An Analysis of DRGs Hospitalization Expenses with Optimized SVM

Chao Liu
Kunming University of Science and Technology, Kunming, Yunnan 650093, China

Zhuohui Liu
Kunming University of Science and Technology, Kunming, Yunnan 650093, China

Chongtao Zhu*
The First People’s Hospital of Yunnan Province, Kunming, Yunnan 650032, China
*Corresponding author(E-mail: 196636170@qq.com)

Abstract
The practice of Diagnosis Related Groups (DRGs) is conducive for avoiding excessive medical treatment and controlling an increase in expenses. The cost-prediction of DRGs illnesses have played an important role in hospitals’ cost control. This paper proposes a cost-predictive method of DRGs illnesses series based with optimized SVM, in order to estimate the costs of illness series accurately with redundant, non-linear and uncertain cost data. The model of cost factors of illness series that relates to generic bill of materials is discussed, which lays the foundation for a cost-predictive model of illness series. The cost-predictive model combines attributes reduction of rough sets and parameters optimization of genetic algorithm, which works with optimized SVM and provides an effective method for predicting illness costs and thus improve the accuracy in cost control of hospitals.

Key words: Diagnosis Related Groups, Illness Series, Support Vector Machines, Generic Bill of Materials, Rough Sets, Genetic Algorithm.

Un Análisis de la Hospitalización de los GRD Gastos con SVM Optimizado

Resumen
La práctica de los Grupos Relacionados con el Diagnóstico (GRD) es propicia para evitar el tratamiento médico excesivo y controlar el aumento de los gastos. La predicción de costos de las enfermedades de GRD ha jugado un papel importante en el control de costos de los hospitales. Este documento propone un método predictivo de costos de series de enfermedades de GRD basadas en SVM optimizado, para estimar los costos de las series de enfermedades con datos redundantes, no lineales e inciertos. Se discute el modelo de la serie de factores de costo de la enfermedad que se relaciona con la lista genérica de materiales, lo que sienta las bases para un modelo de costo de la serie de la enfermedad. El modelo de costo predictivo combina la reducción de atributos de los conjuntos aproximados y la optimización de parámetros del algoritmo genético, que funciona con SVM optimizado y proporciona un método eficaz para predecir los costos de enfermedades y, por lo tanto, mejorar la precisión en el control de costos de los hospitales.

Palabras clave: Grupos relacionados con el diagnóstico, Series de enfermedades, Máquinas de vectores de soporte, Lista de materiales genérica, Conjuntos aproximados, Algoritmo genético

1. Introduction
Governments around the world have adopted drastic reforms in the medical field in the past few years, with the core issue of reforming on medical insurance systems. Investments in medical care make up a certain proportion of the fiscal expenditures. The purposes of medical insurance reform are to ensure that medical institutions meet the nation’s medical needs, and in the meantime to control expense increases. Social security system of China is still under construction, yet about to complete. In the past few years, hospitalization expenses paid by patients and the proportion of medical expenses paid by the third party both grow annually. Controlling health care cost is of vital interest to people as well as national institutions concerning planning and social security. Based on the population all over the world, the key undertaking of how to utilize health resources in a legitimate and efficient way is highlighted by national health departments all over the world. Responding to the
situation, DRGs arise. DRGs change the traditional payment method based on services, to the new one based on illnesses, which categorize hospitalized patients into several groups following International Classification of Diseases, and formulate corresponding payment standards. The practice of DRGs is conducive for avoiding excessive treatment and controlling increases medical expenses. However, due to a shrinkage of profits, hospitals may have to deal with risks of expected return. Through cost-prediction of DRGs illnesses, hospitals are able to adjust therapeutic regimen and reduce costs without losing quality. Furthermore, cost-prediction of DRGs illnesses makes it possible for hospitals to estimate future cost levels correctly and provide standards for cost optimization and control. The timeliness, accuracy and reliability of costs data for prediction are directly connected to decisions on cost control and management strategy made by hospitals.

Scholars home and abroad have composed multiple predictive methods of illness expenses and costs and have researched on the relationship between costs and causes (Paladino,2007), exploiting models of regression, neural network and case reasoning. Methods of detailed estimation such as activity-based costing and marginal costing are experimented to predict illness costs and provide foundations for hospitals’ decision-making and fee-scale establishment. Despite the fact that large amount of studies has been conducted worldwide, the present method is not capable to estimate future costs of illnesses in a satisfactory way. Due to a lack of effective predictive model, studies on cost-prediction are limited to single illness, while DRGs illnesses series are overlooked. With regard to the selection of factors of illness costs, prevalent researches are heavily relied on experts’ opinions instead of a scientific and theoretical basis. Moreover, with regard to the selection of predictive models of illness costs, there are apparent flaws of established models stated as follow. Methods of detailed estimation are highly accurate yet time and capital consuming. Statistical methods are restricted with strict assumptions such as normal distribution which is usually not complied by illness costs data. At the same time, statistical analysis models are generally not accurate when dealing with high-noise conditions and multivariate illness data. Neural network shave shortcomings of being easily trapped in local extremum and low extension. The complexity in the process of illnesses’ treatment brings difficulties in obtaining data and various influence factors which differ in importance and affect each other. Furthermore, the configuration of medical service of illness series needs to satisfy patients’ customized yet sometimes similar needs. Therefore, it is a scientific question with significance of estimating costs of illness series under the data conditions of redundancy, non-linear and uncertainty.

Considering both difference and similarity presented in cases of illness series, it is necessary to introduce Generic Bill of Materials (GBOM) (van Venn and EA,1991) which provides a method to describe a number of cases with limited number of data. On the question of the variety of influence factors and redundancy, Rough Sets (RS)(Xu, J. and Liu, Z.,2013) are advantageous in reducing dimension through eliminating redundancy by using data per se. In contrast with methods of hierarchical analysis and fuzzy comprehensive evaluation, human factors are cut down in RS. Moreover, researches prove that RS performs better in accuracy and speed after reduction. Based on statistical learning theory, Support Vector Machines (SVM) (Han, X and Liu, Z,2016) solves the problem of non-linear and small samples of cost-prediction. Being able to avoid low level of generalization and local extremum, SVM is proved to be more accurate than neural network and logistic regression. However, the selection of parameters of SVM poses relatively enormous influence on results. Genetic Algorithm (GA) (Mi, N. and Fang, M.,2012) has fast global search ability which can be used to search for optimal parameters of SVM. It is proved that SVM optimized by GA has greater predictive accuracy than standard SVM.

Based on the considerations aforementioned, this paper is going to propose a cost-predictive method of DRGs illness series with optimized SVM. First, the paper introduces the model of factors of illness costs based on GBOM, and further proposes a cost-predictive model of illness series with SVM optimized by the combination of attributes reduction by Rough Sets and parameters optimization based on Genetic Algorithm. It can be verified that this method has greater accuracy and speed.

2. Model of Cost Factors of Illness Based on GBOM

2.1. Generic Bill of Materials (GBOM)

All kinds of information of DRGs illness series include the main parameters of cost-prediction. The main parameters need to be able to reflect characteristics of cases and easy to recognize and operate, and have the advantages of being succinct and hierarchical. The selection of factors of illness costs will directly influence the results of cost-prediction. Among all the factors, the table of illness configurative information is the foundation of cost-prediction of illnesses.

The practice of DRGs requires hospitals to design clinical pathways for all kinds of illness cases and be able to react fast to patients’ demands. Correspondingly, Bill of Materials (BOM) (Wang, B. and Liu, Y.,2008) clearly presents medical programs and lists of drugs of certain illness, and is able to organize and manage medical service for patients. However, a variety of types and cases of DRGs illness diagnosis will lead
to data redundancy to a large extent if BOM is constructed within the traditional structure of single illness. It can be reckoned that the core problem to be solved is how to organize treatment programs of illness series accurately and fast. As a result, researchers tend to apply GBOM to avoid redundancy while keeping structural information of BOM. GBOM is a kind of structures of bill of materials family, which consists of a BOM structure and a selection tree. The BOM structure is the hierarchy formed by generic parts, which indicates the general structure of bill of materials, while the selection tree is the hierarchy formed by variables, value thereof and configuration rules, choices between whose branches lead to a detailed bill of materials. Because GBOM based on knowledge description is able to decrease redundancy effectively, which, in the meantime, spends less time in organizing treatment programs and is easy to realize and update, GBOM is widely practiced among the descriptive family of bills of material.

This paper builds the model of factors of costs of GBOM illness series with the hierarchical interpretation aforementioned, as presented in Chart 1. Cases in illness series are equipped with corresponding BOM structures, while differences only exist in the hospital admission, medical quality, treatment configuration on leaf nodes and medical service modules that include this configuration. Cost factors of instances or cases of illness series are interpreted with structural level and instantiation level. The structural level focuses on the information of hospital admission, configuration and medical quality, which form the basic structure of cost factors, while instantiation level maps relative information to cases or instances and builds instance vectors. Cost-prediction is conducted in the same illness series when the information of hospital admission and configuration is treated as the major factors that influence costs.

**Figure 1. The Model of Factors of Costs of GBOM Illness Series**

2.2. Configurative Process of Bill of Materials

The configurative process of bill of materials is the instantiation of bill of materials models in accordance with clients’ demands, GBOM and configurative rules. Based on the GBOM discussed in this paper, the configurative process of bill of materials are showed in steps as follow.

1) Input clients’ demands. Clients input demand parameters of performance of bill of materials on a customized basis.

2) Process and convert clients’ demands. Because parameters input by clients can not be directly used by the system, conversion is necessary. According to the converted clients’ demands, a model of bill of materials can be decided.

3) Decide on class node effectiveness. Overview the selection tree of bill of materials and examine every class node according to clients’ demands and constraints to decide the effectiveness of class node.

4) Instantiate of the model of bill of materials family based on GBOM. The instantiation of the model is implemented by deciding on the value of variables of class nodes. In other words, selected subsets of effective
class nodes should be examined based on clients’ demands, and the sub-nodes with discrepant value of variables should be eliminated.

In the process of instantiation, the decision of the value of variables should be made with certain constraints, i.e. configuration rules. In the structure of bill of materials family, configuration rules mainly mean the logic constraints among parts and components, including compatibility constraints, bundled constraints and exclusion constraints. Compatibility constraints is the symbiotic relationship among parts and components. For example, if part A is selected, part B can also be selected. Bundled constraints describe the dependency relationship. That is to say, if part A is selected, part B has to be selected. Exclusion constraints is also called incompatibility constraints which describe the incompatible relationship. In other words, if part A is selected, part B can never be selected.

3. The Mathematical Representation of Cost Factors of Illness Series

Since the hierarchy expresses factors of costs of illness series in a visual way. A mathematical representation is displayed as follow in a standardized way.

(1) The set of types of factors: $F = H \cup C \cup Q$.

H, C and Q are incompatible. H represents hospitalization information. C stands for configuration information. Q means medical quality information.

(2) The set of types of modules of configuration information: $C = \{ C_1, C_2, C_3, C_4, C_5 \}$. Modules include wards, medical technology programs, surgery anaesthetic, medicine, oxygen therapy and transfusion.

(3) Type $C_i$ of modules is represented by vectors $C_i = [C_{i1}, C_{i2}, \ldots, C_{iN_i}]$, where $N_i$ is the number of submodules in the $i$th type of modules, while mijstands for submodules. There are five types of modules in total, and thus the modular structure can be defined as $C = \{C_1, C_2, \ldots, C_{51}, C_{52}\}$. To satisfy customized demands of patients, the instantiation of submodules in the modular structure is necessary. The process of modular configuration is the process of instantiation.

(4) Submodule $C_{ij}$ is represented by vector $\{ C_{ij} \mid i \in \{1, 2, \ldots, 5\}, j \in \{1, 2, \ldots, N_i\} \}$. Lijstands for the number of parameters of the jth functional module of the ithtype of modules.

(5) Hospitalization information H is displayed by vector $\{ h_a \mid a = 1, 2, \ldots, X \}$. X stands for the number of parameters indicating hospitalization information. Medical quality information Q is displayed by vector $\{ q_b \mid b = 1, 2, \ldots, Y \}$. Y stands for the number of parameters indicating medical quality information.

(6) The space factor of instances’ costs of DRGs illness series is $S = \{ h_1, h_2, \ldots, h_X; c_{111}, c_{112}, \ldots, c_{11L_{11}}; c_{121}, c_{122}, \ldots, c_{12L_1}; c_{ik1}, c_{ik2}, \ldots, c_{ikL}; \ldots c_{ij1}; \ldots; c_{5N_{1}}; c_{5N_{2}}, \ldots, c_{5N_{5}}L_{55N_5}; q_1, q_2, \ldots, q_Y \}$.

The total number of parameters of space as one of the computable factors of costs is $\Sigma_{d=1}^{5} \Sigma_{k=1}^{N_{d}} L_{ik} + X + Y$. When all the parameters of S are assigned value based on the actual conditions of instances, an instantiated expression of cost factors is acquired.

(7) P represents costs of illness series. According to the 6 definitions aforementioned, the configurative space of factors of costs of DRGs illness series can be estimated and instantiated. Factors of costs table can be created to provide initialization and preprocessing mechanism for cost-prediction of DRGs illness series.

4. Cost-Prediction Model of DRGs Illness Series by Optimizing SVM

4.1. The Reduction of Rough Sets and Factors of Costs of Illness Series

In 1982, Rough Sets (RS) was first proposed by Z. Pawlak, when the theory of rough sets (Hao, J. and Xue, H.,2012) was given birth. Rough sets theory is mainly practiced to eliminate redundancy of attributes, reduce attributes, and reduce dimensions of attributes, which are now widely applied to fields such as machine learning, knowledge discovery and artificial intelligence. RS is an effective mathematical instrument to deal with fuzzy and uncertain information. Reduction of attributes is the core for rough sets theory to process information systems. Reduction of attributes reserves the useful and important parts and attributes in knowledge database through deleting redundant ones (equivalence relations). This paper is going to realize reduction of attributes by the algorithm of rough sets. In the SVM-optimized-model of cost-prediction of illness, the key factors of illness costs are selected through RS theory.

If information system is $IS=(U, A)$, U is universe, i.e. a finite set of objects, $U = \{x_1, x_2, \ldots, x_m\}$, while A is the attribute set or characteristic variables, and for every attribute $a \in A$. Information function is defined as $fa: U \mapsto Va$. Va is a set of a’s value, and thus is called the domain of attribute a. To examine the independence of an attribute set, one question should be asked that whether the number of base sets would increase in the
information system after the attributes are deducted one by one. If \( \text{ind} (A) = \text{ind} (A - a_i) \), attribute \( a_i \) can be called redundant; otherwise, attribute \( a_i \) is necessary to \( A \). If the attribute set is not independent, it is possible to find the subset of all the potential minimum attributes, and thus the same number of base sets of the entire attribute set (reduction) and all the necessary attribute sets (core) are obtained.

From the perspective of condition attributes, i.e. factors, and decision attributes, i.e. cost, RS acquires the attribute set of minimum conditions through equivalent relations and equivalence class. The detailed method of the reduction of factors of illness series costs based on RS can be summarized into three steps. First, a decision table is formed with factors and cost decision attributes. Second, discretization algorithm with equal frequency can be utilized to discretize the continuous decision attribute of continuous attribute and costs. Last, the discretized decision table should be reduced in attributes by methods such as genetic algorithm.

4.2. Attribute Reduction Algorithm

At present, prevalent attribute reduction algorithms include discernibility function built with discernibility matrix, heuristic algorithm based on importance of attributes, and the reduction of compound system (He, Y. and Liu, H., 2010). The algorithms mentioned here have advantages and disadvantages in different aspects. This paper is going to apply the attribute reduce dalgorithm based on importance of attributes. This algorithm obtains the core by acquiring the discernibility matrix from the given table. In other words, the algorithm reduces attributes of data sets by calculating importance of attributes with functions, whose range starts with the most important attribute, until the acquired attributes form a reduction, which ensures the final attribute reduction is the minimum reduction. The advantage of this algorithm is simple and straightforward, and able to deal with relatively large data sets. Concrete operation method is displayed below:

Input: a decision table \( S (U, R); R = CUD; C \) is an attribute set of conditions; \( D \) is an attribute set of decisions.

Output: decision rules of decision table \( S \).

1) Eliminate repetitive lines. In a two-dimensional data sheet, lines stand for different objects. When the attributes of two objects are the same, one of the two lines should be eliminated.

2) When two columns are of the same attributes, one of the two columns should be eliminated.

3) Generate the discernibility matrix \( M \) of \( S \):

\[
\text{FOR}(i = 1; i <= m; i++)
\]

\[
\text{FOR}(j = 1; j < I; j++)
\]

\[
M = (C_i)
\]

\( M \) is a lower triangular matrix without main diagonal value.

4) Calculate the core of \( C \), Cored \( (C) \), in relation to \( D \) by \( M \);

\[
\text{Given } C_0 = \text{Cored} (C);
\]

5) Let \( B = C_0 \) and all the condition attributes excluding the core is AR. Then \( AR = C - B \). For attribute \( a_i \in AR \), calculate the importance \( f(a_i) \) and sort AR in descending order according to value.

6) Calculate \( r_{\text{g}}(D) \) and \( r_{\text{c}}(D) \) independently.

7) WHILE \( (r_{\text{g}}(D) = r_{\text{c}}(D)) \) AND \( \text{(AR} \neq \emptyset) \)

\[
(\forall a_i \in AR \ni f(a_i) = \text{MAX}(f(a_i)); B = BU[a_k];AR=AR\setminus[a_k]);
\]

8) \( B \) is a reduction of \( C \) in relation to \( D \).

The pre-processing is set as finished and a decision table obtained. The first column of the table is set as the decision attribute and all the attribute value has been transformed into numeric form.

4.3. Cost-Prediction of Illnesses Based on SVM

According to sample data information, SVM maps indivisible data sets to high-dimension feature space by minimizing structural risk, ensures samples are differentiated correctly in high-dimension space, and thus overcomes the difficulty of differentiation in low-dimension space. Let \( x_i \in E \) be the factor of cost-prediction, and \( y_i \) is the predicted value of costs. Steps of regression forecasting of illness costs by SVM is showed below.

**Step. 1** Given a training data set with the value of 1
\[
\{(x_i, y_i) \mid i = 1, 2, \ldots, 1, x_i \in E, y_i \in R \ (E \text{ is Euclidean space})\}
\]

**Step. 2** The hyperplane in high-dimension space can be interpreted as
\[
\omega x + b = 0
\]

(1)

The main idea of SVM estimating regression function is to map data input in space to a high-dimension feature space through a non-linear mapping \( x \), and conduct linear regression in the high-dimension space.

SVM utilizes this predictive function: \( f(x) = \omega x + b \)

(2)

**Step. 3** In accordance with the principle of Structural Risk Minimization (SRM), SVM measures risks with insensitive value \( \varepsilon \) proposed by Vapnik. Define \( \varepsilon \) as below:
fitness value in the evolution should be output as the optimal solution acquired as P(t+1).

should be calculated based on concrete questions. For different questions, the definitions of fitness function vary. Individual fitness in population P(t) generated M individuals to form a initial population P(0).

set evolutional generation solver which is an individual. N individuals form a population. Genetic algorithm starts iteratively operations such as random selection, crossover and mutation. Initialized populations and evolve them to enter into domains with better search space through genetic operations. Being selected, crossed over and mutated, the next generation of population P(t) is generated from computer simulation of biological system is a random searching algorithm referring to nature selection and natural inheritance, whose operation flow is as follows.

The main operation process of genetic algorithm is as follow.

Coding. Solution data x in solution space is the phenotypic form of genetic algorithm. The mapping from phenotype to genotype becomes the coding. Before searching, genetic algorithm interpreted data of solution space as genotypic string structured data of genetic space. The combinations of string structured data become different dots.

Generation of initial populations. N initial string-structured data are randomly generated, every one of which is an individual. N individuals form a population. Genetic algorithm starts iteration from the N string structures. Set evolutional generation solver t  0; set the maximum evolutional generation T; randomly generate M individuals to form a initial population P(0).

Fitness value assessment. Fitness function indicates the advantages and disadvantages of individuals or solutions. For different questions, the definitions of fitness function vary. Individual fitness in population P(t) should be calculated based on concrete questions.

Genetic operations. Being selected, crossed over and mutated, the next generation of population P(t) is acquired as P(t+1).

Judgement on end condition. If t ≤ T, t  t + 1. Turn to step (2). If t > T, the individual of maximum fitness value in the evolution should be output as the optimal solution. Operation ends.

4.4. SVM Parameters Optimized by Genetic Algorithm

The predictive accuracy of SVM is closely connected to the range of parameters C SVM, σ and ε. It is necessary to adopt optimization algorithm to search for the optimized combination of parameters in certain domains, and thus predictive performance can be improved to optimal status.

Genetic Algorithm (GA) (Yang, K., 2011) that originates from computer simulation of biological system is a random searching algorithm referring to nature selection and natural inheritance, whose operation flow is presented in Chart 2. The algorithm is suitable for optimization in multi-peak space, which starts from any initial populations and evolve them to enter into domains with better search space through genetic operations such as random selection, crossover and mutation.

To sum up, it is necessary to decide on an insensitive value ε, the penalty factor C SVM and kernel function parameters in order to apply SVM. Different numbers of cores influence the accuracy of cost-prediction to a large extent. Since radial base kernel function (RBF) (Zhang, R. and Pei, H., 2009) is able to approach any nonlinear function and perform satisfactorily in predictive accuracy and speed, this paper adopts RBF:

\[
K(x_i, x) = \exp\left(-\frac{\|x_i-x\|^2}{2\sigma^2}\right)
\]

Therefore, σ, ε, C SVM are going to be training parameters of SVM.
5. Conclusions

There are three stages of the predictive model of illness cost based on optimized SVM that is proposed in this paper. First, learning samples are generated according to initial data. Through quantifying and discretizing the value of condition attributes, a decision table of attributes of condition and decision is obtained. By applying rough set theory, the set of minimum condition attribute and the new leaning sample set that stems from corresponding initial data are acquired. Second, the structural parameters of SVM is optimized by genetic algorithm. With optimized parameters, SVM predictive model of illness costs is built. SVM, then, should be trained with training samples reduced by rough sets. Last, the new test sample set that is formed by initial data in correspondence with the minimum condition attribute set should be input, test the system and output predictive results.

This paper fully considers the multi-factor, non-linear characteristics of disease cost and the similarity of medical service allocation in each case of disease series. The cost-forecasting model of DRGs disease series based on GBOM and improved SVM is reusable. Once the influencing factors of the new case are obtained, the cost forecasting can be directly based on the specific value of the disease case of the minimum condition attribute set in the model. The model has broad application prospects in the field of DRGs-oriented disease cost forecasting, and can provide an effective auxiliary means for disease series cost forecasting, which is of great significance.

Acknowledgements

This work was supported by National Natural Science Foundation of China 2013 (NSFC2013).

References


