Expert Systems with Applications 39 (2012) 5784–5791

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

An algorithm designed for improving diagnostic efficiency by setting multi-cutoff values of multiple tumor markers

Qiang Su [a], Jinghua Shi [b], Ping Gu [c], Gang Huang [c,*], Yan Zhu [d]

[a] School of Economics & Management, Tongji University, Shanghai 200092, China
[b] Department of Industrial Engineering and Logistics Management, Shanghai Jiao Tong University, Shanghai 200240, China
[c] Department of Nuclear Medicine, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China
[d] School of Economics and Management, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Colorectal cancer (CRC)
Diagnostic efficiency (DE)
Tumor markers
Cutoff values
Rough set theory (RST)
Genetic algorithm (GA)

ABSTRACT

Currently, tumor markers have been effectively applied for colorectal cancer (CRC) diagnosis. In order to decrease the information loss caused by single cutoff value and improve diagnosis efficiency (DE), we explore the integrative application of multiple tumor markers with multiple cutoff values systematically by developing an optimization algorithm named MVMTM. The effectiveness of the MVMTM is experimentally studied based on a real medical dataset. With MVMTM, the united use of three tumor markers can enhance DE from 0.78 to 0.86. Furthermore, MVMTM has been proved to be better than other baseline machine learning algorithms significantly.

1. Introduction

Colorectal cancer (CRC) is one of the most common cancers whose mortality is ranked third in the world. Colonoscopy and fecal occult blood test (FOBT) are the frequently used screening measures for CRC. Although large-scale clinical trials demonstrate that colonoscopy is effective for the diagnosis of CRC, the comparatively high cost and the painful process seriously limit its application (Lewis, 2000). In terms of FOBT, due to the low diagnostic accuracy, it is not an effective method for CRC diagnosis. In this circumstance, a number of tumor markers have been set forth and utilized in CRC diagnosis. For example, in Renji Hospital, a teaching hospital in Shanghai, China, nine different tumor markers, viz. AFP, CEA, CA 19-9, CA 125, CA 153, CA 50, CA 724, CA 211 and CA 242 can be employed for cancer diagnosis. Among them, the most commonly used tumor markers for colorectal cancer diagnosis are CEA, CA 19-9 and CA 50.

Usually, single cutoff value is applied to separate the value range of the tumor marker into two segments as normal and abnormal. Some studies demonstrate that diagnostic result can varied widely along with the adjustment of the cutoff value (Körner, Söreide, Stokkeland, & Söreide, 2007). Up to now, many efforts have been dedicated to searching for an appropriate cutoff value so as to achieve the highest diagnosis efficiency (DE) (Armitage, Davidson, Tsikos, & Wood, 1984; Carriquiry & Pineyro, 1999; Wichmann, Lau-Werner, Muller, Stierber, & Schildberg, 2000). The commonly used approach is to numerate the possible cutoff values and choose the value at which the highest DE can be achieved (Wan & Zhang, 2007; Weiss, Niwas, Grizzle, & Piyathilake, 2004). For instance, Körner et al. suggested that the optimal cutoff value for CEA was 4 μg/L (Körner et al., 2007). Carpelan-Holmstrom et al. suggested a cutoff value of 5 μg/L for CEA and 20 U/ml for CA 242 (Carpelan-Holmstrom, Haglund, Kuusela, Jarvinen, & Roberts, 1995).

Some researchers found that the combination of different tumor markers can improve DE. Lucarotti et al. stated that the combination use of CEA, CA 19-9, and CA 50 could improve DE significantly in differentiating benign tumor from malignant tumor for the pancreatic cancer (Lucarotti et al., 1991). Moghimi and Ghodosi figured out that the combination use of CEA and CA 19-9 could achieve a higher DE compared with individual usage (Moghimi & Ghodosi, 2007). These findings imply that, instead of being used individually, the different markers should be applied simultaneously and the check results should be considered synthetically.

In almost all existing studies, only a single cutoff value is set for each tumor marker (Duffy, 2001; Moertel et al., 1993; Persijn & Hart, 1981; Wood, Ratcliffe, Burt, Malcolm, & Blumgart, 1980). In this way, some important information obtained from the diagnosis test will be neglected. Taking CEA as an example, its value can range from 0 to 550 μg/L. If its cutoff value is set to 4 μg/L, the test results of 5 μg/L and 500 μg/L will be simply judged as the same (abnormal) although they are very different from each other. From this point of view, it is too rough to simply separate the broad value range of a tumor marker into two segments of normal and abnormal. Hence, multi-cutoff values should be employed to take more advantages of the test result so as to improve DE.

* Corresponding author.
E-mail address: cherry_shi44@163.com (G. Huang).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.11.089

Under this consideration, aiming at improving DE for CRC, a novel diagnosis strategy with multiple tumor markers and multi-cutoff values are studied systematically. An optimization algorithm is designed for setting multi-cutoff values for multiple tumor markers (MVMTM). In this work, three tumor markers, i.e., CEA, CA 19-9, and CA 50, are used simultaneously. And no more than three cutoff values are permitted for each tumor marker. With the algorithm, the optimal cutoff values are calculated out based on a real diagnosis dataset with 124 cases. Furthermore, other 88 cases are employed to validate the effectiveness of the algorithm.

This paper is organized as follows. In Section 2, the evaluation method of DE is addressed and the algorithm related technologies including the rough set theory (RST) and the genetic algorithm (GA) are introduced. The details of the MVMTM are elaborated in Section 3. Then, in Section 4, the experimental study is conducted to demonstrate the effectiveness of the algorithm. Furthermore, some discussions are given in Section 5. Section 6 concludes the paper.

2. Related technologies

2.1. Measurement of diagnostic efficiency (DE) for cutoff values

Suppose a set of sample cases are tested, the diagnosis result for each case can be either normal or abnormal. However, sometimes, wrong decisions can be made in which the normal case is judged as abnormal or vice versa. To measure these errors, sensitivity (*Se*) and specificity (*Sp*) are proposed. Sensitivity (*Se*) represents the probability that an abnormal case is correctly judged as abnormal. Specificity (*Sp*) stands for the probability that a normal case is correctly diagnosed as normal.

Suppose there are seven possible cutoff values, namely, C_1, C_2, C_3, C_4, C_5, C_6, and C_7. For each cutoff value, the corresponding sensitivity (*Se*) and specificity (*Sp*) can be computed. Then, the coordinate point (1-*Sp*, *Se*) can be drawn on the coordinate plane. As shown in Fig. 1, when all the seven coordinate points are obtained, a curve entitled the receiver operating characteristic (ROC) (Wan & Zhang, 2007) can be derived. This curve can be used to assist clinicians in choosing the appropriate cutoff value according to the required levels of *Sp* and *Se*. To take a special example, in Fig. 1, C_1 would be chosen to be the cutoff value if clinicians prefer high *Sp* while do not care about *Se*. In other words, actually, cutoff value is chosen based on the preferred levels of *Sp* and *Se*. Similarly, in our research, multi-cutoff values should also be determined according to the requirements of *Sp* and *Se*. However, the requirements of *Sp* and *Se* are heavily related to the characteristics of the disease, diagnosis cost, and disease prevalence in the given area

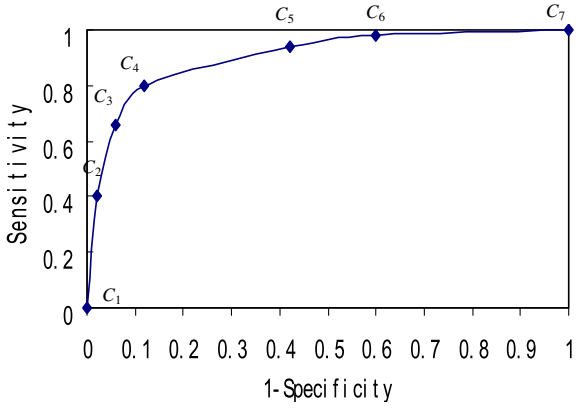

Fig. 1. ROC curve.

(Yu & Xu, 1998), etc. Due to these uncertainties, many researches assign the same weights to *Sp* and *Se* so that the point located at the upper left-hand corner in Fig. 1 is usually chosen as the cutoff values (Streiner & Cairney, 2007). Then, DE in Eq. (1), defined as the average of *Sp* and *Se*, is employed as the criterion for choosing cutoff values.

$$DE = \frac{Se + Sp}{2} \tag{1}$$

2.2. Rough set theory (RST)

The RST is a fairly new methodology developed for extracting useful knowledge from imprecise, uncertain, and vague information. Since being put forward by Pawlak in 1982, the RST has been successfully applied in many fields including machine learning, data mining, and intelligent data analyzing, etc.

RST is suitable for medical data analysis since it does not need any preliminary or additional requirements for the data (An et al., 2007). Wakulicz-Deja and Paszek tried to apply RST in diagnosis of Mitochondrial Encephalomyopathies (MEM) (Wakulicz-Deja & Peszek, 2003). With the capabilities of attribute reduction and decision rule generation, RST can improve the classification quality and diagnosis efficiency remarkably. In order to generate decision rules in medical diagnosis, Ilczuk, Mlynarski, Wakulicz-Deja, Drzewiecka, and Kargul (2005) constructed an algorithm entitled LEM2 based on RST. Fakih and Das (2006) stated that the RST was useful in extracting diagnostic information in the form of rules from the medical databases. In one word, RST offers a useful mechanism for analyzing and distilling essential attributes and rules from medical data (Pattaraintakorn and Cercone, 2008). These research results imply that RST can provide significant assistance for diagnosis problem encountered in this paper.

In RST, an information system can be described as $S = \langle U, A, V, f\rangle$. In which, U is a finite set of objects $\{x_1, x_2, \ldots, x_n\}$; A is a finite set of attributes $\{a_1, a_2, \ldots, a_m\}$; V is the value range of each attribute $\{v_{a_1}, v_{a_2} \ldots v_{a_m}\}$; $f : U \times A \to V$ is called an information function, and $f(x, a) \in v_a$, for $\forall a \in A, \forall x \in U$.

Generally, diagnosis can be regarded as a decision process including two types of attributes, i.e., condition attributes C and decision attributes D. Thus, the corresponding information system can be modeled as $S = \langle U, C \cup D, V, f\rangle$.

Let $P \subseteq C$, x_i, $x_j \in U$. We say that x_i and x_j are *indiscernible* if the following function is satisfied.

$$IND(P) = \{(x_i, x_j) \in U \times U | \forall a \in P, f(x_i, a) = f(x_j, a)\} \tag{2}$$

Indiscernibility is an imperative property in medical diagnosis. Due to the complexity of disease, it is likely that some persons with similar test results have totally different diagnosis results. These cases are called the indiscernible cases and can be properly resolved by RST.

Using RST, a set of *decision rules* can be derived referring to the information system $S = \langle U, C \cup D, V, f\rangle$. And these rules can be expressed as $c_1(x_i), \ldots, c_m(x_i) \to d_1(x_i), \ldots, d_n(x_i)$ or $C_x \to D_x$ where $\{c_1, \ldots c_m\} = C$ and $\{d_1, \ldots d_n\} = D$. To evaluate the effectiveness of a decision rule, three criteria, i.e., support, accuracy, and coverage, are defined as follows.

$$Sup_x(C, D) = \frac{card(C_x \cap D_x)}{card(U)} \tag{3}$$

$$Acc_x(C, D) = \frac{card(C_x \cap D_x)}{card(C_x)} \tag{4}$$

$$Cov_x(C, D) = \frac{card(C_x \cap D_x)}{card(D_x)} \tag{5}$$