Computers in Biology and Medicine 42 (2012) 213–221

Contents lists available at SciVerse ScienceDirect

Computers in Biology and Medicine

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

Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment

You-Shyang Chen [a,*], Ching-Hsue Cheng [b], Chien-Jung Lai [c], Cheng-Yi Hsu [d], Han-Jhou Syu [b]

[a] Department of Information Management, Hwa Hsia Institute of Technology, 111, Gong Jhuan Rd., Chung Ho District, New Taipei City 235, Taiwan
[b] Department of Information Management, National Yunlin University of Science and Technology, 123, Section 3, University Road, Touliu, Yunlin 640, Taiwan
[c] Department of Distribution Management, National Chin-Yi University of Technology, 35, Lane 215, Section 1, Chung-Shan Road, Taiping District, Taichung 411, Taiwan
[d] Lin's ENT Clinic 364, Fuxing Rd., Lugang Township, Changhua County 505, Taiwan

ARTICLE INFO

Article history:
Received 5 June 2011
Accepted 25 November 2011

Keywords:
Customer relationship management (CRM)
Target customer segment (TCS)
K-means clustering algorithm
Rough set theory (RST)
Recency–Frequency–Monetary (RFM) analysis model

ABSTRACT

Identifying patients in a Target Customer Segment (TCS) is important to determine the demand for, and to appropriately allocate resources for, health care services. The purpose of this study is to propose a two-stage clustering-classification model through (1) initially integrating the RFM attribute and *K*-means algorithm for clustering the TCS patients and (2) then integrating the global discretization method and the rough set theory for classifying hospitalized departments and optimizing health care services. To assess the performance of the proposed model, a dataset was used from a representative hospital (termed Hospital-A) that was extracted from a database from an empirical study in Taiwan comprised of 183,947 samples that were characterized by 44 attributes during 2008. The proposed model was compared with three techniques, Decision Tree, Naive Bayes, and Multilayer Perceptron, and the empirical results showed significant promise of its accuracy. The generated knowledge-based rules provide useful information to maximize resource utilization and support the development of a strategy for decision-making in hospitals. From the findings, 75 patients in the TCS, three hospital departments, and specific diagnostic items were discovered in the data for Hospital-A. A potential determinant for gender differences was found, and the age attribute was not significant to the hospital departments.

1. Introduction

Under the policy interventions of the Bureau of National Health Insurance (NHI) in Taiwan, the increasing competition in the health care industry, and the increasing requirements of health care quality, hospitals face many obstacles to strengthening their competitive edge. The growing prospects of information technology (IT) in this rapidly changing industry motivate the use and costly investments in IT, which offer a competitive advantage for meeting the opportunities and challenges in the health care industry. The means and methods to enhance a hospital's competitiveness by further increasing the accessibility and availability of unknown patient information is of great interest and highly valuable.

Recently, the rapid changes in IT have been accompanied by the issue of customer relationship management (CRM), as well as data mining techniques within various industries, which have become the focal point of challenges in marketing strategy and implementation. CRM is a broadly recognized, widely implemented strategy for managing the relationships between business and customers (including potential customers) to fulfill customer requirements and enhance the relationships with customers for improved business [1]. The effective and efficient utilization of IT is a method for implementing a successful CRM strategy to understand customer needs, satisfy customers demands, mine for valuable customer information, identify target and potential customers, realize the maximum customer value, increase customer loyalty, and finally maximize profits [2]. In particular, identifying target customer segments (TCS) in the population is a key process in determining high-value patients to offer best service for them when facing the challenges in hospitals. The ability to target patients for marketing is the specified focus of the hospital's operations. Then, hospitals can accordingly develop a set of strategies to differentiate their offerings and services to this core group of target patients and attract the desired target segment. By specifically targeting and tailoring services for these patients, hospitals can more effectively achieve their goal of high quality health service. Understanding more about high-value patients, including the hospital departments and services that they utilize, will help hospitals appropriately design resource allocation strategies for health care services.

TCS is the process of dividing the population into groups based on their characteristics and similar preferences for specified

* Corresponding author. Tel.: +886 2 8941 5143; fax: +886 2 8941 5142.
E-mail address: ys_chen@cc.hwh.edu.tw (Y.-S. Chen).

0010-4825/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2011.11.010

services. As for specific characteristics, customer value analysis offers an effective method to define customers and discover hidden patient information. A well-known customer value analysis method called the Recency–Frequency–Monetary (RFM) analysis model is used to represent customer behaviors [3]. The RFM model has been applied to various industries, such as the education industry [4], retail industry [5], and service industry [6]; however, it is rarely used in the context of the health care industry, based on our research. This study attempts to fill this knowledge gap.

Over the past decade, statistical methods have been applied to the forecasting-related challenges of the health care industry to assist in improving CRM and obtaining useful results; however, the lack of sufficient information is perceived as the key obstacle to more extensive valuation. This lack of sufficient information is related to the following: (1) the lack of decision rules to explain the experimental data, (2) methods need to obey mathematical distributions for the experimental data, and (3) the lack of methods for identifying determinant attributes. It is therefore recommended that more efficient classifiers be employed as forecasting tools based on artificial intelligence (AI) techniques, such as the rough set theory (RST) [7–9]. The RST has the flexibility to adequately perform in a variety of application areas, including finance, manufacturing, health care, and the service industry, particularly for classification models [10]. To reconcile the current shortcomings in classifications in the health care industry, this study aims to develop a two-stage clustering-classification model that integrates a RFM-based *K*-means clustering algorithm to (1) cluster TCS groups and the global discretization-based RST LEM2 algorithm and (2) classify hospital departments and extract decision rules. Concurrently, it aims to further assess the quality of the RFM-based rough sets classifier in the health care industry. The clustering-classifying combination model is seldom used in the health care industry. Furthermore, the fourth author has extensive experience in working in the health care industry as a physician (medical doctor), spanning over 18 years, and has relevant experiential knowledge. The overall objective of this study is to minimize the set of hidden TCS patients and construct a solid relationship between patients and hospitals.

This paper is organized in the following manner: Section 2 describes the related studies for clustering target customer segment patients and classifying hospitalized departments; Section 3 presents the proposed model and algorithm; Section 4 presents an empirical case study with the experimental results and findings; and Section 5 summarizes the conclusions and offers recommendations for future research.

2. Related works

This section reviews the relevant literature on the theoretical components of the proposed model, including customer relationship management, customer value analysis, target customer segments, the Recency–Frequency–Monetary analysis model, the *K*-means clustering algorithm, the rough set theory, and the LEM2 rule extraction method.

2.1. Customer relationship management

CRM, devoted to improving relationships with customers, focuses on a comprehensive approach to integrate customer values, requirements, expectations, and behaviors by analyzing data from customer transactions [11]. Thus, an excellent CRM can help businesses keep existing customers, attract new ones, and maintain customer equities. Kalakota and Robinson [12] explained that CRM integrates the functions of the fields related to customers, including marketing, sales, services, and technical support, and it utilizes IT to help a business manage relationships with customers in a systematic way, thus improving customer loyalty and increasing overall business profits.

It has been estimated that it costs five times as much to attract a new customer as it does to retain an existing one, according to research by the American Management Association [13], and this relationship is particularly obvious in the service sector [14]. Therefore, instead of attracting new customers, companies would like to maximize business operations in order to keep existing customers and build a long-term customer relationship.

2.2. Customer value analysis and target customer segment

Customer value analysis is an analytical method for determining customers' consuming behaviors and allows for the further analysis and data abstraction of specific customer data, such as who the target customers are and whose contributions are more significant. Kaymak [3] noted that a RFM model is one of the well-known customer value analysis methods for TCS, and the advantage of the RFM model is that it can extract customers' characteristics by using fewer criteria (a three-dimensional variable) to cluster attributes, thereby reducing the complexity of the model of customer value analysis. Moreover, Schijns and Schroder [15] noted that the RFM model is a commonly used method to measure the strength of a customer relationship.

TCS is a population subgroup that shares similar preferences or activities for products or services. In practice, different groups of customers can be segmented by their consumption behaviors via RFM attributes or other characteristics to develop an in-depth understanding of a certain consumer segment, identify a high-yield TCS, and analyze this group's specific needs [16]. By using this method to describe the consumers' characteristics and needs, the cluster standards for customer values and TCS are not subjective, but rather defined objectively based only on the three RFM attributes.

2.3. The Recency–frequency–monetary analysis model

The Recency–Frequency–Monetary analysis model proposed by Hughes [17] is commonly used by many industries, including the manufacturing, retail, and service industries. The RFM analysis model is a marketing technique used to differentiate important customers within large databases and to determine which customers to target by examining how recently a customer has purchased an item (recency), how often they purchase items (frequency), and how much the customer spends (monetary). It is based on the marketing axiom – the 80/20 rule – that 80% of the sales come from just 20% of the customers; therefore, response rates greatly increase when marketing is targeted to the highest-yielding existing customer groups. The detailed definitions for the RFM model are described as follows: (1) Recency (R) refers to the interval between the last consuming behavior and the present. The smaller the *R*-value, the more important this customer is. (2) Frequency (F) refers to the number of transactions in a particular period, for example, two times in one year or two times in one month. The larger the *F*-value, the more important this customer is. (3) Monetary (M) refers to the consumption amount in a particular period. The larger the *M*-value, the more important this customer is.

According to the literature [18], the RFM method is very effective for clustering the TCS. There are two types of studies of the RFM model. Hughes [17] considered that the three variables were equal in their importance; however, Stone [19]

contradicted this and indicated that the three variables have different weights that are dependent on the specific industry.

2.4. K-means clustering algorithm

Clustering [20] is the process of grouping a set of objects or people that are characteristically similar to one another [21]. *K*-means clustering is a method of cluster analysis that aims to partition observations into K clusters, in which each observation belongs to the cluster with the nearest mean. The *K*-means clustering algorithm was originally known as Forgy's method [22] and was first used by MacQueen [23]; it has been used extensively in various fields, including data mining, statistical data analysis, and other business applications. The computational processes for *K*-means clustering are presented systematically, as follows:

Step 1: Partition the items into *K* initial clusters. First, partition the items (*m* objects) into *K* initial clusters that are randomly selected from the dataset. *K* clusters are created by associating every observation with the nearest mean.
Step 2: Proceed through the list of items. Assign an item to the cluster with the closest centroid (distance is computed by using the Euclidean distance with either standardized or un-standardized observations) and re-calculate the centroid for the cluster after adding or losing an item. The centroids for each of the K clusters become the new means.
Step 3: Repeat Step 2 until reassigning is completed.

Rather than starting with all items divided into *K* preliminary groups in Step 1, initial *K* centroids (seed points) are specified before proceeding to Step 2. The final assignment of items to clusters will be, to some extent, dependent upon the initial partition or the initial selection of seed points. Experience suggests that most of the major changes in the assignments will occur with the first reallocation step.

2.5. Rough set theory

RST, first proposed by Pawlak [24], employs mathematical modeling to handle the challenges associated with class data classification and has become a useful tool for decision support systems, especially when hybrid data, vague concepts, and uncertain data were involved in the decision-making process. To use the rough set process, one begins with a relational database, a table of objects with attributes, and attribute values for each object. One attribute is chosen to be the decision attribute, and then the remaining attributes are the condition attributes [24]. Rough sets (RS) address the problem of incomplete, vague, and uncertain data [25] by applying the concept of equivalence to partition the instances according to specified criteria. Two partitions are formed in the mining process. The members of the partition can be formally described by unary set-theoretic operators or by successor functions for lower and upper approximations, from which both possible rules can be easily derived. The RST approach is based on refusing certain set boundaries, implying that every set will be roughly defined using a lower and upper approximation [26].

Let $B \subseteq A$ and $X \subseteq U$ be an information system (or called an attribute-value system) that is a basic knowledge representation framework comprising a information table with columns 'attributes' and rows 'objects,' where $\boldsymbol{A}$ is a non-empty, finite set of attributes, $\boldsymbol{B}$ is a reduced set of attributes, $\boldsymbol{U}$ is a non-empty set of finite objects (i.e., the universe), and $\boldsymbol{X}$ is subset of objects in the approximation space. For $B \subseteq A$, there is an associated equivalence relation denoted $\boldsymbol{IND}(\boldsymbol{B})$ that is called a $\boldsymbol{B}$-indiscernibility relation and the equivalence classes of the $\boldsymbol{B}$-indiscernibility relation are denoted $[x]_B$. The set $\boldsymbol{X}$ is approximated using information contained in $\boldsymbol{B}$ by constructing lower and upper approximation sets:

$$\underline{B}X = \{x \mid [x]_B \subseteq X\} \quad \text{(lower approximation), and}$$
$$\overline{B}X = \{x \mid [x]_B \cap X \neq \emptyset\} \quad \text{(upper approximation),}$$

where x is an object in the universe $\boldsymbol{U}$. The $\boldsymbol{B}$-lower approximation is the union of all equivalence classes in $[x]_B$, which are contained by the target set. The objects in $\underline{B}X$ can be classified as positive members of $\boldsymbol{X}$ by the knowledge in $\boldsymbol{B}$. However, the $\boldsymbol{B}$-upper approximation is the union of all equivalence classes in $[x]_B$, which have non-empty intersection with the target set. The objects in $\overline{B}X$ can be classified as possible members of $\boldsymbol{X}$ by the knowledge in $\boldsymbol{B}$. The set $BN_B(x) = \overline{B}X - \underline{B}X$ is called the $\boldsymbol{B}$-boundary region of $\boldsymbol{X}$, and it consists of those objects that cannot be classified with certainty as members of $\boldsymbol{X}$ with the knowledge in $\boldsymbol{B}$. The set $\boldsymbol{X}$ is called 'rough' (or 'roughly definable') with respect to the knowledge in $\boldsymbol{B}$ if the boundary region is not empty. Rough sets theoretic classifiers apply the concept of rough sets to reduce the number of attributes in a decision table and to extract valid data from inconsistent decision tables [27,28]. Specifically, rough sets also accept discretized (symbolic) input. The details of rough sets can be referenced by the studies of Pawlak [26,27].

2.6. The rule extraction—rough sets LEM2 algorithm

In RST, decision rules are frequently induced from a given decision table that is subsequently transformed into a minimal set of rules. Rough set rule induction algorithms were first implemented in a Learning from Examples based on Rough Sets (LERS) [29] system. The learning system LERS induces a set of rules from examples and classifies new examples using the set of rules previously induced. A local covering is induced by exploring the search space of blocks of attribute-value pairs, which are then converted into the rule set. The Learning from Examples Module, version 2 (LEM2) [30] algorithm works correctly for symbolic attributes and is a part of the LERS data mining system [26]. The LEM2 algorithm is based on calculating a single local covering for each concept from a decision table to generate the decision rules. The large numbers of rules limit the classification capabilities of the rule-set, as some rules are redundant or considered 'poor quality.' Rule-filtering algorithms [31] are thus used to reduce the number of rules. For example, a filtering rule solution may be based on the computed quality indices of rules in a rule-set.

3. Methodology

To develop operating strategies, the health care industry must first understand a patient's characteristics and needs based on patient's consumption behaviors, and then ineffective strategies can be avoided, saving money and time. Thus, a specific TCS becomes the focus for hospital operations, and the hospital attempts to cater to these patients' needs, thus meeting their various demands, solving their problems, increasing their satisfaction, and enhancing their loyalty. Therefore, a methodology is needed to determine the TCS.

3.1. Concept and research framework of the proposed model

The rapid changes within the health care industry in Taiwan have been accompanied by rising consumer demands for better health care and many associated challenges in hospital operations. Implementing an appropriate marketing strategy in the health care industry is in line with the current trend to improve CRM, achieve the goal of high quality health care, and provide

complete, pertinent health care information throughout this customer-centered service. An essential part of the health care industry is directed toward the promotion and awareness of the health care management 'brand' among the TCS. Therefore, determining the TCS will be a prominent component of marketing strategies and will identify target and potential customers in this extremely competitive market. An efficient CRM strategy can be obtained through a RFM analysis model of the TCS, identifying the needs of customers and enhancing the strength of the relationship between the customers and the company. Practically, some shortcomings do exist in the statistical methods used to analyze CRM issues, namely in the assumption of the statistical methods of linear separability, multivariate normality, and independence of the predictive variables [32]; alternative methods have emerged based on AI techniques.

In the literature [33–35], this study notes that an important objective of this research is to build a rule-based AI model that can construct different model representations to explain the dataset and provide reasonable and powerful explanations for interested parties. Additionally, many earlier studies revealed that in comparing the efficiency of these classifiers, they nearly all concluded that the model performances were highly dependent on the context and the data used [36]. Therefore, it is of interest to find more reliable tools applied in the health care industry to address these classification problems [37,38].

RS has become an important tool in AI algorithms and is used to induce decision rules. Therefore, this study aims to propose a rule-based model based on the RS classifier to provide meaningful decision rules as knowledge-based systems [39] from an intelligent perspective, and it offers an alternate method for forecasting credit ratings in the Asian banking industry. The strengths of this study, based on RST, are as follows [40]: (1) it does not require preliminary or additional parameters to describe the data; (2) it is compatible with missing values, as it switches among different reducts and requires little time to generate rules; (3) it can handle large amounts of quantitative and qualitative data; (4) it yields easily understandable decision rules supported by a set of real examples; (5) it models highly nonlinear or discontinuous functional relationships and is a powerful method for characterizing complex and multidimensional patterns; and (6) it discovers important facts hidden in the data and expresses them in the natural language of decision rules. However, one drawback of traditional rough sets is that data must be discretized first to improve the classification accuracy [41]. Thus, the global method of RST is valuable when data mining for discretized continuous attributes and can be employed in this study. Nevertheless, another drawback is that many redundant rules are generated; a rule filter should be employed to combat this limitation. Based on the reasons mentioned above, this study proposes a hybrid two-stage clustering-classification model, combining experiential knowledge, RFM attributes, *K*-means clustering algorithms, global discretization method, RST, and rule filters for assessing the quality of this RFM-based RST AI model in the health care industry. Fig. 1 illustrates the flowchart of the proposed model.

3.2. Algorithms of the proposed model

The algorithms of the proposed model, along with its computational processes for determining the TCS, are outlined below:

Step 1: Data selection and collection
First, based on professional knowledge of the authors, select the dataset that includes the hospital data of patients from the health care industry to be the experimental data and set some preconditions. For example, specify the collection period. Accordingly, collect the practical data from a specific hospital in Taiwan.
Step 2: Data preprocessing
Pre-processing the dataset is needed to make the knowledge discovery process easier. The data of hospitalized patients may be noisy. Thus, delete the records that include inaccurate values and eliminate the redundant attributes that are not used in the study. Accordingly, transform the datum into an EXCEL file that will be more effectively processed by experimental operations.
Step 3: First stage—*K*-means clustering technique
This step clusters TCSs using the *K*-means clustering algorithm. The first stage is divided into five sub-steps, and its detail processes are presented step-by-step:
Step 3-1: Define the three RFM attributes
R represents the differences (i.e., number of days) between the date for latest hospital record and a specific date; *F* represents the total hospitalizations (number) in one

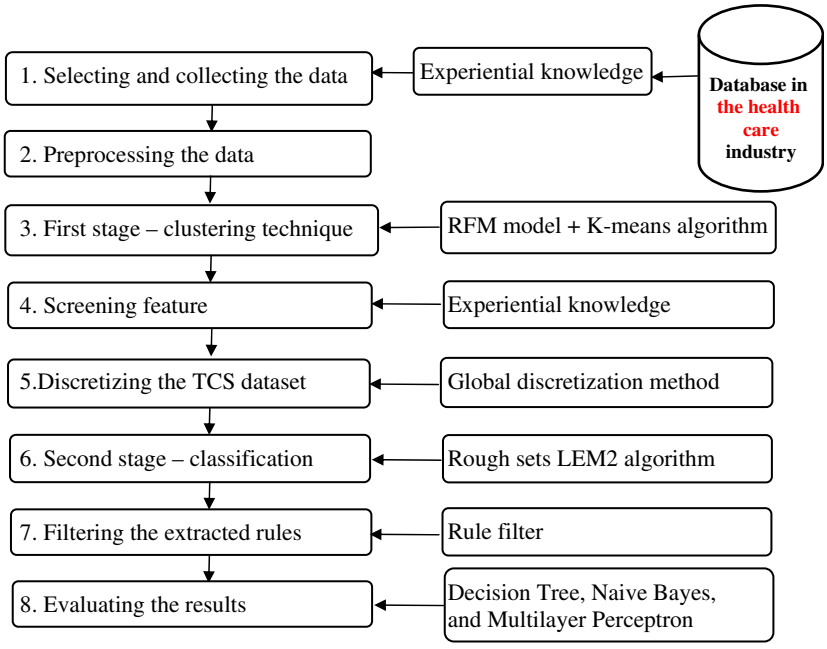

Fig. 1. The flowchart of the proposed model.

year; and M represents average hospital expense in the interval. Then, calculate the quantitative value of the three RFM attributes as input data for clustering TCS via the K-means clustering algorithm.

Step 3-2: Normalize the RFM attributes

Because the unit of measurement for the three RFM attributes is inconsistent, it is necessary to normalize the data. All attributes are standardized to the interval [0,1]—a mean of 0 and variance of 1. The equations for normalization of the RFM attributes are presented below:

$$\text{Recency} : 1-\frac{R_i}{Max\{R_i\}} \tag{1}$$

$$\text{Frequency} : \frac{F_i}{Max\{F_i\}} \tag{2}$$

$$\text{Monetary} : \frac{M_i}{Max\{M_i\}} \tag{3}$$

Step 3-3: Cluster TCS and estimate the number of clusters

With the normalized value of the RFM attributes, partition data (m objects) into K clusters using the K-means clustering algorithm for clustering TCS. To obtain better results, different K (i.e., $k=1, 2, 3\ldots$) values are calculated. The ideal number for K is determined by Eq. (4):

$$Max_K\{Min[C_k(d_{ij})]\},\ k=\text{cluster number} \tag{4}$$

where $1 \le k \le m$, and m denotes all the objects; i denotes the cluster, j denotes the element in cluster, d_{ij} denotes the distance for jth element of C_i, C_i denotes the ith cluster, and $1 \le i \le k$. The Min is related to shortest pair-wise inter-cluster distance, and the Max refers to the maximum number of the cluster.

Step 3-4: Adjust and determine the number of clusters

According to Eq. (4), the best suitable number for K is determined. If it is difficult to determine the TCS, the value of K will be adjusted until the TCS is found.

Step 3-5: Select the TCS customers to target

Determine the TCS with high added-value customers.

Step 4: Screen the data

After the TCS is determined, the condition attributes and decision attributes are chosen based on the relevant experiential knowledge of the authors.

Step 5: Discretize the continuous data using the global discretization method

For improving the classification accuracy and solving the challenge of generating a large number of decision rules in the traditional rough sets model, attributes are granulated first. The global discretization method based on the Boolean reasoning approach [42] is implemented to discretize all continuous attributes.

Step 6: Second stage—the RST LEM2 classification technique

Through the discretization of the condition attributes and the decision attribute, the rough sets LEM2 algorithm is used to generate the decision rules set from the experimental dataset for classifying hospital departments. The experimental dataset is randomly divided into two groups with 66% of the data serving as a training set and the remaining 34% serving as a testing set. The generated rules set supports the 'if…then' rules set as the knowledge-based system to determine the decision-making strategy and intelligently offer explanatory power.

Step 7: Filter the rules to improve quality

Since more rules complicate the prediction, this step implements a rule-filtering process for which rules below the support threshold are eliminated to improve the rule quality.

Step 8: Evaluate the results

To verify the experimental results, the experiments using the Decision Tree [43], Naive Bayes [10], Multilayer Perceptron [44], and the proposed model are repeated ten times with the 66/34% random split using different sampling data, and then the average accuracy is calculated. Evaluate their performance and explore reasons for their differences.

4. Verification and comparisons

This section introduces an empirical case that is used to verify the classification performance of the proposed model and compare the proposed model with other AI techniques for assessing the quality of this hybrid rough set classifier.

4.1. The introduction of Hospital-A case

To identify the TCS, the process is conducted using a hospital that will be referred to as Hospital-A. Hospital-A was founded in Yunlin County in 1999, and it was accredited as a district teaching hospital by the TJCHA (Taiwan Joint Commission on Hospital Accreditation) in Taiwan, indicating that this medical institute is a superior site compared to the other local hospitals in the Yunlin area in terms of its facility, capability, instrumentation, technology, staff, skills, and medical quality. Thus, Hospital-A has become one of the most important and most trusted hospitals in the Yunlin and Changhua areas, which include Hsilo, Erhlun, and Lunpei, and has experienced rising consumer demand for better health care and access to the best health services available in the country.

To serve the public, Hospital-A provides emergency, outpatient, and inpatient services. The facility has a total of 233 beds for patients and has 28 available medical specialties. Hospital-A treats all types of illness and offers over 20 clinical specialty and subspecialty departments, including division of general internal medicine, surgery, obstetrics and gynecology, pediatrics, orthopedic surgery, urology, family medicine, neurology, neurosurgery, physical medicine and rehabilitation, emergency medicine, cardiovascular surgery, and other sub-divisions of specialties. In this case, the marketing strategy and its implementation, along with the challenges in operation management, are dedicated to establishing Hospital-A's brand and to maximizing and achieving its goal of providing high-quality health services. To effectively establish its brand, it is necessary to provide the pertinent health information and patient-centered services for patients and to improve the quality of services, the efficiency of treatment processes, and the detailed and accurate information to the customers. The primary focus of Hospital-A is directed toward improving its brand and identifying potential customers in the target segments. Therefore, it is ideal for Hospital-A to determine the TCS that will be its most important customers and to create more benefits for these customers.

4.2. Computational processes using the Hospital-A dataset

The computational processes of the proposed model using the Hospital-A dataset are expressed systematically, as follows:

Step 1: Data selection and collection

Initially, select the database of hospital patient records in a division of a specific educational hospital (Hospital-A in Taiwan) as the experimental dataset. Collect the data for all the related hospital departments of patients during 1/1/2008-12/31/2008. Consequently, a total of 44 attributes are characterized and based on the data format of the Bureau of NHI

system in Taiwan, for example, "YYMM of payment", "medical care type", "hospital departments", "date", and "date of birth".

Step 2: Data pre-processing

Following data pre-processing, the Hospital-A dataset contained 183,947 records that belonged to 32,804 patients. The average hospitalization frequency in one year for a patient was 5.6, and the average hospitalization expense for a patient was NT$ (New Taiwan Dollar) 1310.

Step 3: First stage—the *K*-means clustering technique

Three RFM attributes are defined as input data in the *K*-means clustering algorithm to group the TCS patients. The detailed processes of the clustering technique are divided into five sub-steps, as follows:

Step 3-1: Define the three RFM attributes

First, *R* is the difference between the latest hospitalization date and December 31, 2008 for a specific patient. Second, *F* represents the total number of times of hospitalization for a specific patient in 2008. Third, *M* represents the average hospitalized expense in 2008. Therefore, the quantitative values of the three RFM attributes are calculated. Table 1 shows their information. To protect patient privacy, the last three ID numbers are replaced by 'XXX.'

Step 3-2: Normalize the RFM attributes

The RFM attributes are then normalized to intervals [0,1] by Eqs. (1)–(3). The normalized results are listed in Table 2.

Step 3-3: Cluster the TCS and estimate the number of clusters

The normalized values of the RFM attributes in Table 2 are partitioned accordingly into *K* clusters. To find the suitable number of *K* clusters, different values of *K* are calculated. For practical means and purposes, the *K* should be within three and nine. For $K=3$, the cluster distance is outlined in Table 3. The general purpose of clustering attempts to maximize the inter-cluster distance (to ensure the clusters are well separated) while minimizing the intra-cluster distance (to ensure compactness of the clusters). Therefore, the value 0.315 defines a shorter distance between Cluster 1 and Cluster 3 in Table 3. The shorter cluster distance for different *K* is further processed and listed in Table 4. The shortest distance used to find the number of *K* clusters is determined by Eq. (4), and thus *K* is three.

Step 3-4: Adjust and determine the number of clusters

Although Table 4 shows that the shortest cluster distance for *K* is three, the TCS patients cannot be determined. The value of *K* should be adjusted to be closer to the shortest cluster distance; thus, the best suitable number for *K* is shifted to four. Consequently, the clustering results are shown in Table 5.

Step 3-5: Select the highest-rated TCS patients

Table 5 shows that Cluster 4 is superior to the other three clusters not only in the number of objects, but also in the monetary attribute. Thus, the TCS patients that would be of high benefit to Hospital-A are found.

Step 4: Feature screening

After determining the TCS for Hospital-A, select conditional and decisional attributes via experiential knowledge. Accordingly, the determined TCS dataset included a total of 75 patients and was characterized by the following 10 attributes: (i) age, (ii) gender, (iii) max-monetary, (iv) recency, (v) frequency, (vi) surgery, (vii) chronic, (viii) grave, (ix) CT&MRI, and (x) department (class). The age, max-monetary, recency, and frequency attributes were continuous data, and the gender, surgery, chronic, grave, CT&MRI, and department attributes were categorical data. The max-monetary, CT&MRI, and department attributes represented the maximum monetary in a time, computer tomography (CT) and magnetic resonance imaging (MRI), and the hospital department, respectively. The CT&MRI is a family-owned and -operated outpatient diagnostic facility used to help diagnose certain spinal disorders. The first nine items are conditional attributes, and the last item, department, is a decisional attribute that is granulated into eight classes. Table 6 shows all attribute information in the TCS dataset.

Step 5: Discretize continuous data using the global discretization method

All continuous attributes in the TCS dataset are granulated using the global discretization method. Table 7 lists the granulated results. For example, the 'age' attribute is transformed into five linguistic values: L_1 (very low), L_2 (low),

Table 1
Information of the RFM attributes in the Hospital-A dataset.

ID	Recency	Frequency	Monetary
P200517XXX	348	4	1181
P100804XXX	13	13	1463
⋮	⋮	⋮	⋮
P121315XXX	13	2	715

Note: ID indicates the identification number.

Table 2
Information of the RFM attributes following normalization.

ID	Recency	Frequency	Monetary
P200517XXX	0.047	0.025	0.033
P100804XXX	0.964	0.080	0.041
⋮	⋮	⋮	⋮
P121315XXX	0.964	0.012	0.020

Table 3
Cluster distance between two clusters for $K=3$.

Cluster	Cluster 1	Cluster 2	Cluster 3
Cluster 1	–	0.695	**0.315**
Cluster 2	0.695	–	0.381
Cluster 3	**0.315**	0.381	–

Table 4
The shortest cluster distance for different *K*.

Number of clusters	The shortest cluster distance
3 Clusters	**0.315**
4 Clusters	**0.307**
5 Clusters	0.159
6 Clusters	0.157
7 Clusters	0.212
8 Clusters	0.183
9 Clusters	0.144

Table 5
Clustering results for 4 clusters.

	Cluster 1	Cluster 2	Cluster 3	Cluster **4**
Number of objects	12,884	10,428	9417	**75**
Recency	0.914	0.222	0.610	0.641
Frequency	0.063	0.013	0.020	0.015
Monetary	0.035	0.030	0.029	**0.440**

L_3 (medium), L_4 (high), and L_5 (very high), based on six cutoff points. For L_1, its value is between 1.0 and 12.0; that is L_1 = [1.0, 12.0).

Step 6: Second stage—the RST LEM2 classification technique

The experimental dataset is randomly divided into two groups, with 66% in the training dataset and the remaining 34% in the testing dataset. Then, the decision rules are extracted using the rough sets LEM2 algorithm from the TCS dataset to classify hospital departments.

Step 7: Filter the rules to improve quality

Generating excessive numbers of rules is not ideal. A rule-filter algorithm is implemented to eliminate support thresholds lower than 2. Consequently, Table 8 lists the extracted decision rules after rule-filter processing. The decision rule 1: "IF (max-monetary = L_4) & (grave = N) ⇒ class = Uro." It indicates when max-monetary ∈ L_4 (between 22,901.5 and 30,120.0) and grave = N (not a catastrophic illness patient) occur simultaneously, then the class (hospital department) is the Division of Urology. A total of 26 real examples support this rule.

Step 8: Evaluate the results

The experiments were repeated ten times with the 66/34% random split (namely, using different sampling data but once again with 66% of the data in the training dataset and 34% in the testing dataset), using four different methods: Decision Tree, Naive Bayes, Multilayer Perceptron, and the proposed model. The average accuracy and standard deviations were calculated. Table 9 shows the performance evaluation results in accuracy with its standard deviation based on ten tests using different methods in the TCS dataset. The proposed model (96.92%) significantly outperformed the other listing AI techniques in accuracy.

4.3. Findings

The analytic results of the TCS dataset yield six interesting findings, based on Tables 5, 8, and 9. They are described below:

(1) *The TCS patients are determined:* Cluster 4 is a special group of patients for Hospital-A. Cluster 4 has the least number of 75 (75/32,804 = 0.23%) patients, while it has the highest monetary rating, 0.440. An implication for this extracted phenomenon is that greater contributions can be obtained with fewer targeted efforts. Furthermore, the experimental results prove that the proposed model demonstrates superiority to other listing models.

(2) *The three hospitalized departments are mined:* The three departments with the most hospitalizations were discovered: the Division of Urology, Division of Cardiovascular Surgery, and Division of Emergency Medicine. This fact shows that most revenues for the hospital arise from three hospital departments, and it implies that Hospital-A should focus their resources on improving these health care services. This

Table 6
Attribute information in the example TCS dataset.

No.	Attributes name	Attribute type	Number of values	Note
1	Age	Numeric	Continuous	Min: 1.0 and Max: 88.0
2	Gender	Symbolic	2	F = Female and M = Male
3	Max-monetary	Numeric	Continuous	Min: 8722.0 and Max: 54,422.0
4	Recency	Numeric	Continuous	Min: 0.0 and Max: 361.0
5	Frequency	Numeric	Continuous	Min: 1.0 and Max: 15.0
6	Surgery	Symbolic	2	Y: Yes and N: No
7	Chronic	Symbolic	2	Y: Yes and N: No
8	Grave	Symbolic	2	Y: Yes and N: No
9	CT&MRI	Symbolic	2	Y: Yes and N: No
10	Department (Class)	Symbolic	8	Emg: Emergency, Neu: Neuro, Cas: Cardio-Surgery, Fam: Family, Uro: Urology, Gas: Gastro, ENT: Ear-Nose-Throat, and Nep: Nephro

Table 9
Comparison results of different methods in the TCS dataset.

Method	Accuracy (%)	Standard deviation (%)
Decision Tree-C4.5	86.69	1.79
Naive Bayes	79.26	10.77
Multilayer Perceptron	86.68	2.64
The proposed model	96.92	3.89

Table 7
Linguistic values of continuous attributes in the example TCS dataset.

Attribute	Cutoff point	Linguistic value
Age	[1.0, 12.0, 21.0, 45.0, 62.0, 88.0]	L_1, L_2, L_3, L_4, L_5
Max-monetary	[8722.0, 9200.0, 10,543.0, 22,901.5, 30,120.0, 32,472.5, 48,575.0, 54,422.0]	L_1, L_2, L_3, L_4, L_5, L_6, L_7
Recency	[0.0, 10.0, 25.0, 49.0, 150.0, 361.0]	L_1, L_2, L_3, L_4, L_5
Frequency	[1.0, 1.5, 7.0, 10.5, 15.0]	L_1, L_2, L_3, L_4

Table 8
Decision rules set extracted by the LEM2 algorithm in the TCS dataset.

No	Decision rule	Support
1	IF (Max-monetary = L_4) & (Grave = N) ⇒ Class = Uro	26
2	IF (Surgery = Y) & (Chronic = N) & (CT&MRI = N) & (Frequency = L_1) & (Grave = Y) ⇒ Class = Cas	9
3	IF (Surgery = Y) & (Chronic = N) & (CT&MRI = N) & (Gender = F) & (Max-monetary = L_6) ⇒ Class = Cas	6
4	IF (Surgery = N) & (Chronic = N) & (Grave = N) & (Frequency = L_1) & (CT&MRI = N) & (Gender = F) ⇒ Class = Emg	6
5	IF (Surgery = N) & (Chronic = N) & (Grave = N) & (Frequency = L_1) & (CT&MRI = N) & (Recency = L_5) & (Gender = M) & (Max-monetary = L_3) ⇒ Class = Emg	5
6	IF (Surgery = N) & (Chronic = N) & (Grave = N) & (Frequency = L_1) & (CT&MRI = N) & (Recency = L_5) & (Gender = M) & (Max-monetary = L_2) ⇒ Class = Emg	5
7	IF (Surgery = N) & (Chronic = N) & (Grave = N) & (Frequency = L_1) & (CT&MRI = N) & (Recency = L_5) & (Gender = M) & (Max-monetary = L_1) ⇒ Class = Emg	4

analytical result matches the professional knowledge of the fourth author.

(3) *The specific diagnostic items are discovered:* ICD codes, international classification of diseases, provide a classification system for assigning codes to diagnoses and procedures associated with diseases. Thus, ICD codes are important to hospitals and physicians for the treatment of patients' diseases and charging the Bureau of NHI in Taiwan for the hospital expenses. Every related health problem is assigned to a unique category and given a code up to six characters long; the code is comprised of three characters to the left of a decimal point and one or two digits to the right of the decimal point. The ICD-9 (9th Edition) was published by the WHO (World Health Organization) in 1977, which has been adopted by the NHI of Taiwan and is found on patient paperwork, including hospital records, physician records, and death certificates. Based on the ICD-9 records collected from the hospital data of TCS patients, the specific diagnostic items are statistically used to explore the root causes for high hospitalization expenses, as follows:
   - *For the Division of Urology:* The most common causes (about 88.5%) of disease are renal colic, urolithiasis, nephrolithiasis, and ureteral stone. For treatment, they all undergo extracorporeal shock wave lithotripsy (ESWL). The charge for ESWL is approximately NT$29,033 for the first occasion and NT$22,983 for the second occasion within 30 day. These high costs explain why the Division of Urology has one of the highest hospital expenses. This fact matches that of the decision rule 1.
   - *For the Division of Cardiovascular Surgery:* All patients in the Division of Cardiovascular Surgery have chronic renal failure in the TCS dataset, and the Bureau of NHI sees chronic renal failure as a serious illness. The view of critical illness matches that of the decision rules 2–3. Thus, there is a need to perform surgery on these patients, which results in the hospitalization expenses of NT$30,000–50,000. Moreover, the treatments of cardiovascular surgery include vascular exploration, insertion of a cannula for hemodialysis or other purposes (vein-to-vein), insertion of arterio-venous cannula (external Scribner type), creation of arterio-venous cannula shunt with Gore-Tex graft, and repair and anastomosis of peripheral vessels.
   - *For the Division of Emergency Medicine:* Many illnesses occur in this division because of unfortunate incidents, injuries, or other events that happen unexpectedly and unintentionally, such as traffic accidents and occupational accidents. Although the symptoms of diseases in the Division of Emergency Medicine are inconspicuous and diversiform, an important fact is found when exploring the decision rules. When the five attributes, 'surgery', 'chronic', 'grave', 'frequency', and 'CT&MRI', do not all appear in the decision rules, the case belongs to Division of Emergency Medicine. The treatments performed by cardiovascular surgery include endotracheal intubation, fracture reduction, operation of traumatic injury, and procession of serious wounds. Hospitalization expenses for these patients are approximately NT$9000–25,000 and exceed the average values of other divisions. These interesting facts provide Hospital-A with information for the resource allocation process.

(4) *Two prevalent diseases are noted:* Two prevalent diseases are noted from the analysis of Hospital-A. The first disease is urolithiasis, which occurred in about 85% of patients in the Division of Urology. The second is chronic renal failure, which occurred in nearly 100% of patients in the Division of Cardiovascular Surgery. These diseases should be further examined in subsequent research.

(5) *A potential determinant is found:* From the analytical results, the attribute 'gender' is a potential determinant of the hospital department in the TCS dataset. This information could be explained by the fact that males and females have varying concerns about different diseases and desires for improving their health [45]. The role of gender in many diseases has been reported by numerous earlier studies [46,47].

(6) *A redundant attribute is defined:* This problem is noteworthy of Hospital-A. The 'age' attribute is not significant to the hospital department in Hospital-A, and it is never used to influence the hospital department, based on Table 8. Clearly, it is a redundant attribute and thus can be removed from the TCS dataset. A reasonable explanation for this phenomenon is that larger hospitals are not found in Yunlin County, thus various age groups seek treatment at Hospital-A.

5. Conclusions

There is a need for the hospitals to explore patient contributions to assist in the appropriate allocation of health care resources to patients. The motivation for this study is rooted in the support of efforts to prevent health care squandering, to properly allocate suitable resources for hospitals, and to enhance the quality of health care to ensure that all patients are entitled to a high quality of life (QOL) [48,49]; thus, it is imperative that hospitals construct an intelligent method for determining their TCS.

This study has presented a hybrid clustering-classification approach to address the classification problems encountered by hospital management. From the empirical results (Table 9), the proposed model demonstrated better accuracy than other methods, and the generated decision rules describe the types of patients that contribute more revenue to specific departments in a hospital. As for the 'managerial' contribution, 75 TCS patients, the hospitalization rates of the Divisions of Urology, Cardiovascular Surgery, and Emergency Medicine, specific diagnostic items, and the gender and age characteristics were discovered in this dataset. The proposed model is a useful tool to help Hospital-A objectively focus on the highest-yield TCS patients in order to establish excellent relationships and improve consumer satisfaction. Future research should address the various weight attributes for the three RFM clustering variables and their relationships and effects on the example case.

Conflict of interest statement

No conflicts of interest reported.

Acknowledgments

The authors would like to thank the Editor-in-Chief, associate editor, and unknown referees for their useful comments and suggestions, which were very helpful in improving this study.

References

[1] B. Thompson, D. Sims, CRM improving demand chain intelligence for competitive advantage, Bus. Week 3804 (2002) 75–82.
[2] Y.G. Joo, S.Y. Sohn, Structural equation model for effective CRM of digital content industry, Expert Syst. Appl. 34 (1) (2008) 63–71.

[3] U. Kaymak, Fuzzy target selection using RFM variables, in: Proceedings of the IFSA World Congress and 20th NAFIPS International Conference, vol. 2, 2001, pp. 1038–1043.
[4] D. Bin, S. Peiji, Z. Dan, Data mining for needy students identify based on improved RFM model: a case study of university, in: Proceedings of the International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII), vol. 1, 2008, pp. 244–247.
[5] Y.L. Chen, M.H. Kuo, S.Y. Wu, K. Tang, Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data, Electron. Commer. Res. Appl. 8 (5) (2009) 241–251.
[6] J. Ranjan, R. Agarwal, Application of segmentation in customer relationship management: a data mining perspective, Int. J. Electron. Customer Relat. Manage. 3 (4) (2009) 402–414.
[7] Y.S. Chen, C.H. Cheng, Forecasting PGR of the financial industry using a rough sets classifier based on attribute-granularity, Knowl. Inf. Syst. 25 (1) (2010) 57–79.
[8] T. Herawan, M.M. Deris, J.H. Abawajy, A rough set approach for selecting clustering attribute, Knowl.-Based Syst. 23 (3) (2010) 220–231.
[9] K. Kaneiwa, Y. Kudo, A sequential pattern mining algorithm using rough set theory, Int. J. Approx. Reason. 52 (6) (2011) 881–893.
[10] M.H. Dunham, Data Mining: Introductory and Advanced Topics, New Jersey: Prentice Hall, Upper Saddle River, 2003.
[11] J. Peppard, Customer relationship management (CRM) in financial services, Eur. Manage. J. 18 (3) (2000) 312–327.
[12] R. Kalakota, M. Robinson, e-Business Roadmap for Success, 1st ed., Addison Wesley Longman Inc., New York, USA, 1999, pp. 109–134.
[13] P. Kotler, Marketing Management: Analysis, Planning, Implementation, and Control, Prentice-Hall, New Jersey, 1994.
[14] C.T. Ennew, M.R. Binks, The impact of service quality and service characteristics on customer retention: small businesses and their banks in the UK, Br. J. Manage. 7 (3) (1996) 219–230.
[15] J.M.C. Schijns, G.J. Schroder, Segment selection by relationship strength, J. Direct Mark. 10 (3) (1996) 69–79.
[16] R.T. Rust, V.A. Zeithaml, K.N. Lemon, Customer-centered brand management, Harv. Bus. Rev. (2004) 1–9.
[17] A.M. Hughes, Strategic Database Marketing, Probus Publishing Company, Chicago, 1994.
[18] F. Newell, The New Rules of Marketing: How to Use One-To-One Relationship Marketing to be the Leader in Your Industry, McGraw-Hills Companies, New York, 1997.
[19] B. Stone, Successful Direct Marketing Methods, NTC Business Books, Lincolnwood, IL, 1995, pp. 37–57.
[20] G. Kerr, H.J. Ruskin, M. Crane, P. Doolan, Techniques for clustering gene expression data, Comput. Biol. Med. 38 (3) (2008) 283–293.
[21] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2001.
[22] E. Forgy, Cluster analysis of multivariate data: efficiency versus interpreability of classifications, Biomertrics 21 (1965) 768–780.
[23] J.B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1967, pp. 281–297.
[24] Z. Pawlak, Rough sets, Inf. J. Comput. Inf. Sci. 11 (5) (1982) 341–356.
[25] L. Feng, T. Li, D. Ruan, S. Gou, A vague-rough set approach for uncertain knowledge acquisition, Knowl.-Based Syst. 24 (6) (2011) 837–843.
[26] Z. Pawlak, A. Skowron, Rudiments of rough sets, Inf. Sci. 177 (1) (2007) 3–27.
[27] Z. Pawlak, Rough Sets, Theoretical Aspects of Reasoning About Data, Kluwer, Dordrecht, The Netherlands, 1991.
[28] H. Sakai, M. Nakata, On rough sets based rule generation from tables, Int. J. Innov. Comput. Inf. Control 2 (1) (2006) 13–31.
[29] J.W. Grzymala-Busse, LERS—a system for learning from examples based on rough sets, intelligent decision support, Handbook of Applications and Advances of the Rough Sets Theory, , 1992, pp. 3–18.
[30] J.W. Grzymala-Busse, A new version of the rule induction system LERS, Fundam. Inf. 31 (1) (1997) 27–39.
[31] H.S. Nguyen, S.H. Nguyen, Analysis of stulong data by rough set exploration system (rses), in: P. Berka (Ed.), Proceedings of the ECML/PKDD Workshop, 2003, pp. 71–82.
[32] A. Ravi, H. Kurniawan, P.N.K. Thai, P.Ravi Kumar, Soft computing system for bank performance prediction, Appl. Soft Comput. 8 (1) (2008) 305–315.
[33] M. Bani Amer, Fuzzy-based framework for diagnosis of acid–base disorders, Comput. Biol. Med. 41 (9) (2011) 737–741.
[34] Y.S. Chen, J.F. Chang, C.H. Cheng, Forecasting IPO returns using feature selection and entropy-based rough sets, Int. J. Innov. Comput. Inf. Control 4 (8) (2008) 1861–1875.
[35] P. Ravi Kumar, V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques—a review, Eur. J. Oper. Res. 180 (1) (2007) 1–28.
[36] M.W. Kattan, R.B. Cooper, A simulation of factors affecting machine learning techniques: an examination of partitioning and class proportions, Omega 28 (5) (2000) 501–512.
[37] S. Jabbari, H. Ghassemian, Modeling of heart systolic murmurs based on multivariate matching pursuit for diagnosis of valvular disorders, Comput. Biol. Med. 41 (9) (2011) 802–811.
[38] N.E. Saeedi, F. Almasganj, F. Torabinejad, Support vector wavelet adaptation for pathological voice assessment, Comput. Biol. Med. 41 (9) (2011) 822–828.
[39] X. Wang, E.C.C. Tsang, S. Zhao, D. Chen, D.S. Yeung, Learning fuzzy rules from fuzzy samples based on rough set technique, Inf. Sci. 179 (20) (2007) 4493–4514.
[40] Z. Pawlak, Rough set approach to knowledge-based decision support, Eur. J. Oper. Res. 99 (1997) 48–57.
[41] F.E.H. Tay, L. Shen, Economic and financial prediction using rough sets model, Eur. J. Oper. Res. 141 (2002) 641–659.
[42] J.G. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak, J. Wr'oblewski, Rough set algorithms in classification problem, in: L. Polkowski, S. Tsumoto, T.Y. Lin (Eds.), Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, Physica-Verlag, 2000, pp. 49–88.
[43] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[44] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, Parallel distributed processing: explorations in the microstructure of cognition 1 (1986) 318–362.
[45] A. Redondo-Sendino, P. Guallar-Castillón, J.R. Banegas, F. Rodríguez-Artalejo, Gender differences in the utilization of health-care services among the older adult population of Spain, BMC Public Health 6 (2006) 155–164.
[46] K. Okamoto, Y. Momose, A. Fujino, Y. Osawa, Gender differences in the relationship between self-related health (SRH) and 6-year mortality risks among the elderly in Japan, Arch. Gerontol. Geriatr. 47 (2008) 311–317.
[47] J. Shi, M. Liu, Q. Zhang, M. Lu, H. Quan, Male and female adult population health status in China: a cross-sectional national survey, BMC Public Health 8 (2008) 277–286.
[48] M.J. Fraga, S.A. Cader, M.A. Ferreira, T.S. Giani, E.H.M. Dantas, Aerobic resistance, functional autonomy and quality of life (QoL) of elderly women impacted by a recreation and walking program, Arch. Gerontol. Geriat. 52 (2011) e40–e43.
[49] J.G. Sonati, D.M. Modeneze, R. Vilarta, E.S. Maciel, E.M.A. Boccaletto, C.C. da Silva, Body composition and quality of life (QoL) of the elderly offered by the "University Third Age" (UTA) in Brazil, Arch. Gerontol. Geriatr. 52 (2011) e31–e35.