Construction of a Disease Prediction Model Using Support Vector Machine Based on Electronic Medical Records: A Case Study of Early Warning for Severe Acute Pancreatitis (Postprint)

Authors: Zhang Ye, Zhang Han, Yin Bincan, Zhao Yuhong

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

Abstract

Objective: To construct a disease prediction model, using early warning of severe acute pancreatitis as an example, and propose a method based on support vector machines. Methods: Using LIBSVM3.11 support vector machine with a classifier generated by an optimized radial basis function kernel, combined with statistical univariate and multivariate Logistic regression analysis for feature variable selection, a simple and feasible early warning model for severe acute pancreatitis was proposed. Results: The constructed early warning model achieved an accuracy of 70.37%. The final variables included in the model were white blood cell count, serum calcium ion, serum lipase, systolic blood pressure, diastolic blood pressure, and pleural effusion. Limitations: The sample size was limited, and support vector machine was primarily employed to construct the disease prediction model; a system could be established in the future to highlight clinical application value. Conclusion: Support vector machine can construct optimal models for disease prediction, and further system establishment can assist clinical decision-making.

Full Text

Building Disease Prediction Models Using Support Vector Machines Based on Electronic Medical Records: A Case Study of Early Warning for Severe Acute Pancreatitis

Zhang Ye[1], Zhang Han[1], Yin Bincan[1], Zhao Yuhong[2]

[1] Department of Medical Informatics, China Medical University, Shenyang 110122, China

[2] Shengjing Hospital of China Medical University, Shenyang 110004, China

Abstract: [**Objective**] This study proposes a methodology for constructing disease prediction models using support vector machines, with early warning for severe acute pancreatitis as a case study. [**Methods**] Based on LIBSVM 3.11, we employed an optimized radial basis function kernel classifier combined with univariate and multivariate logistic regression analysis for feature selection to develop a simplified early warning model for severe acute pancreatitis. [**Results**] The constructed severe acute pancreatitis warning model achieved an accuracy of 70.37%. The final variables included in the model were white blood cell count, serum calcium, serum lipase, systolic blood pressure, diastolic blood pressure, and pleural effusion. [**Limitations**] The sample size was limited, and the study primarily used support vector machines to construct the disease prediction model. Future work should establish a systematic implementation to enhance clinical application value. [**Conclusion**] Support vector machines can construct optimal disease prediction models, and further system development can assist clinical decision-making.

Keywords: Support Vector Machine; Severe Acute Pancreatitis; Early Warning; Clinical Decision

Classification Numbers: TP393; G35

Electronic Medical Records (EMR) are computer-based patient records that enable electronic storage, management, transmission, and reproduction of medical data. EMRs primarily include outpatient, inpatient, and emergency records, with inpatient EMRs comprising admission notes, progress notes, surgical records, and laboratory reports. EMRs provide accurate and complete medical information that can alert and remind healthcare providers while offering clinical decision support. The core value of EMRs lies in clinical decision support—applying statistical analysis and data mining methods to assist clinical decisions and enable early disease warning or monitoring of specific outcome events.

With the development of healthcare informatization, demand for EMR-based clinical decision support functions has increased. Establishing clinical decision support systems can reduce misdiagnosis and missed diagnosis in clinical practice while decreasing healthcare resource utilization and addressing medical overcrowding.

Current clinical decision support research primarily includes disease diagnosis and prediction of risk factors or recurrence. Examples include developing gold standard diagnostic criteria for heart failure, predicting Alzheimer' s disease progression, forecasting cardiopulmonary arrest or death events, and creating infectious disease symptom surveillance systems. Although such research has developed rapidly in recent years, most studies aim to establish clinical standards or compare and evaluate existing prediction methods without fully integrating with clinical practice. True clinical decision support involves not merely building prediction models or evaluating methods, but improving physician work quality by shortening diagnosis time, avoiding overtreatment, and reducing medical

errors.

Common decision-making methods include machine learning, statistical analysis, and rule induction. Machine learning primarily employs support vector machines (SVM) and artificial neural networks. Previous studies have used these methods to develop preoperative prediction models for advanced prostate cancer, breast cancer recurrence prediction models, and colorectal cancer liver metastasis prediction models. Support vector machines, proposed by Vapnik et al., represent a robust classification algorithm. Comparative studies across different domains and data types (medical, financial, biological) have demonstrated SVM's advantages in classification performance, generalization ability, and computational efficiency. SVM defines an optimal linear hyperplane and, through Mercer kernel expansion theorem and nonlinear mapping , transforms sample space into a high-dimensional or even infinite-dimensional feature space where linear learning methods can solve highly nonlinear problems. With the embedding of kernel functions, increasing numbers of clinical decision researchers in the medical field have applied SVM to build prediction models. Disease prediction modeling is essentially a classification problem characterized by limited sample sizes and high dimensionality—conditions well-suited for SVM. Unlike traditional statistical methods based on large number laws or other machine learning methods such as artificial neural networks that require large samples and suffer from issues like difficult network structure determination, overfitting, underfitting, and long training times, SVM is specifically designed for small samples. This study constructs a severe acute pancreatitis prediction model based on EMR first-visit records and laboratory test data using SVM, aiming to propose a simplified early warning model for further system development to assist clinical decision-making.

Research Framework and Methods

Disease prediction primarily involves diagnosis, progression, or recurrence prediction, which are essentially binary classification problems of "yes" or "no." The specific workflow for building disease prediction models based on SVM is shown in Figure 1 [Figure 1: see original paper].

The process involves five key steps: (1) Data collection and download: Determine research and outcome variables. Query and download corresponding disease EMRs based on ICD codes from EMR front pages, including first-visit records and laboratory reports. Then include cases meeting inclusion/exclusion criteria and divide them into training and test sets. (2) Data preprocessing: If downloaded records include first-visit text, perform Chinese word segmentation using the NLPIR software from the Institute of Computing Technology, Chinese Academy of Sciences. After initial segmentation, adjust for negation recognition based on rules and professional dictionaries if needed. Extract adjusted symptom feature words (keywords) using NLPIR. (3) Disease prediction model establishment: Using MATLAB 2010a platform and LIBSVM 3.11 toolbox, read dataset Excel files into matrix format using the xlsread function. The

required data format includes sample outcome labels, research variable indices, and variable attribute values. Normalize data to [-1,1] using mapminmax to unify variable scales and simplify calculations. SVM solves nonlinear problems through kernel function selection and parameter tuning. Adjust kernel types (linear, polynomial, radial basis, sigmoid) via parameter t, use grid search to find optimal kernel parameters (c, g), train classification models with optimal parameters, and calculate prediction accuracy using leave-one-out cross-validation. (4) Feature variable selection: Perform univariate statistical analysis (independent samples t-test, two-sample Kolmogorov-Smirnov test, or chi-square test) based on data distribution for preliminary variable screening. Include variables with $P < 0.2$ in logistic regression analysis to select final high-performance predictors, with $P < 0.05$ considered statistically significant. (5) Model reconstruction: Rebuild the disease prediction model based on selected feature variables using SVM and compare whether prediction performance improves.

Classification Model Construction

Study Subjects

We randomly extracted 323 de-identified EMRs with a primary diagnosis of acute pancreatitis from a Liaoning hospital between January 2013 and March 2015, including 203 non-severe cases and 120 severe cases. Inclusion criteria were: inpatient cases with acute pancreatitis as the primary discharge diagnosis and abdominal pain onset within 30 days. Exclusion criteria included transferred or readmitted cases, chronic pancreatitis, pancreatic tumors, and cases with missing research variables.

Research Variables

Referencing risk factors for acute pancreatitis severity listed in the UpToDate clinical database, we selected 20 variables: age, white blood cell count, hematocrit, urea, creatinine, K+, Na+, Ca2+, serum amylase, serum lipase, body temperature, heart rate, respiratory rate, blood pressure (systolic/diastolic), abdominal pain onset time, and presence of consciousness, organ failure, pancreatic necrosis, and pleural effusion.

Outcome Variable

The outcome variable was a discharge diagnosis of severe acute pancreatitis.

Data Preprocessing

Among the included variables, four categorical variables (“consciousness,”“organ failure,”“pancreatic necrosis,”and “pleural effusion”) appeared as text descriptions in first-visit records requiring Chinese word segmentation. Samples were divided into training (n=242) and test (n=81) sets at a 3:1 ratio. For the training set, NLPIR software performed initial segmentation of case characteristics from first-visit records, including present illness, past history, physical examination, and auxiliary examinations. We added user dictionaries to NLPIR containing ICD-10 terms, Chinese MeSH subject headings, and corresponding subject terms from Chinese databases, marking predictive indicator words as “Key.” After initial

segmentation, we adjusted for negation recognition based on rules (manually formulated) and professional dictionaries (standard terms from AP diagnosis and treatment guidelines). NLPIR then extracted adjusted disease severity feature words describing the four categorical variables and marked diagnosis results (SAP or not). For the test set, the same procedures applied except that feature words extracted from the training set were additionally added to the NLPIR user dictionary.

Classification Prediction Model Establishment

We built classification models linking the 20 variables to the outcome variable using training data and calculated prediction accuracy on the test set. Using default parameters, we tested different kernel functions with performance shown in Table 1 . The table shows cross-validation folds (v), support vector count (sv), boundary support vector count (bsv), training accuracy (trA), and test accuracy (teA) for radial basis (RBF) and sigmoid kernels.

Using the RBF kernel with grid search and cross-validation for optimal parameter selection, performance results are shown in Table 2 . According to the latest revised Atlanta classification system, acute pancreatitis is classified into three levels: mild (MAP), moderately severe (MSAP), and severe (SAP).

Feature Variable Selection

We performed statistical analysis using SPSS 19.0 to select high-performance predictive variables. Univariate analysis results are shown in Table 3 . For continuous variables following normal distribution, we used t-tests; for non-normally distributed continuous variables, Kolmogorov-Smirnov tests; and for categorical variables, chi-square tests. Variables with $P < 0.2$ in univariate analysis were included in logistic regression: white blood cell count (WBC), systolic blood pressure (SBP), urea (Urea), sodium (Na+), calcium (Ca2+), temperature (T), heart rate (P), diastolic blood pressure (DBP), serum amylase (AMY), serum lipase (LPS), abdominal pain onset time, organ failure (OF), pancreatic necrosis (PN), and pleural effusion (PE). Logistic regression analysis identified the final feature variables: white blood cell count, serum calcium, serum lipase, systolic blood pressure, diastolic blood pressure, and presence of pleural effusion. The logistic regression equation is shown in formula (1):

$$P = 1/\{1+\exp[-(4.767+0.126\times\mathrm{WBC}-3.142\times\mathrm{CA}+0.001\times\mathrm{LPS}-0.027\times\mathrm{SBP}+0.059\times\mathrm{DBP}-2.157\times\mathrm{PE})]\}$$

Reconstruction of Classification Prediction Model

Using the RBF kernel with grid search and cross-validation, the optimal parameters were c=2 and g=1, with 180 support vectors. Training and test accuracies were 65.29% and 70.37%, respectively. The final decision function is:

$$\mathrm{Predict}_y = \mathrm{sign}\left(\sum_{i=1}^{n} w_i \exp(-\gamma\|x_i - x\|^2) + 0.388\right)$$

where $\|x_i - x\|^2$ is the Euclidean distance, n represents the number of support vectors (180), w_i = model.sv_coef(i) are the support vector coefficients, and x_i = model.SVs(i, :) is the support vector matrix. X is the sample to be predicted, and gamma is parameter g.

Classification Model Performance Evaluation

We objectively evaluated model accuracy using test set data, with SAP cases as positive and non-SAP as negative classes. Accuracy was calculated as:

$$\text{Accuracy}(A) = (\text{True Positives} + \text{True Negatives})/\text{Total Samples}$$

In the 81 test samples (30 positive, 51 negative), there were 14 true positives, 16 false negatives, 8 false positives, and 43 true negatives, yielding an accuracy of 70.37%.

Comparative Experiments

We compared SVM with logistic regression, decision trees, and artificial neural networks using identical training and test sets. WEKA 3.7.13 software package was used for implementation. WEKA is a Java-based data mining and machine learning software including complete data processing tools, learning algorithms, and evaluation methods. For classification problems, different algorithms (Bayesian, decision trees, multilayer perceptron, support vector machines) can build different classifiers. We selected "Logistic," "J48," and "Multilayer Perceptron" under the "Classifier" menu with default parameters, using leave-one-out cross-validation on the training set. Table 4 compares prediction accuracies of the three methods with SVM before and after feature selection.

Results and Discussion

Table 1 shows that different kernel functions yield varying prediction accuracies. For the test set, polynomial and RBF kernels achieved higher accuracies with ideal support vector counts. We selected SVM for building the severe acute pancreatitis early warning model because SVM' s kernel-based dimension elevation solves nonlinear classification and regression problems. The final decision function depends only on a few support vectors, making computational complexity depend on support vector count rather than variable number. Compared with other methods, SVM requires less human intervention, ensuring model objectivity. SVM commonly uses RBF kernels for nonlinear problems due to their suitability for nonlinear relationships, favorable model complexity, and numerical implementation feasibility. Grid search with cross-validation may not yield theoretically optimal results but achieves satisfactory optimization. Table 2 demonstrates that parameter optimization not only improves accuracy but also reduces support vector count, simplifying the prediction function.

The SAP prediction model based on SVM achieved satisfactory accuracy, demonstrating SVM' s applicability for disease prediction modeling. Kernel selection and parameter optimization can improve prediction performance. Although SVM models show good performance, not all variables are necessarily highly correlated with outcomes, making feature selection crucial post-modeling.

We employed different univariate analysis methods based on data distribution to preliminarily screen variables, eliminating non-highly correlated or redundant variables while reducing sample size requirements for logistic regression. Selected variables reflect inflammatory responses, characteristic enzyme changes, and complications affecting AP severity progression. Rebuilding the prediction model with selected variables improved accuracy, demonstrating that feature selection simplifies dimensionality, eliminates irrelevant data, reduces training sample requirements, and enhances model prediction performance.

In comparative experiments (Table 4), logistic regression, decision trees, and neural networks showed higher training set accuracies but lower test set accuracies than SVM. This occurs because these three methods minimize empirical risk, potentially causing overfitting. Disease prediction modeling ultimately seeks robust classification models. SVM, based on structural risk minimization, balances average prediction error with model complexity, minimizing the sum of empirical risk and confidence intervals to produce robust models with ideal test set accuracy.

With hospital informatization development, EMR' s core value—clinical decision support—will become a future development direction. This study constructed a disease prediction model based on SVM using severe acute pancreatitis as an example. Study characteristics include: (1) using both textual and numerical medical data to establish early warning models for potential decision system development; (2) combining SVM with statistical analysis for feature selection before building the final model, improving accuracy while simplifying the model. Limitations include insufficient clinical application. Future research will create decision systems based on the established model to highlight clinical value. Additionally, increasing sample sizes and applying decision methods to build disease prediction models that serve clinical needs represent directions for clinical decision support researchers.

References

[1] Lei Jianbo. Clinical Decision Support and the Core Value of Electronic Medical Record [J]. China Digital Medicine, 2008, 3(3): 26-30.

[2] Byrd R J, Steinhubl S R, Sun J, et al. Automatic Identification of Heart Failure Diagnostic Criteria, Using Text Analysis of Clinical Notes from Electronic Health Records [J]. International Journal of Medical Informatics, 2014, 83(12): 983-992.

[3] Ye J, Farnum M, Yang E, et al. Sparse Learning and Stability Selection for Predicting MCI to AD Conversion Using Baseline ADNI Data [J]. BMC Neurology, 2012. DOI: 10.1186/1471-2377-12-46.

[4] Alvarez C A, Clark C A, Zhang S, et al. Predicting out of Intensive Care Unit Cardiopulmonary Arrest or Death Using Electronic Medical Record Data [J]. BMC Medical Informatics and Decision Making, 2013. DOI: 10.1186/1472-6947-13-128.

[5] Matheny M E, Fitzhenry F, Speroff T, et al. Detection of Infectious Symptoms from VA Emergency Department and Primary Care Clinical Documentation [J]. International Journal of Medical Informatics, 2012, 81(3): 143-156.

[6] Kim S Y, Moon S K, Jung D C, et al. Pre-Operative Prediction of Advanced Prostatic Cancer Using Clinical Decision Support Systems: Accuracy Comparison between Support Vector Machine and Artificial Neural Network [J]. Korean Journal of Radiology, 2011, 12(5): 588-594.

[7] Kim W, Kim K S, Lee J E, et al. Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine [J]. Journal of Breast Cancer, 2012, 15(2): 230-238.

[8] Lv Yi, Wang Qing. A Probability Calibration and Ensemble Learning Based Colorectal Cancer Liver Metastasis Prediction Model [J]. Computer Applications and Software, 2011, 28 (9): 48-51.

[9] Wang Xing, et al. Big Data Analysis: Methods and Applications [M]. Beijing: Tsinghua University Press, 2013: 68-90.

[10] Chen Yongyi, Xiong Qiufen. Application of Support Vector Machines Tutorial [M]. Beijing: China Meteorological Press, 2011: 6-10.

[11] ICTCLAS 2014 [EB/OL]. [2015-03-25]. http://ictclas.nlpir.org/.

[12] LIBSVM—A Library for Support Vector Machines [EB/OL]. [2015-03-25]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[13] Up To Data [EB/OL]. [2015-03-25]. http://www.uptodate.com/contents/search.

[14] Banks P A, Bollen T L, Dervenis C, et al. Classification of Acute Pancreatitis–2012: Revision of the Atlanta Classification and Definitions by International Consensus [J]. Gut, 2013, 62(1): 102-111.

[15] Yuan Meiyu. Data Mining and Machine Learning—WEKA Application Technology and Practice [M]. Beijing: Tsinghua University Press, 2014: 2.

[16] Liu Kan, Zhu Huaiping, Liu Xiuqin. Detection of Internet Deceptive Opinion Based on SVM [J]. New Technology of Library and Information Service, 2013(11): 75-80.

Author Contributions:
Zhang Ye: Designed the research protocol, conducted experiments, collected and analyzed data, drafted and revised the manuscript.

Zhang Han: Proposed research ideas, revised the manuscript.
Yin Bincan: Collected and analyzed data.
Zhao Yuhong: Proposed research ideas, revised the manuscript.

Conflict of Interest Statement: All authors declare no conflicts of interest.

Supporting Data:
The supporting data is self-archived by the authors and can be obtained via email: 1332457636@163.com.
[1] Zhang Ye, Zhao Yuhong. ap.xls. Variable data of included acute pancreatitis EMR samples.
[2] Zhang Ye, Zhao Yuhong. apdic.txt. Acute pancreatitis professional dictionary.

Received: 2015-09-21
Revised: 2015-12-07

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