. data quality verification, data preparation or data modeling values) in order to retrofit the case to reflect the state of the current DM problem. As a result we have attempted to enrich our DM assistant with complementary knowledge (OWL ontology concepts, individuals and rules) in order to provide the user with adaptation or validation knowledge to complete her DM task.

5.1 Case Facts

Our system is essentially data driven and employs a forward chaining rule based inference engine (JESS). A user basically interacts with the DM Assistant Interface (the DM Assistant Information Grid in Figure 2) by entering or modifying a series of DM case attribute values. The abbreviated DM Assistant Information Grid represents the state of the "working memory" of the system. As the user changes the state of the working memory, the SWRL rules come into play to provide automatic responses (modifying facts) and advice in the form textual messages. The main purpose of the textual messages is to actively assist and empower the user to provide acceptable fact values. Typically having chosen a basis case to work with for a given DM task, a user progresses through the CRISP DM methodology by answering facts.

5.2 Initial Bootstrap Advice

Under ideal circumstances, the state of the initial working memory should be adequately specified from automatically provided facts (i.e. "Ratio of missing values", etc.) to allow the firing of certain rules (and subsequent automatic responses and/or advice) to move the DM process forward. Nevertheless, there are circumstances when user input is required (i.e. identification of incomplete or erroneous values, table joins performed on the data). When such facts are required directly from the user, initial textual messages (bootstrap advice) are given to the user, explaining how to acquire the missing information. This approach is analogous to traditional AI interview or conversational techniques used for soliciting tacit information from the user.

5.3 Detailed Ontology Knowledge

We have currently implemented an OWL DL ontology (using the Protégé editor) of approximately 200 data mining concepts comprising of methodology knowledge (CRISP-DM) and detailed DM knowledge (Data Preparation Advice, Data Modeling advice, DM algorithms, etc.). The concepts illustrated in the ontology segment of Figure 2 (starting from the root Data Prep Advice concept), represent important potential data mining problems that can have a significant impact on the final quality of a generated model. For instance, some algorithms can perform poorly if the quantity of examples becomes large while other machine learning algorithms can be Fairly easily mitigate these problems significantly affected by the curse of dimensionality (too many attributes). An experienced data miner can by applying a supplementary procedure (i.e. aggregation, a cost sensitive learning method, etc.). Specific advice for a given problem is represented by ontology individuals (as indicated by dashed ovals). Complemented with the rule based reasoning approach discussed above, the advice can contain specific textual attributes or associations to more specialized individuals that provide further advice as order to target varying levels of ontology knowledge detail depending on the user's level of DM expertise. For instance, an expert DM user may be satisfied with getting general advice such as "Apply a Dimensionality Reduction technique", while a less experienced user may wish a specific recommendation for using a particular technique such PCA (Principal Component Analysis). The detailed DM knowledge (in the form of SWRL rules) was mainly elicited from introductory data mining texts ([21], [22]), the Weka mailing list [23], scientific articles (too numerous to mention in this paper) and our own DM experiences.

DM Assistant Information Grid

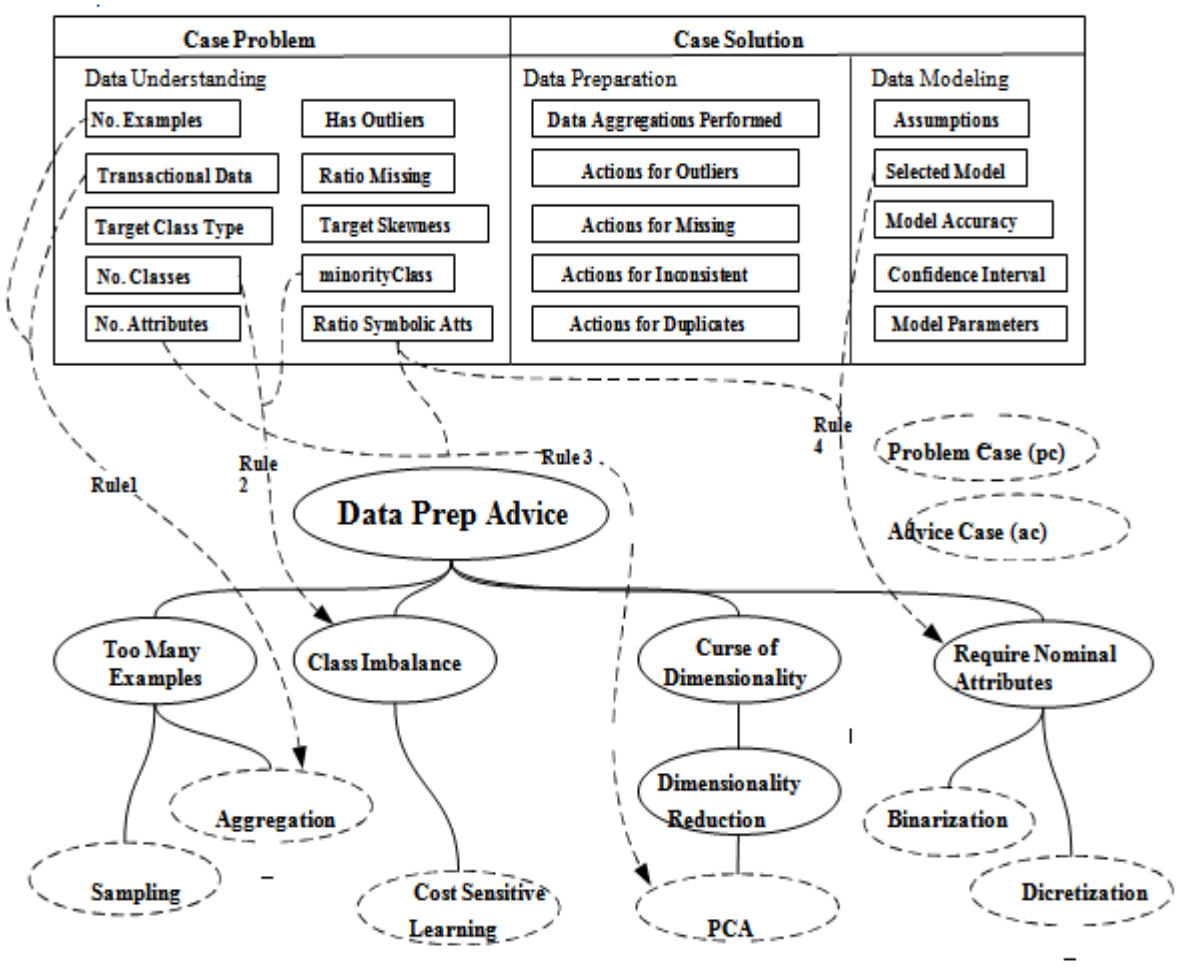

Figure 2. Abstract View of the CBR (case facts), Ontology Knowledge and SWRL Rules

5.4 The Synergistic SWRL Rule Set

The SWRL rules provide two important benefits for the realization of our system, (1) a convenient method for expressing domain knowledge as a set of antecedent consequent pairs and (2) provide an integration mechanism for bridging knowledge from two disparate knowledge bases (CBR and ontology). We have currently defined a preliminary rule set of approximately 50 SWRL rules that span the three main phases of the CRISP-DM methodology. Due to space limitations the examples herein shall only be limited to the Data Preparation phase (some rules are illustrated as dashed lines in Figure 2. For instance, the following rule represents detailed knowledge that may be required for successfully performing the data preparation phase:

Rule1 := NoExamples (pc, ?x1) ^ swrlb:greaterThan(?x1, 30000) ^ transactionData(pc, True) => advice(ac, aggregation)

The above SWRL rule asserts that if the problem case (pc) has an "example count" greater than 30000 and the dataset is of a "transactional type", the user should consider performing an aggregation operation over the dataset. An arbitrary "adaptation case" (ac) individual is used for holding advice values. Furthermore, Rule2 below essentially expresses that if a binary class problem has its minority class represented by less than 15%, a class imbalance problem may be eminent:

Rule2 := numOfClasses(pc, 2) ^ minorityClass(pc, ?x1) ^ swrlb:lessThan(?x1, 0.15) => advice(ac, classImbalance)

Hence, the class Imbalance individual provides advice by offering a cost sensitive learning algorithm (to attempt to improve overall model performance). In addition, the following rule asserts that if the quantity of attributes is greater than 20 (but less than 50 as PCA can be computationally prohibitive) and the "symbolic attributes ratio" is zero(only numerical values)then the system would recommend specifically using the PCA dimensionality reduction technique:

Rule3 := noAttributes(pc, ?x1) ^ swrlb:greaterThan(?x1, 20) ^ swrlb:lessThan(?x1, 50) ^ ratioSymbAttributes(pc, 0) => advice(ac, PCA)

5.5 Current Focus and Limitations

Since the area of data mining is a highly knowledge rich environment (i.e. data refining, feature transformation, algorithms, parameters, evaluation, etc.) it is impossible to foresee capturing all the DM knowledge that is required to support users under all conceivable circumstances. Hence our current prototype detailed ontology knowledge is currently constrained to the following:

(1) Support the data preparation phase for handling common data quality and model input requirements.

(2) Support for common classification models (i.e. linear/logistic regression, naïve bayes, most decision trees, support vector machines.

(3) Common data modeling issues (i.e. class imbalance, curse of dimensionality, basic model over fitting avoidance)

(4) General knowledge for model evaluation (i.e. P-values, cross validation, ROC curves).

(5) Specific tool dependent knowledge is only available for the Weka environment.

Hence more advanced topics such as Meta learning, feature selection, massive datasets, model comparison methods and intricate classifier parameter details are not yet covered. Realistically, our objective has been to elicit a "first pass" to capture common DM knowledge (as is pertinent to our application domain - see Section 6) and subsequently evolve our ontology as the needs arise (i.e. to handle specialized and exceptional DM process conditions).

VI. DISCUSSION

A prototype version of our DM assistant has recently been deployed to support a strategic decision support department. The business objectives consist of analyzing large amounts of student academic details and deriving various predictive and explanatory models using a range of data mining tools (i.e. Oracle Data Miner, SAS Data Miner and Weka). During our early attempts at soliciting detailed DM knowledge using our OWL DL based ontology, we quickly encountered several problems when attempting toimplementproceduralor rule like knowledge,

(1) DL based ontologies are declarative in nature.

(2) Attempting to use existing ontology query languages (i.e. SPARQL, RDQL) for emulating reasoning mechanisms was deemed unmanageable. Hence, this problem was resolved by making use of SWRL rules and an external rule based inference engine (JESS). Several recent research activities in the area of the semantic web have demonstrated a similar need for integrating a rule base and associated reasoner with ontologies ([24], [19]).

In addition, it is worth noting that the SWRL rules could have been implemented purely using propositional rules (without using ontology concepts and individuals). Nevertheless, we believe that the formal capture and representation of detailed DM knowledge within ontology provides some important benefits,

(1) It provides a more explicit form of knowledge representation that is more amenable to human interpretation.

(2) It may provide a knowledge representation format that can be readily exploited by other reasoner (i.e. DL reasoner).

(3) Unlike traditional rule bases where the relationships between the rules tend to be "opaque", the explicit representation of linguistic variables as formal ontology concepts facilitates rule set maintenance.

Overall the proposed approach provides an additional benefit in that knowledge management efforts can be performed in several independent stages. For instance, declarative DM knowledge can first be elicited, and subsequently another domain expert can make use of this knowledge to craft a set of SWRL rules for expressing procedural DM knowledge.

VII. CONCLUSIONS AND FUTURE WORK

Although much work remains to be done, we have presented hybrid architecture for an intelligent data mining assistant. The following are some of the novelty features that our intelligent DM assistant attempts to provide. First by combining both declarative and procedural ontology knowledge, the system addresses the need for supporting the non expert data miner throughout the key phases of the DM process. In addition, the evolution of both knowledge based components (CBR and ontology) provides an effective means to foster knowledge reuse.

Furthermore the use of the DM ontology provides a natural extension to our existing CBR for addressing the need for "deeper" knowledge to empower the data miner. Though our prototype currently only supports classification and regression activities, plans for future research are underway to include the support for clustering and association mining. The next steps will lie in conducting more DM activities in order to increase the size of our case base and elicit more relevant detailed DM knowledge. We are highly optimistic that this synergistic combination of DM ontology knowledge (both declarative and procedural) and the case based reasoning paradigm can significantly empower a non expert data miner for effectively carrying out data mining activities.

References

[1] S. Delisle, Integrating Data Mining and Decision Support via Computational Intelligence, PMKD-DEXA Workshop. Copenhagen, Denmark, 2005
[2] C. Giraud-Carrier, Reporting on the Data Mining Advisor. Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2005.
[3] A. Kalousis and M. Theoharis, NOEMON: Design, Implementation and Performance Results of an intelligent Assistant for Classification Selection. Intelligent Data Analysis, Elsevier, 1999
[4] MetaL, A Meta Learning Assistant for Providing Data Mining Support, www.metal-kdd.org.
[5] G. Linder and R. Studer, AST: Support for Algorithm Selection with a CBR Approach. Recent Advances in Meta Learning and Future Work, 1999, 38-47
[6] A. Suyama, N. Negishi, and T. Yamaguchi,CAMLET: A Platform for Automatic Composition of Inductive Applications Using Ontologies.Progress in Discovery Science, 2002
[7] J.Phillips and B. Buchanan, Ontology-Guided Knowledge Discovery in Databases, International Conference on Knowledge Capture. Victoria, Canada, 2001
[8] A. Bernstein, F. Provost, and S. Hill, Intelligent Assistance for the Data Mining Process: An Ontology based Approach, IEEE Transactions on Knowledge and Data Engineering, 2005.
[9] M. Bauer and S. Baldes, An Ontology Based interface for Machine Learning, Intelligent User Interfaces. San Diego, California, 2005
[10] M. Cannataro and C. Comito, A Data Mining Ontology for Grid Programming, 1st Int. Workshop on Semantics in Peer to Peer and Grid Computing 2003
[11] CRISP-DM1.0, A Step by step Data Mining Guide, www.crisp-dm.org, 2000
[12] A. Aamodt, Knowledge-Intensive Case Based Reasoning in CREEK, Advancesin Case Based Reasoning, 7th European Conference. Madrid, Spain, Springer, 2004
[13] J. Bello-Thomas, P. Gonzalez Calero, and B. Diaz-Agundo, JColibri: An Object-Oriented Framework for Building CBR Systems, Advances in CBR,7th European Conference. Madrid, Spain, Springer, 2004
[14] I. Bichindaritz, Mémoire: Case Based Reasoning Meets the Semantic Web in Biology and Medecine. Advances in Case based Reasoning,7th European Conference, LNAI, Spain, 2004
[15] M .Charest, S. Delisle, O. Cervantes,and Y. Shen, Intelligent Data Mining Assistance via CBR and Ontologies, to appear in DEXA-PMKD Workshop. Poland, 2006
[16] C. Castliello, G. Castellano, and A. Fanelli, Meta Data: Characterization of Input Features for Meta Learning. Modeling Decisions for Artificial Intelligence, LNAI, 2005, 457-468
[17] D. Aha, D. Kibler, and M. Albert, Instance Based Learning Algorithms. Kluwer Academic Publishers, 1991. 6, 37-66
[18] SWRL SemanticWebRule Language, http://www.w3.org/Submission/SWRL/
[19] M. O'Connor, H. Knublauch, S. Tu, B. Grosol, M. Dean, W. Grosso, nd M. Musen, Supporting Rule System Interoperability on the Semantic Web with SWRL, ISWC, LNCS 3729. Berlin, 2005
[20] JESS,Java Expert System Shell, http://herzberg.ca.sandia.gov/jess/
[21] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining (Boston, MA: Addison-Wesley, 2005).
[22] T. Hastie, R. Tibshirani, and J. Friedman, The Element of Statistical Learning, Data Mining, Inference and Prediction (Berlin, GmbH: Springer, 2001).
[23] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. 2 ed (San Francisco, CA: Morgan Kaufman, 2005).
[24] C. Golbreich, Combining Rule and Ontology Reasoners for the Semantic Web, Rules and Rule Markup Languages for the Semantic Web Springer, 2004.