International Journal of Hybrid Information Technology
Vol.8, No.4 (2015), pp. 193-200
http://dx.doi.org/10.14257/ijhit.2015.8.4.22

Application of Data Mining on Influence Factors of Medical Expense Analysis: Taking Hepatitis A as an Example

Pei Shen and Jikai Zhang

[1]School of Informatics
Guangdong University of Finance & Economics
Guangdong, China
[2]Guangdong Provincial Institute of Biological Products and Material Medica
Guangdong China
11126332@qq.com[1], jnuzjk@163.com[2]

Abstract

In this study, we used three popular data mining techniques (decision trees, artificial neural networks and support vector machine) to analyze the risk factors of medical expense of patients with hepatitis A in Guangdong Province in 2008. We compared the three methods to find out an effective method to predict medical expense and extract main influence factors of the medical expense. The results showed that support vector machine is the most accurate predictor.

Keywords*: medical expense; data mining; support vector machine*

1. Introduction

Along with the rapid develop of social economy and advances in medical technology, the high medical expense become a serious problem for governments all around the world. The rapidly ageing population also makes medical expense a trouble in China. The growth rate of medical expense has exceeded the residents' income growth levels. It has serious effect on fairness, efficiency and accessibility of basic medical services. How to control health care expense has become one of the most important issues for our government.

Hepatitis A is an infectious disease. Once infected with hepatitis A one must be isolated and treated for at least 21 days. Full recovery needs for about six months. Pathogenesis of hepatitis is hepatitis virus replication in the liver, resulting in liver cell damage followed by a series of symptoms, such as fever, tired of the oil, anorexia, abdominal pain, diarrhea, and jaundice. Hepatitis A causes significant economic and social consequences and emotional burden for the patients.

In this study, we collected medical expense information of patients with hepatitis A in seven hospitals in six cities in Guangdong Province in China. We used three popular data mining techniques, decision trees, artificial neural networks (ANNs) and Support Vector Machines (SVMs) to develop prediction models for risk factors of medical costs of hepatitis A. Potential factors affecting medical costs included length of stay in a hospital, hospital grade, city, discharge statue, gender, payer type, disease type, age and diagnostic control. The total medical costs of hepatitis A can be reduced through controlling the key influence factors, so it is important to identify the relevant risk factors. We also tried to find the best methods to support the analysis for influence factors of medical expense by comparing the three data mining methods.

The results showed that SVM might be the best approach with a test accuracy of 97.39%, followed by ANNs with an accuracy of 93.04%, and decision trees with an accuracy of 91.30%.

ISSN: 1738-9968 IJHIT

2. Previous Research

Data mining is a type of machine learning and adaptive computation. Data mining is the analysis of large data sets to reveal the implicit, previously unknown and potentially information and to describe the data in novel ways that are both understandable and useful to the data owner. Data mining is a decision support process, which is mainly based on artificial intelligence, machine learning, pattern recognition, statistical, database, visualization techniques, find potential model to help policymakers adjust marketing strategies, reduce risk and make the right decisions.

In the study of the factors influencing medical expense, the most widely used methods are traditional statistical methods, such as regression analysis. For these methods, the variables need to have variable normality, homogeneity of variance and other applicable assumptions. However, the complex, multi-dimensional and incomplete medical data usually do not meet these requirements. Excellent data mining method can solve this problem [1-2].

Data mining can be used in medical decision support system, treatment project selection and evaluation, distribution of medical resources, evaluation of drug therapy effect and other medical fields [3]. Walter and Mohan (2000) proposed an algorithm applied to the diagnosis of breast cancer. This algorithm extracted classification rules from the trained neural network [4]. Zupan and Demsar (2000) proposed a classification method, including machine learning classifier for prostate cancer survival analysis to forecast survival time of patients after prostate cancer [5]. Bellazzi *et al.* (2001) used data mining tool to extract a predictive model of hepatic carcinoma from the past pathogenicity results of liver cancer [6]. Land (2001) explored a new neural network to improve the diagnosis of breast cancer by the breast X-rays. Furthermore, a simple data preprocessing can increase the accuracy of the neural networks and the classification rules extracted from the networks [7]. Dursun (2009) used the popular statistic method, logistic regression and three data mining techniques, including decision tree, artificial neural network and support vector machines to develop survival prediction models for prostatic cancer. The results showed that support vector machines are the best predictor with the highest accuracy and logistic regression is the worst [8].

Although data mining is a new field of medical expense analysis, there are a lot of literatures and researches focusing on application of data mining in influencing factors of medical expense studies. Holloway (1996) estimated the relationship between medical expense and a number of demographic and administrative variables with different cerebrovascular events. This research also found that length of stay as a measure of resource use was strongly predictive of medical cost, explaining 72% to 82% of the variation in medical cost. Demographic variables (age, gender, race, insurance status) revealed virtually little predictive power [9]. Jayadevappa (2005) compared the direct medical expense of black and white prostate cancer patients with regression analysis in USA. That study found that race had little influence on the medical expense after controlling for the patients' age, complications and the stage of cancer [10]. Tsann (2004) investigated the medical expense of diabetic patients participating in national health insurance in Taiwan from 1998 to 1999. They analyzed the influence factors of these patients' medical expense and forecasted the future medical expense with multiple regression analysis. The results showed that age, gender had little influence on the medical expense of diabetic patients. But medical expenses were significantly affected by the diabetic complication [11]. Agnes (2009) selected the direct medical expense of Alzheimer's disease from the Taiwan national health insurance research database. They found that age and gender had little influence on medical expense [12]. But the current data mining methods used in medical expense research were still undeveloped in China. There is little complete study result.

3. Research Method

3.1. Data Understanding and Data Preparation

This is a retrospective survey. First, we selected the hepatitis A cases reported from disease monitoring system in 2008, and then checked the hospital records and expense data containing inpatient charges reported by the corresponding hospitals. Finally, we interviewed the recorded patients in a face-to-face manner. After questionnaire data collection, we got medical information of 195 patients from 7 hospitals in 6 cities of Guangdong Province. The seven hospitals included primary hospital, secondary hospital and tertiary hospital. The process of medical data collection was complex, and it's hard to avoid data missing, mistake or even duplication. In order to reduce the influence of these noises, we cleaned the data under the guidance of medical experts. After excluding records without hepatitis A, duplicated records and the abnormal value that seriously deviated from the group in the data source, we got 115 useful cases. The survey collected the medical expense and influence factors of these hepatitis A patients.

The data collected from the interview were imported in SPSS for initial data exploration and understanding. Then we used Clementine platform to analyze the data and develop predictive models. Clementine is a data mining platform developed by Integral Solutions Limited Company. In 1999 SPSS purchased ISL Company, and redesign the Clementine products. Now Clementine becomes another bright spot of SPSS Company. As a data mining platform, Clementine can perform predictive models quickly, help people improve the decision making process. Compared with the other data mining tools that focus only on model and ignore the application of data mining in the whole business process, Clementine has more powerful data mining algorithms.

The distribution of total medical costs was shown in Figure 1. Obviously, medical costs actually did not follow normal distribution.

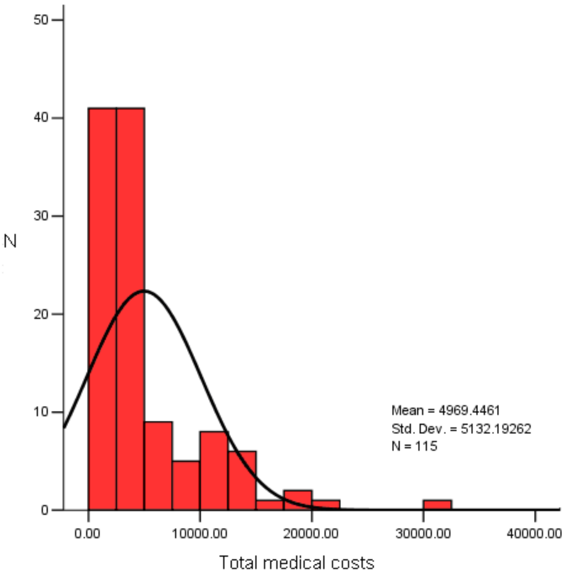
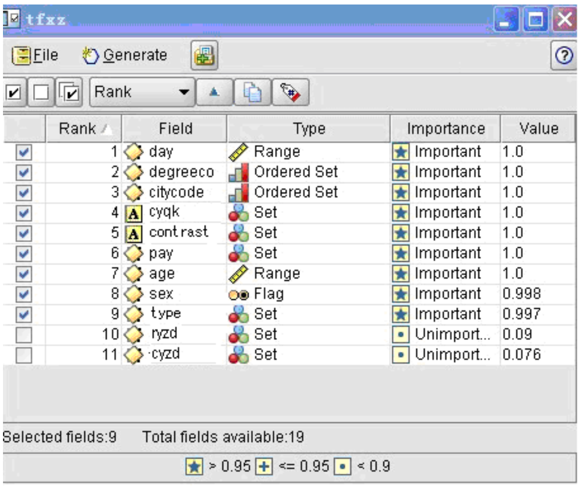


Figure 1. The Distribution of Total Medical costs

According to previous studies and the suggestion of epidemiologists, this study collected 19 attributes of medical expense, such as gender, age, medical record number, occupation, educational background, location, name of hospital, hospital grade, type of disease, admission departments, length of stay, diagnosis control and so on. Put these attributes into the feature selection model of SPSS Clementine, run model and select 9 attributes having important influence on the medical expense. Their importance are greater than 0.95. Major factors affecting medical expense are length of stay, hospital grade, region, discharged situation, gender, payment model, diagnostic control, age and type of disease.



	Rank	Field	Type	Importance	Value
☑	1	day	Range	Important	1.0
☑	2	degreeco	Ordered Set	Important	1.0
☑	3	citycode	Ordered Set	Important	1.0
☑	4	cyqk	Set	Important	1.0
☑	5	contrast	Set	Important	1.0
☑	6	pay	Set	Important	1.0
☑	7	age	Range	Important	1.0
☑	8	sex	Flag	Important	0.998
☑	9	type	Set	Important	0.997
□	10	ryzd	Set	Unimport...	0.09
□	11	·cyzd	Set	Unimport...	0.076

Selected fields:9 Total fields available:19

★ > 0.95 + <= 0.95 • < 0.9

Figure 2. The Distribution of Total Medical Costs

3.2. Methods

Decision tree is a typical classification method. After data processing, decision tree algorithm generates readable rules and trees, and then the consequences are used in the new data analysis. In essence, decision trees are often used to classify the data through a series of rules. Typical decision tree algorithms include ID3, C4.5, C5.0 and CART. According to different splitting rule, the mathematical algorithms of decision tree include information gain, the Gini index and the $\chi 2$ test [13]. Decision tree algorithm has many advantages. First, users don't need to have a lot of background knowledge. Secondly, decision tree model is an efficient classifier. It achieves high performance in training large dataset. Finally, the decision tree is a simple and intuitive tree structure. Furthermore, the decision tree has higher classification accuracy on many issues.

Artificial neural networks are a lot of simple processing units connected by a wide range of intelligent computing systems to simulate biological neural networks. Generally speaking, neural network is a set of similar processing units of neurons, each of which is associated with a weight. In the learning phase, by adjusting these weights, the neural network can predict the correct class label [14]. The neural network takes a long training time. Neural network also requires a lot of parameters, usually determined mainly by experience, such as network topology. The neural network was always criticized for its poor interpretability. However, neural network has high tolerance of the noise data and strong classification ability of data pattern without training.

Support vector machines are firstly proposed by Vapnik in 1995 which is a new general learning method based on statistical learning theory. SVMs have many advantages in solving small sample, nonlinear and high dimensional pattern recognition problems. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [8].

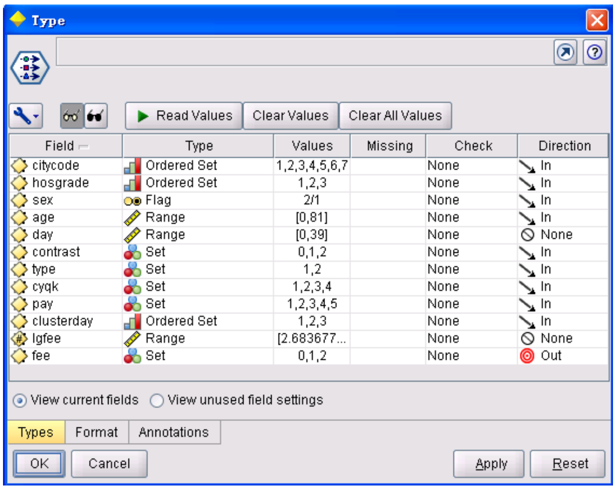
3.3. Models and Results

Firstly, we classified the medical costs of the hepatitis A patients into two categories, as the dependent variables. Secondly, 9 attributions were selected to participate in the data mining model as the independent variables. The distribution of major factors affecting medical expense is shown as Table 1.

Table 1. The Distribution of Independent Variables

Variables	Variables explanation	N	%
Gender	1=Male 2=Female	71 44	61.74 38.26
Age	continuous variables		
City	1=City 1 2= City 2 3= City 3 4= City 4 5= City 5 6= City 6	46 13 15 15 6 20	40.00 11.30 13.04 13.04 5.22 17.39
Hospital Degree	1=Third-class hospital 2=Second-class hospital 3=First-class hospital	7 69 39	6.09 60.00 33.91
Diagnostic control	1=Different 2=Same	42 73	36.52 63.48
Length of stay	continuous variables		
Discharged statue	1=Others； 2=Not cure； 3=On the mend； 4=Recovery	11 12 44 48	9.57 10.43 38.26 41.74
Payer type	1=Social medical insurance； 2=New rural cooperative medical 3=Government welfare insurance ； 4=Self paid； 5=others	4 5 4 80 22	3.84 4.35 3.48 69.57 19.13
Disease type	1=Hepatitis A 2= Hepatitis A with other diseases	67 48	58.26 41.74

Secondly, build up the data mining model with the dependent variables and the independent variables by decision tree, ANNs and SVMs to ranking the influence factors according to their importance.



Field	Type	Values	Missing	Check	Direction
citycode	Ordered Set	1,2,3,4,5,6,7		None	In
hosgrade	Ordered Set	1,2,3		None	In
sex	Flag	2/1		None	In
age	Range	[0,81]		None	In
day	Range	[0,39]		None	None
contrast	Set	0,1,2		None	In
type	Set	1,2		None	In
cyqk	Set	1,2,3,4		None	In
pay	Set	1,2,3,4,5		None	In
clusterday	Ordered Set	1,2,3		None	In
lgfee	Range	[2.683677...		None	None
fee	Set	0,1,2		None	Out

● View current fields ○ View unused field settings

Figure 3. Screen Shot for the Independent Variables and Dependent Variable

Compare the predicting accuracy and the ranking reasonability of the three data mining methods. The results showed that SVMs is the most accurate model with the predicting accuracy of 97.39%, followed by artificial neural networks and decision tree.

Table 2. Results of Predicting Accuracy

Model	True	False	Predicting accuracy
Decision Threes	105	10	91.3%
ANNs	107	8	93.04%
SVMs	112	3	97.39%

Then compare the three algorithms by the reasonableness of influence factors. Sorted according to importance of the influencing factors by the three data mining methods are shown in Figure 4, Figure 5 and Figure 6. The importance factors selected by the decision trees are hospital grade, age and length of stay.

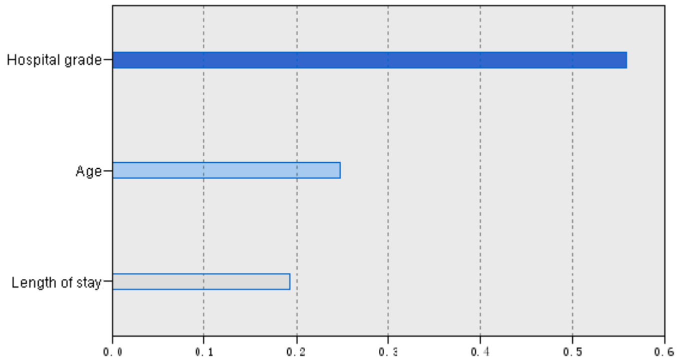


Figure 4. The Variable Importance Ranking for Decision Trees Model

The importance factors ranked by the artificial neural network are payer type, length of stay, hospital degree, city, gender, discharge statue, diagnostic control, disease type and age.

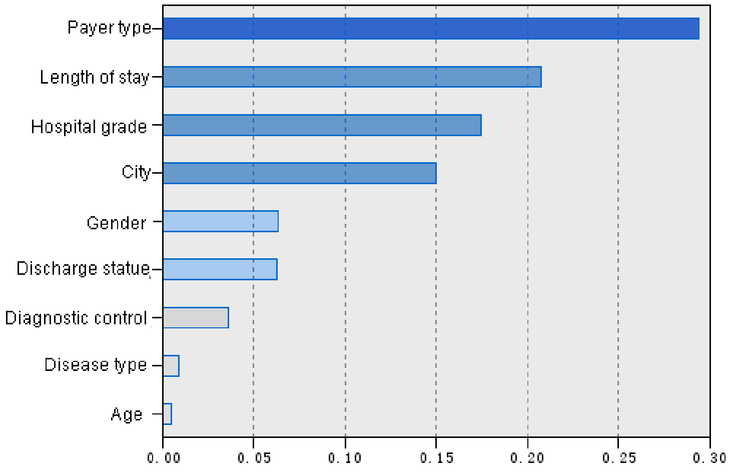


Figure 5. Screen Shot for the Variable Importance Ranking of ANNs Model

The importance factors ranked by SVMs are hospital degree, length of stay, payer type, city, diagnostic control, gender, age, disease type and discharge statue.

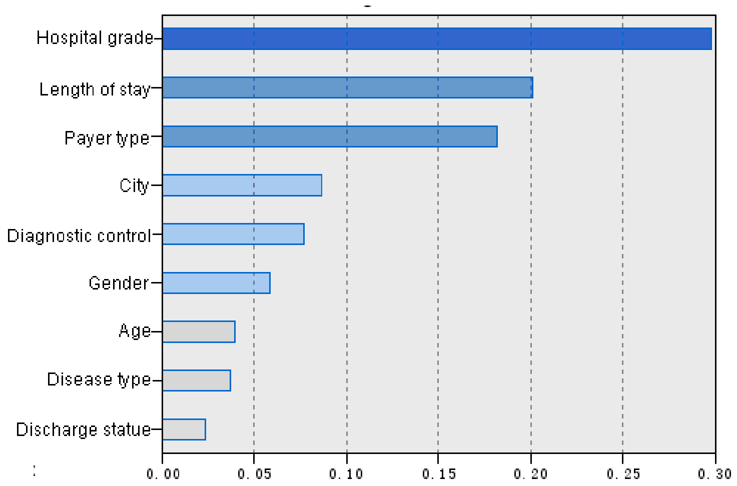


Figure 6. Screen Shot for Variable Importance Ranking of SVMs Model

So SVMs model performs best with an accuracy measure of 97.39% (followed by ANNs and decision trees) and also selects the most reasonable influence factors of medical expense with the confirmation of epidemiological experts.

4. Conclusion

Traditional statistical method was generally based on the asymptotic theory that the sample size tends to infinity. But in practical situation, the sample is often limited [15]. In this study, we found that SVMs showed the unique advantage of high prediction accuracy in solving small sample, nonlinear and high dimensional pattern recognition problem. We believe support vector machines will be more widely used in medical data mining.

Although data mining methods can extract patterns and relationships hidden deep in large medical data sets, without the cooperation and feedback from the medical experts the results would be useless [16]. The results found by data mining methods should be evaluated by medical experts who have years of experience in the relative field. They can decide whether the patterns found by data mining was logical, actionable and novel to new biological and clinical research.

In summary, data mining is not aiming to replace medical experts and researchers, but to complement their invaluable efforts to save more human lives.

References

[1] K J. Cios and G. W. Moore, “Uniqueness of medical data mining”, Artificial Intelligence in medicine, vol. 26, (**2002**), pp. 1-24.

[2] I.-N. Lee, S.-C. Liao and M. Embrechts, “Data Mining Techniques Applied to Medical Information”, Medical Information, vol. 2, no. 25, (**2000**), pp. 81-102.

[3] V. Karthikeyani and I. P. Begum, “Comparison a Performance of Data Mining Algorithms in Predicion of Diabetes Disease”, International Journal on Computer Science and Engineering, vol. 3, no. 5, (**2013**), pp. 205-210.

[4] D. Walter and C. K. Mohan, ClaDia, “A fuzzy classifier system for disease diagnosis. Proceedings of the 2000 Congress on Evolutionary Computation, (**2000**) Fubary; New York, USA.

[5] B. Zupan and J. Demsar, “Machine learning for survival analysis: a case study on recurrence of prostate cancer”, Artificial Intelligence in Medicine, vol. 1, no. 20, (**2000**), pp. 59-75.

[6] R. Bellazzi, I. Azzini and G. Toffolo, “Mining data from a knowledge management perspective: an application to outcome prediction in patients with resectable hepatocellular carcinoma”, Proceedings of AIME,(2001); Berlin: Springer

[7] Land, "New results in breast cancer classification obtained from an evolutionary computation/adaptive boosting hybrid using mammogram and history data", Proceedings of the 2001 IEEE Mountain Workshop on Soft Computing in Industrial Applications, (2001); New York.

[8] D. Delen, "Analysis of cancer data: a data mining approach", Expert Systems, The Journal of Knowledge Engineering, vol. 1, no. 26, (**2009**), pp. 100-112.

[9] R G. Holloway, D M. Witter and KB. Lawtan, "Inpatient costs of specific cerebrovascular events at five academic medical centers", Neurology, vol. 3, no. 46, (**1996**), pp. 854.

[10] R. Jayadevappa, S. Chhatre, M. Weiner, B. S. Bloom and S. Bruce Malkowicz, "Medical care cost of patients with prostate cancer", Urologic Oncology: Seminars and Original Investigations, vol. 3, no. 23, (**2005**), pp. 155-162.

[11] T. Lin, P. Chou, S.-T. Tsai, Y.-C. Lee and T.-Y. Tai, "Predicting factors associated with costs of diabetic patients in Taiwan", Diabetes Research and Clinical Practice, vol. 63, (**2004**), pp. 119-125.

[12] A. L. F. Chan, T.-M. Cham and S.-J. lin, "Direct medical cost in patients with Alzheimer's disease in Taiwan: A population-based study", Current therapeutic research, vol. 1, no. 70, (**2009**), pp. 10-18.

[13] A. Khemphila and V. Boojing, "Comparing performance of logistic regression, decision tree and neural network for classifying heartdisease patients", Proceedings of International Conference on Computer Information System and Industrial Management Application, (**2010**).

[14] P. K. Srivathsa, "Knowledge Discovery in medical mining by using genetic algorithms and artificial neural networks", Proceedings of the 2nd International Conference on Methods and Models in Science and Technology, (**2011**).

[15] K. Srinivas, B. Kavitha Rani and A.Govrdhan, "Applications of datamining techniques in health care and prediction heart attacks", International Journal on Computer Science and Engineering, vol. 2, (**2010**), pp. 250-255.

[16] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective", Artificial Intelligence in Medicine, vol. 23, (**2001**), pp. 89-109.

Authors

Pei Shen, She received her PhD (management) in 2012 from Huazhong University of Science and Technology. She has 11 years of teaching experience in Guangdong University of Finance & Economics. Her research interests are E-commerce, Data Mining and Supply Chain

Jikai Zhang, He is an associate chief physician. He received his MA (medicine) from Jinan University in 2002. Prior to his appointment at Guangdong Provincial Institute of Biological Products and Material Medica in 2013 he has worked for the Guangdong Provincial Center for Disease Control and Prevention of for ten years. His interests are the area of health statistics and epidemiology.