Incoporating Data Mining Applications into Clinical Guildelines

Reza Sherafat Kazemzadeh and Kamran Sartipi
McMaster University
Department of Computing and Software
1280 Main Street West, Hamilton, Ontario, Canada
{sherafr, sartipi}@mcmaster.ca

Abstract

Computer based clinical guidelines have been developed to help caregivers in practicing medicine. GLIF3 (Guideline Interchange Format 3) is one of several standards that specify the structure for defining guideline models. The decision making steps within the GLIF3 guidelines are limited to evaluation of basic logic expressions. On the other hand, data mining analyses aim at building descriptive or predictive mining models that contain valuable knowledge. However, this knowledge can not be represented using the current guideline specification standards. In this paper, we focus on encoding and sharing the results obtained from a data mining study in the context of clinical care and hence to make it available at the point of care. For this purpose, we investigate available standards to encode the mining results in an interoperable manner and then elaborate on how to incorporate them in the context of guideline-based clinical decision support systems, and in particular GLIF3 models. Finally, we demonstrate the proposed approach using a developed prototype tool for modeling and execution of knowledge-assisted guideline.

KEYWORDS: Data Mining; Knowledge Management; Healthcare Informatics; Clinical Decision Support; Clinical Guideline; GLIF.

1. Introduction

Today's Healthcare professionals are overwhelmed with a huge amount of information generated from different sources. In this context, preventable medical errors are estimated to be the cause of thousands of deaths and loss of billions of dollars per year only in the United States. To reduce the human related errors, clinical decision support systems have been developed to help physicians and caregivers in practicing medicine. These systems provide a variety of mechanisms to encode medical knowledge that can be readily accessed at the point of care to help and guide the practitioner through possibly many steps of care delivery. In recent years, many studies in health informatics literature have investigated the effectiveness of the clinical decision support systems and concluded that these systems are indeed helpful [5]. On the other hand, data mining technologies have also been extensively applied on clinical data in order to extract new medical knowledge in the form of useful and non-trivial patterns. The mining techniques attempt to fit a model to the data and describe specific characteristics that allow prediction of new cases based on previous data. This extracted knowledge is valuable for decision making at the point of care for diagnosis, treatment planning, risk analysis, and predictions. It is usually the case that the mined knowledge is only accessible to the institutions that originally performed the study. This valuable knowledge is not readily available to other host institutions to be integrated with their local information systems. Clinical guidelines seem to be the best mechanism to deliver the mining results to the Healthcare personnel; however, the current guideline standards are unable to express the complex nature of the mining models and the extracted rules.

In this paper, we adopt standards and techniques from the fields of data mining and clinical decision support systems in order to provide a way for dissemination of mined knowledge into the clinical practice. We describe the details of the encoding, sharing and exchange of the mined knowledge using the *Predictive Model Markup Language* (PMML) specification and define new guideline modeling constructs, i.e., data mining decision nodes for this purpose. An execution environment can finally access the knowledge source and patient data as the guideline flows through a data mining decision node. By interpreting the patient data using the mined knowledge within the context of the clinical guidelines, the Healthcare personnel will be assisted to make a guided decision about the clinical case at hand. Moreover, they may receive recommendations that help to direct the flow of the guideline.

The structure of the remaining of the paper is as follows: Section 2 addresses the current trends in clinical guide-

Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)
0-7695-2517-1/06 $20.00 © 2006 IEEE

line modeling, and Section 3 describes application of data mining techniques in Healthcare. In Section 4 we explain our approach to integrating data mining results into clinical practice guidelines. Section 5 presents an execution environment that has been developed to run data mining aware clinical guidelines. Section 6 discusses the related work, and finally Section 7 concludes the paper.

2. Clinical guidelines

This section serves as an introduction to the approaches that capture clinical knowledge in the form of best practice computer-readable clinical guidelines with focus on *Guideline Interchange Format 3* (GLIF3) standard. Using GLIF3 specification, we define a flowchart-like modeling diagram with five basic modeling constructs, as follows:

- *Decision step*: determines the direction of the flow based on a decision criterion written in an expression language. For example, the age of the patient might be compared to a specific age as a decision criterion.
- *Activity step*: performs an action; prescribes medications; orders tests; or recommends treatments.
- *Patient state step*: designates a specific patient's condition or symptom; previous treatments; or diagnoses.
- *Branch step*: generates two or more concurrent decision making guideline-flows; for example, order a laboratory test and prescribe medication both at the same time.
- *Synchronization step*: merges two or more concurrent guideline flows into a single guideline flow; for example, receive the laboratory test report and observe the result of prescribed medications in order to proceed to the next step.

A GLIF3 guideline retrieves data items from patient medical record systems of the corresponding institutions either through standard or ad-hoc interfaces. Consequently, the guideline model leads the Healthcare personnel through different steps and activities that are defined in its flowchart-like diagram. Execution of clinical guidelines may trigger different actions, such as generating warnings, alerts, or recommendations for the Healthcare personnel.

3. Applications of different data mining techniques in Healthcare

There are various types of data mining techniques that have applications in Healthcare. In this section, we consider three categories of data mining techniques, such as: *classification; cluster analysis*; and *association rules mining*, along with a brief description and examples in clinical care. One characteristic that is common to almost all of these knowledge discovery methods is that they are time consuming activities. Furthermore, the process of knowledge extraction involves several steps, including: data selection, data cleansing, dealing with incorrect or missing values, data transformations, execution of the mining algorithm, evaluation, interpretation, and validation of the results [8]. In the following subsections, the applications of different data mining techniques in the Healthcare domain are briefly discussed.

3.1. Classification

Classification is defined as training a function to map or classify a data item into one of several pre-defined classes [8]. The most apparent application of classification in clinical care is the process of diagnosis of diseases. A classification model, e.g., a decision tree, can be trained over some historical patients' medical data who have been diagnosed with disease X, as opposed to the othe patients who are not diagnosed with disease X.

Vlahou et al. [13] and Duch et al. [2] have applied decision trees classification for diagnosis of ovarian cancer and Melanoma skin cancer, respectively.

3.2. Cluster analysis

Cluster analysis is another interesting topic in the data mining field. In clustering, the aim is to identify groups in the studied population (here, the patient data) along with a mapping function from each new case to one of the identified groups. The clustering function minimizes a disctance measure between an item to be clustered and those items already in the cluster. As opposed to classification which is supervised, clustering is usually performed unsupervised. An application of clustering in the clinical care domain might be to perform risk analysis and assess patients' risk factors. Churilov et al. [4] have carried out a similar study to assign patients to three disjoint clusters of high, intermediate, and low risk patients.

3.3. Association rule mining

Association rule mining was initally defined in the context of market baskets that contain a number of items, where the association rule $A \Rightarrow B$ indicates that $x\%$ of the baskets that contain the group of items A also contain the group of items B. In this form, x is called the rule's *confidence*.

An application of association rule mining in clinical care is to idenify hidden patterns within the training data set. For example, idenifying the risk of infection by disease Z after

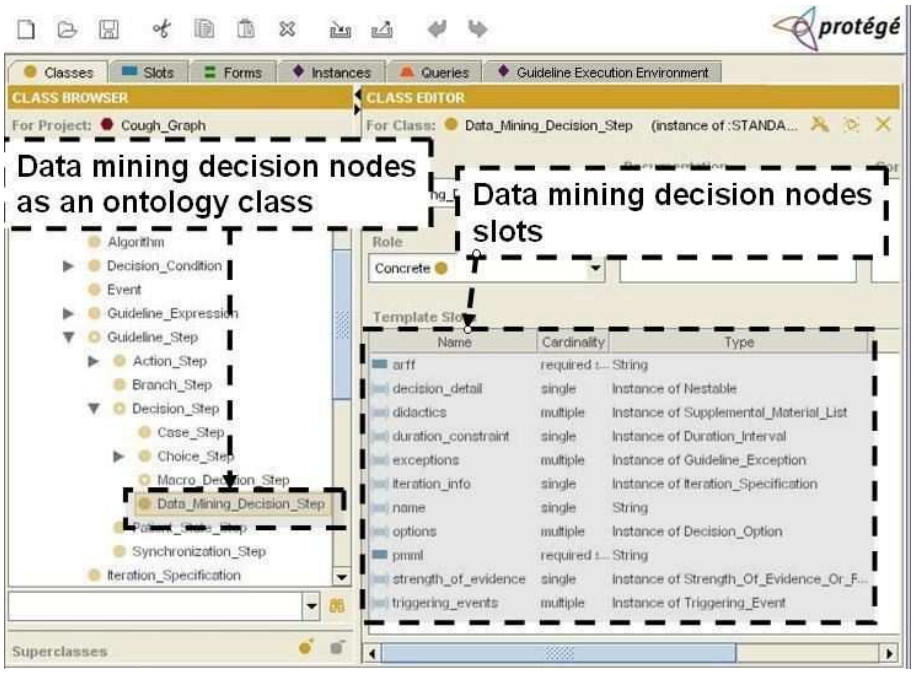

Figure 1. A data mining decision node class (in a hierarchy of classes) and its attributes (slots).

being infected by diseases X and Y. An association rule for the above example is represented as:

$$\{\text{Patients with } X \text{ and } Y\} \Rightarrow \{\text{Patients with } Z\}$$

Roughly speaking, the rule's support is the total number of patients having X, Y and Z together, and the rule's confidence is the ratio of the number of patients having all X, Y and Z to the number of patients having X and Y. In particular, a high confidence and support show that there is high risk of infection to Z and a large population has been identified as having this condition; whereas a low confidence and support might designate a weak association between occurrences of Z and X and Y. The knowledge extracted in the form of an association rule can possibly be used to inform the physician about the hidden relations between the patient's current symptoms and some ignored symptoms that may signal for a different diagnosis or treatment. As another example Ordonez et al. [10] have used association rules to predict heart disease.

4. Incorporating data mining results in clinical guidelines

The motivation to incorporate data mining results into clinical guidelines arises from the lack of a common approach to deal with the results of clinical data mining studies which are carried out and used only within the same Healthcare institution. As a result, the applications of many data mining studies on the Healthcare data are restricted to the original institutions that performed the analysis.

This restriction can be alleviated by adopting the required standards to define, share, exchange, and use the mined clinical knowledge. In this section, we elaborate on a possible solution comprising of two standards from both the data mining and the clinical decision support fields.

We extended GLIF3 guideline modeling standard in order to incorporate data mining results into clinical best-practice guidelines. To achieve this goal, the modeling aspect of GLIF3 has been extended to incorporate special types of decision nodes that are capable of interpreting the knowledge extracted by a data mining analysis. We refer to these extended decision nodes as *data mining decision nodes* which inherit the basic attributes (*slots* in ontology terminology) associated with a simple decision node class, and also define new attributes to access the knowledge base in order to apply the data mining model[1] to the patient data. Figure 1 illustrates a snapshot of a data mining decision node class within the class hierarchy of GLIF3, and its attributes in the *Protégé* ontology editor tool. The new attributes specify the name of a file that contains the patient

[1] A *data mining model* refers to the data structure where the results of application of a *specific* data mining algorithm are stored. For example, in case of data decision tree classifier algorithms the data mining model represents a tree structure; and for association rules mining algorithms the data mining model represents a set of association rules.

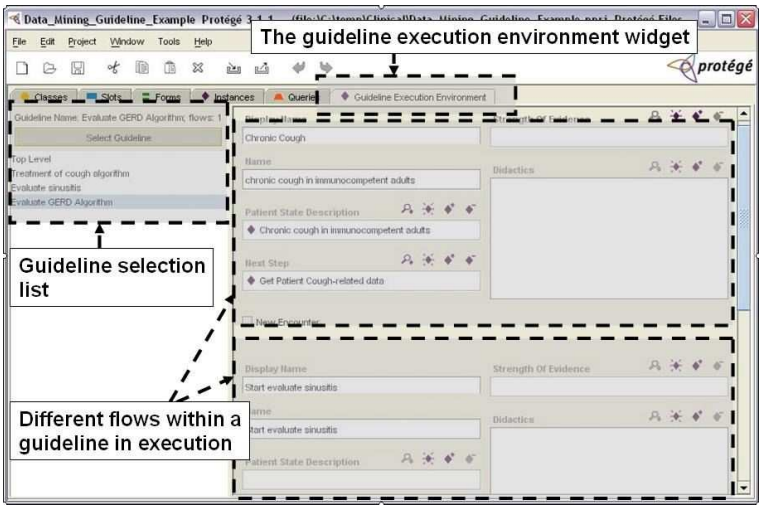

Figure 2. The guideline execution environment within the *Protégé* ontology editor.

data and the name of the files from a knowledge base repository that contains the data mining models that are to be interpreted.

In order to share, exchange and interpret data mining results, the models are encoded into XML files using Predictive Model Markup Language (PMML). PMML [6] files may contain one or more mining models along with their data dictionaries, model schemas, and data transformations. The data dictionary of a mining model defines the required attributes and their types. On one hand, these attributes are referenced by their assigned names within the PMML file and on the other hand these attributes are the access points that should be bound to data items that are external to the PMML file. In this case the external data items are the patient data. Many data mining models need special treatment of the raw input data in the form of appropriate transformations; e.g., normalization, discretization, value mapping, and function aggregation. The data transformation section of the PMML file specifies these details. Finally, the model schema is the place where the actual input and output of the model are defined. The input might be either an attribute that is readily defined in the data dictionary or one that has gone through some transformations.

A wide range of data mining models can be defined using the PMML encoding, such as: association rules; clusters; regression results; and tree models. In the proposed approach in this paper, the mining models are encoded in PMML files and interpreted by the guideline execution engine to direct the execution flow and eventually assist the Healthcare personnel in making appropriate decision on the patient data.

5. The guideline execution environment

There are mainly two approaches in executing a guideline model [7]. In the first approach, a new software application is developed for each guideline instance that must interconnect the nodes of the guideline as appropriate. This approach has many drawbacks as it wastes the resources. Also, small changes in the model may require considerable recoding and hence the necessary flexibility is obviously missing. Therefore, this is not considered as a favourable approach.

On the other hand, we can think of an environment with an engine that receives a guideline model as input and interprets the model according to the guideline specification. In this environment, we define a set of software modules that are responsible for performing the necessary actions as determined by the node's specification. The environment's execution engine is capable to interpret the guideline model and invoke the corresponding modules. During the execution of the guideline, the environment keeps track of the guideline's execution flow and provides the required data and knowledge retrieval facilities.

We adopted the second approach and implemented an environment and its execution engine to automatically interpret and execute a clinical guideline that has been defined according to the GLIF3 specification. We used the GLIF3 modeling ontology in *Protégé* and added a set of new data mining decision support classes and attributes (slots) as it was described in the previous section. To interpret the PMML files, we adopted the above module-based approach. After the engine identifies that the currently active node is of type "data mining decision node", it determines the module responsible for retrieving the result from a specified data

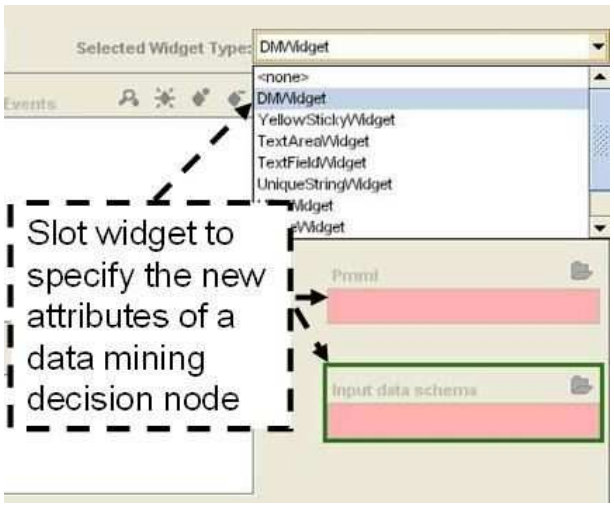

Figure 3. The slot widget for defining data mining decision nodes parameters.

mining model. In our implementation, these modules are Java programs that describe how to access PMML files (for data mining model) and also access the patient data. The module returns the result of data mining operation to the execution engine that will control the execution flow accordingly. Also, since the host institution's data format does not necessarily comply with that of the organization that built the PMML model, the Java module might need to perform some additional transformations prior to the application of the model.

To integrate our tool's environment with the *Protégé* graphical user interface, a tab widget and a slot widget have been developed. In Figure 2 the tab widget is responsible for invoking the execution engine. This widget allows to select a guideline instance from the list of guidelines defined in the environment and consequently runs the guideline. The active nodes of each execution flow is displayed on the right part of the screen. The slot widget in Figure 3 is used to define a guideline. As mentioned earlier, the data mining decision nodes require additional attributes that are supported by this slot widget. It also checks to see if the data dictionary of the specified PMML model is consistent with the attribute names and types of the input patient's data.

To access the specified PMML file during the execution, the execution environment connects to a local or distributed repository of guidelines and retrieves the PMML file. Since we also intended to deploy the environment on mobile platforms, i.e., PDA devices, as an extra computational capability we support on-the-server execution of data mining decision nodes. This is mainly due to the fact that PDAs have limited storage and processing resources.

6. Related Work

The purpose of clinical guidelines is to provide the best available scientific evidence [15] and many guideline standards have been developed so far. Arden Syntax [1] is a guideline specification standard used in many Clinical Decision Support Systems. The idea is to define a simple, yet powerful enough procedural language to encode medical logic. Knowledge is captured in separate *Medical Logic Modules* (MLMs) that specify the decision making logic in the form of some *if-then* rules. The *Evidence-Based Guidelines And Decision Support System* (EGADSS) [3] is an implemented decision support system that is based on Arden Syntax. PRO*forma* [14] is another guideline specification standard similar to GLIF3 that captures clinical knowledge in the form of a set of tasks. *Cocoon* [16] is a clinical decision support system that is based on PRO*forma*. Typical guidelines built using the above standards are unable to encode the knowledge that is extracted from a data mining study.

On the other hand, clinical data mining studies have built valuable knowledge that is rarely subject to widespread use. The systems that incorporate and use this kind of knowledge are usually separately developed applications and hence lack interoperability requirement to work with other health information systems. The models are usually built over some training data and then tested on some other data set. Examples of such studies are numerous [12, 9, 11].

In contrast to the standards, projects and studies that were mentioned in this section, our approach aims to tackle the knowledge interoperability problems involved with complex data mining models through the use of standards from both the guideline modeling and data mining research fields. Different standards have been integrated to provide additional decision making support. Also, based on application of the knowledge to the patient's data the flow of the guideline models determines the type of actions to take and hence it is not confined to reminders or alerts that is provided by Arden Syntax.

7. Conclusion

In this paper, we described the process of incorporating the extracted knowledge as a part of data mining research into the clinical best-practice guidelines. The knowledge extraction phase can take place at a remote site and the results can be encoded as PMML files for the purpose of stage, sharing, and exchange among different institutions. At the application site, the guideline execution environment will dispatch data mining specific modules to retrieve the stored data and mined knowledge from a local or distributed repository. The execution engine then interprets the data and knowledge within a guideline node. The proposed

guideline execution environment in this paper can be used to provide valuable recommendations to the Healthcare personnel.

References

[1] The Arden Syntax for Medical Logic Systems, URL = http://cslxinfmtcs.csmc.edu/hl7/arden/.

[2] Rules for melanoma skin cancer diagnosis, URL = http://www.phys.uni.torun.pl/publications/kmk/.

[3] Evidence-based guidelines and decision support system (egadss). URL = http://egadss.org/.

[4] L. Churilov, A. M. Bagirov, D. Schwartz, K. A. Smith, and M. Dally. Improving risk grouping rules for prostate cancer patients with optimization. In *HICSS*, 2004.

[5] E. W. K. Cynthia M. Farquhar and J. R. Slutsky. Clinicians' attitudes to clinical practice guidelines. *Medical Journal of Australia*, Aust 2002; 177: 502-506.

[6] DMG.org. Pmml version 3.0 specification. URL = http://www.dmg.org/pmml-v3-0.html.

[7] S. W. T. A. A. B. O. O. Q. Z. R. A. G. V. L. P. Dongwen Wang, Mor Peleg and E. H. Shortliffe. Design and implementation of the glif3 guideline execution engine. *Journal of biomedical informatics*, 2004 Oct;37(5):305-18.

[8] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[9] T. Lee, G. Juarez, and E. Cook. Ruling out acute myocardial infarction: a prospective multicenter validation of a 12-hour strategy for patients at low risk. *N Engl J Med*, 324:1239–1246, 1991.

[10] C. Ordonez, E. Omiecinski, L. de Braal, C. A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczynska, and E. V. Garcia. Mining constrained association rules to predict heart disease. In *ICDM*, pages 433–440, 2001.

[11] C. Otto and A. Pearlman. Doppler echocardiography in adults with symptomatic aortic stenosis. *Arch Intern Med*, 148:2553–2560, 1988.

[12] O. Pahlm, D. Case, G. Howard, J. Pope, and W. Haisty. Clinical prediction rules for ecg diagnosis of inferior myocardial infaction. *Comp Biomed Res*, 23:332–345, 1990.

[13] A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman. Diagnosis of ovarian cancer using decision tree classification of mass spectral data. *Journal of Biomedicine and Biotechnology*, 2003:5 (2003) 308-314.

[14] Proforma. URL = http://www.acl.icnet.uk/lab/proforma.html.

[15] URL = http://openclinical.org/guidelines.html.

[16] Cocoon. URL = http://www.cocoon-health.com/.