Forecasting: Adopting the Methodology of Support Vector Machines to Nursing Research

Huey-Ming Tzeng, RN, PhD

In nursing studies, linear statistical methods are commonly used to analyze data on subjective perceptions. Most of the collected data include participants' subjective perceptions and are usually not collected in a controlled environment. For linear regression analyses, there is a common understanding that the distributions of some studied variables might not be close to linear in nature. However, when using the support vector machine (SVM), there is no violation concern for the regression assumption of linearity.

SVM, as an artificial intelligence learning method, is based on the structural risk minimization principle that relies on the computational learning theory. It is a powerful technique for nonlinear regression or pattern recognition, which can learn complex patterns or trends from input data (independent variables) and create outputs (dependent variables). A trained learning machine can then generate output for unseen data. The quality of SVM was determined by the machine's training error rate and generation capability. Choosing the approaches of nonlinear regression or pattern recognition is based on the characteristics of the dependent variable (Cristianini & Shawe-Taylor 2000; Figure 1).

How does SVM differ from the statistical methods? How could nursing researchers adopt SVM to address nursing issues? In this commentary, these two questions are answered. The process of developing SVM and three example studies using SVM are introduced in the following.

ADVANTAGES OF UTILIZING SVM

Characteristics of SVM

When adopting the method of SVM, it is not necessary to create dummy variables while dealing with categorical variables. In addition, there is no limitation in the number of independent variables for SVM. The developed SVM will be influential and the accurate forecasting rate will be improved only when meaningful predictors are selected.

Moreover, SVM allows a nonlinear mapping of the input space to a higher dimensional feature space; the input data can be constructed into a multidimensional space. For example, if there are five independent variables and one dependent variable, this SVM would be constructed into a five-dimensional space.

High Accuracy Rates in Forecasting

Previous studies in engineering and biology have demonstrated that SVM is an effective tool with over 70% prediction accuracy rate and can capture the underlying structure of the data in a nonlinear fashion (Tzeng et al. 2004). These developed SVMs usually have good generalization capability and need only limited samples.

PROCESS OF TRAINING AND VALIDATING SVMs

The process of developing an SVM includes three major steps: data preprocessing, variable selection, and training a learning machine. A brief overview of these three steps is as follows (Figure 1).

Step 1: Data Preprocessing

Since SVMs are unable to deal with data information with missing values, cases that have any missing values have to be eliminated. As a common practice, the original dataset usually has to be divided into two groups: (1) the training dataset for training the SVM and (2) the testing dataset for validating the learning machine trained by the training dataset.

Step 2: Variable Selection

SVMs do not have any limitations on the types of data and the number of included independent variables. It is important to select the independent variables that have demonstrated some relational trends with the dependent

Huey-Ming Tzeng, Associate Professor, Division of Nursing Business and Health Systems, School of Nursing, The University of Michigan, Ann Arbor, MI
Address correspondence to Huey-Ming Tzeng, Division of Nursing Business and Health Systems, School of Nursing, The University of Michigan, 400 N. Ingalls, Room 4170, Ann Arbor, MI 48109-0482; tzeng_hueyming@yahoo.com

Accepted 12 May 2006.

1545-102X1/06

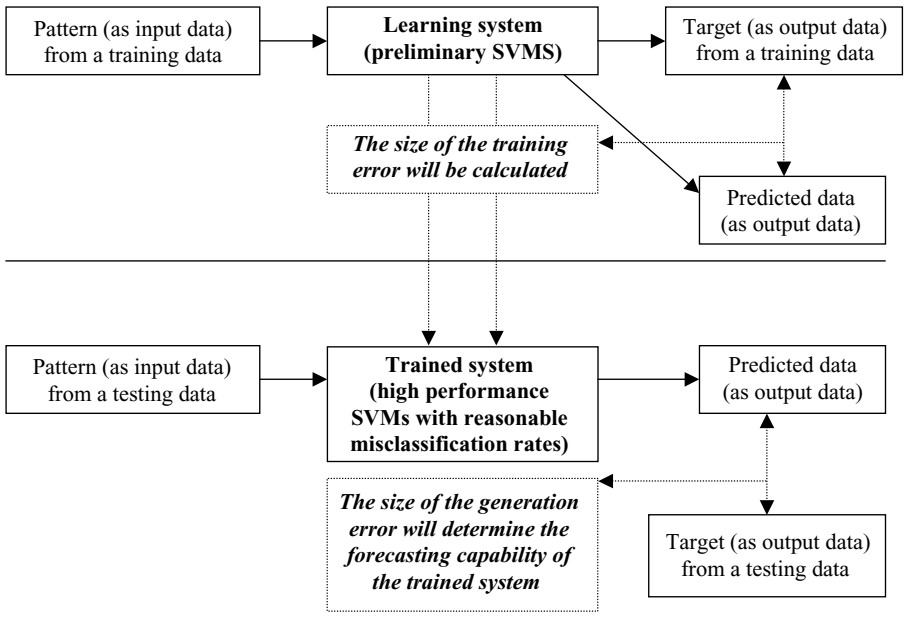

Figure 1. The procedures of training a support vector machine and validating the trained learning machine. This process is applicable to either classification or regression problems.

variable through prior univariate or multivariate statistical analysis.

Step 3: Training a Learning Machine

The power of SVMs comes from the kernel representation. The choice of a proper kernel function is crucial for a better regression/classification result and achieving a low test-error rate. In the training process, after a certain kernel function is chosen, the kernel width *C* requires tuning.

For instance, in a binary classification problem, for a large value of *C*, the classifier in a classification model might not be able to separate the dataset without errors; for a small width, the dataset might be practically memorized; and for an intermediate width, the classifier might allow some training errors (Cai et al. 2002). Both the kernel parameters and the value of *C* often are chosen using the procedure of cross-validation (Schölkopf & Smola 2002).

In order to divide data into training and testing datasets, it would be appropriate to first split the dataset into several, say *m* parts of approximately equal size, and then perform *m* training runs. Each time, one of the *m* parts is left out and is used as an independent validation set for optimizing the parameters. Taking into consideration both the average training and testing error rates, the parameters with acceptable results on average over the *m* runs would be chosen. Then, the SVM would be trained on the full training set, using these average parameters (Schölkopf & Smola 2002).

THREE EXAMPLE STUDIES THAT ADOPTED THE METHOD OF SVM

Example 1: The Occurrence of Violent Behavior Tendencies for Schizophrenic Outpatients

The study by Tzeng et al. (2004) used the method of SVM to build a forecasting model that explored the predictive value of disease insight toward the violent behavior (yes/no) in a group of 63 Taiwanese schizophrenic outpatients over a one-year period. This SVM was a classification problem for pattern recognition. The binary logistic regression model was built first, which explained 65.2% of the variance of patients' violent behavior tendency (Nagelkerke R^2 value). Then, an SVM was developed resulting in a 76.2% predictive power.

Example 2: The Occurrence of Suicidal Tendencies for Schizophrenic Outpatients

Using the same dataset as the first example, this study trained and validated an SVM for predicting the incidence of suicidal tendencies as a classification problem (see Table 1 for predictors). The binary logistic regression model was built first, which explained 46.4% of the variance of patients' suicidal tendency (Nagelkerke R^2 value).

An SVM was also acquired; there were no misclassifications in the training data, while the average percentage of misclassification in the testing data was 25.4%, resulting in a 74.6% predicting power. The patient's insight rating, medication compliance, and demographic characteristics

TABLE 1
Definitions of the Variables

VARIABLE	DEFINITION
Duration of illness	Number of years since being diagnosed as schizophrenia until the first interview
Time of hospital admission	Number of hospital admissions since being diagnosed as schizophrenia until the first interview
Education	Number of years of education
Age	Age at the follow-up interview in years
Gender	0 = male 1 = female
Marital status	1 = married and living together, unmarried but living together with his/her partner, married but being separated, remarried 2 = single, married but being separated, divorcee, widower
Occupation	1 = no job 2 = being employed, student, serving military obligation, retired
Religion	1 = no religion 2 = having a religion
Medication compliance	The measurement of this concept was translated from the one developed by Blackwell as cited in the study of Kemp and David (1996). Subjects were asked to respond to their compliance from the aspects of right medication, right time to take medication, right dosage, missing dosages, stopping taking medication, changing dosage, and changing frequency and time for taking medication without psychiatrist's advice. These seven items were measured on a 4-point Likert scale (1 = never; 2 = sometimes; 3 = usually; 4 = always). The mean of these seven items was calculated
Insight 1: Awareness and description of psychotic symptoms	Insight 1 includes two items: (1) Is the patient aware of his/her psychotic symptoms? (2) Can the patient describe his/her psychotic symptoms?
Insight 2: Ability to recognize and respond appropriately to early symptoms of relapse	Insight 2 is composed of two items: (1) Can the patient recognize his/her early symptoms of relapse? (2) Can the patient respond appropriately to the early symptoms of relapse?
Insight 3: Awareness and etiology attribution of having schizophrenia	Insight 3 has two items: (1) Is the patient aware that he/she has schizophrenia? (2) Can the patient have reasonable etiology attribution of having schizophrenia as related to his/her social, cultural and educational background?
Insight 4: Awareness of achieved effect of treatment and likely compliance with treatment	Insight 4 includes two items: (1) Does the patient believe that the treatment (such as medication and hospitalization) has effect on himself/herself? (2) Does the patient believe that it is necessary for him/her to receive medication continuously?
Insight 5: Awareness of the change in life after having schizophrenia	Insight 5 has only one item: (1) Is the patient aware of the relationship between the change in life and having schizophrenia? The change in life involves being involuntarily hospitalized, deterioration of occupational or social function, poor interaction with families, and poor daily ability to take care of himself/herself
(Target variable) Having suicidal behaviors or ideas	The Chinese version (Ho et al. 1995) of the VASA (Feinstein & Plutchik 1990) was used to assess suicidal tendencies over the previous year. Patients were divided into those who had tried or planned to commit suicide and those who had not (1 = who had exhibited suicidal tendencies; 0 = who had not)

Note. The information for Insight 1 to Insight 5 was collected through a semi-structured interview by a psychiatrist, with a 4-point Likert scale (1 = true insight; 2 = somewhat insight; 3 = somewhat denial; 4 = complete denial; missing = unable to evaluate or missing value). The mean of included items was calculated.

of these schizophrenic outpatients explained 74.6% of the total variance of the subsequent incidence of suicidal tendencies over a one-year period.

Example 3: Predicting Nurses' Intention to Quit

Tzeng et al. (2004) built up an SVM as a regression problem for predicting nurses' intention to quit, using working motivation, job satisfaction, and stress level. The ordinal logistic regression model was built first in an earlier study, which explained 41.0% of the variance of nurses' intention to quit (Nagelkerke R^2 value; Tzeng 2002).

For cross-validation, Tzeng and colleagues randomly split cases into four groups of approximately equal sizes and performed four training runs. After the training process, the average percentage of misclassification in the training data

was 0.86%, concurrently, which on the testing data resulted in 89.20% predicting power.

HOW SVM MIGHT BE HELPFUL FOR EVIDENCE-BASED PRACTICE-RELATED ISSUES

Utilizing SVMs in an evidence-based practice has been considered as an efficient approach for processing clinical problems of nonlinear regression or pattern recognition. Choosing the approaches of nonlinear regression or pattern recognition is based on the nature of the dependent variable. This SVM approach is interesting and intriguing to social science researchers. However, its mathematical nature and the SVM languages are less familiar to the nursing society than the statistical programs such as SPSS and SAS.

Limitations

In nursing studies, for testing the predicting ability of statistical models, linear or nonlinear regression analysis (including log linear, logic or categorical regression models, LISREL, hierarchical linear modeling, etc.) are commonly used. These regression analyses illustrate the predicting directions from independent variables to the predetermined dependent variables, and the amount of variances being explained.

When using multiple regression analyses, researchers' focus, however, is not usually on forecasting the values of the dependent variables by unseen data. In addition, prior to carrying on these analyses, researchers usually have to examine the characteristics of included variables for any violation to regression assumptions that include linearity of the phenomenon measured, constant variance of the error terms, independence of the error terms, and normality of the error term distribution. The linearity of the relationship between the dependent variable and the independent variable is usually assumed in linear regression analyses. Remedies through transforming the data to achieve linearity might be used. If appropriate, it is also possible to choose nonlinear regression analyses.

Much data collected in nursing studies include participants' subjective perceptions, which were usually not collected in a controlled or laboratory environment. It is possible that the distribution of our data might not be even close to linear in nature. For example, an independent variable might have a curvilinear effect or more complex nonlinear relationship with the dependent variable.

Opportunities

The finding of Example 2 is used for illustrating the implications for nursing and evidence-based practice using SVM. The developed SVM results in a 74.6% predicting power, which is in equilibrium to R^2 in regression models; thus, the binary logistic regression model explained only 46.4% of the variance of patients' suicidal tendency.

As shown in Example 2, based upon human experts' judgment, patient demographic, and clinical data, such an SVM approach would help in establishing an early-warning, evidence-based mechanism for detecting suicidal tendencies of community-based schizophrenic patients. Community nurses as well as the nurses working in hospital-based outpatient services could continuously improve such a kind of SVM model and use the SVM model as an expert-artificial intelligence system on a regular basis in their practice. The caring experience of these nurse experts should be documented as input data to increase the predicting power of an SVM model for early detection of suicidal tendencies of community-based schizophrenic patients.

CONCLUSIONS

For nursing research, the highly predictive power of SVM and its forecasted result might be critical and helpful in making appropriate and in-time treatment or administrative decisions in clinical settings. Documentation of evidence-based caring experiences is important in converting nursing knowledge into an artificial intelligence system, that could later on serve as an *ad hoc* system in monitoring.

Thus, the method of SVM should be adopted along with univariate or multivariate analysis in the process of variable selection for the best model performance. Despite the fact that SVM could generate an artificial intelligence system with a high predictive power for forecasting, this method would not replace conventional statistical analyses.

References

Cai, Y.D., Liu, X.J., Xu, X.B. & Chou, K.C. (2002). Prediction of protein structural classes by support vector machines. *Computers & Chemistry*, *26*(3), 293–296.

Cristianini, N. & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge: Cambridge University Press.

Feinstein, R. & Plutchik, R. (1990). Violence and suicide risk assessment in the psychiatric emergency. *Comprehensive Psychiatry*, *31*, 337–343.

Ho, H., Yin, C.C., Hwu, H.G. & Tsuang, M.M. (1995). Violence and suicide assessment scale: A reliability and validity study. *Chinese Psychiatry-R.O.C.*, *9*, 122–128.

Kemp, R. & David, A. (1996). Insight and compliance. In Blackwell B. (eds), *Treatment Compliance and the Therapeutic Alliance*. Newark, NJ: Gordon and Breach Publishing Group, pp. 61–84.

Schölkopf, B. & Smola, A.J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Tzeng, H.M. (2002). The influence of working motivation and job satisfaction on intention to quit: An empirical investigation on nurses. *International Journal of Nursing Studies*, *9*(8), 867–878.

Tzeng, H.M., Hsieh, J.G. & Lin, Y.L. (2004). Predicting the influence of working motivation and job satisfaction on nurses' intention to quit with a support vector machine: A new approach to set up an early warning in human resource management. *Computers, Informatics, Nursing*, *22*(4), 232–242.

Tzeng, H.M., Lin, Y.L. & Hsieh, J.G. (2004). Forecasting violent behaviors for schizophrenic outpatients using their disease insights: Development of a binary logistic regression model and a support vector model. *International Journal of Mental Health*, *33*(2), 17–31.