Comparison of Three Data Mining Algorithms for Knowledge Discovery in Electronic Medical Records (Postprint)

Authors: Mou Dongmei, Ren Ke

Date: 2017-10-11T00:00:00+00:00

Abstract

【Purpose】To identify disease risk factors from heterogeneous electronic medical record data, offering insights for data mining and knowledge discovery. 【Method】Clinical electronic medical record data encompassing multiple structural types were selected, and three data mining algorithms—decision tree, logistic regression, and neural network—were employed to establish disease risk factor prediction models, respectively. Comparative analysis and statistical evaluation of the three prediction models were conducted. 【Results】The decision tree prediction model demonstrated superior precision and recall compared to logistic regression and neural network, with decision tree achieving optimal overall performance, though differences among the three models were minimal. 【Limitations】Optimized selection of electronic medical record attributes was not performed. 【Conclusion】Decision tree outperforms logistic regression and neural network in risk factor discovery and disease prediction. This study establishes a knowledge discovery framework for heterogeneous data sources based on data mining algorithms, offering valuable insights for future domain knowledge discovery, knowledge base construction, and data mining algorithm selection.

Full Text

Three Data Mining Algorithms for Knowledge Discovery in Electronic Medical Records

Mu Dongmei[1], Ren Ke[2] [1](School of Public Health, Jilin University, Changchun 130021, China) [2](School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract

[**Objective**] This empirical study aims to identify disease risk factors from heterogeneous electronic medical record (EMR) data, providing insights for data mining and knowledge discovery. [**Methods**] We selected clinical EMR data encompassing various structures and employed three data mining algorithms—decision tree, logistic regression, and neural network—to construct disease risk factor prediction models. A comparative analysis and statistical evaluation of the three prediction models were conducted. [**Results**] The decision tree prediction model achieved higher precision and recall rates than logistic regression and neural network models, demonstrating optimal overall performance, though the differences among the three were not substantial. [**Limitations**] We did not optimize the selection of EMR attributes. [**Conclusions**] Decision trees outperform logistic regression and neural networks in discovering risk factors and predicting diseases. This study establishes a knowledge discovery framework for heterogeneous data sources based on data mining algorithms, offering guidance for future domain knowledge discovery, knowledge base construction, and algorithm selection.

Keywords: Knowledge discovery; Electronic medical record; Data mining algorithms; Prediction model

Classification Number: G202

1. Introduction

With the emergence of the Big Data concept and the advent of the Big Data era, the research scope of information science has increasingly exhibited the typical characteristics of Big Data. The "4V" features of Big Data—large volume, high velocity, diverse data types, and low value density—present new challenges to information science. In particular, the wide variety of data types, diverse structures, and uneven quality require information processing in the field of information science to continuously expand toward data cleaning, standardized integration, and consolidation. American management scientist Russell Ackoff constructed the DIKW (Data-Information-Knowledge-Wisdom) hierarchy, Zeleny distinguished the elements within the DIKW system, CIO Era Network analyzed its content and value, and Wang Yuefen argued that bibliometric and content analysis methods are key algorithms for achieving DIKW transformation. The DIKW system provides vast development space for information science while clarifying its research purpose and connotation. Information science needs to perform data standardization and normalization through information science methods such as natural language processing and concept mapping based on data cleaning, and then utilize diverse data analysis algorithms including content analysis, scientometrics, and social network analysis to extract implicit knowledge through data mining, achieving knowledge discovery and providing embedded, personalized, and precise services for users.

Currently, medical data represent the most complex data, best exemplifying

the characteristics of Big Data with multiple types, sources, and uses. This study selected clinical Electronic Medical Record (EMR) data and, guided by the information science knowledge discovery framework, utilized data mining algorithms including decision tree, logistic regression, and neural network to construct disease risk factor prediction models and evaluate the three prediction models. This research standardizes the process of information science methods for knowledge discovery in the medical field, explores effective associations between knowledge and optimal algorithms for knowledge discovery from complex data, and provides guidance for future data processing and knowledge discovery. Additionally, it can provide data support for clinical diagnosis, visual evidence for disease prevention and control personnel, and scientific research data support for the entire process of "prevention-diagnosis-treatment-prognosis" for pregnancy-induced hypertension. Applying data mining methods to study disease risk factors can enhance the development and utilization of medical Big Data information.

2. Knowledge Discovery Framework for Heterogeneous Data Sources Based on Data Mining Algorithms

Knowledge discovery based on data mining algorithms for heterogeneous data sources follows the knowledge discovery research within the logical framework of scientific domains. The knowledge processing flow emphasizes data standardization, achieving data semantic normalization based on the fusion of heterogeneous domain ontologies from different sources, and then deeply exploring topic models, association data analysis, and machine learning methods. This represents an essential path for efficient domain knowledge discovery, with a four-step process as shown in [Figure 1: see original paper].

[Figure 1: see original paper] Knowledge Discovery Framework for Heterogeneous Data Sources Based on Data Mining Algorithms

(1) Data Collection Using Database Technology. This involves multiple data sources, such as diagnostic reports from physician workstations, patient social characteristic data from nursing stations, image data stored in imaging departments, structured laboratory examination data stored in laboratories, and medication and monitoring data reports from operating rooms. Data from different sources present various structures, requiring the structuring of differently structured data for storage in databases.

(2) Data Cleaning. This step completes data de-identification, data type standardization, missing value processing, natural language processing, and semantic annotation, with natural language processing and semantic annotation being the key technologies.

(3) Prediction Model Construction. This employs supervised learning methods from machine learning to construct disease risk factor prediction models, mining valuable intelligence from large-scale multidimensional data to analyze the knowledge behind the data. Data mining technology includes numerous

algorithms, which are divided into supervised learning algorithms, unsupervised learning algorithms, and special algorithms based on whether the training data have labels. This study used the open-source software R to establish data mining models. In R, decision tree, logistic regression, and neural network algorithms were applied to process and analyze relevant data, including removing missing values, discovering outliers, performing unique data processing, and conducting association analysis on relevant categories, ultimately establishing reasonable and effective data mining models. R software functions were used for model visualization and display, and the model was used to predict data to obtain effective processing results.

(4) Model Evaluation. Statistical methods were used to evaluate prediction models, with evaluation metrics including precision, recall, accuracy, and F-value.

3. Empirical Research

3.1 Data Source The research data were derived from EMRs of a tertiary Grade A hospital in Changchun, comprising information from 31,443 pregnant women who visited the hospital between January 1, 2014, and April 30, 2015. Information center personnel extracted the data to establish an Excel database (see [Figure 2: see original paper]). The data included: patient basic information (department, age, registration number, gender, ethnicity, occupation, education level, marital status, income); lifestyle and work habit information (smoking status, alcohol consumption, work pressure, and mental stress); medical history information (past medical history, family history); routine physical examination data (height, weight) and laboratory examination data (systolic blood pressure, diastolic blood pressure, total cholesterol, triglycerides, low-density cholesterol, high-density cholesterol, fasting blood glucose, hemoglobin); and diagnostic results. Each patient was strictly diagnosed according to medical diagnostic standards, and detailed descriptions in natural language format were provided in past medical history, family history, and diagnostic results following the EMR format.

3.2 Data Cleaning **(1) Extraction of Effective Attribute Columns.** Since some attributes in the data have no or minimal impact on prediction models, their inclusion in analysis might create noise (e.g., admission date, registration number). During the extraction of effective attribute columns, noise attribute columns were removed while meaningful ones were retained. This study primarily employed manual extraction to enhance accuracy and effectiveness.

(2) Natural Language Processing. Unstructured information described in natural language within EMRs, including past medical history, family history, admission diagnosis, and discharge diagnosis, was processed. First, binary classification was performed, and past medical history and family history were extracted using punctuation marks as separators. Disease name data separated

out and disease names from admission and discharge diagnoses underwent concept mapping to the International Classification of Diseases ICD10 under the Unified Medical Language System (UMLS) to facilitate effective data recognition by subsequent data mining models.

(3) Text Data Numericalization. In data mining models, neural networks can only process numerical variables. Therefore, to facilitate model establishment, qualitative data were converted to numerical variables in this stage. For example, in the "marital status" attribute column, "divorced" was set to 1, "married" to 2, "unmarried" to 3, "widowed" to 4, and "other" to 5.

(4) Missing Value Processing. Due to non-standard EMR recording, some patient records were incomplete. These records would affect final model establishment and mining, but since these missing values were not numerous, R software was used to remove data containing missing values to present better mining results.

Through the above steps, basic data preparation was completed, rendering the data processable and enabling clearer and more concise representation. Ultimately, 29,901 data entries were obtained, as shown in [Figure 3: see original paper].

3.3 Construction of Risk Factor Prediction Model for Pregnancy-Induced Hypertension An empirical study was conducted on mining risk factors for Pregnancy-Induced Hypertension (PIH) using the aforementioned algorithms. After completing the data preparation and processing stages, and to ensure research consistency and rigor, all three data mining algorithms used the same training and test sets. The data were divided into training and test sets at a 7:3 ratio, with 70% (20,910 entries) used as training data for model establishment and PIH risk factor mining, and the remaining 30% (9,433 entries) as test data for algorithm performance testing. Subsequently, missing values in both training and test sets were removed in R, resulting in 20,940 training set entries and 8,961 test set entries.

(1) Decision Tree Model for PIH Risk Factor Mining. As a supervised learning method, decision trees can be used for classification and prediction. In their tree structure, each node and branch has specific meaning: decision trees divide complex data into several types through continuously refined branches (i.e., classification criteria), represented by leaf nodes, thus enabling intuitive and clear data classification. This study employed the ID3 algorithm to construct the decision tree model. The key to constructing the smallest possible decision tree lies in selecting appropriate attributes for branching. The core of the ID3 algorithm is to select attributes that best classify samples using information gain.

Let $E = D_1 \times D_2 \times \cdots \times D_n$ be an n-dimensional finite vector space, where D_j is a finite discrete symbol set. Elements in E, $e = \langle v_1, \ldots, v_n \rangle$, are examples, where $v_j \in D_j$, $j = 1, 2, 3, \ldots, n$. Let $S = \{s_1, \ldots, s_m\}$ be a set of m examples

from E. Assuming the size of these m examples in vector space E is S, ID3 is based on the following two assumptions:

1. A correct decision tree on vector space E has the same probability of classifying any example as the probability of these m examples in E.
2. The amount of information required for a decision tree to make a classification judgment on an example is given by:

$$\text{Entropy}(s_1, s_2, \dots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where p_i is estimated by s_i/s.

If attribute A is used as the root of the decision tree, and A has v values, it divides E into v subsets $\{E_1, E_2, \dots, E_v\}$. Assuming E_i contains S_i examples $(i = 1, 2, \dots, m)$, the expected information required by subset E_i is $\text{Entropy}(S_i)$. Therefore:

$$\text{Entropy}(A) = \sum_{i=1}^{v} \frac{S_i}{S} \text{Entropy}(S_i)$$

The information gain using attribute A as the root is:

$$\text{Gain}(A) = \text{Entropy}(s_1, s_2, \dots, s_m) - \text{Entropy}(A)$$

ID3 selects the attribute A^* that maximizes $\text{Gain}(A)$ as the root node, recursively applying this process to the v subsets E_i of E corresponding to different values of A^* to generate child nodes of A^*, thereby constructing a tree.

Using R software with the rpart and rpart.plot packages, risk factors were mined. The final diagnosis of PIH in the training set (i.e., "yes" or "no") was used as the final classification result (i.e., root node), with patient physical examination attribute variables analyzed as classification conditions to display risk factors affecting the final diagnosis and their data ranges through decision tree visualization. Since overly complex decision trees with too many branches are prone to overfitting and lose predictive meaning for test set data, the Complexity Parameter (CP) was used for pruning. CP decreases as decision tree complexity increases. When the change in classification accuracy caused by adding a node is less than CP times the change in decision tree complexity, the node should be pruned. Generally, the CP value corresponding to the minimum misclassification rate is selected for pruning. At CP = 0.0048, a decision tree was obtained that could both fit the training set well and predict the test set well. The final decision tree emphasized four attributes: "systolic blood pressure," "diastolic blood pressure," "fasting blood glucose," and "triglycerides." According to the decision tree path: when systolic pressure > 138 mmHg and diastolic pressure > 92 mmHg and triglycerides > 1.7 mmol/L, these are primary risk factors. When systolic pressure > 138 mmHg but diastolic pressure < 92 mmHg, if fasting blood glucose < 5 mmol/L, diastolic pressure > 86 mmHg, and triglycerides > 2.6 mmol/L, these also constitute PIH risk factors.

(2) Logistic Regression Model for PIH Risk Factor Mining. The Logistic Regression (LR) model most commonly uses gradient descent to obtain the minimum value of the cost function, providing better classification boundaries under certain optimization conditions. Due to its simple structure and easily understandable results, logistic regression is widely applied in disease prevention and represents a typical data mining method in the medical field.

Let P be the probability of an event occurring, with a value range of $[0, 1]$, and $1 - P$ be the probability of the event not occurring. Taking the natural logarithm of $P/(1 - P)$, i.e., the logit transformation of P, denoted as $\text{logit}P$, yields a value range of $(-\infty, +\infty)$. Using P as the dependent variable, a linear regression equation is established:

$$\text{logit}P = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

$$P = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)}{1 + \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)}$$

This model is the logistic regression model, a generalization of ordinary multiple linear regression, but with error terms following a binomial distribution rather than a normal distribution. In the model, α is a constant and β_i $(i = 1, \ldots, m)$ are logistic regression coefficients.

Using R software with the glm function and MASS package, risk factors were mined. Since the logistic regression model has no parameters, parameter adjustment is unnecessary. To obtain a model that fits the training set well and predicts the test set well, appropriate attributes must be selected—these are the risk factors mined by the model. Through reasonable variable selection, an appropriate logistic regression model and its visualization map were obtained. The logistic regression model underwent 8 Fisher scoring iterations, screening out significant attribute variables: "age," "weight (kg)," "systolic pressure (mmHg)," "diastolic pressure (mmHg)," and "fasting blood glucose." Using these five attributes, the most reasonable model for PIH diagnosis was established in R. If these five attributes are represented by X_1-X_5, and Y is the patient' s final disease status (Y can only be 0 or 1), the final logistic regression formula can be expressed as:

$$\text{logit}P = -25.45 + 0.05X_1 + 0.03X_2 + 0.17X_3 + 0.01X_4 + 0.21X_5$$

Whether a patient truly has PIH is calculated according to this formula. By reconstructing the logistic regression model to include these five attributes, both a logistic regression model based on training set data was established and risk factors affecting PIH were identified.

(3) Neural Network Model for PIH Risk Factor Mining. Neural Network is a data mining method comprising an input layer, hidden layer, and output layer. The essence of the neural network method is that results are not related to feature values in the input layer but are closely related to the hidden layer

method. Neural network models can quickly learn arbitrary feature items. Data mining software typically uses backpropagation to minimize the cost function. Neural networks can be applied to classification and regression problems with strong fault tolerance and robustness.

Each neuron in a neural network is a simple computing device whose characteristics are described by simple mathematical functions. Neuron i receives input information from other neurons, performs weighted averaging according to the summation function net_i, and generates output information according to the transfer function f_i, which is then passed to the next neuron according to the network' s topology. Each connection arc is assigned a certain value representing the connection strength. Positive weights indicate increased influence, while negative weights indicate decreased influence. In feedforward networks, neurons are connected forward, neurons in the same layer are not connected, and information can only propagate in one direction. The connection pattern of feedforward networks is represented by the weight vector W. In the network, the weight vector determines how the network responds to any input in the environment. Similarly, the network completes its entire learning process by continuously adjusting weights.

The function proposed by McClelland et al. in 1986 was applied:

$$net_i = \sum_j w_{ij} I_j + Q_i$$

$$x_i' = f(net_i)$$

where I_j is the input to neuron j, w_{ij} is the connection weight between neurons i and j, x_i' is the output of neuron i, and Q_i is the threshold of neuron i.

When using the neural network mining algorithm to process the training set data, the nnet and mlbench packages in R were applied. Through continuous experiments, the number of hidden layers and thresholds were changed to optimize the neural network model. Finally, a neural network model with 10 hidden layers and a threshold of 0.01 was obtained.

(4) PIH Risk Factor Mining Results. Risk factors play a crucial role in diagnosing PIH. The research data contained 16 attributes, but not all contributed to PIH occurrence. This study identified truly effective risk factors through three data mining models: decision tree, logistic regression, and neural network. The specific mining results are shown in .

Comparison of PIH Major Risk Factor Mining Effects

As shown in the table, decision trees can extract attribute combinations and values of risk factors; logistic regression can only analyze risk factor attributes; neural networks cannot obtain attributes or values. Therefore, decision trees demonstrate the best intuitiveness in PIH risk factor mining. Decision trees use the fewest attributes to determine whether a patient has the disease, indicating

the strongest representativeness. Mining these risk factors can assist clinical diagnosis and guide PIH prevention and prognosis.

4. Evaluation and Analysis

4.1 Evaluation Metrics In Big Data analysis, the three data mining models were used to predict test set data. Based on the fourfold table, four metrics were employed to evaluate algorithm performance: Precision, Recall, Accuracy, and F-value.

The specific meanings of each metric are:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$F\text{-value} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In the medical field, TP (True Positive) represents the number of cases where both physician diagnosis and data mining results indicate PIH; TN (True Negative) represents cases where both machine diagnosis and physician diagnosis indicate no PIH; FN (False Negative) represents cases where machine diagnosis indicates PIH but physician diagnosis does not; FP (False Positive) represents cases where machine diagnosis indicates no PIH but physician diagnosis does. Precision reflects algorithm sensitivity, Recall reflects specificity, Accuracy reflects precision, and a higher F-value indicates better overall algorithm performance.

4.2 Prediction Results of Three Models Using the established three data mining models and processed test set data, predictions were made on PIH occurrence. The fourfold table data were used to calculate Precision, Recall, Accuracy, and F-value for model evaluation, as shown in -.

Decision Tree Model Prediction Results for PIH Counts

Logistic Regression Model Prediction Results for PIH Counts

Neural Network Model Prediction Results for PIH Counts

4.3 Performance Comparison of Different Data Mining Algorithms Through comparative research on the different characteristics of decision tree, logistic regression, and neural network algorithms in R during operation, modeling, and prediction, their performance in TP, FP, FN, TN, Precision, Recall,

Accuracy, and F-value was evaluated to provide a basis for algorithm selection in PIH applications, as shown in .

Comparison of Performance Metrics of Three Data Mining Algorithms

As shown in , for Precision, the performance ranking is: Decision Tree > Logistic Regression > Neural Network, which is also their sensitivity ranking. For Recall, the ranking is: Decision Tree > Neural Network > Logistic Regression, which is also their specificity ranking. For Accuracy, the ranking is: Decision Tree > Neural Network > Logistic Regression, which is also their precision ranking. Since Precision and Recall are mutually exclusive metrics that cannot individually evaluate overall algorithm performance, the F-value was used for comprehensive evaluation, yielding: Decision Tree > Logistic Regression Neural Network. Overall, decision trees demonstrate the best performance, neural networks perform slightly better than logistic regression, but the difference is minimal. All three supervised learning algorithms show very strong performance.

4.4 Results Analysis Based on the above mining model establishment and evaluation:

(1) For disease risk factor research, decision trees can extract attribute combinations and values of risk factors, while logistic regression can only analyze risk factor attributes to determine disease status according to formulas, and neural networks cannot provide predictive risk factor possibilities due to their black-box nature. Therefore, decision trees demonstrate the best intuitiveness in PIH risk factor mining. Decision trees use the fewest attributes to determine patient disease status, indicating the strongest representativeness.

(2) For PIH prediction, comprehensive metrics indicate that decision trees perform best for PIH diagnosis, prevention, and prognosis, followed by neural networks, with logistic regression performing worst, possibly due to the binary classification performance of logistic regression and the black-box characteristics of neural networks.

(3) The decision tree algorithm uses the fewest attributes to obtain the optimal model, making it the most suitable algorithm for PIH risk factor mining and final disease diagnosis.

5. Conclusion

Data mining algorithms that extract useful knowledge from Big Data to support decision-making have become one of the most cutting-edge research directions in the knowledge discovery field internationally. Combining data mining algorithms with theories and technologies such as natural language processing, concept mapping, and ontology, the heterogeneous data source knowledge discovery framework established through data collection, data cleaning, model construction, and model evaluation can rapidly achieve intelligence collection and

analysis. As a tool for knowledge discovery from complex information, data mining is no longer limited to pure technical research but is increasingly cross-integrated with other applied disciplines. Therefore, information professionals should embed themselves in disciplines to achieve embedded disciplinary services. The study also found that different data mining algorithms have different effects on different knowledge discoveries, and selection should be targeted to better support decision-making for relevant domain personnel.

References

[1] Zeng Jianxun, Wei Lai. The Changes of Information Science in Big Data Era[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(1): 37-44.

[2] Ackoff R L. From Data to Wisdom[J]. Journal of Applies Systems Analysis, 1980(16): 3-9.

[3] Bellinger G, Castro D, Mills A. Data, Information, Knowledge, and Wisdom[EB/OL]. [2015-11-24]. http://www.systems-thinking.org/dikw/dikw.htm.

[4] Zeleny M. Human Systems Management: Integrating Knowledge, Management and Systems[M]. Singapore: World Scientific, 2005: 15-16.

[5] CIO Network Era. DIKW: Pyramid Hierarchy of Data, Information, Knowledge, Wisdom[EB/OL]. [2014-11-24]. http://www.ciotimes.com.

[6] Wang Yuefen. The Source and Basis of the Methodology of Synthetic Research with Bibliometric Method and Content Analysis[J]. Information Studies: Theory & Application, 2009, 32(2): 21-26.

[7] Wang Liwei, Li Mei, Mu Dongmei, et al. A Knowledge Service-oriented Domain Knowledge Discovery Process[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(1): 45-52.

[8] Xu Ge, Wang Houfeng. The Development of Topic Models in Natural Language Processing[J]. Chinese Journal of Computers, 2011, 34(8): 1423-1436.

[9] He Qing, Li Ning, Luo Wenjuan, et al. A Survey of Machine Learning Algorithms for Big Data[J]. PR&AI, 2014, 27(4): 327-336.

[10] Tang Huifeng, Tan Songbo, Cheng Xueqi. Research on Sentiment Classification of Chinese Reviews Based on Supervised Machine Learning Techniques[J]. Journal of Chinese Information Processing, 2007, 21(6): 88-94, 108.

[11] Hou Yajun. On the Application of R Language in Data Mining[J]. Journal of Jincheng Institute of Technology, 2014, 7(2): 63-65.

[12] Yang Jing, Zhang Nannan, Li Jian, et al. Research and Application of Decision Tree Algorithm[J]. Computer Technology and Development, 2010, 20(2): 114-116, 120.

[13] Hong Jiarong, Ding Mingfeng, Li Xingyuan, et al. A New Algorithm of Decision Tree Induction[J]. Chinese Journals of Computers, 1995, 18(6): 470-474.

[14] Xing Qiuju, Zhao Chunyong, Gao Kechang. Logical Regression Analysis on the Hazard of Landslide Based on GIS[J]. Geography and Geo-Information Science, 2004, 20(3): 49-51.

[15] Wu Lun, Liu Yu, Zhang Jing, et al. Geographical Information System—Theory, Method, Application[M]. Beijing: Science Press, 2001.

[16] Wang Chunfeng, Wan Haihui, Zhang Wei. Credit Risk Assessment in Commercial Banks Using Neural Networks[J]. System Engineering Theory and Practice, 1999(9): 24-32.

[17] McClelland J L, Rumelhart D E, Hinton G E. Parallel Distributed Processing: Explorations in the Microstructure of Cognition[M]. Cambridge, MA: MIT Press, 1986.

[18] Zhang Y, Cui H, Burkell J, et al. A Machine Learning Approach for Rating the Quality of Depression Treatment Web Pages[C]. In: Proceedings of iConference 2014.

[19] Manning C D, Schutze H, Raghavan P. Introduction to Information Retrieval[M]. Translated by Wang Bin. Beijing: Posts & Telecom Press, 2010: 105-107, 196-200.

[20] Zhao Ying. Bayes Analysis of Conditional Logistic Model for Paired Fourfold Table Data[J]. Journal of Mathematical Medicine, 2010, 23(5): 505-506.

Author Contributions

Mu Dongmei: Conceived the research idea, designed the research plan and technical route, and wrote and revised the paper.
Ren Ke: Implemented the research process, performed data cleaning and analysis, and wrote the paper.

Conflict of Interest Statement

Mu Dongmei and Ren Ke used electronic medical records from Changchun Maternity Hospital as supporting data in this study.

Supporting Data

Supporting data [1] can be found in the journal' s online version at htttp://www.infotech.ac.cn; supporting data [2-3] are self-archived by the authors, E-mail: moudm@jlu.edu.cn.

[1] Mu Dongmei, Ren Ke. prog_code.rdf. Experimental environment, program code, and results for disease prediction models.

[2] Mu Dongmei, Ren Ke. trainingData.csv. Training data for pregnancy-induced hypertension prediction models.
[3] Mu Dongmei, Ren Ke. testingData.csv. Testing data for pregnancy-induced hypertension prediction models.

Received Date: 2016-02-19
Revised Date: 2016-03-26

Note: Figure translations are in progress. See original paper for figures.