EUROPEAN JOURNAL OF BIOMEDICAL AND PHARMACEUTICAL SCIENCES
http://www.ejbps.com

ISSN 2349-8870
Volume: 3
Issue: 12
316-320
Year: 2016

ROUGH SET BASED RULE INDUCTION APPROACH FOR SURVIVAL ANALYSIS

***Perumal Venkatesan**

Department of Community Medicine Sri Ramachandra Medical College & Research Institute Sri Ramachandra University, Chennai, India.

***Corresponding Author: Dr. Perumal Venkatesan**
Department of Community Medicine Sri Ramachandra Medical College & Research Institute Sri Ramachandra University, Chennai, India.

Article Received on 27/09/2016 **Article Revised on 17/10/2016** **Article Accepted on 07/11/2016**

ABSTRACT
The boundaries in medical diagnosis and treatment are usually vague and imprecise. With vague information on signs and symptoms the physicians diagnose a patient and decide on the best way to cure them. Rough set is a leading soft computing method and its theory provides methods for knowledge extraction from imperfect data. This paper deals with a rough set based new approach for survival analysis. The decision problem is formulated over database on spinal tuberculosis patients treated under a randomized trial. The rule based method expresses the difference between the expected outcomes under the different treatment groups. The rules for different survival tendencies are also formulated using a prognostic index. The new frame work is combined with semi-parametric survival models to identify survival patterns. The usefulness of the method is demonstrated using outcomes of the trial.

KEYWORDS: Rough set, rule induction, survival analysis, Kaplan-Meier, Cox model, spinal tuberculosis.

INTRODUCTION
Pawlak[1] introduced mathematical rough set theory in the early 1980's. For many research areas such as machine learning, pattern recognition, data mining, inductive reasoning and expert systems this theory is fundamental importance.[2-4] The rough set approach is applied in many fields including bioinformatics, medicine, computer networks and other fields.[5-8] The theory was based on the discernibility of objects. Rough set theory provides methods to handle uncertainty. If a concept is 'not definable' in a given knowledge base, rough sets can 'approximate' with respect to that knowledge. Rough set was developed to approximate vague knowledge efficiently and is a form of inductive reasoning which is appropriate to deal with incomplete data problems such as survival analysis.[9]

Rough Set Theory
We describe the fundamental theory of rough sets from Pawlak[1-2], Pattaraintakorn and Cercone.[10] Given a finite set $U \neq \emptyset$ (universe) of objects, any subset $X \subseteq U$ of the universe is called a concept in U and any family of concepts in U is referred to as knowledge. A family of classifications over U is called a knowledge base over U. Rough set methodology endeavours to discover the variety of data sources while requiring integration of other approaches to handle extensibility of data sets. Let R→X × X be an equivalence relation over U. Then R is reflexive (x Rx), symmetric (if x Ry then yRx) and transitive (if xRy and yRz then xRz). Define U/R as the family of equivalence classes of R and let $[x]_R$ denote a category in R containing an element XϵU. Given a knowledge base K = (U, R) if P⊆ R and P≠ φ then there is an equivalence relation IND(P) called the indiscernibility relation over P.

We define with XϵU and Rϵ IND(K)
x= RX if and only if $[x]_R \subseteq X$
$x=\overline{R}(X)$ if and only if $[x]_R \cap X \neq \varphi$ (1)
called R-lower approximation and R-upper approximation of X respectively. The degree of completeness using card R is given by

$$\alpha_R(X) = \frac{\text{card}\,\underline{R}}{\text{card}\,\overline{R}} \quad \text{where } X \neq \emptyset. \tag{2}$$

One can measure the inexactness using the expressions given below:

(1) If $\underline{R}X \neq \emptyset$ and $\overline{R}X \neq U$, then X is *roughly R-definable.*
(2) If $\underline{R}X = \emptyset$ and $\overline{R}X \neq U$, then X is *internally R-undefinable.*
(3) If $\underline{R}X \neq \emptyset$ and $\overline{R}X = U$, then X is *externally R-undefinable.*
(4) If $\underline{R}X = \emptyset$ and $\overline{R}X = U$, then X is *totally R-undefinable.* (3)

Using discernibility matrix, we calculated relative discernibility function which gave the minimum set of attributes necessary to differentiate a given class from others. To do this concatenated the cells with logical "AND"s and concatenated the attributes within each cell with logical "OR"s.[11] We were looking for the minimum set of attributes necessary to attributes necessary to differentiate each class. The importance of

the attributes were determined using relative reduct which was calculated by taking the relative discernibility function and removing superfluous attributes. The rules were derived from the reducts by binding the conditional attribute values of the equivalence class to the corresponding attributes reduct.[12-13]

For vague class we used the rough membership function which looks at the distribution of different decision attributes in a particular vague class and assigned to the class with maximum probability. Each class has a lower approximation-the set definitely [belongs to the class and an upper approximation- the set which possibly belong to the class and a boundary region –the set whose class could not be proven either way.[14]

Rule-Based Methods

In medical data analysis rule induction was applied mainly to problems concerning disease diagnosis.[15-16] Recent works attempts to apply rule induction and rough set theory for identification of important features for patients survival time.[9,10] The rough sets were used for identification of main factors affecting survival time of patients. The survival was considered as a discrete variable with predefined values.

The rough sets were also applied for rule induction. Before applying rule based induction, a prognostic index was calculated for each sample based on Cox's PH model. The PI values was divided in to intervals so that survival curves determined for observations with PI belong to different intervals were statistically different. The LASD algorithm and LAD algorithm for induction of survival are used.[17-18]

Survival Analysis

The Kaplan-Meier product limit estimator is a providing the means for construction of survival function S(t) given the complete and censored observations ordered in time. It returns the cumulative proportions of cases surviving up to time t.[19]

$$S(t)=\pi(N-J)/(NJ+1)^{\delta(t)} \quad (4)$$

Where N denotes the total number of patients and

$$\Delta(t)= \{1 \text{ if uncensored} \quad (5)$$
$$\{0 \text{ if censored}$$

The Cox model is a semi-parametric model essentially based on the assumption that the survival distribution is exponential. It attempts to estimate the instantaneous risk of death of each particular patient $u \epsilon U$ at time t by

$$h(t;x(u)= h_0(t) \exp(\Sigma\beta X) \quad (6)$$

Where Xs are attributes, β's are regression coefficients, $h_0(t)$ is the baseline hazard. This model can be linearized to get Prognostic Index(PI).[18]

$$PI(u)= \ln \{ h(t;x(u))/ h_0(t)\} = \Sigma\beta X) \quad (7)$$

PI is the linear combination of the attributes. The task is to find the collection of attributes that at the same time have clear clinical interpretation and allows the PI to be applied efficiently to making prognosis.

Spinal Tuberculosis Database

The database consists of three hundred and four spinal tuberculosis patients involving the thoracic or lumbar spine allocated to one of three treatment regimens.[20] In two of the series the patients were ambulant from the start of chemotherapy and received isoniazid plus rifampicin for either 6 months (Regimen2) or 9 months (Regimen3). In third series all patients had isoniazid plus rifampicin for 6 months and in addition underwent a modified 'Hong Kong' operation of radical resection of the lesion and insertion of autologous bone grafts (Regimen1).

Progress was assessed weekly/monthly until the end of chemotherapy, then every 3 months until 30 months, then at 6-monthly intervals until 5years and yearly up to 15 years. In brief, out of 304 patients, 44 patients were excluded for various reasons and here we consider 260 patients as given in Table 1. Other details of the study can be found in ICMR/BMRC.[20] We have considered a total 16 variables (demographical, clinical, radiological and bacteriological) including the response variable for the analysis and the coding of these variables are given in Table2.

Table 1. Patients pre-treatment characteristics

PreRx Characteristic	No.	%
Total	260	100
Age (yrs)<14	87	34
14-34	107	41
35+	66	25
Sinus/Abscess Present	51	20
Kyphosis	219	84
Lim. movement	246	95
Mylopathy	19	7
Site of Lesion		
Thoracic/Thoraco-lumbar	132	51
Lumbar/Lumbo-Sacral	129	49
Radiographic Activity of Disease	250	96
Mediastinal/Psoas	149	57

Allocated Treatment		
Rad6	85	33
Amb6	83	32
Amb9	92	35

Table 2: Variables coding

Variable	Codes
Treatment	1-Amb6, 2-Amb9, 3-Rad6,
Age	1-(<14), 2-(14-34), 3-(35+)
Sinus/Abscess	1-Present, 0-Absent
Kyphosis	Angle in degrees
Lim.Movement	1-Yes, 0-No
Myelopathy	1-Present, 0-Absent
Site of Lesion	1-Thoraci, 2-Thoraco-Lumber, 3-Lumbar, 4-Lumbo-Sacral
Mediastinal/psoas	1-Present, 0-Absent
Radiographic Activity	1-Active, 0-Not active
Number of vertebra involved	Exact number
Total vertebral body Loss	Number
Hospital Stay	Days
Deviation from allocated treatment	1-Yes, 0-No
Regularity in treatment taking	1-Yes, 0-No
Response	1-Favourable, 0-Unfavourable
Response Time	Exact Months
Status	1-Uncensored, 0-Censored

Here the main task was to show, whether the response of ambulatory treatments of 6 and 9 months are not lower than in case of treatment plus radical operation in terms of response. It is an important factor, since ambulatory treatments do not involve surgery and giving less side effects. The five year survival period was considered for analysis. At the end of five years we had three classes of patients: favourable response, doubtful response and failures.

We considered a new decision attribute with three values, corresponding to the above classes. This attribute was created by basing on time interval. All conditional indiscernibility classes contained objects with all three decision values. Hence, probabilistic analysis was applied based on Bazan et al.[9] Probabilistic decision distributions objects from favourable and unfavourable classes were discerned. The doubtful class were merged with favourable for constructing decision tables in the final analysis. The main comparison was between the risks of surgical and ambulatory treatments. The Rosetta package was used to carry out RST calculation.[21]

RESULTS

The Kaplan-Meier estimates for the three treatments along with their cumulative hazard function are given in Fig.1. From the figures we note that the estimates are similar indicating that the modified Hongkong surgery seems to be not contributing much to the reduction of response time. Using rough set theory, the reduct set was formed. The Prognostic index was calculated using six variables namely: surgery, age, site of lesion, number of vertebra involved, total loss of vertebra and adherence to allocated treatment. The choice of the attributes help us to identify the interval of values of PI which corresponds to the group of patients with different survival pattern. Some examples of the rules used are listed below:

Rule1: If (surgery no & level 1 & vertebra< 3) =>PI<0.1

Rule2: If (surgery no & level 2 & vertebra ≥3) => 0.1<PI<0.6

Rule3: If (surgery yes & vertebra ≥3) => PI>0.6

There are 21 rules satisfying the requirements mentioned above. Each rule is labelled with probability of favourable response conditioned on other variables in the reduct. A two dimensional statistics related to surgery and response time. The probabilities of favourable response under ambulatory and surgical treatments over its indiscernibility class were considered. The KM curves and corresponding cumulative hazard function for the three groups from the Cox PH model are given in Fig1. The KM curves based on PI derived from Cox's PH model and objects matching the decision rules are given in Fig.2.

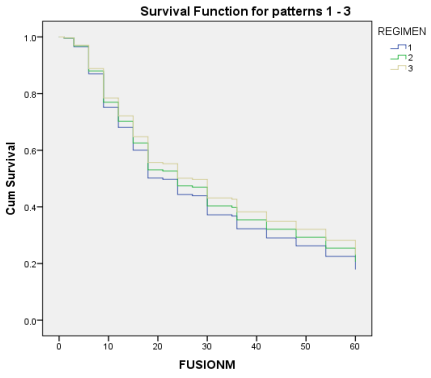
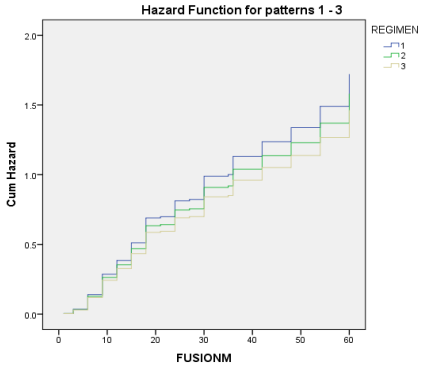


Fig1. (a): Kaplan-Meier for Regimens



(b): Cumulative Hazard Function for Regimens

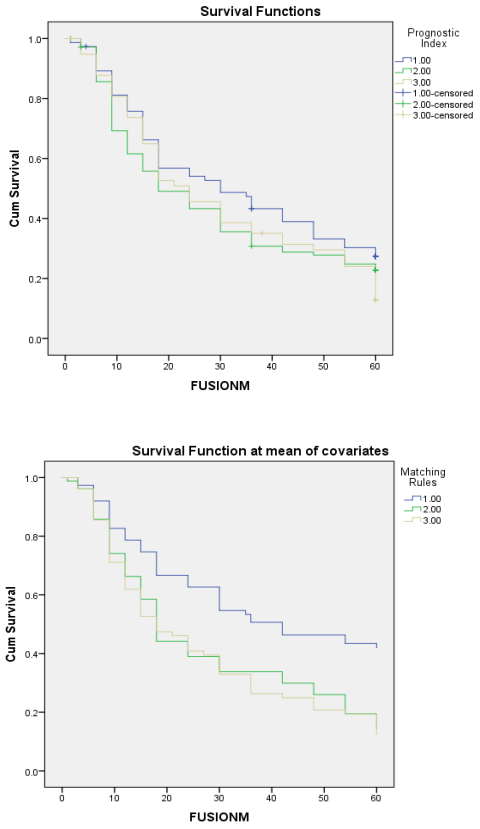


Fig2: Results from the Cox's proportional hazard model: (a) Kaplan-Meier estimate three grouped defined by the values of PI; (b) Kaplan-Meier estimate objects matching decision rules targeting these intervals.

The KM estimates are very close between the three treatments (Fig.1). The choice of the attributes proved to be proper as the plots were very similar. We can identify the intervals of values of PI which correspond to patient groups with different response patterns (Fig.2). We followed the pattern of Bazan et al.[9] The RST methodology provides a useful methodology to extract medical diagnosis and treatment rules to identify a group of patients who does not require costly invasive surgery to treat spinal tuberculosis patients and ambulatory treatment is cost-effective. These results are in confirmation with the main findings of the trial.

CONCLUSIONS

The response time under two ambulatory treatments are compared with the surgical treatment. The expected differences under the applications of the treatments with different response tendencies are compared. The Kaplan-Meier curve and the Cox models are combined with rough set framework for devising optimal decision rules. The rough set approach has many advantages over other approaches in handling imprecise and vague survival information, data reduction and framing decision rules that can be easily interpreted. Further work in needed in this direction to handle different types of censoring.

REFERENCES

1. Pawlak Z (1982) Rough sets. International Journal of Information and Computer Sciences. 11: 341-356.
2. Pawlak Z (1991) Rough sets-Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Boston.
3. Pawlak Z (2002) Rough sets and intelligent data analysis. Information Science, 147: 1-12.
4. Pawlak Z, Skowron A (2007) Rudiments of rough sets. Information Sciences, 177: 3-27.
5. Midelfart H, Komorowski J, Nørsett K, Yadetie F, Sandvik AK, Lægreid A (2004) Learning roughset classifiers from gene expression and clinical data. Fundamenta Informaticae, 2: 155-183.
6. Hvidsten TR, Wilczynski B, Kryshtafovych A, Tiuryn J, Komorowski J, Fidelis K (2005) Discovering regulatory binding-site modules using rule-based learning. Genome Research, 15: 856-866.
7. Hirano S, Tsumoto S (2005) Rough representation of a region of interest in medical images. International Journal of Approximate Reasoning, 40: 23-34.
8. Cai Z, Guan X, Shao P, Peng Q, Sun G (2003) A rough set theory based method for anomaly intrusion detection in computer network systems. Expert Systems, 20: 251-25.
9. Bazan J, Osmo´lski A, Skowron A, ´Sle¸zak D, Szczuka M and Wro´blewski J (2002) Rough Set Approach to the Survival Analysis. In J.J. Alpigini et al. (Eds.): RSCTC 2002, LNAI 2475, pp. 522–529, 2002. Springer-Verlag Berlin Heidelberg.
10. Pattaraintakorn P, Cercone N (2008) A foundation of rough set s theoretical and computational hybrid intelligent systems for survival analysis. Computers and Mathematics with Applications, 56: 1699-708.
11. Skowron A, Pawlak Z, Komorowski J, Polkowski L (2002) A rough set perspective on data and knowledge. In: W. Kloesgen, J. ˙Zytkow (eds), Handbook of KDD. Oxford University Press, pp. 134–149.
12. Lakshmi RA, Geetha PV Venkatesan P (2013) Rough Set theory based approach for attribute reduction. Indian Journal of Automation and Artificial Intelligence, 1: 70-80.
13. Anitha K, Venkatesan P. (2013) Feature extraction by rough set quick reduct algorithm. International

journal of innovative Research in Science Engineering and Technology, 2: 3989-93.
14. Anitha S, Venkatesan P (2015) Biological datamining-A novel attributable reduction based on rough set theory. International Journal of Applied Engineering Research, 10: 180-82.
15. Tsumoto S (2000) Automated Discovery of positive and negative knowledge in medical databases. IEEE Engineering in Medicine and Biology, 19: 56-62.
16. Tsumoto S (2004) Mining diagnostic rules from clinical data bases using rough sets and medical diagnostic model. Information Sciences, 162: 65-80.
17. Kronek LP, Reddy A. (2008) Logical analysis of survival data: Prognostic survival models by detecting high degree interactions in right censored data. Bioinformatics, 24: 248-53.
18. Sikora M, Wrobel L, Mielcarek M, Kalwak K (2013) Application of rule induction to discover survival factors of patients after bone-marrow transplantation. Journal of Medical Informatics and Technologies, 22: 35-53.
19. Elisa LT, John WW. (2003) Statistical methods for survival analysis. John Wiley NY.
20. ICMR/BMRC Working Party Study (1989) A controlled trial of short-course regimens of Chemotherapy in patients receiving ambulatory treatment or undergoing radical surgery for tuberculosis of the spine. Indian Journal of Tuberculosis, 36(suppl): 1-21.
21. http://www.lcb.uu.se/tools/rosetta/.