Intelligent Clinical Decision Support Systems Based on SNOMED CT

Ewelina Ciolko, BHSc, Fletcher Lu, PhD and Amardeep Joshi, BSc, BHSc

Abstract—The decision support systems that have been developed to assist physicians in the diagnostic process often are based on static data which may be out of date. We present a comprehensive analysis of artificial intelligent methods which could be applied to documents encoded by SNOMED CT. By mining information directly from SNOMED CT encoded documents, a decision support system could contain timely updated diagnostic information, which is of significant value in fast changing situations such as minimally understood emerging diseases and epidemics. Through a high level comparison of many AI methods it is found that a TAN-Bayesian method could be the most suitable to apply to SNOMED CT data.

I. INTRODUCTION

The concept of a Clinical Decision Support System (CDSS) is common; however, many physicians are reluctant to use one [1]. It is possible that physicians will be more likely to take advantage of CDSS if it is based on real patient data history instead of static data and concepts which are already known to the user. This could be possible with an Intelligent CDSS (ICDSS).

A decision support system which can learn the relationships between patient history, diseases in the population, symptoms, pathology of a disease, family history and test results, would be useful to physicians and hospitals. This ICDSS method would be created from standardized machine readable medical data which is representative of patient journeys.

The most suitable basis for an ICDSS would be a database which is made of standardized codes. Among these codes is SNOMED CT. In the following section we will illustrate the suitability of SNOMED CT for such a system and suggest a learning support structure to organize its encoded information into a decision support system.

II. BACKGROUND

SNOMED CT is a standardized clinical terminology system. It includes a comprehensive coverage of diseases, clinical findings, therapies, procedures and outcomes and provides core terminologies to encode into an electronic health record [2].

It contains over 993,000 English language descriptions of synonyms. This provides a robust database for physicians to search queries related to pathology of diseases.

Although SNOMED is comprehensive on its own, it also maps to other medical classifications and terminologies which are widely used in the medical world. Among these is ICD-9. SNOMED CT ensures that the codes do not overlap and at the same time allow ICD-9 codes to represent the billing and administrative parts of the patients journey [3].

Brown et. al. have shown coding in SNOMED CT to have a sensitivity of 92.3% and a positive predictive value of 99.8% [2]. They measure sensitivity as when SNOMED CT correctly assigns a code to a clinical term. The positive predictive value is the ability of SNOMED CT to tie certain clinical terms to non-clinical terms correctly.

SNOMED CT simplifies the search for diseases and symptoms; however if a physician wishes to use this system, they can find only simple hierarchies of SNOMED CT codes, which they must then interpret. These hierarchies are only meant for logical organization of the data. There is no indication of which path would be the best to follow for the patient in the current situation, or correlations between top level SNOMED CT codes.

Structure of SNOMED CT: SNOMED CT code identifies medical data as concepts which it distributes into concept hierarchies which create a context for the concepts [4]. There are also brief descriptions included with the machine code which are readable terms. Each concept may have one or more related descriptions and possible synonyms. Each description is represented by a code and is organized in a description table.

Each relationship has an ID and is also distributed into a relationship table. A relationship identifies two related
concepts and the type of relationship they have. Relationships also have a group, characteristic type, and a list of devices which may have been used in the medical intervention. Figure 1 illustrates the hierarchical organization of tables and how information is related [5].

A limitation with the existing SNOMED CT diagnostic software is the lack of a learning component. It is simply a standard clinical terminology made up of lists and tables of concepts [6]. This causes the system to have mistakes which are difficult to correct. By incorporating a learning component, the robust data may be mined for new relations and verification on current ones.

III. MACHINE LEARNING METHODS

We will briefly introduce the standard artificial intelligence methods chosen, and their current applications in healthcare:

- Decision trees (DTs) & Random forests (RFs)
- Artificial neural networks (ANNs)
- Bayesian networks
- Gaussian processes (GPs)

A. Decision Trees and Random Forests

The algorithms used in Decision Tree learning search for the descriptive attribute that is chiefly related to a target variable. As the name states the output is a tree-shaped model that represents a small set of variables that together have a high predictive power for the target variable. ‘If-then’ rules can also be used to describe the tree, as each rule can be identified by following the branches from the root node to a terminating node [7]. Random forests are models that are a collection of decision trees (Figure 2). The result is a dataset with some duplicates and some examples left out. The RFs model is a set of trees for each dataset averaged to get a final prediction. At times there are small deviations in a dataset that can have influence on a learned tree. RFs are a way of avoiding these unnecessary influences [7].

DTs have been found to be beneficial in Neonatal ICUs as they are used to classify streams of received physiological signals and detect artifacts, thereby reducing the high number of false alarms [8]. DTs are easy to understand and sharp in labeling errors and noise. This feature of DTs would be very useful in an ICDSS because there are a number of possible errors which can be made in a patient’s record. It has been found that compared to DT, the RF has consistently been shown to perform better in the ICU prediction tasks [9]. The feature of RFs which allows for duplications and missing data however would not be effective on a SNOMED database as some information is already duplicated, and multiple duplications are undesirable.

B. Artificial Neural Networks

ANNs are made up of simple processing units or nodes, interconnected to increase the computational power over any single unit. They have various input nodes, which represent observations that will be used for predictions. This includes symptoms and test results. These input nodes connect to one or several output nodes or predictions, which would be represented as diseases and interventions. Between the input and output nodes, there are many intermediate nodes also known as hidden nodes. These are organized as layers. These nodes are calculated from the values of the input and are then used to calculate the values for the output [10].

ANNs are known to be sharp in detecting errors in training data; therefore they are a good method for learning from noisy data inputs [7]. This characteristic would be effective on a SNOMED CT database. The main characteristics of ANN are: a black box structure, flexibility against noise, interpolation capability, and learning by example or training. This structure has allowed ANNs to be used to assist with the medical diagnosis of patients. The main drawback of ANNs is in the long time to train the system compared to other learning algorithms such as DTs [7]. The black box structure is also undesirable in this case because physicians would not see any process, just a final diagnosis.

C. Bayesian Networks

A Bayesian network is a probabilistic graphical model that uses a set of random variables to specify a joint probability distribution. There are two fundamental components: a directed acyclic graph showing dependencies & independencies between variables, and a set of probability distribution tables. Within supervised learning there are two
classes of Bayesian networks: Naïve Bayesian networks (NB) and Tree-Augmented Naïve Bayesian networks (TAN) (Figure 4). In NB the non-target variables are independent of other non-target variables; and there is a link from each non-target variable to the target variable. The TAN, in contrast, takes into account dependencies between non-target variables by having links between them [11]. TAN is more similar to the current organization of SNOMED CT.

Bayesian networks have been used in medicine since the 1970s. They have been shown to successfully deal with uncertainties which are commonly present in clinical practice. Their robust capabilities allow the Bayesian models to be used in a wide range of complex domains such as medicine [12].

D. Gaussian Processes

Gaussian Processes give a prior probability to every possible function, with higher probabilities for the functions that are more likely. GPs allow for multi-dimensional inputs, have a small number of tunable parameters and result in full predictive distributions as opposed to the point predictions typical or other methods. GPs are found to consistently outperform more conventional methods such as ANNs in different regression tasks [13].

The use of GPs for regression has recently begun in the intensive care domain. GPs outperform other modeling methods for the analysis of electroencephalograph (EEG) signals to detect neonatal seizures. Another application of GPs has allowed the classification of patients according to the time frame in which they can be weaned from mechanical ventilation [8]. These are more physiological streams of data as opposed to the pathology streams seen in SNOMED CT.

IV. DISCUSSION

To decide which of these methods is best suited for our purpose we compared them directly by assessing their advantages and disadvantages in terms of their utility, restrictions on their domains and previous applications in the health care field or with SNOMED CT. We then decided which AI methods would work effectively with SNOMED CT. The methods which would be applicable to SNOMED CT are the ones which have a bullet in the SNOMED CT column of the table. The analysis is summarized in (Table 1).

Many comparisons between artificial intelligence methods have been made, however none of these comparisons included which method would best search SNOMED CT data. Many of these comparisons have alluded to which AI methods would be best to apply to health concepts that are represented in different ways. In a recent study done on the prediction of gastric cancer, it was found that neural networks were the better predictors of gastric cancer than the decision tree model [14]. This suggests that ANN’s are more sensitive when finding diagnoses.

Another comparison study was performed on three different machine learning techniques: Naïve Bayes, neural networks, and decision trees. They were used to predict the likelihood of a patient developing heart disease based on medical profiles such as age, sex, blood pressure and blood sugar. The study showed high performance by all three models, but the Naïve Bayes model appeared to outperform the other two as it gave the highest number of correct predictions [15].

This suggests that a Naïve Bayes method may be the most effective with SNOMED CT. The most appropriate approach would utilize a Tree-Augmented Bayes (TAN) approach in contrast to using a Naïve Bayes method. The problem with a Naïve Bayes approach would be failure to capture interrelationships between concepts. This essentially would create a black box approach which is not desirable for diagnostic processes. A TAN approach would represent such inter-relations with both target and non-target variables. For a decision support system it is important to see how all of a patient’s symptoms, history and even occupation would relate to each other. This is also why a black box-structure would not work for an ICDSS.

TAN Bayes is a supervised learning method which uses

<table>
<thead>
<tr>
<th>Type</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Uses in Healthcare</th>
<th>Snomed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree (DT)</td>
<td>Easy to understand</td>
<td>Robust in labeling errors &amp; noise</td>
<td>Small deviations in datasets can greatly affect the resulting tree.</td>
<td>Detects artifacts in streams of physiological signals</td>
</tr>
<tr>
<td>Random Forests (RF)</td>
<td>Doesn’t get influenced by slight variations in data</td>
<td>Popular in labeling rather than producing new knowledge</td>
<td>Used for ICU data &amp; prediction tasks</td>
<td></td>
</tr>
<tr>
<td>Artifical Neural Networks (ANNs)</td>
<td>Detects errors &amp; learns from noisy examples Used in data analysis, pattern recognition, &amp; prediction</td>
<td>Long training time Black box structure</td>
<td>Currently being used in neonatal intensive care physiological data mining</td>
<td></td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>Used in complex domains such as medicine to derive relationship probabilities and pathways</td>
<td>Not good at overlooking mistakes which exist in databases</td>
<td>Deals well with uncertainties commonly present in clinical practice</td>
<td></td>
</tr>
<tr>
<td>Gaussian Processes (GPs)</td>
<td>Allows for multi-dimensional inputs Outperforms ANN in different regression tasks</td>
<td>Small number of tunable parameters</td>
<td>Recently used in intensive care domain Proficient in reading EEGs</td>
<td></td>
</tr>
</tbody>
</table>
classification tasks to produce a disease diagnosis. Initially it will be necessary to allow the TAN Bayes to run on SNOMED CT encoded data which includes the target variables defined by the user.

The system would learn old and new relationships within the SNOMED CT database and apply probabilities to any relationships creating a Bayesian network. This system would have the potential to improve any important decisions within a patient’s journey through a hospital stay.

For example, after shadowing a neonatal physician at Sick Kids hospital in Toronto, Ontario it became apparent that some sort of reliable decision support system based on data of past patients is in fact a desired tool. Situations arose where physicians were at a stand-still on which direction a patient’s care should proceed. In one particular case there was only partial family history available. The history included a sibling of the patient who had similar symptoms years ago and their outcome was known, however nothing was known about their care process. A decision support tool which was based on real patient data could have helped in this situation because it is likely that former patients at other hospitals had similar patient journeys. This system could have the potential to fill in empirical evidence where there are gaps of knowledge about related patient history.

The novelty in such a system would lie in the fact that it would use available resources which many hospitals already have in order to make a tailored decision support system. Since the ICDSS would be based on a standardized terminology it would be relevant for many years with updates being made regularly. There would be no need for a different coding terminology to be put in place. It would be easy to update the system with new probabilities; to update one would just need to run the Bayesian process over again on new patient records.

As with any AI system it would be necessary to perform some adjustments during the initial phase in order to get relevant results. This would be the most complicated part of the system as it is important to get meaningful data. TAN Bayes, as a supervised learning method, would require data that is logically sound. This is well suited to the fairly well organized SNOMED CT structure. With SNOMED CT encoded data one would merely have to decide on inclusion and exclusion criteria, and the questions & answers. Possible inclusion criteria could encompass any code relating to the diagnoses, or major decisions in a patient journey. For instance, this will exclude any redundancies which exist within SNOMED CT code, such as since “abdomen” is considered a part of “trunk of body” it may not be a necessary fact in the diagnosis of patients, but more for computer usability. Many such instances exist in SNOMED CT therefore an appropriate method of finding such instances would be useful.

V. CONCLUSIONS

This paper introduces the concept of an ICDSS. After a comprehensive overview of different artificial intelligence methods and how they may be applied to SNOMED CT our analysis suggests that a Tree-Augmented Network Bayes may be the best general method to use, however it is possible that other methods are suitable under specific situations. The dilemma of potential inclusion and exclusion criteria used in the creation of such a system should also be examined further. There are many potential applications for such a system within health care and within the budding health informatics profession. A patient journey could be more successful with the use of such a system and physicians may be more apt to trust and use a system based on real patient data.

REFERENCES