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

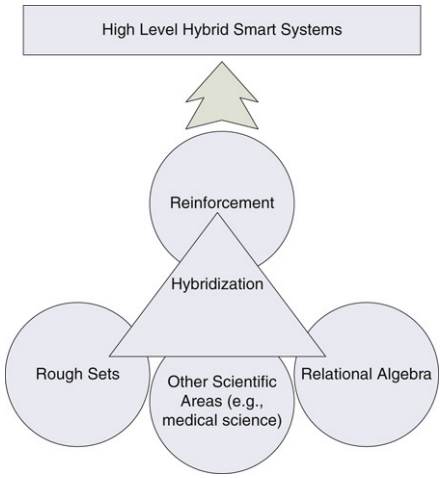

Fig. 1. A perspective of how to build a high level rough sets hybrid smart system.

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].