34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Towards a Classification Model to Identify Hospice Candidates in Terminally Ill Patients

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

Abstract— **This paper presents a Rough Set Theory (RST) based classification model to identify hospice candidates within a group of terminally ill patients. Hospice care considerations are particularly valuable for terminally ill patients since they enable patients and their families to initiate end-of-life discussions and choose the most desired management strategy for the remainder of their lives. Unlike traditional data mining methodologies, our approach seeks to identify subgroups of patients possessing common characteristics that distinguish them from other subgroups in the dataset. Thus, heterogeneity in the data set is captured before the classification model is built. Object related reducts are used to obtain the minimum set of attributes that describe each subgroup existing in the dataset. As a result, a collection of decision rules is derived for classifying new patients based on the subgroup to which they belong. Results show improvements in the classification accuracy compared to a traditional RST methodology, in which patient diversity is not considered. We envision our work as a part of a comprehensive decision support system designed to facilitate end-of-life care decisions. Retrospective data from 9105 patients is used to demonstrate the design and implementation details of the classification model.**

I. Introduction

A. *Hospice referral criteria*

Hospice is designed to provide comfort and support to terminally ill patients and their families. According to Medicare regulations, a patient should be referred to hospice if his/her life expectancy is approximately 6 months or less [1]. However, most patients are not referred to hospice in a timely manner [2, 3] and therefore they do not reap the well-documented benefits of hospice services. A premature hospice referral translates to a patient losing the opportunity to receive potentially effective treatment, which may prolong their life. Conversely, a late hospice referral may deprive patients and their families of enjoying the benefits offered. Therefore, accurate prognostication of life expectancy is of vital importance for terminal patients as well as for their families and physicians.

B. *Prognostic models for estimating survival of terminally ill patients*

Survival prognostic models range from traditional statistical and probabilistic techniques [4-10], to models based on artificial intelligence such as neural networks [11, 12], decision trees [13, 14] and rough set methods [15, 16]. The primary goal of survival prognostic models is to provide accurate information regarding life expectancy and/or determine the association between prognostic factors and survival. Typically, the information derived by prognostic models is presented in terms of probability of death within a time period. Recent systematic reviews [17, 18] have highlighted the necessity of prediction models that can be easily integrated into clinical practice and facilitate end-of-life clinical decision-making.

Several important issues demand particular consideration when developing clinical classification models: First, clinical data, representing patient records that include symptoms and clinical signs, are not always well defined and are represented with *vagueness* [19]. Therefore, it is very difficult to classify cases in which small differences in the value of an attribute may completely change the classification of a patient and, as a result, the treatment decisions [20]. Second, clinical data may present *inconsistencies*, which means that it is possible to have more than one patient with the same description but with different outcomes. Third, the results of prognostic models should be readily interpretable to enable practical and posteriori inspection and interpretation by the treating physician or an expert system [21]. Finally, prognostic models should consider the heterogeneity in clinical data, i.e. the existence of patient diversity presented in terms of risk of disease and responsiveness to treatment [22, 23]. This consideration will enable a prognostic model to identify possible subgroups of patients for which certain covariates do not influence their classification. The practical implications of such considerations are associated with the ability to customize the prognostic model for each subgroup of patients (e.g. expensive and/or potentially harmful tests may be avoided for particular subgroups).

Rough Set Theory (RST) [24], a mathematical tool for representing and reasoning about vagueness and inconsistency in data sets, has been used in a number of applications dealing with modeling medical prognosis [15, 16, 25-28]. For example, Tsumoto et al. [25], provide a framework to model medical diagnosis rules showing theoretically that the characteristics of medical reasoning reflect the concepts of approximation established in RST. Komorowski et al. [26], 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. Recently, [28] highlighted features of RST for integrating into medical applications. For example, RST has the ability to handle imprecise and uncertain information and provides a schematic approach for analyzing data without initial assumptions on data distribution.

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. 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)

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

In our previous work [29], we proposed the use of RST to predict the life expectancy of terminally ill patients using a *global reduction* [30] methodology to identify the most significant attributes for building the classification model. However, we found that the number of attributes used in the model was barely reduced and therefore produced long decision rules. Moreover, considering the number of discretization categories associated with each attribute, the generated decision rules were built to describe each object in the training set and therefore, they were poorly suited for classifying new cases.

Here, we propose the use of an alternative attribute reduction methodology that aims to identify groups of patients that share common characteristics that distinguish them from the rest of the patients. As a result, we obtain subgroups of patients from which different sets of significant attributes are identified. The decision rules generated in this manner contain fewer attributes and therefore are more suitable to classify new patients. Moreover, by studying each subgroup, we can reason about how a different rule-set is applied to a particular patient.

The rest of the paper describes details of the proposed RST based 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 [31] software is used to perform the analysis described in the remainder of the paper.

II. Methodology

A. Data Set

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 [30]. We consider all variables used in the SUPPORT prognostic model [3] as condition attributes, i.e. the 10 physiologic variables along with the diagnosis groups, age, number of days in the hospital before entering the study, presence of cancer, and neurologic function. Data collection and patient selection procedures are detailed in [3]. 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 prognosis model dataset:

- death: 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 "Yes". Otherwise, it is implicit that a patient survived the 6-month period; hence, "deceased_in_6months" is "No".

B. Rough Set Theory Data Representation

Based on RST, the data set is represented as:

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

TABLE I. Condition Attributes

Name	Description
alb	Serum albumin
bili	Bilirubin
crea	Serum creatinine
hrt	Heart rate
meanbp	Mean arterial blood pressure
pafi	Arterial blood gases
resp	Respiratory rate
sod	Sodium
temp	Temperature (Celsius)
wblc	White blood cell count
dzgroup	Diagnosis group
age	Patient's age
hday	Days in hospital at study admit
ca	Presence of cancer
scoma	SUPPORT coma score based on Glasgow coma scale

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* is a non-empty finite set of attributes called the condition attributes. In our case, an object designates a terminally ill patient and an attribute $a \in A$ designates each of the fifteen condition attributes that describe a patient (Table I). For every attribute, 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.

C. Development of the Classification Model

This process typically involves numerous steps, such as data preprocessing, discretization, reduction of attributes, rule induction, classification and interpretation of the results. Details on the data preprocessing and data discretization for this data set are described in [29]. The ultimate goal of this process is to generate decision rules, which are used to classify each patient as surviving or not surviving within the defined period of time. A decision rule has the form: *if A then B (A → B),* where *A* is called the condition and *B* the decision of the rule.

Here, we are focusing on an alternative method of reducing the attribute dimensions and identify different subgroups of similar patients in the data set. In [32], two types of reducts are defined:

1) Global Reducts:

Consists of the minimal set of attributes that preserve the structure of the entire data set. A set $B \subseteq A$ is called a global reduct if the indiscernibility relation using attributes B is equal to the indiscernibility relation using all the condition attributes A, i.e.:

$$IND(B) = IND(A), \text{ where,}$$

$$IND(B) = \{(u_i, u_j) \in U^2 \colon \forall\, a_k \in B, \mathrm{a_k}(\mathrm{u_i}) \neq \mathrm{a_k}(\mathrm{u_j})\}$$

As an example, consider the following global reduct obtained from the data set containing 12 condition attributes:

G_RED = {age, dzgroup, scoma, ca, meanbp, wblc, hrt, resp, temp, bili, crea, sod}

Using *G_RED*, few patients will have exactly the same attribute-value combinations because the number of discretization categories associated with each attribute is high. Thus, the decision rules generated are too specific to the cases in the training set and therefore may not be able to classify new cases accurately. Moreover, the fact that global reducts represent the entire data set makes it difficult to detect the presence of heterogeneous groups in the data meaning that the causes of diversity between the patient outcomes will remain unknown.

2) *Object related reducts (ORR):*

Represents the minimal attribute subsets that discern an object $u \in U$ from the rest of objects belonging to a different decision class. Mathematically, an ORR $R_u \subseteq A$ is defined as:

$$\forall u_i \in U : d(u_i) \neq d(u_j) \Rightarrow \exists a_k \in R_u : a_k(u_i) \neq a_k(u_j), \; where\ u_i \neq u_j .$$

An ORR is the minimal and vital information that is used to partition the universe of objects into smaller, homogeneous subgroups, where objects within a subgroup are related by means of information described by the ORR. Decision rules generated by this scheme will usually contain fewer attributes and are more suitable to classify new cases. Some decision rules contain a different set of attributes applicable for a particular subgroup of patients.

III. RESULTS

The two methods for dimensionality reduction produce a set of reducts. The number of reducts and decision rules obtained are presented in Table II. Based on the decision rules generated, patients are classified as surviving or not surviving the six-month period. A standard voting algorithm [30] is used for this purpose. Table III, presents the performance of two classification models based on each type of reduct generation described. The performance of each classification model is represented in terms of *sensitivity*, *specificity*, *Area under the Receiver Operating Characteristic curve* (AUC) and *coverage* of the model. A 5-fold cross validation procedure was applied to estimate the performance of each classification model, where, the entire data set is randomly divided into five subsets (folds). Then, each fold (20% of the data set) is used once as a testing set, while the remaining folds (80%) are used for training. The process is repeated five times and the results are averaged to provide an estimate for the classifier performance.

Compared to the Global reduct approach, the ORR approach has enhanced the classification performance in terms of AUC and sensitivity. Moreover the decision rules generated are able to classify all new cases.

TABLE II. NUMBER OF REDUCTS AND DECISION RULES GENERATED – GLOBAL VS. ORR

Method	**Number of reducts**	**Number of rules**
Global reducts	99	647,223
ORR	11,894	68,492

TABLE III. CLASIFICATION RESULTS – GLOBAL VS. ORR

Method	**Sensitivity**	**Specificity**	**AUC**	**Coverage**
Global reducts	73.67%	44.05%	61.8%	86.43%
ORR	86.92%	39.2%	71.9%	100%

IV. DISCUSSION

Analyzing the information obtained from the ORR, we can identify groups of patients for whom it is possible to evade costly, invasive or even unnecessary tests required by the prediction model. For example, the following two ORRs generate rules independent of the *Pafi* score (associated with the patient's blood gases), without reducing the classification accuracy. The importance of such finding becomes apparent considering that in clinical practice *Pafi* is not collected routinely for patients outside the Intensive Care Unit (ICU).

- ORR = {Age, dzgroup, meanbp} generates the following decision rules:
  - if age= [45, 60) AND dzgroup = (Lung Cancer) AND meanbp=[60, 70) then: Survive = 22.86%, Die = 77.14%.
  - if age= [45, 60) AND dzgroup = (CHF) AND meanbp=[100, 120) then: Survive = 82.93%, Die = 17.07%.
  - if age= [70, 75) AND dzgroup = (COPD) AND meanbp=[80,100) then: Survive = 84.21%, Die = 15.79%.
- ORR = {Age, dzgroup, hrt, crea} generates the following decision rules:
  - if age= [45, 60) AND dzgroup = (CHF) AND hrt=[100,110) and crea[1.95, *] then: Survive = 83.33%, Die = 16.67%.
  - if age= [75,85) AND dzgroup = (CHF) AND hrt=[50,110) and crea[0.5, 1.5) then: Survive = 82.19%, Die = 17.81%.

Consequently, the use of *Pafi* test in patients that belong to one of those groups defined by the ORR's will not improve the prognostication accuracy.

Our approach demonstrates features that make it particularly suitable for use in clinical decision-making. It is a patient-centric methodology which is able to predict without the use of unnecessary, expensive and/or invasive procedures for certain subgroups of patients. Consequently, selection of attributes upon which a decision is to be made is critical to minimizing healthcare costs and maximizing the quality of patient care. Finally, considering that more than one ORR could discern each patient, the information acquired offers several options dependent on the attribute values available for each individual patient.

V. FUTURE WORK

The number of ORR and the decision rules generated depends on the number of condition attributes and its categories. For clinical datasets, which contain large numbers of condition attributes, the number of ORRs and decision rules generated can be extremely large to be evaluated directly by human experts. Therefore, the interpretation and analysis of the ORRs and their decision rules requires the use of a well-defined methodology.

Compared to our previous work [29], the results presented in this paper show an improvement in the classifier performance. However, further research need to be conducted in order to achieve a reliable prognostic model.

REFERENCES

[1] L.R. Aiken and NetLibrary Inc., "Dying, death, and bereavement," in Book Dying, death, and bereavement, *Series Dying, death, and bereavement*, 4th ed. Lawrence Erlbaum Associates, 2000.

[2] N.A. Christakis, "Timing of referral of terminally ill patients to an outpatient hospice.," *J Gen Intern Med*, vol. 9, (no. 6), pp. 314-20, Jun 1994.

[3] A. Tsalatsanis, L.E. Barnes, I. Hozo, and B. Djulbegovic, "Extensions to Regret-based Decision Curve Analysis: An application to hospice referral for terminal patients," *BMC Med Inform Decis Mak*, vol. 11, pp. 77, 2011.

[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, and A. Damiano, "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults," *Chest*, vol. 100, (no. 6), pp. 1619-1636, December, 1991.

[5] W.A. Knaus, F.E. Harrell, J. Lynn, L. Goldman, R.S. Phillips, A.F. Connors, N.V. Dawson, W.J. Fulkerson, R.M. Califf, N. Desbiens, P. Layde, R.K. Oye, P.E. Bellamy, R.B. Hakim, and D.P. Wagner, "The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults," *Annals of Internal Medicine*, vol. 122, (no. 3), pp. 191-203, February, 1995.

[6] D.W. Hosmer and S. Lemeshow, *Applied survival analysis regression modeling of time to event data*, New York, NY: Wiley, 1999.

[7] J.R. Beck, S.G. Pauker, J.E. Gottlieb, K. Klein, and J.P. Kassirer, "A convenient approximation of life expectancy (the "DEALE"): II. Use in medical decision-making," *The American Journal of Medicine*, vol. 73, (no. 6), pp. 889-897, 1982.

[8] I. Hyodo, T. Morita, I. Adachi, Y. Shima, A. Yoshizawa, and K. Hiraga, "Development of a Predicting Tool for Survival of Terminally Ill Cancer Patients," *Japanese Journal of Clinical Oncology*, vol. 40, (no. 5), pp. 442-448, May 1, 2010.

[9] D. Porock, D. Parker-Oliver, G. Petroski, and M. Rantz, "The MDS Mortality Risk Index: The evolution of a method for predicting 6-month mortality in nursing home residents," *BMC Research Notes*, vol. 3, (no. 1), pp. 200, 2010.

[10] P.K.J. Han, M. Lee, B.B. Reeve, A.B. Mariotto, Z. Wang, R.D. Hays, K.R. Yabroff, M. Topor, and E.J. Feuer, "Development of a Prognostic Model for Six-Month Mortality in Older Adults With Declining Health," *Journal of Pain and Symptom Management*, vol. 43, (no. 3), pp. 527-539, 2012.

[11] G. Clermont, D.C. Angus, S.M. DiRusso, M. Griffin, and W.T. Linde-Zwirble, "Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models," *Critical Care Medicine*, vol. 29, (no. 2), 2001.

[12] R. Dybowski, V. Gant, P. Weller, and R. Chang, "Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm," *The Lancet*, vol. 347, (no. 9009), pp. 1146-1150, 1996.

[13] M.R. Segal, "Features of Tree-Structured Survival Analysis," *Epidemiology*, vol. 8, (no. 4), pp. 344-346, 1997.

[14] S.S. Hwang, C.B. Scott, V.T. Chang, J. Cogswell, S. Srinivas, and B. Kasimis, "Prediction of Survival for Advanced Cancer Patients by Recursive Partitioning Analysis: Role of Karnofsky Performance Status, Quality of Life, and Symptom Distress," *Cancer Investigation*, vol. 22, (no. 5), pp. 678-687,2004.

[15] J. Bazan, A. Osmólski, A. Skowron, D. Ślçezak, M. Szczuka, and J. Wróblewski, "Rough Set Approach to the Survival Analysis - Rough Sets and Current Trends in Computing," vol. 2475, *Lecture Notes in Computer Science*, J. Alpigini, J. Peters, A. Skowron and N. Zhong eds.: Springer Berlin / Heidelberg, pp. 951-951, 2002.

[16] P. Pattaraintakorn, N. Cercone, and K. Naruedomkul, "Hybrid rough sets intelligent system architecture for survival analysis," in Transactions on rough sets VII, W. M. Victor, O. Ewa, owska, S. Roman, owinski and Z. Wojciech eds.: Springer-Verlag, 2007, pp. 206-224.

[17] F. Lau, D. Cloutier-Fisher, C. Kuziemsky, F. Black, M. Downing, and E. Borycki, *A systematic review of prognostic tools for estimating survival time in palliative care*, Montreal, CANADA: Centre of Bioethics, Clinical Research Institute of Montreal, 2007.

[18] P. Glare, C. Sinclair, M. Downing, P. Stone, M. Maltoni, and A. Vigano, "Predicting survival in patients with advanced disease," *European Journal of Cancer*, vol. 44, (no. 8), pp. 1146-1156, 2008.

[19] P. Simons, "VAGUENESS" *International Journal of Philosophical Studies*, vol. 4, (no. 2), pp. 321-327, Sep 1996.

[20] B. Djulbegovic, "Medical diagnosis and philosophy of vagueness-uncertainty due to borderline cases," *Annals of Internal Medicine*, 2008.

[21] J.C. Wyatt and D.G. Altman, "Commentary: Prognostic models: clinically useful or quickly forgotten?," *BMJ*, vol. 311, (no. 7019), pp. 1539-1541, 1995.

[22] R.L. Kravitz, N. Duan, and J. Braslow, "Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages," *Milbank Quarterly*, vol. 82, (no. 4), pp. 661-687, 2004.

[23] P. Schlattmann, "Introduction - Heterogeneity in Medicine Medical Applications of Finite Mixture Models," *Statistics for Biology and Health*: Springer Berlin Heidelberg, 2009, pp. 1-22.

[24] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Norwell, MA, 1992.

[25] S. Tsumoto, "Modelling Medical Diagnostic Rules Based on Rough Sets- Rough Sets and Current Trends in Computing," vol. 1424, *Lecture Notes in Computer Science*, L. Polkowski and A. Skowron eds.: Springer Berlin / Heidelberg, 1998, pp. 475-482.

[26] J. Komorowski and A. Øhrn, "Modelling prognostic power of cardiac tests using rough sets," *Artificial Intelligence in Medicine*, vol. 15, (no. 2), pp. 167-191, 1999.

[27] P. Grzymala-Busse, J.W. Grzymala-Busse, and Z.S. Hippe, "Melanoma prediction using data mining system LERS," in *Proc,, COMPSAC,* 2001, pp. 615-620.

[28] P. Pattaraintakorn and N. Cercone, "Integrating rough set theory and medical applications," *Applied Mathematics Letters*, vol. 21, (no. 4), pp. 400-403, 2008.

[29] E. Gil-Herrera, A. Yalcin, A. Tsalatsanis, L.E. Barnes, and D. B, "Rough Set Theory based Prognostication of Life Expectancy for Terminally Ill Patients," in *Conf Proc IEEE Eng Med Biol Soc*, 2011, pp. 6438-6441.

[30] J. Bazan, H. Nguyen, S. Nguyen, P. Synak, J. Wroblewski, L. Polkowski, S. Tsumoto, and T. Lin, "Rough Set Algorithms in Classification Problem," in Rough set methods and applications: new developments in knowledge discovery in information systems: Physica-Verlag, 2000, pp. 49-88.

[31] Ø. Alexander and J. Komorowski, "ROSETTA: A Rough Set Toolkit for Analysis of Data," in *Proc. Third International Joint Conference on Information Sciences*, 1997, pp. 403-407.

[32] J.G. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak and J. Wróblewski, "Rough set algorithms in classification problem," in *Rough set methods and applications*: Physica-Verlag GmbH, 2000, pp. 49-88.