Computers and Mathematics with Applications 56 (2008) 1699–1708

Contents lists available at ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

A foundation of rough sets theoretical and computational hybrid intelligent system for survival analysis

Puntip Pattaraintakorn [a,*], Nick Cercone [b]

[a] Department of Mathematics and Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand
[b] Faculty of Science and Engineering, York University, ON, Canada M3J 1P3

ARTICLE INFO

Article history:
Received 11 March 2008
Accepted 17 April 2008

Keywords:
Intelligent system
Rough sets
Survival analysis
Soft computing

ABSTRACT

What do we (not) know about the association between diabetes and survival time? Our study offers an alternative mathematical framework based on rough sets to analyze medical data and provide epidemiology survival analysis with risk factor diabetes. We experiment on three data sets: geriatric, melanoma and Primary Biliary Cirrhosis. A case study reports from 8547 geriatric Canadian patients at the Dalhousie Medical School. Notification status (dead or alive) is treated as the censor attribute and the time lived is treated as the survival time.

The analysis result illustrates diabetes is a very significant risk factor to survival time in our geriatric patients data. This paper offers both theoretical and practical guidelines in the construction of a rough sets hybrid intelligent system, for the analysis of real world data. Furthermore, we discuss the potential of rough sets, artificial neural networks (ANNs) and frailty index in predicting survival tendency.

1. Introduction

Survival analysis [1] is a branch of statistics that studies time-to-event data. Death or failure is called an *event* in the survival analysis literature. Survival analysis attempts to answer questions such as: is diabetes a significant risk factor for geriatric patients? What is the fraction of patients who will survive past a certain time? Survival analysis is called *reliability analysis* in engineering, and *duration analysis* in economics. Presently, survival data in existence worldwide highlights the need for further comprehensive and systematic analysis to improve overall health outcomes. Much data analysis research has been conducted in several areas [2–5]. The aim of such data analysis techniques is to use the collected data for training in a learning process, and then to extract a hidden pattern by model construction. However, a successful technique involves far more than selecting a learning algorithm and running it over data sets. Successful data analysis requires in-depth knowledge of data. The challenges in real world problems are the complexity and unique properties of the survival data at hand. In many practical situations, survival data sets are vague and come with redundant and irrelevant attributes. The inclusion of these attributes in the data causes some difficulties in discovering the knowledge. To avoid these troubles, it is essential to precede the learning task with an attribute selection process to delete redundancy records, uncertainty attributes and overwhelming data. To this end, we create an attribute subset large enough to include all of the important attributes, but small enough for our learning system to handle easily.

Another issue in survival data analysis is the desire for automatic analysis processes [1]. Classical approaches are designed theoretically, automation is then increasingly challenging. Traditional data analysis is not adequate (e.g., Dempster–Shafer theory, grade of membership [7]), and methods for efficient mathematical and computer-based analysis, e.g., *rough sets*, are

* Corresponding author. Tel.: +011 662 326 4339; fax: +011 662 326 4354.
E-mail address: kppuntip@kmitl.ac.th (P. Pattaraintakorn).

0898-1221/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2008.04.030

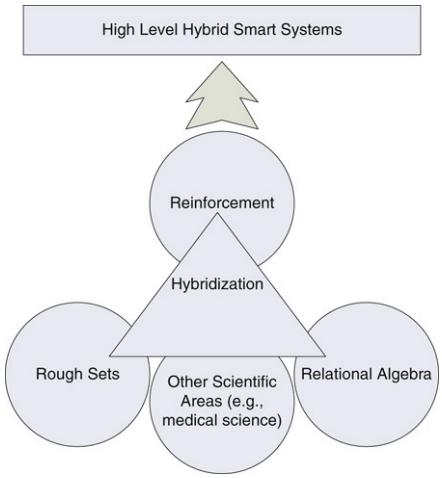

Fig. 1. A perspective of how to build a high level rough sets hybrid smart system.

indispensable. Rough set theory was developed by Zdzislaw Pawlak [6–10]. It provides system designers with the ability to compute with imperfect data. If a concept cannot be defined in a given knowledge base (*vagueness*), rough sets can approximate that knowledge efficiently. While logic is *deductive* and hardly applies to real situations, rough sets is in the form of *inductive reasoning* that widens the scope of the research to deal with real world data [8]. Rough sets do not require a specific model that can fit the data to be used in the analysis process. This ability provides flexibility in real situations. Rough sets provide a semi-automatic approach to data analysis and can be combined with other complementary techniques. Thus, current research tends to hybridize diverse methods of soft computing [4]. In this paper, we offer an approach based on rough sets with the capability to reason and to distil useful knowledge for survival data (e.g., risk factor, survival prediction model).

This article is organized as follows. We introduce in Section 2 preliminaries of rough sets, relational algebra and other scientific areas along with hybridization of these approaches. In Section 3, we propose our rough sets hybrid intelligent system and new CDispro algorithm. We demonstrate the applicability of our system by several experiments over a range of data sets reported in Section 4. In Section 5, the evaluation results are presented and also a brief comparison to another case studies in Sections 6 and 7. We conclude in Section 8.

2. Preliminaries and notations

2.1. The role of soft computing

One data analysis technique can generate very accurate results for one data set and poor results for another data set. Moreover, each technique has underlying advantages and disadvantages. The amount of real world data requires such techniques to have tractable time complexity, and simultaneously provide satisfactory outcome. Research in *soft computing* has demonstrated successes. Soft computing works synergistically with other data analysis methods to provide flexible analytical tools in real situations. Medsker [4] stated that soft computing differs from traditional computing in that it is tolerant of imprecision, uncertainty and partial truth. This guiding principle of soft computing can be used to achieve tractability, robustness and low cost solutions.

Rough sets is a leading soft computing approach. Works on hybrid rough sets based approaches have been conducted in [8–14] and in our previous studies to relational algebra [15,16], to flow graphs [17], to Cox proportional hazard model [18] and to medical applications [19,20]. However, the new generation of such research needs to understand the problem and to increase the intelligence of the system. This new generation of research can fulfill this objective by combining several related research areas. We introduce the new perspective of hybrid rough sets based approach (Fig. 1). The components we integrate into our hybrid intelligent system are rough sets, relational algebra and other scientific areas. Afterwards, the reinforcement step increases the intelligence of the system to a high level hybrid intelligent system, such as optimization approaches.

2.2. Rough sets

The rough sets philosophy relies on theoretical mathematics to extract significant attributes or rules from the data. In [22], the authors experimented on data sets from the UCI [23] and an actual cardiac care data set. The results of using rough sets are comparable with those obtained by using other systems under a wide variety of domains (c.f. [12]). Rough sets is advancing but the initial studies have focused on information retrieval and business tasks. Systematic developments for integrating rough sets to other scientific areas are at an initial stage. We recall the fundamental rough set theory from [7,12].

Definition 1. Let $K = (U, \mathbf{R})$ be a knowledge base. Given a finite set $U \neq \emptyset$, the universe of objects, any subset $X \subseteq U$ of the universe is called a *concept* in U.

Definition 2. Given a finite set $U \neq \emptyset$, the universe of objects and a concept $X \subseteq U$ of the universe we are interested in, any family of concepts (or category) in U is referred to as *knowledge* about U.

Let R be an equivalence relation over U, define U/R as the family of all equivalence classes of R and let $[x]_R$ denote a concept in R containing an element $x \in U$.

Definition 3. Given $K = (U, \mathbf{R})$ if $\mathbf{P} \subseteq \mathbf{R}$ and $\mathbf{P} \neq \emptyset$, then there is an equivalence relation $IND(\mathbf{P})$ called the *indiscernibility relation* over $\mathbf{P}$.

2.3. Other scientific areas: a medical science example

In this study, we give several case studies that for survival analysis focus on scientific medical data. We provide a brief example of how to hybridize this scientific area and our intelligent system. Many people have at least some of their medical information in an electronic medical database. It is essential to carefully analyze the data while considering domain knowledge. We provide an example of survival analysis, the time that patients are admitted to the study until the time to death as well as the time to particular events.

Example 1. In survival analysis, it often happens that the study does not span enough time in order to observe the event for all patients. Thus conclusions are difficult using traditional statistical models (e.g., multiple linear regressions). Moreover, if any patient leaves the study for any reason, censor variable is required. To properly address censoring, modeling techniques must take into account that for these patients the event does not occur during the follow-up period. Thus, the inclusion of domain knowledge is important to the analysis.

2.4. Relational algebra and hybridization

Traditional rough sets approaches in real applications are time-consuming, thus rendering rough sets less efficient for large scale data unless heuristics are included (c.f. [12]). One reason for this phenomenon is that the data resides in flat files most of the time. Thus, studies to reduce time complexity remain necessary. In [24], rough sets are redefined using relational algebra and elaborated in [25]. The computational time is improved and automatic analysis is achieved. Let us assume that a *decision table* is denoted by $T(U, C, D)$, where C is the set of *condition attributes* and D is (singleton set) the *target function*. *Card* and $\prod$ denote the Count and Projection operations. $Card(X)$ counts the number of elements in the set X. Unary operation $\prod(Y)$ is defined as the set of examples in the decision table which are restricted to the attribute set Y.

Definition 4. An attribute c_i is a core attribute if

$$Card\left(\prod(C - \{c_i\} + D)\right) \neq Card\left(\prod(C - \{c_i\})\right).$$

Definition 5. An attribute $c_i \in C$ is a *dispensable attribute* with respect to D if

$$Card\left(\prod(C - \{c_i\} + D)\right) = Card\left(\prod(C - \{c_i\})\right).$$

Definition 6. The *degree of dependency*, $K(R, D)$, between the attribute $R \subseteq C$ and attribute D in decision table $T(U, C, D)$ is

$$K(R, D) = \frac{Card\left(\prod(R + D)\right)}{Card\left(\prod(C + D)\right)}.$$

Definition 7. The subset of attributes $RED \subseteq C$ is a *reduct* of attributes C with respect to D if

$$K(RED, D) = K(C, D) \quad and \quad K(RED, D) \neq K(RED', D) \quad \forall RED' \subset RED.$$

Definition 8. The *merit* value of an attribute $c_i \in C$ is defined as

$$\text{merit}(\{c_i\}, C, D) = 1 - \frac{Card\left(\prod(C - \{c_i\} + D)\right)}{Card\left(\prod(C + D)\right)}.$$

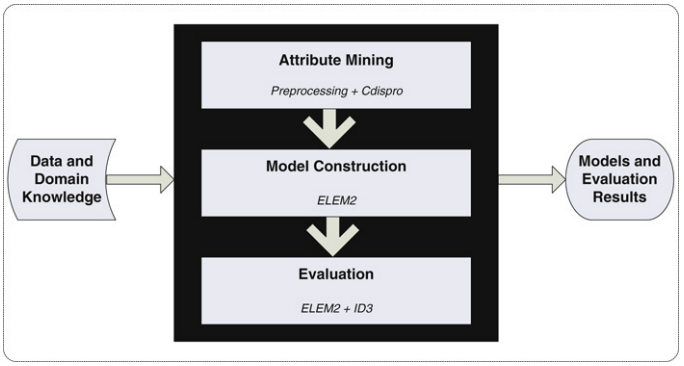

Fig. 2. Our proposed rough sets hybrid intelligent system architecture.

The core attributes are the necessary condition attributes, while the dispensable attributes are the unnecessary ones. Reducts are minimal condition attributes that are constructed from the core attribute. The merit is the tool to analyze the relationship between the condition and the decision attributes. The inclusion of domain knowledge in our hybrid intelligent system is the significant contribution in this study. Thus, we use two concepts from [15] as:

Definition 9. A *probe attribute* $P \in C$ corresponding to $T(U, C, D)$ is defined as an attribute of concern in $T(U, C, D)$ for each domain by an expert.

Definition 10. A *probe reduct* corresponding to decision table $T(U, C, D)$ is defined as a reduct which contains a probe attribute.

In a medical data table, each patient record represents an example. Each field is treated as an attribute. Each column contains patients' information (e.g., patients' symptoms, clinical information) which can be treated as the attribute values provided in an attribute set. Such a data set is treated as the training set. The known outcome (target function or decision) is an analysis goal, e.g., classification and prediction. The classification task can be to classify each example to a certain disease outcome, or to recommend a treatment regimen. The prediction task can be to predict a patient's survival time. The notion of *probe attribute* and *probe reducts* can be described with the following example.

Example 2. In survival analysis, {survival time} is the target function while {patient's symptoms}, {surgery type} and so on describe the condition attributes. If we want to know about the survival time for each patient, the risk of radical surgery or mild surgery becomes significant. Hence, we consider the {surgery type} attribute as a probe attribute. The probe reducts are the reducts constructed from the probe attribute.

3. Methodology

Our rough sets hybrid intelligent system architecture of rough sets, relational algebra and medical science is in Fig. 2. It consists of three modules: attribute mining, model construction and evaluation. Our system serves to build a high level rough sets hybrid intelligent system by the inclusion of domain knowledge in the analysis process. Real world survival data and domain knowledge are input in *Attribute Mining* and are analyzed with the preprocessing step and new CDispro algorithm. The outcome is the essential and informative attributes (significant risk factors). All of these acquired attributes will be input to *Model Construction*. In this module, ELEM2 (Version 3) [26] derives a rough sets model in the form of decision rules for survival prediction. These rules are passed on to the last module. Then the ELEM2 and ID3 [27] (WEKA software [28]) are used to validate and obtain the evaluation results to guarantee the correctness of the rules. The final output are rules and their evaluation results. Rough sets has led to many interesting extensions to data mining [8–10] and feature selection [29].

3.1. CDispro algorithm

In this section, we present the main algorithm in the first module of our rough sets hybrid intelligent system. CDispro stands for "**C**ore-**Dis**pensable Attributes and **Pro**be Reducts Extraction Algorithm". It was first proposed in [15] and is revised in this study. CDispro is able to discover essential information from a data set using core attribute and reducts/probe reducts in Definitions 4–10. It comprises two main steps. CDispro discovers core attributes in the first step and provides two kinds of reducts in the following step; (i) traditional reducts R, if no probe, and (ii) user defined probe reducts PR. CDispro takes domain knowledge into account by using an input probe attribute P. Users can identify a probe attribute to produce the probe reducts PR. The probe attribute is an attribute known to be important for a particular data set. Our previous study of probe reducts can be found in [16]. This revision of CDispro is also more efficient in computing reducts.

Skowron et al. [31] showed that the lower and upper approximations, positive regions, short reducts, etc. can be computed in a straightforward manner from the discernibility matrix with $O(kn^2)$ time complexity where n is the number of examples and k is the number of attributes of the data set, which is not feasible for large data sets. Nguyen et al. [24,32] proposed several algorithms that do not require storing the discernibility matrix in the calculation step. Their algorithm for generating *short reducts* by using Johnson strategy has $O(k^2 nlogn)$ time complexity. This algorithm is an efficient way to compute reducts without using a discernibility matrix. The computational time of CDispro is $O(n \log n)$ where $n = |\{C - Core\}|$. However, on average $|\{C - Core\}|$ is less than $n/2$.

Another key feature of CDispro is to analyze both the relationship among condition attributes and the relationship between condition attributes and its target function. In addition, CDispro combines core finding and reducts generation in the same algorithm which differs, but improves on similar studies [25,30].

Algorithm 1 New Core-Dispensable attributes and probe reducts extraction algorithm.

INPUT : A decision table: $T(U, C, D)$
A probe attribute: P
OUTPUT: Core attribute: *Core*
Dispensable attribute: *Dis*
Probe reducts/Reducts: *PR*/*R*

Set Core = ∅, Dis = ∅, PR/R = ∅.
//Construct Core, Dispensable attributes and Reducts or Probe reducts
For each attribute $c_i \in C$ {
if $Card(\prod(C - c_i + D)) < Card(\prod(C - c_i))$ **then**
  $Dis = Dis \cup c_i$
else if $Card(\prod(C - c_i + D)) > Card(\prod(C - c_i))$ **then**
  $Core = Core \cup c_i$
end if
if $P = \emptyset$ **then**
  $R = C$ and $Reducts = R$
else
  PR = C and Core = Core ∪ Probe and Reducts = PR
end if
//CDispro generates probe reducts if the user enters a predefined probe attribute. Otherwise, traditional reducts will be generated
}
For each $c_j \in \{C - Core\}$ {
//Measure the merit of each condition attribute and compare it to the other condition attributes to generate reducts or probe reducts
Find $merit(c_j, C, D)$
Sort c_j in decreasing order of $merit(c_j, C, D)$
}
Set $V = \{C - Core\}$
while K(R,D) $\neq 1$ **do**
  Select largest c_j by merit in the list
  Reducts = Reducts ∪ c_j
  $V = V - c_j$
end while

4. Experiments

4.1. Data and materials

We applied our rough sets hybrid intelligent system to the survival analysis data sets in Table 1. Table 2 shows description of geriatric data from Dalhousie Medical School (Canada) collected during 2002–2003.[1] We consider this geriatric data as two independent data sets followed by two separate target functions i.e. $geriatric_{nStatus}$ and $geriatric_{sTime}$, respectively. $geriatric_{nStatus}$ contains 8547 patient records with *notification status* as the target function. The objective is to develop a model to predict the *notification status* for new patient. $Geriatric_{nStatus}$ describes each patient with 44 condition attributes, e.g., *age at investigation*, *Parkinson's disease*. $geriatric_{sTime}$ has the target function *survival time* (in months). $Geriatric_{sTime}$ has

[1] Collection of personal data creates privacy issues. Stronger privacy assurance anonymously requires specific model, and is outside the scope of this work.

Table 1
Experimental data sets

Data sets	Number of condition attributes	Number of example
$geriatric_{nStatus}$	44	8547
$geriatric_{sTime}$	44	8546
melanoma [1]	7	30
PBC[a]	17	424

[a] Stands for Primary Biliary Cirrhosis collected from Mayo Clinic during 1974–1984 [23].

Table 2
The geriatric data description

Attribute	Description	Attribute	Description
edulevel	Education level	*hbp*	High blood pressure
eyesight	Eyesight	*heart*	Heart
hear	Hearing	*stroke*	Stroke
eat	Eat	*arthriti*	Arthritis or rheumatism
dress	Dress and undress yourself	*parkinso*	Parkinson's disease
takecare	Take care of your appearance	*eyetroub*	Eye trouble
walk	Walk	*eartroub*	Ear trouble
getbed	Get in and out of bed	*dental*	Dental
shower	Take a bath or shower	*chest*	Chest
bathroom	Go to the bathroom	*stomach*	Bladder
phoneuse	Use the telephone	*kidney*	Kidney
walkout	Get places out of walking distance	*bladder*	Stomach or digestive
shopping	Go shopping for groceries etc.	*bowels*	Bowels
meal	Prepare your own meals	*diabetes*	Diabetes
housew	Do your housework	*feet*	Feet
takemed	Take your own medicine	*nerves*	Nerves
money	Handle your own money	*skin*	Skin
health	Health	*fracture*	Fractures
trouble	Trouble	*age*	Age group
livealo	Live alone	*studyage*	Age at investigation
cough	Cough	*sex*	Gender
tired	Tired	*livedead*	Notification status
sneeze	Sneeze	*survivaltime*	Survival time

Table 3
Core attributes and reducts results generated from CDispro

Data sets	CDispro core attributes	CDispro reducts
$geriatric_{nStaus}$	edulevel hear housw health livealo eyetroub heart eartroub dental chest diabetes studyage sex hbp	edulevel hear housw health livealo eyetroub heart eartroub dental chest diabetes studyage sex hbp
$geriatric_{sTime}$	edulevel eyesight hear shower phoneuse meal shopping housew money tired sneeze trouble livealo cough sex arthriti eyetroub hbp heart bladder stroke dental stomach kidney age chest bowels diabetes feet nerves skin health fracture	edulevel eyesight hear shower phoneuse shopping housew money tired sneeze trouble livealo cough sex arthriti eyetroub hbp heart bladder stroke dental stomach kidney age chest bowels diabetes feet nerves skin health fracture
melanoma	age sex trt	age sex trt
PBC	none	none

notification status attribute (dead or alive) as the censor attribute (c.f. [16]). The purpose of the study is to develop a model to predict the *survival time* for each patient based on the training set, then cross-fold validate the model on the test set.

4.2. Attribute mining

In the preprocessing step, since data inconsistency is an issue in rough sets and affects discernibility of the data, we performed a data cleaning step to obtain consistent data. A study of computing with inconsistency can be found in [33]. Subsequently, the consistent data is discretized [16]. Next, all preprocessed data sets are analyzed with the CDispro algorithm.

The results in Table 3 illustrate the selection of core attributes and generation of reducts. {*diabetes*} is the core attribute for both $geriatric_{nStatus}$ and $geriatric_{sTime}$, which means that {*diabetes*} is the significant risk factor for notification status and survival time of our Canadians geriatric data. The previous studies, CDispro and ROSETTA [11], can be found in [15].

Our CDispro algorithm produces dispensable attributes, depicted in Table 4. The absence of these attributes does not decrease the predictive ability from the original data set. In the medical domain, the adoption of dispensable attributes can

Table 4
Dispensable attributes results generated from CDispro

Data sets	CDispro dispensable attributes
$geriatric_{nStatus}$	eyesight shopping trouble cough sneeze arthriti stomach bladder feet nerves skin fracture
$geriatric_{sTime}$	eartroub walk
melanoma	none
PBC	none

minimize the expensive series of laboratory tests or drop high risk treatments. Several factors, for example, {*eyesight*}, {*ear trouble*} are not significant risk factor for predicting notification status and survival time of the geriatric data.

4.3. Model construction

All acquired attributes from the first module are passed to the second module, *Model Construction*. ELEM2 generates decision rules in the form: "*If* C_1 *is* c_1 *and* C_2 *is* c_2 *then* D *is* d_1" where c_1, c_2 and d_1 are possible values corresponding to attribute C and D, respectively. This rule can be used to predict the outcome in new data such as survival time of new elderly patient. Among over 800 rules of geriatric data, example rules of $geriatric_{sTime}$ are:

Decision Rule 1: *IF* (*health*$>$ 0.25) *and* (*hear* $=$ 0) *and* (*nerves* $=$ 0) *and* (*feet* $=$ 0) *and* (*heart* $=$ 0) *and* (*dental* $=$ 0) *and* (*stomach* $=$ 0) *and* (*hbp* $=$ 0) *and* (*diabetes* $=$ 0) *and* (*age* $\leq$ 2) *THEN* (*survival time* $=$ 7–18 *months*)

Decision Rule 2: *IF* (*sex* $=$ 0) *and* (*edulevel* $=$ 2) *and* (*eyesight* $>$ 0) *and* (0 $<$ *health* $\leq$ 1) *and* (0 $<$ *hear* $\leq$ 0.25) *and* (*diabetes* $=$ 1) and (*tired* $=$ 0) *and* (*feet* $=$ 0) *THEN* (*survival time* $=$ 56–73 *months*).

The first medical diagnosis rule can be interpreted as:

- If the patient is unhealthy and
- has severe hearing damage and
- nerves problem and
- foot problem and
- heart disease and
- dental disease and
- stomach disease and
- high blood pressure and
- especially those who experience diabetes
- then the patient has a tendency of survival time around 7–18 months after being admitted to our study.

The second rule is interpreted as:

- If a female patient has a low education level and
- an eyesight problem from low to serious type and
- a health problem from low to serious type and
- can hear quite well and
- does not have diabetes experience and
- is easily tired and
- has foot problems
- then the patient is likely to have a survival time between 56–73 months.

As these rules show, the {*diabetes*} affects the survival time significantly. Our previous studies on univariate analysis have shown that {*diabetes*} is a very significant risk factor by using the Kaplan–Meier method and Log rank test [16]. From decision rule 1, we see that *combinations* of risk factors possibly affect the survival time. Thus, we perform multivariate analysis and {*diabetes, heart, trouble, getbed, walk, age, sex*} are the significant risk factors by using the Cox method [18]. Furthermore, these rules result in easy interpretation of survival prediction rules and can be read without prior expert knowledge. Next is an example rule from PBC:

Decision Rule 3: *IF* (age $>$ 2) *and* (*biliru* $\leq$ 3) *and* (*albumi*$>$ 3) *and* (*alkal*$>$ 2) *and* (*sgot*$>$ 1) *and* (*prothr* $>$ 3) *THEN* (*survival time* $=$ 1361–1781 *days*).

5. Evaluation

The improvements of all rule quality compared to rule constructions from entire data and from reducts/probe reducts are depicted in Fig. 3. Almost all rule quality outcomes are improved (except the average number of geriatric survival prediction rules). The rule quality generated from geriatric data is improved on average 24.47% for all outcomes. The rule qualities generated from melanoma and PBC data are improved on average by 28.45% and 73.77% for both outcomes respectively. Further, the average number, length and running time of the rules is improved an average of 52.45%, 21.03% and 51.20% for all data sets respectively.

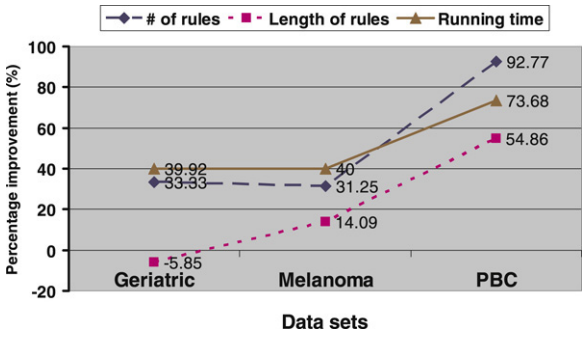

Fig. 3. Improved quality of the generated rules.

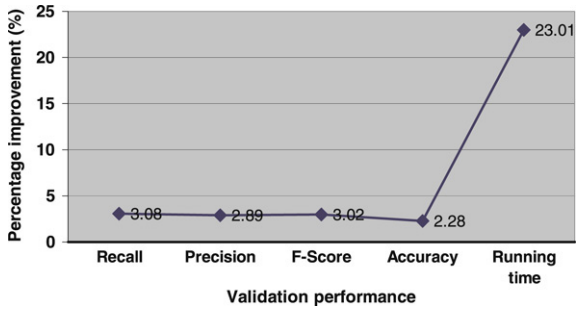

Fig. 4. Improved rule measurements from 10-fold cross validation by ID3.

After generating the survival prediction model with the previous paradigm, we illustrate the quality of the rules by using the optional validation process. We run 10-fold cross validation with ID3 to illustrate the utility of derived rules. We use rule quality measurements: recall, precision, F-score and accuracy to gauge the quality of rules. We then compare the improvement of rules generated from the entire data to those generated from reducts/probe reducts. Our validation process demonstrates the improvement for all measurements in Fig. 4. The validation results illustrate an average improvement of 2.28% while the running time improved on average 23.01%.

6. A case study: comparison to ANNs and frailty index

The same geriatric data was analyzed by other methods [34] to evaluate the potential of rough sets, artificial neural networks (ANNs) and the frailty index in predicting survival time. For the ANNs, randomly selected participants formed the training sample to derive relationships between the 40 variables and survival. An ANN's output was generated for each subject and a separate testing sample was used to evaluate the accuracy of prediction. An individual frailty index score was calculated as the proportion of deficits experienced (c.f. [34]). The output of the rough sets rules, an ANN's model and an unweighted frailty index in predicting survival patterns were measured using the accuracy rate. The accuracy rate of rough sets rules from validation was 83.79%–90.57% [16]. At the optimal receiver operating characteristic (ROC) value, the accuracy of the frailty index was 70.0%. The ANNs accuracy rate over 10 simulations in predicting the probability of individual survival was $79.2 \pm 0.8\%$.

This $geriatric_{sTime}$ data was analyzed from different points of view: (i) The unweighted frailty index captured the relationship of the $geriatric_{sTime}$ data successfully, (ii) ANNs are able to automatically discover the non-linear characteristics in this data, (iii) rough sets offer the capability to handle vagueness and illustrated its usefulness on this data. Due to vagueness, redundancy and irrelevant attributes in the data, we conclude that rough set theory and its discernibility relation can improve the analysis process efficiently and effectively.

7. A case study: Recommender system

Recommender rules were generated by using the same geriatric data [21]. Rule priority, recommendation score and rule-based expert systems can be used to construct recommend clinical examinations for patients. For example, the patients's group that has critical survival time (7–18 months) were selected for providing recommendations. Then, the recommended clinical examinations {*sneeze, high blood pressure, eye trouble, feet, nerves*} were recommended for any patients who trigger the rule: *IF* (*edulevel*! $= 2$ *or* 4) *and* ($0 <$ *shopping* < 0.5) *and* (*meal* ≤ 0) *and* (*trouble* ≥ 0) *and* (*livealo* ≥ 0) *and* (*sneeze* ≤ 0) *and* (*hbp* ≤ 0) *and* (*eyetroub* ≤ 0) *and* (*feet* ≤ 0) *and* (*nerves* ≤ 0) *and* (*sex* > 1) *THEN* (*survival time* $= 7$–18 *months*).

8. Concluding remarks and future works

Starting from mathematical rough set theory the central theme of this study is to invent the hybrid intelligent system. Our rough sets hybrid intelligent system is useful for survival analysis and extracting the most informative and useful knowledge. We created our system to have the following features. Our system was designed to provide comprehensive survival data analysis tasks; preprocessing, analyzing process and postprocessing. We amalgamated rough sets and other techniques in soft computing to be able to make the analyzing process tolerant to imprecise and uncertain data. We ensured the correctness of rules by designing automatic validation processes. Furthermore, the computation times were improved significantly by using database operations. The experimental results show how our rough sets hybrid intelligent system could be employed to quickly process. Clinical diagnosis questions can be answered successfully, e.g., is diabetes a significant factor for survival time of geriatric patients? Analysis results show that it has significant impact on the survival time of geriatric patients. Decision rules described particular tendencies for survival outcomes of patients by using decision rules that are straightforward and simple to use.

In the future, from theoretical viewpoint, we will pay more attention to many advances in rough sets e.g., rough mereology, rough inclusion, decision logic or dissimilarly analysis. Our results offer alternative choices to the patients or anyone concerned by the outcome of medical treatments or the progression of diseases. We have illustrated that pursuing further research in this relatively young area of mathematics, *rough set theory*, is a worthwhile aim.

Acknowledgement

This research was supported by the KMITL Research Fund, Thailand and NSERC, Canada. Thanks are also due to Arnold Mitnitski, Greg M. Zaverucha, Kanlaya Naruedomkul, Manas Sangworasil and the anonymous reviewers for their helpful comments.

References

[1] L.T. Elisa, W.W. John, Statistical Methods for Survival Data Analysis, 3rd ed., John Wiley and Sons, New York, 2003.
[2] I. Cohen, A. Garg, T.S. Huang, N. Sebe, Machine Learning in Computer Version, Springer, Berlin, 2005.
[3] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, 5th ed., Prentice Hall, New Jersey, 2002.
[4] M.R. Larry, Hybrid Intelligent System, Kluwer Academic Publishers, Boston, 1995.
[5] S.J. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 2nd ed., Prentice Hall, New York, 2002.
[6] P. Zdzislaw, Rough sets, Int. J. Inf. Comput. Sci. 11 (5) (1982) 341–356.
[7] P. Zdzislaw, Rough Sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.
[8] Z. Pawlak, A. Skowron, Rudiments of rough sets, Inform. Sciences 177 (1) (2007) 3–27.
[9] Z. Pawlak, A. Skowron, Rough sets: Some extensions, Inform. Sciences 177 (1) (2007) 28–40.
[10] Z. Pawlak, A. Skowron, Rough sets and Boolean reasoning, Inform. Sciences 177 (1) (2007) 41–73.
[11] O. Alexander, Discernibility and Rough Sets in Medicine: Tools and Applications, Dissertation, Norwegian University of Science and Technology, Norway, 1999.
[12] J. Komorowski, L. Polkowski, A. Skowron, Rough sets: A tutorial, in: S.K. Pal, A. Skowron (Eds.), Rough Fuzzy Hybridization: A New Trend in Decision-Making, Springer-Verlag, Berlin, 1998, pp. 3–98.
[13] J.F. Peters, et al. (Eds.), Transactions on Rough Sets VI: Journal Subline, in: Lect. Notes Comp. Sci., vol. 4374, Springer, Heidelberg, 2007.
[14] J.F. Peters, et al. (Eds.), Transactions on Rough Sets VII: Journal Subline, in: Lect. Notes Comp. Sci., vol. 4400, Springer, Heidelberg, 2007.
[15] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Hybrid intelligent systems: Selecting attributes for soft-computing analysis, in: Proceedings of the 29th Annual International Computer Software and Applications Conference, IEEE Computer Society, Edinburgh, Scotland, 2005, pp. 319–325.
[16] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Selecting attributes for soft–computing analysis in hybrid intelligent systems, in: D. Slezak, et al. (Eds.), Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Regina, Canada, in: Lect. Notes. Artif. Int., vol. 3642, Springer-Verlag, Berlin, 2005, pp. 698–708.
[17] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Rule learning: ordinal prediction based on rough set and soft-computing, Appl. Math. Lett. 19 (12) (2006) 1300–1307.
[18] P. Pattaraintakorn, N. Cercone, Hybrid rough sets-population based system, in: J.F. Peters (Ed.), Transactions on Rough Sets VII: Journal Subline, in: Lect. Notes Comp. Sci., vol. 4400, Springer, Heidelberg, 2007, pp. 190–205.
[19] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Hybrid rough sets intelligent system architecture for survival analysis, in: J.F. Peters (Ed.), Transactions on Rough Sets VII: Journal Subline, in: Lect. Notes Comp. Sci., vol. 4400, Springer, Heidelberg, 2007, pp. 206–224.
[20] P. Pattaraintakorn, N. Cercone, Integrating rough set theory and medical applications, Appl. Math. Lett. 21 (4) (2008) 400–403.
[21] P. Pattaraintakorn, G.M. Zaverucha, N. Cercone, Web based health recommender system using rough sets, survival analysis and rule-based expert systems, in: A. An, et al. (Eds.), Proceedings of the 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Toronto, Canada, in: Lect. Notes. Artif. Int., vol. 4482, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 491–499.
[22] T.E. McKee, T. Lensberg, Genetic programming and rough sets: A hybrid approach to bankruptcy classification, Eur. J. Oper. Res. 138 (2002) 436–451.
[23] C. Blake, E. Keogh, C. Merz, UCI Repository of Machine Learning Databases, http://www.ics.uci.edu, University of California, 2005.
[24] S.H. Nguyen, H.S. Nguyen, Some efficient algorithms for rough set methods, in: Proceedings of the Sixth International Conference on Information Processing and Management of Uncertainty Knowledge Based Systems, Granada, Spain, 1996, pp. 1451–1456.
[25] X. Hu, T.Y. Lin, J. Han, A new rough sets models based on database systems, Fund. Inform. 59 (2–3) (2004) 1–18.
[26] A. An, N. Cercone, ELEM2: A learning system for more accurate classifications, in: R.E. Mercer, E. Neufeld (Eds.), Advances in Artificial Intelligence (Proceedings of the 12th Biennial Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, Vancouver, BC, Canada), in: Lect. Notes. Artif. Int., vol. 1418, Springer-Verlag, Berlin, 1998, pp. 426–441.
[27] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[28] I. Witten, E. Frank, Practical Machine Learning Tools and Techniques with JAVA Implementations, Morgan Kaufmann, San Francisco, 2000.
[29] R. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recogn. Lett. 24 (6) (2003) 833–849.
[30] X. Hu, N. Cercone, Discovery of decision rules in relational databases: A rough set approach, in: Proceedings of the International Conference on Information and Knowledge Management, ACM, 1994, pp. 392–400.
[31] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, in: R. Slowinski (Ed.), Intelligent Decision Suppport - Handbook of Applications and Advances of the Rough Sets Theory, Kluwer, Dordrecht, 1992, pp. 331–362.

[32] H.S. Nguyen, Approximate Boolean reasoning approach to rough sets and data mining, in: D. Slezak, et al. (Eds.), Proceedings of the 10th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Regina, Canada, in: Lect. Notes. Artif. Int., vol. 3642, Springer-Verlag, Berlin, 2005, pp. 12–22.
[33] J. Bazan, A. Skowron, D. Slezak, J. Wroblewski, Searching for the complex decision reducts: The case study of the survival analysis, in: N. Zhong, et al. (Eds.), Proceedings of the International Syposium on Methodologies for Intelligent Systems, Maebashi City, Japan, in: Lect. Notes. Artif. Int., vol. 2871, Springer-Verlag, Berlin, 2003, pp. 160–168.
[34] X. Song, A. Mitnitski, C. MacKnight, K. Rockwood, Assessment of individual risk of death using self-report data: An artificial neural network compared to a frailty index, J. Am. Geriatr. Soc. 52 (2004) 1180–1184.