Project Proposal

Data mining in bioinformatics – Eleazar Gil-Herrera

I. Introduction

This project entails to the application of a novel data mining methodology that relies on Rough Set Theory [1] to predict the life expectancy of terminally ill patients in an effort to improve the hospice referral process. Life expectancy prognostication is particularly valuable for terminally ill patients since it enables them and their families to initiate end-of-life discussions and choose the most desired management strategy for the remainder of their lives. According to Medicare regulations, a patient should be referred to hospice if his/her life expectancy is less than 6 months.

To this end, we utilize retrospective data from 9105 patients from which we want to demonstrate the design and implementation details of a series of classifiers based on Rough Set Theory, developed to identify potential hospice candidates, i.e., we want to identify patients who were not able to survive a 6 months period from whom a referral decision of deriving to hospice should be taken.

II. Technical significance

Approaches for developing prognostic models for estimating survival for seriously ill patients range from the use of traditional statistical and probabilistic techniques [2]-[5], to models based on artificial intelligence techniques such as neural networks, decision trees and rough set methods [6]-[11]. A recent systematic review of prognostic tools for estimating survival in palliative care is presented in [12].

Both statistics based techniques and *AI* based models rely on data that are **precisely well defined**. However, medical information, which represents patients records that include symptoms and clinical signs, **is not always well defined** and, therefore, the data is represented with **vagueness** [13]. Particularly, for this kind of information, it becomes very difficult to classify borderline cases in which very small differences in the value of a variable of interest may completely change c

ategorization and therefore the following decisions can changes dramatically [14]. Moreover, the data set is presented with inconsistencies in the sense that it is possible to have more than one patient with the same description but showing different outcomes, as we can see in Table 1:

Patient 1						
Age	**Meanbp**	**Resp**	**Temp**	**Pafi**	**Alb**	**Die**
[55-70]	[210-250]	[95-110]	[99-102]	[130-150]	[30-45]	**Yes**
Patient 2						
Age	**Meanbp**	**Resp**	**Temp**	**Pafi**	**Alb**	**Die**
[55-70]	[210-250]	[95-110]	[99-102]	[130-150]	[30-45]	**No**

In this work we propose the use of Rough Set Theory (RST) [1] to deal with vagueness and inconsistency in the representation of the data set. RST provides a mathematical tool for representing and reasoning about vagueness and inconsistency. Its fundamentals are based on the construction of similarity relations between data set objects from whom approximates yet useful

solutions are provided.

III. Methodology

For this project we describe a RST based knowledge discovery methodology to provide a classifier that properly discriminates patients into two groups, those who survive at least 180 days after evaluation for hospice referral and those who do not.

Based on RST, we can formally define the prognostication problem as:

$$\square=(\square,\square \cup\{\square\}) \quad (1)$$

where T represents the data set in the form of a table. Each row represents an object and each column represents an attribute. U is a non-empty finite set of objects and the set A represents a non-empty finite set of attributes called the condition attributes. In our case, an object designates a terminally ill patient and an attribute designates each of the fifteen condition attributes that describe a patient (Table I). Also, for every attribute $a \in A$, the function a: $U \rightarrow V_a$ makes a correspondence between an object in U to an attribute value V_a which is called the value set of a. The set T incorporates an additional attribute $\{d\}$ called the decision attribute. The system represented by this scheme is called a *decision system*.

Data set

The data set used in this study consists of the 9105 cases from the SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments) prognostic model data set [15]. We consider all variables used in the SUPPORT prognostic model as condition attributes, i.e. the physiologic variables along with the diagnosis groups, age, number of days in the hospital before entering the study, presence of cancer, and neurologic function. Attributes' names and descriptions are listed in Table 2.

TABLE 2
CONDITION ATTRIBUTES

Name	Description
meanbp	Mean arterial blood pressure Day 3
wblc	White blood cell count Day 3
hrt	Heart rate Day 3
resp	Respiratory rate Day 3
temp	Temperature (Celsius)
alb	Serum Albumin
bili	Bilirubin
crea	Serum Creatinine
sod	Sodium
pafi	Pa02 / (.01 * FiO2)
ca	Presence of cancer
age	Patient's age
hday	Days in hospital at study admit
dzgroup	Diagnosis group
scoma	SUPPORT coma score based on Glasgow coma scale

As the decision attribute in the decision table, we define a binary variable (Yes/No) "decesead_in_6months" using the following two attributes from the SUPPORT dataset:

- *"death"* which represents the event of death at any time up to NDI date (National Death Index date: Dec 31, 1994).
- *"D.time":* number of days of follow up

The values of the decision attribute are calculated converting the *"D.time"* value in months and comparing against the attribute *"death"* as follows:

- If "D.time" < 6 months and "death" is equal to 1 (the patient died within 6 months) then "deceased_in_6months" is equal to "Yes"
- If "D.time" > 6 months and "death" is equal to 1 (the patient died after 6 months) then "deceased_in_6months" is equal to "No"
- If "D.time" > 6 months and "death" is equal to 0 (the patient did not died after 6 months) then "deceased_in_6months" is equal to "No"

IV. Expected Results

A set of decision rules will be generated. This set can be used directly by the physician or by a computer based system that interpret the decision rules and help the physician for taking the decision of referring the patient to hospice.

Our intention is to provide a classifier based on Rough Set Theory that improve the accuracy of the standard models used in prognostication of terminally ill patients. Therefore we want to provide a reliable methodology to predict life expectancy.

References

[1] Z. Pawlak, "Rough Sets: Theoretical Aspects of Reasoning about Data," *Kluwer Academic Publishers*, Norwell, MA. 1992

[2] D. W. Hosmer Jr., S. Lemeshow, "Applied Survival Analysis: Regression Modeling of Time to Event Data," *John Wiley & Sons*, Chichester, 1999.

[3] W. A. Knaus, F. E. Harrell Jr, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors Jr, et al, "The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults," *Ann Intern Med*. 1995, pp. 191-203.

[4] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P.G. Bastos, C.A Sirio, D.J Murphy, T. Lotring, A. Damiano, "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults," *Chest*, vol. 100, no. 6, 1991, pp. 1619-1636.

[5] J. R. Bech, S. G. Pauker, J. E. Gottlieb, K. Klein, J. P. Kassirer, "A convenient approximation of life expectancy (The "D.E.A.LE")," Use in medical decision-making, *Am J Med*. 1982, pp. 889-97.

[6] K. J. Cios, J. Kacprzyc, "Medical Data Mining and Knowledge Discovery," *Studies in Fuzziness and Soft Computing 60,* Physica Verlag, Heidelberg, 2001.

[7] J. F. Lucas-Peter, A. Abu-Hanna, "Prognostic methods in medicine," *Artificial Intelligence in Medicine,* vol. 15, no. 2, Feb. 1999, pp. 105-119.

[8] J. Bazan, A. Osmolski, A. Skowron, D. Slezak, M. Sacauka and J. Wroblewski. "Rough Set Approach to the survival Analysis," *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing series* , 2002, pp. 522-529.

[9] J. P. Grzymala- Busse, J. W. Grzymala-Busse, Z. S. Hippe, "Prediction of melanoma using rule induction based on rough sets," In: *Proc of SCI'01*, 2001, vol. 7, pp. 523-527.

[10] S. Tsumoto, "Modelling Medical Diagnostic Rules Based on Rough Sets," in *Proceedings of the First International Conference on Rough Sets and Current Trends in Computing (RSCTC '98),* Lech Polkowski and Andrzej Skowron (Eds.). Springer-Verlag, London, UK, 1998, pp. 475-482.
[11] J. Komorowski and A. Øhrn, "Modeling prognostic power of cardiac tests using rough sets," *Artificial intelligence in medicine*, Feb. 1999, vol. 15, no. 2, pp. 167-191.
[12] F. Lau, D. Cloutier-Fisher, C. Kuziemsky, et al. "A systematic review of prognostic tools for estimating survival time in palliative care," *Journal of Palliative Care*, 2007, vol. 23, no. 2, pp. 93-112.
[13] T. Williamson, "Vagueness," London, Routledge, 1994.
[14] B. Djulbegovic, "Medical diagnosis and philosophy of vagueness – uncertainty due to borderline cases," *Ann Intern Med*. 2008.
[15] Support Datasets Archived At ICPSR