The Knowledge Engineering Review, Vol. 25:1, 49–67. © Cambridge University Press, 2010
doi:10.1017/S0269888909990361

Toward an integrated knowledge discovery and data mining process model

SUMANA SHARMA and KWEKU-MUATA OSEI-BRYSON

Department of Information Systems, the Information Systems Research Institute, Virginia Commonwealth University, Richmond, VA 23284, USA;
e-mail: Sharmas5@vcu.edu, KMOsei@.vcu.edu

Abstract

The knowledge discovery and data mining (KDDM) process models describe the various phases (e.g. business understanding, data understanding, data preparation, modeling, evaluation and deployment) of the KDDM process. They act as a roadmap for implementation of the KDDM process by presenting a list of tasks for executing the various phases. The checklist approach of describing the tasks is not adequately supported by appropriate tools, which specify 'how' the particular task can be implemented. This may result in tasks not being implemented. Another disadvantage is that the long checklist does not capture or leverage the dependencies that exist among the various tasks of the same and different phases. This not only makes the process cumbersome to implement, but also hinders possibilities for semi-automation of certain tasks. Given that each task in the process model serves an important goal and even affects the execution of related tasks due to the dependencies, these limitations are likely to negatively affect the efficiency and effectiveness of KDDM projects. This paper proposes an improved KDDM process model that overcomes these shortcomings by prescribing tools for supporting each task as well as identifying and leveraging dependencies among tasks for semi-automation of tasks, wherever possible.

1 Introduction

Data have emerged as a new found source of competitive advantage in an era in which traditional bases of competition have largely evaporated (Davenport & Harris, 2007). This competitive advantage is based on the knowledge gained from the analysis of data and has catapulted to the forefront fields like data mining and knowledge discovery, which offer techniques and processes for extracting this knowledge. Knowledge discovery is widely acknowledged as an interactive and iterative multi-step process ranging from the development of business (or domain) understanding, data understanding, data preparation, modeling (or data mining), evaluation and ultimately deployment (consolidation) of discovered knowledge. This process is embodied in the form of knowledge discovery and data mining (KDDM) process models that describe the steps/phases and tasks involved in the knowledge discovery process.

Our review of existing KDDM process models (Fayyad *et al.*, 1996a; Berry & Linoff, 1997; Anand & Buchner, 1998; Cabena *et al.*, 1998; Cios *et al.*, 2000; Han & Kamber, 2001; CRISP-DM 2003; Cios & Kurgan, 2005) reveals that they suffer from certain common deficiencies such as they often (i) present the complex knowledge discovery process in a checklist manner; and (ii) present a fragmented view of the KDDM process and do not explicate the various dependencies existing in the knowledge discovery process. A disadvantage of the latter is that it hinders the potential for semi-automation of tasks, affecting the efficiency with which the KDDM projects can be

carried out. Some other limitations include the lack of support for execution of the various tasks (Charest *et al*., 2006) and lack of attention toward the business understanding (BU) phase (Sharma & Osei-Bryson, 2008b), both of which have been highlighted in the literature.

This paper aims to address the deficiencies of existing models (Fayyad *et al*., 1996a; Berry & Linoff, 1997; Anand & Buchner, 1998; Cabena *et al*., 1998; Cios, Teresinska *et al*., 2000; Han & Kamber, 2001; CRISP-DM, 2003) through an improved integrated model that addresses the limitations of the existing process models thereby improving the efficiency and effectiveness with which KDDM projects are currently carried out. The scope of the integrated model includes all phases of the KDDM process, except the deployment phase.

This paper is organized as follows: we provide an overview of KDDM process models and a discussion of limitations of KDDM models that this paper addresses through the proposed solution in section 2. In section 3, we present the proposed solution and discuss how it achieves the set objectives. Section 4 summarizes the objectives and contribution of the research.

2 Overview of relevant literature

In this section, we discuss (a) KDDM process models and their instances to describe the lifecycle of KDDM projects; and (b) the common deficiencies in existing process models and the effect of these deficiencies on the efficiency and effectiveness with which KDDM projects are currently being executed.

2.1 KDDM process models and examples

Several KDDM process models have been proposed (Fayyad *et al*., 1996a; Berry & Linoff, 1997; Anand & Buchner, 1998; Cabena *et al*., 1998; Cios *et al*., 2000; Han & Kamber, 2001; CRISP-DM, 2003; Cios & Kurgan, 2005), which, while differing in their level of detail, describe the same sequence of phases from BU to deployment to describe the lifecycle of a KDDM project. (e.g Table 1). For a survey and comparison of different KDDM process models, please refer to Kurgan & Musilek (2006). Figure 1 shows the schematic of the CRISP-DM process model, where CRISP-DM (2003) is an acronym for Cross Industry Standard Process (CRISP) for data mining (DM).

Table 1 Description of Phases of CRISP-DM

Phase	Description
Business understanding	Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives
Data understanding	Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information
Data preparation	Covers all activities to construct the final modeling data set (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools
Modeling	Relevant modeling techniques are selected and applied and their parameters are calibrated to optimal values
Evaluation	Consists of thoroughly evaluating the model and reviewing the steps executed to construct the model to make sure that it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached
Deployment	This phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise

CRISP-DM, Cross Industry Standard Process (CRISP) for data mining.

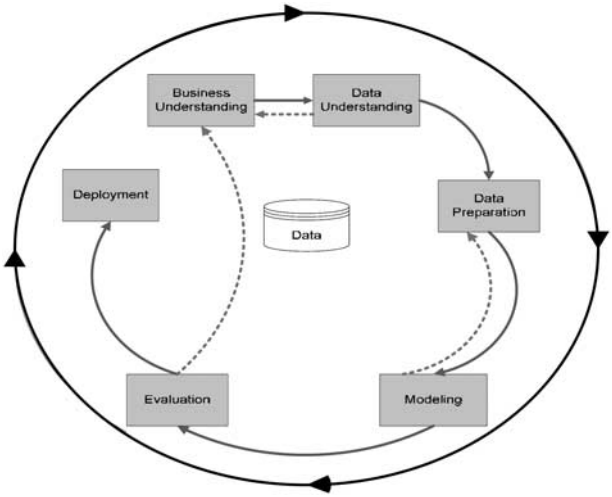

Figure 1 Cross Industry Standard Process model (CRISP-DM, 2003)

2.2 Limitation of KDDM process models

Our analysis of existing KDDM process models reveals that they suffer from certain common deficiencies. We discuss these various limitations and their effect on the efficiency and effectiveness with which KDDM process models are currently executed.

2.2.1 Description of the KDDM process in a checklist manner

Existing KDDM process models typically describe the complicated KDDM process in terms of a valid list of steps (sometimes also referred to as tasks or activities) in the checklist, which while providing a broad guideline, could still be perceived as being very cost prohibitive to implement. For example, CRISP-DM (2003) recommends executing a total number of 288 activities, which, when presented in a checklist approach, are likely to be intimidating to personnel involved in executing the project.

2.2.2 Fragmented view of the KDDM process

Existing KDDM process models do not capture or highlight the important dependencies (i.e. interrelationships between the various steps, or between the various phases and tasks) existent in a typical KDDM process. Consider Figure 1, which also represents the typical flow of any KDDM process. The dependency that is most obvious from this model is the phase–phase dependency resulting from the ordering of phases proposed by the model. These dependencies are critical, as they cannot be reversed without leading to detrimental effects or even incapability of executing a particular phase. Since a phase really comprises of various tasks, then its output of a phase really comprises of the output of the diverse array of tasks that lie within it. Therefore, it is important to explicate and highlight the task-level dependencies also, but these are not shown in Figure 1. This issue has not been addressed in existing KDDM process models, although CRISP-DM briefly alludes to the capturing of dependencies, but does not incorporate them in the design of the model.

The repercussion of not explicating various dependencies existent in the context of a KDDM project could lead to inefficient/ineffective implementation of projects. For example, the selection of a modeling algorithm without clearly formulating the business objective(s) first is an important task–task dependency, which, if neglected, can lead the project to take a completely different direction than is appropriate.

2.2.3 Fragmented view: a hindrance to semi-automation

Although there was the assumption that only the task of implementation of DM methods (modeling phase) was amenable to automation (Berry & Linoff, 2000) for some time, recently

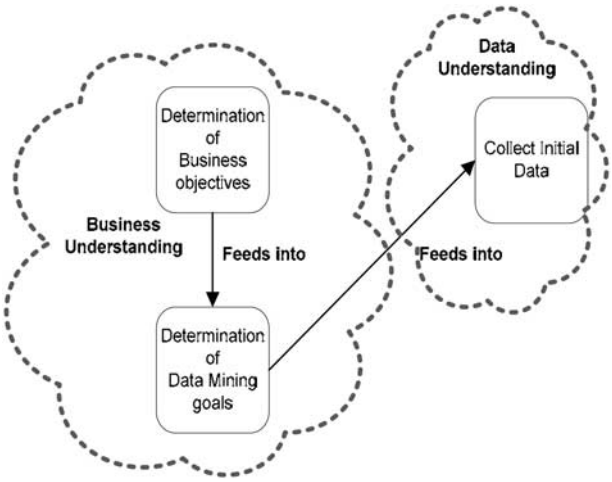

Figure 2 Explicating of dependencies as a first step toward enabling semi-automation

however, researchers have realized that some of the other tasks, such as the selection of appropriate modeling techniques or algorithms (Bernstein *et al.*, 2005), are also amenable to full or partial automation. Support for automation of relevant aspects of the DM process, however, requires the specification of an integrated process model in which task–task dependencies have been explicated.

Continuing with the example presented in the above section, we argue that the identification of dependency between two tasks such as a business and DM objective should be leveraged to drive the execution of the latter task. For instance, effort should be made to examine whether the output of business objects can be used to semi-automate tasks, such as determination of DM objectives, that utilize it as its input (see Figure 2).

2.2.4 Lack of support for the end-to-end KDDM process

Existing KDDM models do not provide enough support to 'how' to implement the long list of tasks and activities suggested by them (Charest *et al.*, 2006). Charest *et al.* (2006) note that existing process models 'only provide general directives, while what a non-specialist really needs are explanations, heuristics and recommendations on how to effectively carry out the particular steps of the methodology'. Therefore, it is necessary that the process models should be complemented with appropriate tools and techniques for carrying out the various tasks in order to prevent or at least minimize the non-execution of relevant tasks during the knowledge discovery process.

Although overtly, it may appear that this issue is less problematic in case of the modeling phase that has benefitted from the rapid advancement in development of plethora of DM techniques. As noted by Simoudis *et al.* (1996) that a single DM technique is often insufficient for extracting knowledge from a data set, the modeling phase requires careful selection of the techniques is required if the objectives of the project are to be accomplished (Pyle, 2003). Therefore, support is needed to aid the user in selecting these techniques and the order in which they should be used if the KDDM project is to be effectively executed.

2.2.5 Lack of adquate attention toward BU phase

The importance of BU phase, which includes making determinations about business and DM objects, assessing resources and generating a project plan for the remainder of the project, cannot be overemphasized. However, our review of published DM case studies reveals that that the BU phase of KDDM projects is often implemented in an ad hoc manner (Sharma & Osei-Bryson, 2008b). We believe that the reason for such an unstructured approach is the general lack of support toward the manner in which (how) the tasks of this phase can be implemented.

This issue has been highlighted and somewhat addressed by Pyle (2003) who describes how the real-world business problems (to be addressed through DM) can be modeled. While the author has not based his approach on any particular DM methodology, he discusses various tools to carry out many (though not all) of the activities prescribed under the BU phase of the CRISP-DM methodology. However, they are only presented in a linear fashion, with the description of each activity followed by a brief description of a proposed tool. The overall framework that consists of nested sequences of action boxes, discovery boxes, technique boxes and example boxes is complicated to navigate, and may appear to be cumbersome or even cost prohibitive to actors involved in carrying out the critical BU phase.

3 Design of an integrated KDDM process model

The objective of this study is to design an improved KDDM process model that addresses the deficiencies of existing models. Given that the design is a goal-oriented activity (Simon, 1996), the requirements that the proposed model should meet must be clearly outlined. These are also necessary for adequate evaluation of the proposed solution against the set requirements. The requirements that the proposed solution must address are described in Table 2 below.

Next we describe how we designed the proposed solution in the form of the improved KDDM process model. The design of the proposed model incorporated treating each phase and its constituent tasks to understand the task–task dependencies existing among the various tasks of the same phase. The next step was to integrate the various phases together by linking the task–task dependencies existing among the tasks of the various phases. The final step was to carefully analyze all the task–task dependencies (same phase and between different phases) to identify opportunities for leveraging the dependencies identified through semi-automation. A simultaneous consideration was also given to prescribe approaches and or tools for implementing each task of the KDDM process. In Table 3, we present the chief tasks of all the phases of the KDDM process, their output, and the tools that can be used for implementing the given task and an indication as to whether a given task can be a candidate for semi-automation.

Table 2 Design requirements for the integrated KDDM model

Issues Identified (as-is situation)	Design requirements for the integrated model (to-be situation)
Description of the KDDM process in a checklist manner	Present a user-oriented coherent description of the KDDM process
Fragmented view of the KDDM process	Develop an integrated view of the KDDM process by explicating the various phase–phase and task–task dependencies
Emphasis on feedback loops before the complete understanding of the primary sequencing of phases and tasks in a KDDM process	Explicate sequencing of the various phases and their tasks before identifying feedback loops and establishing conditions under which the loops would get triggered
Fragmented view serves as a hindrance to building an integrated process model and 'semi-automating' tasks	Leverage the dependencies explicated in the integrated process model to drive semi-automation of tasks
Lack of support for the end-to-end KDDM process	Prescribe approaches for offering decision support to all tasks described in the integrated KDDM model
Insufficient discussion and conspicuous lack of support toward execution of business understanding phase—the foundational phase of a KDDM process	Discuss significance of the business understanding phase and uses the tasks recommended for this phase as the basis for developing the integrated model

KDDM, knowledge discovery and data mining.

Table 3 KDDM process: phases, tasks, output, tools and opportunities for semi-automation

	Methods/tools	Repositories/sources	Output
Business understanding phase			
Creation of business objectives (selection among a set of competing objectives)	Goal mapping, cognitive mapping AHP	DM projects base	Business objectives
Business objectives to business success criteria	GSS	DM projects base	Business success criteria
Business objectives to DM objectives	VFT, GSS	DM projects base	DM objectives
DM objectives to DM success criteria	GQM method	Cross reference matrix (tables 6 and 7)	Evaluation measures and thresholds for modeling phase
Determination of preference function**	Preference function elicitation tool (e.g. AHP)	Domain experts	Preference function to be used in evaluation phase (e.g. weights for evaluation measures)
Determination of value functions for relevant evaluation measures**	Domain expert	Value function repository	Value function (s)
Identification of applicable data resources		Domain experts metadata repository	List of required data sets
Verification of data		Business rules base	List of available data
Identification of relevant personnel**		Ontologies, organization charts, skills/competency base	List of available personnel
Clarification of business requirements	Requirement elicitation tools	Domain experts	List of business requirements
Business objectives to financial constraints	GSS		
DM objectives to relevant modeling techniques		DM cross reference matrix (table 9)	Relevant modeling techniques
Identifying benefits and risks from a business perspective**	GSS	DM projects base domain experts	Statement of expected benefits (tangible and intangible) and risks
Creation of contingency plans	GSS	DM projects base domain experts	Contingency plan for each risk situation
Estimation of data collection, implementation and operational costs	Project management cost estimation tools	External data sources	Statement of expected costs
Cost–benefit analysis	Automated cost–benefit analysis tools	Domain experts DM projects base	Statement of Costs and benefits
Data understanding phase			
Analyze data for anomalies, missing values and outliers	DM software, basic statistical analysis	Domain expert(s)	Data quality report
Analyze data to explore relationships among variables and exploring potential for derived attributes		Domain expert(s), metadata repository	Data exploration report

Table 3 *(Continued)*

Data preparation phase			
Create data set for analysis		Domain expert(s)	Integrated data set containing the relevant data
Format data in accordance with the first modeling technique**		Software tools base, domain expert(s)	Formatted modeling data set
Modeling phase			
Run models using the first technique in the array of applicable techniques	DM software		Output of modeling technique
Run models using all applicable techniques	DM software	Cross reference matrix (tables 9 and 10)	Output of all relevant modeling techniques
Compare results of models' output from different modeling techniques against DMSC setup earlier**	MS Excel	DMSC, domain experts	Model results assessed with respect to each DMSC

KDDM, knowledge discovery and data mining; DM, data mining; DMSC, DM success criteria; AHP, analytic hierarchy process; GSS, group support systems; VFT, value-focused thinking; GQM, goal question metric.
** Candidate tasks for semi-automation.

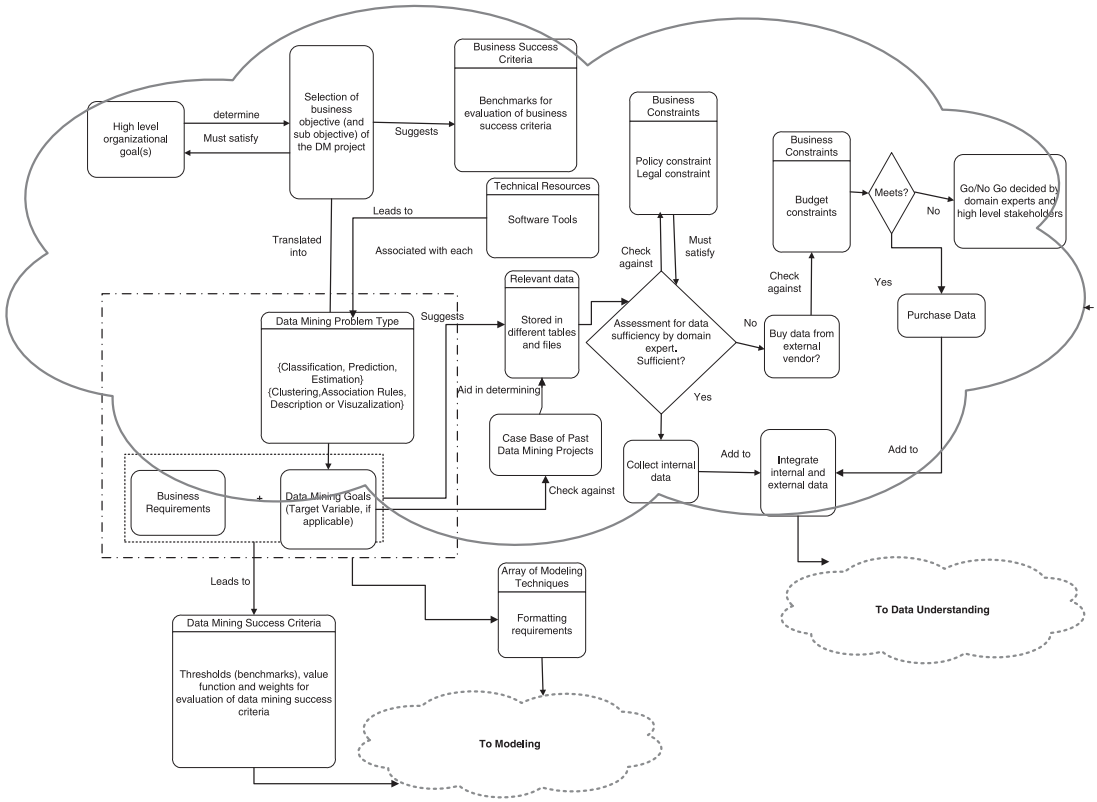

Figure 3 Business understanding Phase

Table 4 SMART criteria for evaluating business objectives

Criterion	Description
Specific	The business objective must clearly describe the objective of the project and should relate to one or more higher level organizational objectives
Measurable	Concrete, clearly defined criteria should be laid down for measuring the attainment of the proposed business objective. These criteria are referred to as business success criteria and are described in the next task
Attainable	The business objective must be agreed upon by the key stakeholders involved in the project
Realistic	The business objective must be achievable within the constraints of the available resources, knowledge and time
Time-bound	There should be clear deadlines for achievement of the business objective

3.1 Business understanding phase

In Figure 3, we present our explication of dependencies among the various tasks of this phase. In the remainder of this subsection, we describe how the various tasks of this phase could be implemented.

3.1.1 Formulation of business objectives

The formulation of business objectives is a multi-step process and requires collaboration among various high-level business stakeholders of the company. Doran (1981) proposed the SMART criteria (see Table 4) for evaluating the quality of business objectives. Of the five SMART criteria, the measurable and realistic criteria are implemented through separate KDDM tasks, namely setting up of business success criteria and assessment of inventory of resources, respectively.

This serves as a reminder that the setting up of business objectives is not a one-step process and needs to be revisited to finalize the well thought through but preliminary business objectives that are set up at the end of the completion of this first task.

Consider the following illustrative example that involves Financial Services Company and Global Credit that wishes to revise one of its outdated credit-scoring models as a means of meeting one of its organizational goals of improving profit.

> The business objective of a DM project launched by the Credit Risk Division of Global Credit is 'to improve profits over the Financial Year 2008–2009, by improving approval rates of sub prime customers by 5% while maintaining better or similar loss rates'. Based on past data analysis, it has been observed that a 5% increase in approval rate while maintaining similar loss rates leads to approximately $5 million increase in net profit assuming everything else remains constant.

This business objective satisfies the specificity criteria, as it relates to at least one high-level organizational objective, here, improvement in profits. Lack of association between a business objective of a DM project and organizational objectives makes the business objective vague and ill formulated. The stated objective also specifies the timely criterion and specifies the time frame during which the business objective must be accomplished for the project to be considered successful.

If available, an exploration of the DM projects base (which is a repository of past DM projects) could be used to determine whether a similar project has been conducted in the past and note must be taken of the various details of the project such as entities involved, solution chosen, benefits and contingencies, findings, etc.

3.1.2 Setting up of business success criteria

The goal question metric (GQM) approach (Basili & Weiss, 1984) can be used to formalize business success criteria that specify how the outcome of implementation of the business objective can be evaluated. The GQM approach proposes refining the overall goal (business objective in the case of a DM project) into a set of questions, and then refining the questions into a set of metrics, which could be objective or subjective (see Figure 4).

The metrics help to implement the measurability criterion associated with well-formulated business objective. The metrics can be objective or subjective in nature. We use the example presented above to show how the GQM approach can be used for setting up business success criteria. The metrics describe the business success criteria and must meet the threshold values specified in the statement of objectives. For instance, in case of the example shown in Figure 3, (a) Δ dollar profits should be $\geq$ $5 million; (b) Δ approval rate $\geq$ 5%; and (c) Δ loss rate ≤ 0, as the objective is to maintain better or similar loss rates.

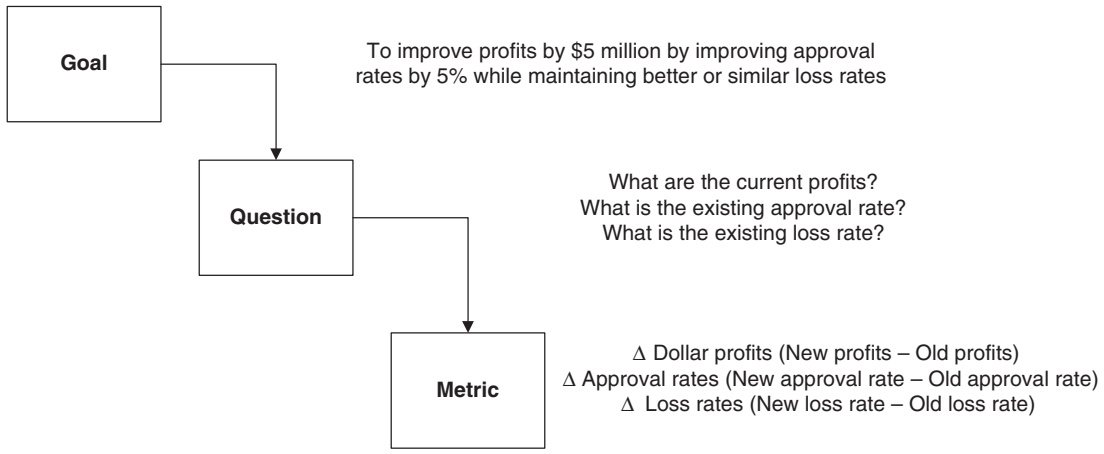

Figure 4 Implementation of GQM approach (Basili & Weiss, 1984) for setting up of business success criteria

3.1.3 Analysis of inventory of business personnel and other business resources

This task ensures that the business personnel, key high-level stakeholders, domain experts and other organizational actors who will be part of the project team are available for the duration of the project. An organization ontology (Fox *et al.*, 1998; Sharma & Osei-Bryson, 2008a) or organizational chart can play an important role in identifying the relevant personnel for DM projects. An organizational chart can be used to locate particular individuals and their role in the hierarchy of an organization. With an organizational ontology, such as that of Sharma & Osei-Bryson (2008a), identification of agents or relevant personnel can be accomplished by simply navigating through the various links in the organizational ontology. Once identified, the domain experts and other key business personnel can be used to elicit information about relevant business resources such as business glossary and business metadata associated with the project.

3.1.4 Clarification of business requirements

All requirements of the project must be clarified through consultations with relevant business personnel. When the business objective is related to creating or refining models, an important aspect of requirement analysis should entail establishing details about whether or not an explanatory model is to be produced through the DM project. Requirement elicitation tools (Laguna *et al.*, 2001) can be used to aid the execution of this task.

3.1.5 Clarification of business constraints

Assessment of business constraints, such as policy constraints, legal and budgetary constraints, as well as availability of business personnel and business resources (described above) must be undertaken, as the potential solutions designed during the succeeding phases, such as data preparation and modeling, as well as tasks, such as identification of necessary data that are performed during the BU phase, must be in accordance with the business rules laid down by the organization. Legal constraints may prohibit an organization from using certain variables in a certain manner and must be satisfied in the naming of solutions. Budgetary constraints are also an important type of business constraints and must present details about the budget allocated to the given project. The business and technical personnel can assess whether or not their needs in the form of resources (personnel, data, tools, etc.) can be satisfied within the confines of the allocated budget.

3.1.6 Determination of DM objective

A DM objective is often defined as the technical translation of the business objective; but this definition by itself does not provide the user with enough guidance regarding the creation of a well-formulated DM objective. We propose using the technique of the value-focused thinking (VFT) (Keeney, 1996) to move from business objectives to DM objectives. VFT includes three types of objectives: fundamental objectives, means objectives and strategic objectives. Whereas fundamental objectives concern the ends that decision-makers value in a particular decision context, means objectives are the methods to achieve the ends. In the context of KDDM process, the fundamental objectives are the business objectives, whereas the means objectives are the DM goals. The DM goals (e.g. development of a more accurate classification model) are methods to achieve the ends, that is, the business objective (e.g. improvement in profits). Within this context, it is important to determine whether any of the means objectives can be addressed using DM methods, for if that is not the case then DM techniques are not appropriate for addressing the given decision-making problem. Once it is established that the means objective is amenable to the use of DM techniques, a formal process toward generation of a well-formulated DM objective should be employed. The first step of the process is to select the problem type best representing the proposed project. Each problem type involves certain features that must be taken into account to lead to a well-formulated DM project (see Table 5). The features of a well-formulated DM problem serve to confirm that the user has selected the correct DM problem type.

Table 5 Creation of well-formulated data mining objectives

Type of learning	Problem types	Features of a well-formulated data mining objective
Supervised (or directed)	Classification—if goal is to classify unseen records into predefined classes	Entity to be classified, name of target variable (types: binary, nominal, ordinal and interval/continuous)
	Estimation—if goal is to estimate the value of a continuous target variable	Name of target variable (type: continuous)
	Prediction—if the goal is to classify or estimate but based on some future behavior or estimated future value	If classification: entity to be classified, name of target variable (types: binary, nominal, ordinal and interval/continuous) If estimation: name of target variable (type: continuous)
Unsupervised (or undirected)	Clustering—if goal is to divide records into several clusters or segments	Records or subset of records to be divided into clusters
	Association rules—if goal is to study implicative co-occurring relationship between two sets of binary-valued transactional database attributes	Item set
	Description or Visualization—if goal is to explore or visually analyze the relationship between two of more variables	Set of variables to be explored using description or visualization techniques
Combination of undirected and directed with the output of undirected data mining being used to drive directed data mining	Output of clustering can be used to drive directed data mining efforts such as classification, estimation or prediction	Clustering: records or subset of records to be divided into clusters. Output of clustering to be used as input to classification or estimation If classification: entity to be classified, name of target variable (types: binary, nominal, ordinal and interval/continuous) If estimation: name of target variable (type: continuous)
	Output of association rules can be used to drive a directed data mining effort such as classification	Association rules: item set Classification: entity to be classified, name of target variable (types: binary, nominal, ordinal and interval/continuous)

Continuing with the example of Global Credit, how the DM objective can be formulated is given below:

> The business objective was set up as improving approval rates of subprime customers by 5% while maintaining better or similar loss rates. The business and technical personnel involved in the project realize that this is a prediction problem that requires creation of a classification model that improves the rank ordering of credit card applicants as compared to the existing model. The data mining objective therefore becomes to 'predict (problem type) the probability of charge off (target variable) of subprime credit card applicants (entity) within 12 months from the point of booking'.

3.1.7 Setting up of DM success criteria (DMSC)

The DMSC are used to evaluate the results of the implementation of modeling techniques. These criteria must be defined before the implementation of the modeling phase. We suggest to move from the DM objectives to the DM success criteria using the GQM approach of Basili & Weiss (1984). In this case, the GQM approach can help translate the DM objective into a set of questions, which can then be refined into a set of objective or subjective metrics. These metrics are the evaluation criteria that can be used for assessing the results of the modeling phase to establish whether or not the selected model was helping to accomplish the DM objectives of the project. DMSC influence the critical decision of whether or not a model should be deployed. Technical personnel in consultation with business users must be involved in setting up these criteria. Table 6 shows relevant evaluation criteria in the context of the directed DM. We present only classification and estimation as instances of the directed DM problems as prediction can be modeled as either of these problems (Berry & Linoff, 2000). Table 7 shows relevant evaluation criteria in the context of the undirected DM problems. The criteria presented here are discussed in (Redpath & Srinivasan, 2003). The criteria associated with clustering (Osei-Bryson, 2006) and association rules (Choi *et al.*, 2005) can be used for evaluating the results from these modeling techniques.

3.1.8 Elicitation of preference functions and creation of a value function

Techniques from the field of decision analysis can be adapted here. For example, given the need to evaluate generated models in the modeling phase, a composite score could be calculated for each

Table 6 Data mining (DM) success criteria for directed DM

Problem type	DM success criteria
Classification	Accuracy, precision, recall, profit and loss, lift, simplicity*, stability, speed, training time and memory usage
Estimation	Mean square error, variance (std.dev.), simplicity*, stability, speed, training time and memory usage

* Simplicity is not relevant in case of non-explanatory, black box models.

Table 7 Data mining (DM) success criteria for undirected DM

Problem type	DM success criteria
Clustering	Normalized cluster mean, variable importance vectors, outliers and usefulness
Association rules	Lift, simplicity (rule length), support, confidence, recall, precision, interest factor, expected monetary factor, incremental monetary factor
Description or visualization	Number of instances in data set, number of dimensions, overlapping data instances, ability to reveal patterns in data set, ability to reveal clusters of two or three dimensions, number of clusters present, amount of background noise, variance of clusters, ability to manipulate display automatically, ease of use

Table 8 DMSC for classification problems (BusReq = explanatory)

Applicable DMSC (description)	Value function	Thresholds	Weights
Accuracy (correctly classified proportion)	1: Test misclassification rate	>0.75	0.60
Profit and loss (unequal misclassification costs)	(Average worst possible loss−average loss of model)/(average worst possible loss−average best possible loss)	>0.75	
Lift (cumulative %captured response at the k^{th} decile)	(Model baseline)/(exact baseline)	>0	0.20
Stability (visual inspection of the non-cumulative %response lift chart)	Stability is binary, with 1 indicating a stable model and 0 indicating an unstable model	>0	0.15
Simplicity score (SIMPL) based on the number of rules (NR)	SIMPL = 0, if NR $\geqslant$ 2 or $\leqslant$13; SIMPL = (NR 2)/3, if NR $\in$ (2, 5); SIMPL = 1, if NR $\in$ (5, 8); SIMPL = (NR 13)/5, if NR $\in$ (9, 12)		0.05
Speed (run time)	Number of minutes	<25	
Training time (time taken to train the model)	Number of hours	<5	
Formula for creating composite score	(0.60 × accuracy score) + (0.20 lift score) + (0.15 × stability score) + (0.05 × simplicity score)		

DMSC, data mining success criteria.

model based on preference function (e.g. the weighted sum of measures). Osei-Bryson (2004) proposed an approach for comparing and selecting the 'optimal' decision tree (DT) model based on the preference and value functions specified by the domain expert(s). A technique such as the analytic hierarchy process (AHP) of Saaty (1991) could be used to determine the relevant weights based on the input of domain experts. In Table 8, we present an example of the DMSCs and corresponding value functions and weights for the classification problem (where business requirement is to produce an explanatory model) to illustrate the concepts of value functions, weights, thresholds and composite scores that are involved in the evaluation framework of Osei-Bryson (2004).

3.1.9 Analysis of applicable data resources

The business objective and DM objective provide a glimpse into the applicable data resources. The DM projects base can also be used to identify applicable data by searching for similar past projects. It is important to note that as business situations change, new variables may needed based on the set DM objectives. Data on these new variables may be available to the organization or may need to be purchased from an external data vendor. In the former case, there will be a cost associated with extracting the data and ensuring that it will be available to relevant personnel for the duration of the project. In the latter case, there will be a cost associated with buying the data from an external vendor. The costs in both instances should be analyzed in accordance with the budget and should be approved before proceeding to the next task.

3.1.10 Analysis of other technical resources (personnel and tools)

During this task, the lead technical personnel must analyze the availability of other technical resources such as personnel and tools for implementing the problem type selected in the previous task. An organizational ontology, organizational chart, or a skill and competency base can aid the technical stakeholders in identifying the technical personnel most suited for the project quickly. Analysis of tools can be simplified by storing the problem types supported by the DM tools (such as SAS Enterprise Miner, SPSS Clementine, etc.) available to the organization. If no available tool supports the selected problem type, then the relevant actors may propose sourcing of a relevant tool to the project sponsor

Table 9 Applicable modeling techniques for various data mining problem types

Problem type Target variable	Classification	Prediction	Estimation
Binary	Logistic regression, classification tree, k-nearest neighbor; Naïve Bayes*, Neural Network*, Support Vector Machines*, Genetic Algorithm*	Logistic regression, classification tree, k-nearest neighbor, Naïve Bayes*, Neural Network*, Support Vector Machines*, Genetic Algorithm*	Not applicable
Ordinal	Ordinal logistic regression, classification Tree, k-nearest neighbor, Naïve Bayes*, Neural Network*, Support Vector Machines*, Genetic Algorithm*	Ordinal logistic regression, classification tree, k-nearest neighbor, Naïve Bayes*, Neural Network*, Support Vector Machines*, Genetic Algorithm*	Not applicable
Nominal	Multinomial logistic regression, classification tree, k-nearest neighbor, Naïve Bayes*, Neural Network*, Support Vector Machines*, Genetic Algorithm*	Multinomial logistic regression, classification tree, k-nearest neighbor, Naïve Bayes*, Neural Network*, Support Vector Machines*, Genetic Algorithm*	Not applicable
Interval	Prompt user to discretize the target variable, then apply any of the classification techniques based on the number of bins	Regression, regression tree, k-nearest neighbor, Naïve Bayes*, Neural Network*, Support Vector Machines*	Regression, regression tree, k-nearest neighbor, memory-based reasoning, Neural Networks*

* Cannot be a final stage modeling technique, if business requirement demand is an explanatory model.

or other key high-level stakeholder who can then make the decision about whether or not the budget would support the purchase of a new tool and ensuing training and implementation costs.

3.1.11 Initial assessment of applicable modeling techniques

Generation of the DM model can involve the use of a single-modeling technique, and/or an unsequenced combination of modeling techniques; and/or sequence(s) of modeling techniques. For first two cases, characteristics of the DM project (i.e. its problem type, data type of the target variable, the business requirement of whether or not an explanatory model is desired) can be used for identifying modeling techniques that are applicable (see Table 9). For the case that involves the use of sequence of modeling techniques (e.g. Neural Network followed by DT), the penultimate technique would be a black box technique (e.g. Neural Network, Support Vector Machines and Genetic Algorithm) whose output could feed into the final technique of the pair (e.g. DT and logistic regression).

The identification of relevant modeling techniques could be done in the following manner:

1. Given characteristics of the DM project that were identified in the BU phase, use a cross-reference table such as Table 9 to:
   (a) identify modeling techniques relevant for the single technique and non-sequenced combinations approaches;
   (b) techniques identified in (1a) are the ones that are relevant for the final stage technique of the sequenced techniques approach.
2. Identify directed learning techniques that are applicable for the penultimate stage sequenced techniques approach by selecting DM techniques (e.g. Neural Network, Support Vector Machines and Genetic Algorithm) whose output could feed into the techniques identified in (1b).

The approach proposed above indicates that this task of generating a list of applicable techniques can be semi-automated. Our approach is different from that of Bernstein *et al.* (2005) who start at the level of the data itself and propose that the data type can be used for making decisions about the applicable

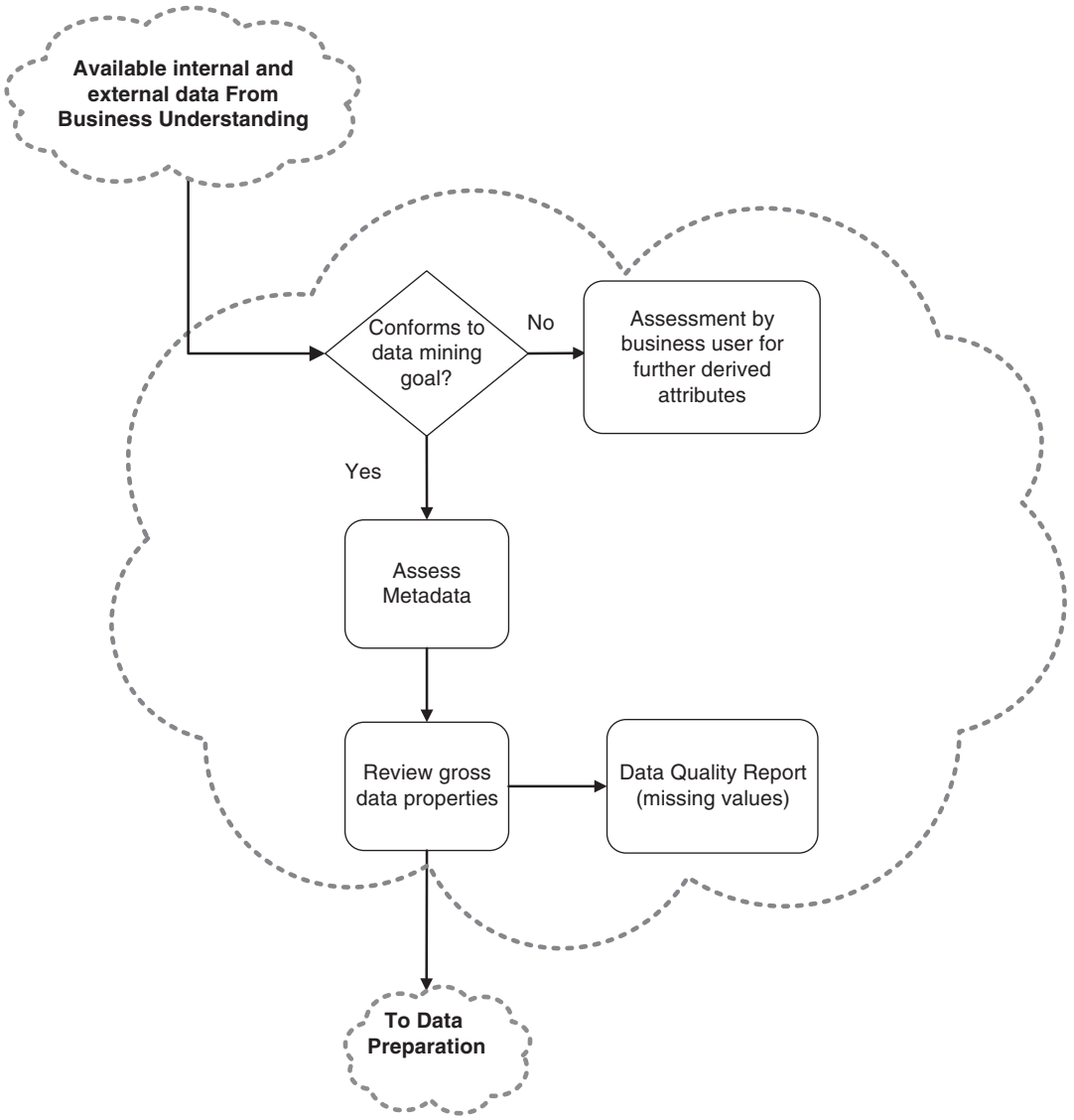

Figure 5 Data-understanding phase

techniques. Use of their approach can result in enumeration of those techniques that clash with the business requirement. So, even if these techniques were tried, the results would not eventually be accepted, resulting in inefficient usage of resources. In addition, their approach results only in enumeration of single techniques, and the combination of techniques is not accommodated in their approach.

3.2 Data understanding phase

During this phase, the integrated data set (consisting of internal and or external data) is to be explored and analyzed in order to gain an understanding of gross properties of the data; identify data quality issues; and to assess whether the available data are adequate to address the DM goals. A metadata base could be used to identify relevant derived attributes. Figure 5 shows a schematic of the data-understanding phase. Note that it also highlights that the phase received input from the BU phase and that the output of the phase is fed into the data preparation phase.

3.3 Data preparation phase

During this stage, the final data set is constructed from the raw initial data. The data set constructed should be evaluated by the domain expert for appropriateness, and the need for any

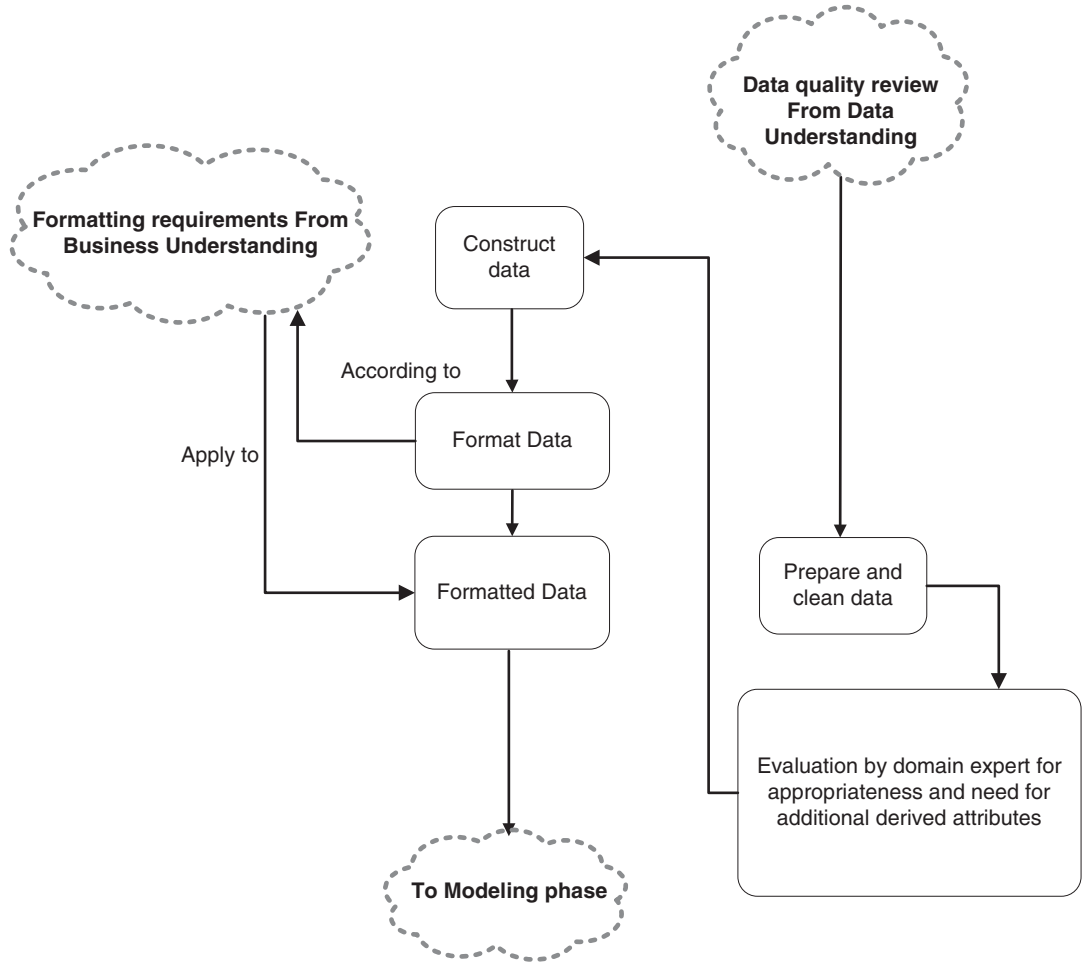

Figure 6 Data preparation phase

additional derived attributes should be reconsidered in light of the integrated data. Figure 6 shows a schematic of the data preparation phase, its relations with two preceding phases, namely business and data understanding and its output to the modeling phase. The *Modeling Techniques Base* that stores the formatting requirements for the various techniques should be used to format the data in accordance with the modeling techniques generated during the BU phase.

3.4 Modeling phase

During this phase, each modeling technique (or their combinations) would be applied to the formatted data. The results of the DM model application would then be assessed to find whether or not it needs the DM success criteria. Then, those models that meet the criteria should be stored in the list of acceptable models and the next modeling technique should be executed. Figure 7 shows a schematic of the modeling phase, its relation to two preceding phases, namely BU and data preparation, and its output to evaluation phase.

3.5 Evaluation phase

In the evaluation phase, the top-ranked models should be identified based on a composite score comprising of value functions and weights for the evaluation of the DM success criteria. This step for the selection of the best model can be semi-automated.

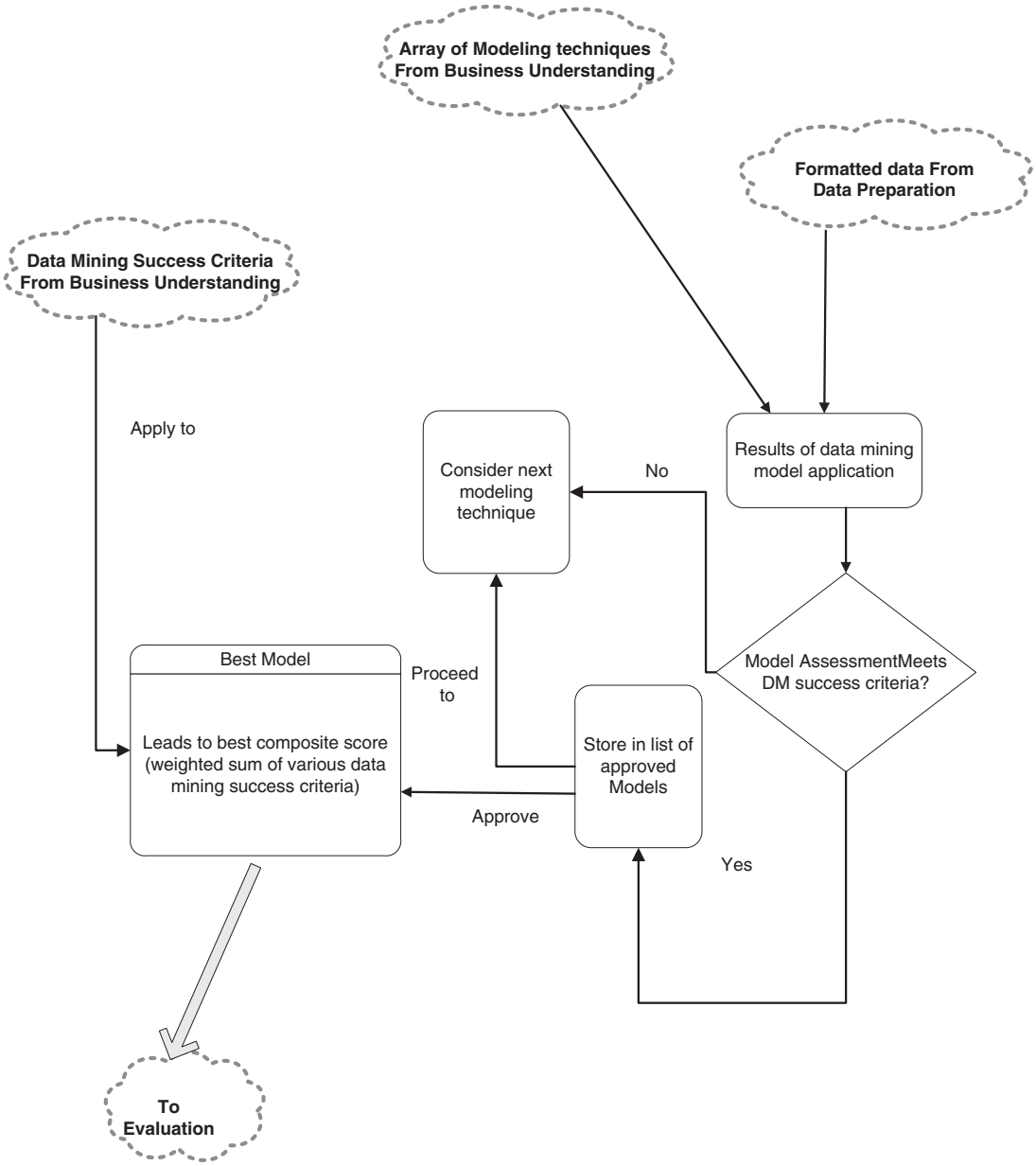

Figure 7 Modeling phase

3.6 Schematic of integrated KDDM process model

In Figure 8, we present the schematic of our integrated KDDM process model. The figure shown here explicates some of the tasks (and their proposed execution) based on the discussion provided above.

4 Summary

Kurgan & Musilek (2006) who conducted a detailed review of the existing KDDM models noted that the future of KDDM process models lies in achieving the integration of the whole process. This paper addresses an important research objective, creation of an improved KDDM process model that is relevant to both academicians and practitioners. We identified significant limitations of existing KDDM process models (Fayyad *et al.*, 1996a; Berry & Linoff, 1997; Anand & Buchner, 1998; Cabena *et al.*, 1998; Cios *et al.*, 2000; Han & Kamber, 2001; CRISP-DM, 2003), and

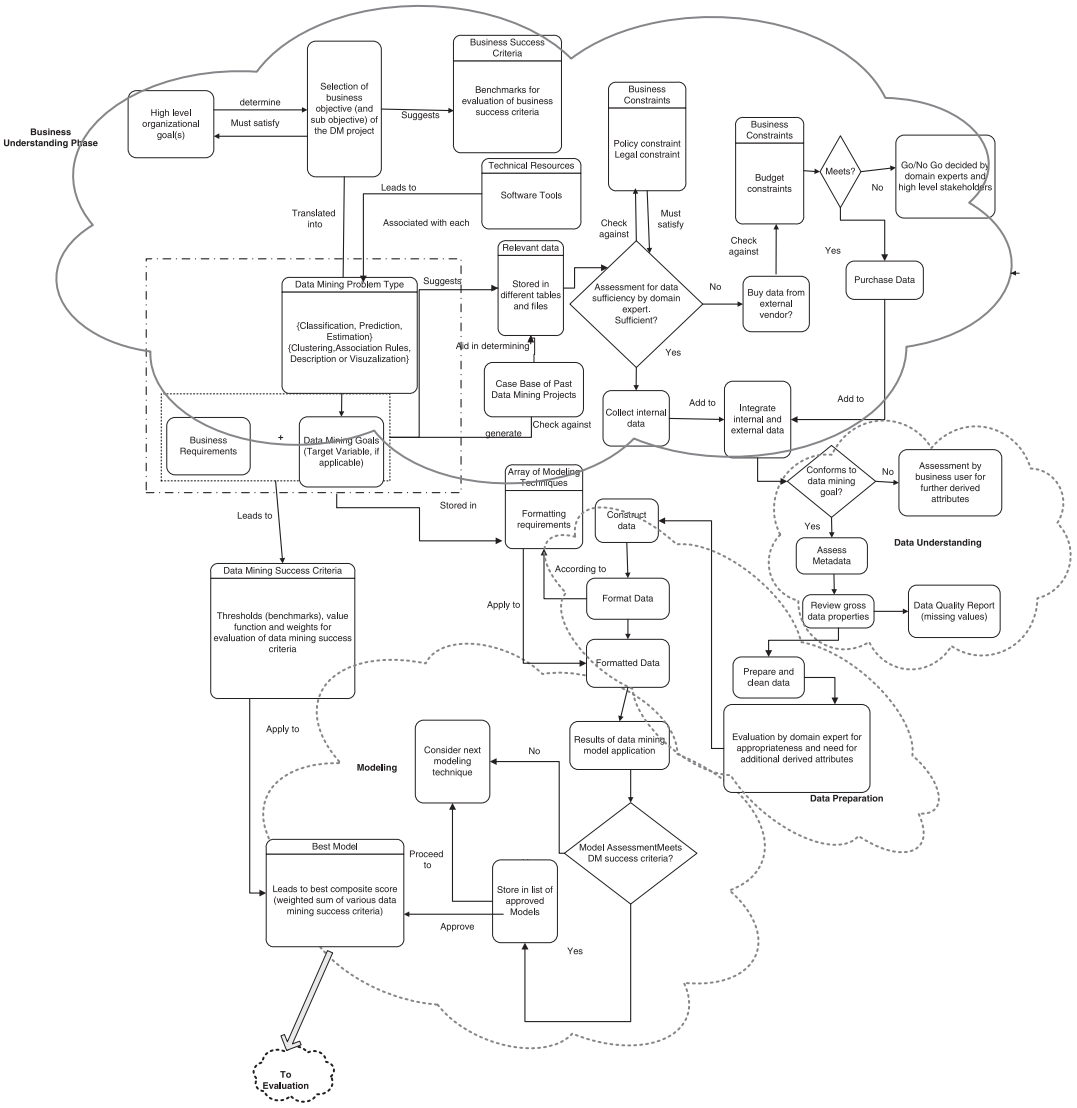

Figure 8 Toward an integrated knowledge discovery and data mining process model

designed an integrated KDDM process model to address these limitations. We discuss how the dependencies highlighted in the integrated model can be used for semi-automating the execution of six different tasks belonging to the BU through the modeling phases. The semi-automation of proposed tasks is likely to result in more efficient and effective implementation of the knowledge discovery process. Further, we also propose techniques that can be used for providing decision support in the form of appropriate tools and techniques for the various tasks (excluding tasks belonging to deployment phase) belonging to the integrated KDDM process model. The identification and description of relevant techniques can serve to ensure that all the tasks of the process model are executed and no task is inadequately executed due to the lack of support toward its implementation.

The proposed integrated KDDM process architecture can be used as a platform for executing different knowledge discovery projects across an organization. Future research should focus on identifying more candidate tasks for semi-automation, improvements to the KDDM artifact, development of an architecture to support the implementation of KDDM process and implementation of the artifact in an organizational setting.

References

Anand, S. & Buchner, A. 1998. *Decision Support Using Data Mining*. London: Financial Times Pitman Publishers.

Basili, V. R. & Weiss, D. M. 1984. A methodology for collecting valid software engineering data. *IEEE Transactions on Software Engineering* **10**(6), 728–738.

Bernstein, A., Provost, F. & Hill, S. 2005. Toward intelligent assistance for a data mining process: an ontology-based approach for cost-sensitive classification. *IEEE Transactions on Knowledge and Data Engineering* **17**(4), 503–518.

Berry, M. & Linoff, G. 1997. *Data Mining Techniques for Marketing, Sales and Customer Support*. John Wiley and Sons.

Berry, M. & Linoff, G. 2000. *Mastering Data Mining: The Art and Relationship of Customer Relationship Management*. John Wiley and Sons.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. 1998. *Discovering Data Mining: From Concepts to Implementation*. Prentice Hall.

Charest, M., Delisle, S., Cervantes, O. & Shen, Y. 2006. Intelligent data mining assistance via CBR and ontologies. In *Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06)*.

Choi, D. H., Ahn, B. S. & Kim, S. H. 2005. Ranking discovered rules from data mining with multiple criteria by data envelopment analysis. *Expert Systems with Applications* **29**(4), 867–878.

Cios, K. & Kurgan, L. 2005. Trends in data mining and knowledge discovery. In *Advanced Techniques in Knowledge Discovery and Data Mining*. Pal, N. & Jain, L. (eds). Springer, 1–26.

Cios, K., Teresinska, A., Konieczna, J. & Sharma, S. 2000. Diagnosing myocardial perfusion from PECT bull's-eye maps—a knowledge discovery approach. *IEEE Engineering in Medicine and Biology Magazine, Special Issue on Medical Data Mining and Knowledge Discovery* **19**(4), 17–25.

CRISP-DM. (2003). *Cross Industry Standard Process for Data Mining 1.0: Step by Step Data Mining Guide*. `http://www.crisp-dm.org/` accessed October 1, 2007.

Davenport, T. H. & Harris, J. G. 2007. *Competing on Analytics*. Harvard Business School Press.

Doran, G. T. 1981. There's a S.M.A.R.T. way to write management goals and objectives. *Management Review (AMA Forum)*, 35–36.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthuruswamy, R. (eds). 1996a. *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT Press.

Fox, M. S., Barbuceanu, M. & Gruninger, M. 1998. An organization ontology for enterprise modeling. *Simulating Organizations: Computational Models of Institutions and Groups*. AAAI/MIT Press, 131–152.

Han, J. & Kamber, M. 2001. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.

Keeney, R. 1996. *Value focussed thinking: a path to creative decision-making*, Harvard University Press.

Kurgan, L. A. & Musilek, P. 2006. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review* **21**(1), 1–24.

Laguna, M. A., Marqués, J. M. & Garcia F. 2001. A user requirements elicitation tool. *ACM SIGSOFT Software Engineering Notes Archive* **26**(2), 35–37.

Osei-Bryson, K.-M. 2004. Evaluation of decision trees. *Computers and Operations Research* **31**, 1933–1945.

Osei-Bryson, K.-M. 2006. Class Notes: Clustering Info 614: Graduate Course in Data Mining Virginia Commonwealth University.

Pyle, D. 2003. *Business Modeling and Data Mining*. Morgan Kaufmann Publishers.

Redpath, R. & Srinivasan, B. 2003. Criteria for a comparative study of visualization techniques in data mining. *IEEE 3rd International Conference On Intelligent Systems Design and Application, Tulsa, USA*. Springer-Verlag.

Saaty, T. L. 1991. Response to Holder's comments on the analytic hierarchy process. *The Journal of the Operational Research Society* **42**(10), 909–914.

Sharma, S. & Osei-Bryson, K.-M. 2008a. *Organization-Ontology Based Framework for Executing the Business Understanding Phase of Data Mining Projects*. Hawaii International Conference on Systems Sciences.

Sharma, S. & Osei-Bryson, K.-M. 2008b. Framework for formal implementation of the business understanding Phase of data mining projects. *Expert Systems with Applications* **36**(2), 4114–4124.

Simon, H. A. 1996. *The Sciences of the Artificial*. MIT Press.

Simoudis, E., Livezey, B. & Kerber, R. 1996. Integrating inductive and deductive reasoning for data mining. In *Advances in Knowledge Discovery and Data Mining*. Fayyad, U., Paitetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. (eds). AAAI Press/MIT Press.