33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Rough Set Theory based Prognostication of Life Expectancy for Terminally Ill Patients

Eleazar Gil-Herrera, Ali Yalcin, Athanasios Tsalatsanis, Laura E. Barnes and Benjamin Djulbegovic

Abstract—**We present a novel knowledge discovery methodology that relies on Rough Set Theory 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. We utilize retrospective data from 9105 patients to demonstrate the design and implementation details of a series of classifiers developed to identify potential hospice candidates. Preliminary results confirm the efficacy of the proposed methodology. We envision our work as a part of a comprehensive decision support system designed to assist terminally ill patients in making end-of-life care decisions.**

I. Introduction

ACCORDING to Medicare regulations, a patient should be referred to hospice if his/her life expectancy is less than 6 months [1]. However, despite the well-documented advantages of hospice services, terminally ill patients do not reap the maximum benefits of hospice care with the majority of them being referred to hospice either prematurely or too late. In general, premature hospice referral is translated to patients losing the opportunity to receive potentially effective treatment, which may have prolonged their lives. Conversely, late hospice referral reduces the quality of life for patients and their families. It is apparent that accurate prognostication of life expectancy is of vital importance for all parties involved in the hospice referral process (e.g. patients, their families, and their physicians).

Here, we propose a novel knowledge discovery methodology developed to identify terminally ill patients with life expectancy less than 6 months. The core of the proposed methodology is Rough Set Theory [2]. The rest of this paper describes implementation details, reports results, and discusses limitations and future directions of our work.

II. Methodology

A. Literature Review

Approaches for developing prognostic models for estimating survival for seriously ill patients range from the use of traditional statistical and probabilistic techniques [3]-[6], to models based on artificial intelligence techniques such as neural networks, decision trees and rough set methods [7]-[11]. A recent systematic review of prognostic tools for estimating survival in palliative care highlighted the lack of accurate end-of-life prognostic models [13].

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 are represented with vagueness [14]. 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 categorization and therefore the following decisions can changes dramatically [15]. Moreover, the dataset 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.

In this work we propose the use of Rough Set Theory (RST) [2] to deal with vagueness and inconsistency in the representation of the dataset. RST provides a mathematical tool for representing and reasoning about vagueness and inconsistency. Its fundamentals are based on the construction of similarity relations between dataset objects from which approximate yet useful solutions are provided. In RST, the knowledge extracted from the data set is represented in the form of "if-then" decision rules where an explanation of how the final decision was derived can be traced. Clinical credibility in prognosis models depends on the ease with which practitioners and patients can understand and interpret the results [16]. Therefore, the if-then decision rule representation offers a significant advantage over "black box" modeling approaches such as neural networks.

RST has been used in a number of applications dealing with modeling medical prognosis [9]–[12]. For example, Tsumoto et al. [11], provides a framework to model medical diagnosis rules showing theoretically that the characteristics of medical reasoning reflect the concepts of approximation established in Rough Set Theory. Komorowski et al. [12], show that RST is useful to extract medical diagnosis rules to identify a group of patients for whom performing a test that is costly or invasive is redundant or superfluous in the prognosis of a particular medical condition.

In this paper 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. ROSETTA [17] software is used to perform the analysis described in the remainder of the paper.

Manuscript received March 26th, 2011. This work was supported in part by the Department of Army under grant #W81 XWH-09-2-0175.

E. Gil-Herrera and A. Yalcin are with the Department of Industrial and Management System Engineering, University of South Florida, Tampa, FL 33620, USA(e-mail: eleazar@mail.usf.edu, ayalcin@eng.usf.edu).

A. Tsalatsanis, L. E. Barnes and B. Djulbegovic are with the Center for Evidence Based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL 33612, USA (e-mail: atsalats@health.usf.edu, lbarnes@health.usf.edu, bdjulbeg@health.usf.edu).

B. Dataset

The dataset 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 dataset [18]. We consider all variables used in the SUPPORT prognostic model [4] 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 I.

As the decision attribute, we define a binary variable (Yes/No) "deceases_in_6months" using the following two attributes from the SUPPORT dataset:

TABLE I
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

- *"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 "deceases_in_6months" is equal to "Yes"
- If "D.time" > 6 months and "death" is equal to 1 (the patient died after 6 months) then "deceases_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 "deceases_in_6months" is equal to "No"

C. Rough Set Theory

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

$$T = (U, A \cup \{d\}) \quad (1)$$

where T represents the dataset 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*.

D. Rough Set Theory Based Knowledge Discovery Process

RST based knowledge discovery process requires sequential and parallel use of various mathematical, statistical and soft computing methodologies with the objective of identifying meaningful relationships between condition and decision attributes.

The selection of specific methodologies for knowledge discovery is largely dependent on the considered dataset. We have taken the following steps in our approach:

1) Data preprocessing: If the selected table contains "holes" in the form of missing values or empty cell entries; the table may be processed in various ways to yield a completed table in which all entries are present. The data completion process for SUPPORT dataset in [18] is adopted in this work. After the preprocessing phase, the number of patients with missing information is reduced by 2 cases. Therefore, there are 9103 complete cases.

The next step in preprocessing is the discretization process. 13 out of 15 of the conditional attributes are continuous; therefore we transformed them into categorical variables. The discretization process is based on the searching of cuts that determine intervals. This process enables the classifier in obtaining a higher quality of classification rules. We found that using cut-off defined by medical experts is the best alternative for the discretization process. We consider the APACHE III Scoring System [5] for determining the cut-off for the physiologic variables along with the age variable. The remaining variables, not defined in [5] are discretized using Boolean Reasoning Algorithm [19] implemented in the ROSETTA software.

Finally, the dataset is divided randomly into training and testing sets containing 500 and 8603 cases, respectively. The training set is used in the discretization process to obtain the cut-off for the numerical attributes.

2) Reduct Generation: This step reduces the dimensionality of the dataset with the intention of removing redundant information and consequently decreases the complexity of the mining process. A reduct is the minimal set of attributes that enable the same classification as the complete set of attributes without loss of information. There are many algorithms for computing reducts for which the effect to the classification performance is critical. Since the computational complexity of the reduct generation problem is NP-hard [19], various suboptimal techniques have been proposed. In this work the dynamic reduct approach ([20-21]) is used for reduct generation.

2.1) Dynamic Reducts

Dynamic reducts algorithm aims at obtaining the most

stable sets of reducts for a given dataset by sampling within this dataset. Random samples of the testing set are selected iteratively and reducts for the samples are computed using genetic algorithms [22-23]. The reducts that most frequently appear in the samples are the most stable.

Based on the principle of the dynamic reducts technique, we have randomly selected 100 subdivisions of the training set to use for reduct generation. The actual number of patient profiles included in each subdivision of the training set varies between 50% and 90% of the training dataset. Using this approach, 229 reducts were obtained from which the set of decision rules are generated.

2.2) Using the decision attribute as condition attribute

Typically only the condition attributes are used to generate reducts. As an alternative, we included the decision attribute d in the set of condition attributes and calculated the reducts based on this scheme.

The decision attribute (deceases_in_6_months) used as a condition attribute is intended to represent the physician's estimate of life expectancy expressed in terms of the decision classes defined for this problem. Survival prognosis models that incorporate physician estimates are shown to improve both predictive accuracy and the ability to identify patients with high probabilities of survival or death [4]. In this case, 549 reducts were obtained. The next step is the induction of decision rules.

3) Rule Induction. The ultimate goal of the RST based knowledge discovery methodology is to generate decision rules, which will be used in classifying each patient as surviving or not surviving within the defined period of time. A decision rule has the form: *if A then B* ($A \rightarrow B$), where A is called the condition and B the decision of the rule. Decision rules can be thought of as a formal language for drawing conclusions from data.

The decision rules were generated based on the two aforementioned sets of reducts. After the process of reducts generation, the decision table is presented in a compact shape from which the decision rules are generated

4) Classification. Based on the set of rules generated, we can classify patients as surviving or not surviving the six-month period. However, not all rules are conclusive. Patients with profiles identical to the conditions of the rules are not decisively classified. In addition, there are situations of contradictory rules, e.g. one or more rules classify a patient as surviving and some other rules classify the same patient as dying. To overcome these problems a *standard voting* algorithm [19] is used which allows all rules to participate in the decision process and classify a patient based on majority voting.

III. Results

This section compares the performance of the classification processes where, the patients in the training dataset are classified as *survive, not survive* or *undefined* based on the induced rules and the classification process described. The results are presented in a confusion matrix form.

The accuracy of each classification model is reported in terms of Area under the Receiver Operating Characteristic curve (AUC). The best possible classification is achieved when AUC is equal to 1, while no classification ability exists when AUC is equal to 0.5.

Table 2 presents the confusion matrix for the classification model based on reducts generated on only the original condition attributes (without including the decision attribute). Table 3 shows the confusion matrix for the alternative case where the decision attribute is included in the set of condition attributes.

TABLE 2
CONFUSION MATRIX. THE REDUCTS ARE BASED ON SET A. THE CLASSIFIER PRESENTS AUC EQUAL TO 0.55 INDICATING WEAK DISCRIMINATION ABILITY.

		Predicted		
		Not survive	Survive	Undefined
Actual	Not survive	1395	1953	677
	Survive	1410	2542	626

Sensitivity = 0.64
Specificity = 0.42
AUC = 0.55

TABLE 3
CONFUSION MATRIX. THE REDUCTS ARE BASED ON SET $A = A \cup \{d\}$. THE CLASSIFIER PRESENTS AUC EQUAL TO 0.90 INDICATING GOOD DISCRIMINATION ABILITY.

		Predicted		
		Not survive	Survive	Undefined
Actual	Not survive	1999	471	1555
	Survive	312	3245	1021

Sensitivity = 0.91
Specificity = 0.81
AUC = 0.90

The dynamic reducts approach without using the decision attribute as a condition attribute shows a weak discrimination ability. However, it demonstrates a fairly high level of coverage, being able to classify around 85% of the test cases. As shown in Table 3, the classification performance in terms of AUC when using the decision attribute as a part of the condition attributes is approximately 0.90. Both the specificity and sensitivity scores are tremendously improved. However, the classification coverage in this case is reduced to 70%.

The described classification process was repeated 10 times using randomly selected samples from the dataset (again 500 cases for training and the remainder 8603 cases for testing). The overall classification performance is obtaining by averaging the AUC from each iteration. Using the original set of attributes, the overall AUC is 0.56 (SD = 0.01). Following the same, we obtained an AUC of 0.85 (SD = 0.065) for the case where the decision attribute is used as a condition attribute.

IV. Conclusions and Future Work

The SUPPORT model is the "gold standard" model for prognostication of terminally ill patients. The AUC for prediction of survival for 180 days in the SUPPORT study is 0.79, and 0.82 when SUPPORT model is combined with physician's estimates [4].

This initial exercise in applying knowledge discovery methodologies based on rough set theory shows promise in developing a reliable methodology to predict life expectancy. The baseline model using dynamic reducts presents several opprotunities for improvement:

1. Due to the limitations of the ROSETTA software, the size of the training set was limited to 500. The size of the training set may be a limiting factor to obtaining better classification accuracy and coverage considering the high number of categories associated with each attribute.
2. One area that needs to be explored is the appropriate weighting of the condition attributes in terms of their impact on the decision variable. The baseline case assumes that all physiological attributes are weighed equally. We believe that a careful weighting of the attributes by consulting an expert will greatly improve the classification accuracy of the approach.

Including the physician's estimate in the prognostication process is an important component of our future work. The classifier which uses the decision attribute as a condition attribute is intended to incorporate the professional opinion of the physician. This classifier performed much better than the baseline model and its accuracy exceeded that of the SUPPORT model. However we note that, in this approach only 70% of the test cases could be classified and more research is required to minimize the number of *undefined* cases. Furthermore, our model used the decision attribute from a retrospective study for which the decision was known with 100% accuracy. Ideally this approach should be tested on a prospective dataset and its performance compared to other soft models based on AI techniques which are a part of our future work.

Finally, it is important to remember that regardless of the accuracy of any classifier, medical decisions must take into account the individual patient preferences towards alternative forms of treatments[24]. Therefore, our intent is to incorporate our methodology into a patient-centric decision support system to facilitate the hospice referral process.

References

[1] L. R. Aiken, "Dying, Death, and Bereavement," *Allyn and Bacon*, 1985, p. 214.

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

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

[4] 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. s

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] T. Williamson, "Vagueness," London, Routledge, 1994.

[15] B. Djulbegovic, "Medical diagnosis and philosophy of vagueness – uncertainty due to borderline cases," *Ann Intern Med*. 2008.

[16] A. Hart and J. Wyatt, "Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks," *Medical informatics,* 1990 vol. 15, no. 3, pp. 229-236.

[17] A. Øhrn, J. Komorowski, "ROSETTA: A Rough Set Toolkit for Analysis of Data," *Proc. Third International Joint Conference on Information Sciences, Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97),* Durham, NC, USA, 1997, March 1-5, vol. 3, pp. 403-407.

[18] Support Datasets Archived At ICPSR (http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets)

[19] J. G. Bazan, H. S. Nguyen, P. Synak, J. Wroblewski, "Rough set algorithms in classification problem," In: L. Polkowski, S. Tsumoto, T.Y Lin, (Eds.), "Rough set methods and applications: new developments in knowledge discovery in information systems. Studies in Fuzziness and Soft Computing," *Physica-Verlag*, Heidelberg, Germany, 2000, pp. 49-88.

[20] J. Bazan, A. Skowron, P. Synak, "Dynamic reducts as a tool for extracting laws from decision tables," *Proceedings of the Eighth International Symposium on Methodologies for Intelligent Systems. Lecture Notes in Artificial Intelligence 869*, Berlin, Springer-Verlag, 1994, pp. 346-355.

[21] J. Bazan, "Dynamic Reducts and Statistical inference," In *Sixth International conference, Information Procesing and Management of Uncertainty in Knowledge-Based Systems*, Granada, Spain, Universidad de Granda, 1996.

[22] J. Wroblewski, "Finding minimal reducts using genetic algorithms," In *Proc. Second International Joint Conference on Information Sciences*, 1995, pp. 186–189.

[23] D. E. Goldberg, "GA in search, optimization, and machine learning," *Addison-Wesley,* 1989.

[24] A.Tsalatsanis, I. Hozo, A. Vickers, B. Djulbegovic, "A regret theory approach to decision curve analysis: A novel method for eliciting decision maker's preferences and decision making," *BMC Medical Informatics and Decision making*, 2010, vol. 10, issue 51.