HIR
Healthcare Informatics Research

Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients

Youn-Jung Son, RN, PhD[1], Hong-Gee Kim, PhD[2], Eung-Hee Kim, ME[2], Sangsup Choi, PhD[2], Soo-Kyoung Lee, RN, Doctoral Candidate[2]

[1]Department of Nursing, Soonchunhyang University, Cheonan; [2]Biomedical Knowledge Engineering Laboratory, Seoul National University, Seoul, Korea

Objectives: Heart failure (HF) is a progressive syndrome that marks the end-stage of heart diseases, and it has a high mortality rate and significant cost burden. In particular, non-adherence of medication in HF patients may result in serious consequences such as hospital readmission and death. This study aims to identify predictors of medication adherence in HF patients. In this work, we applied a Support Vector Machine (SVM), a machine-learning method useful for data classification. **Methods:** Data about medication adherence were collected from patients at a university hospital through self-reported questionnaire. The data included 11 variables of 76 patients with HF. Mathematical simulations were conducted in order to develop a SVM model for the identification of variables that would best predict medication adherence. To evaluate the robustness of the estimates made with the SVM models, leave-one-out cross-validation (LOOCV) was conducted on the data set. **Results:** The two models that best classified medication adherence in the HF patients were: one with five predictors (gender, daily frequency of medication, medication knowledge, New York Heart Association [NYHA] functional class, spouse) and the other with seven predictors (age, education, monthly income, ejection fraction, Mini-Mental Status Examination-Korean [MMSE-K], medication knowledge, NYHA functional class). The highest detection accuracy was 77.63%. **Conclusions:** SVM modeling is a promising classification approach for predicting medication adherence in HF patients. This predictive model helps stratify the patients so that evidence-based decisions can be made and patients managed appropriately. Further, this approach should be further explored in other complex diseases using other common variables.

Keywords: Heart Failure, Medication, Patient Adherence, Support Vector Machine

I. Introduction

The support vector machine (SVM) is a relatively new classification or prediction method developed by Cortes and Vapnik [1] in the 1990s as a result of the collaboration between the statistical and the machine learning research community. SVM tries to classify cases by finding a separating boundary called hyperplane. The main advantage of the SVM is that it can, with relative ease, overcome 'the high dimensionality problem', i.e., the problem that arises when there is a large number of input variables relative to the number of available observations [2]. Also, because the SVM approach is data-driven and possible without theoretical framework, it

may have important discriminative power for classification, especially in cases where sample sizes are small [3]. This technique has recently been used to improve methods for detecting diseases in clinical settings [4,5]. Moreover, SVM has demonstrated high performance in solving classification problems in bioinformatics [6,7]. In this study, we use SVM to predict medication adherence of HF patients using common variables that can be obtained with relative ease.

Over the last several decades, HF has increased in both incidence and prevalence worldwide [8]. HF is a condition characterized by unpleasant symptoms, high mortality, recurrent hospitalization and significant cost burden [9,10]. HF is one of the many chronic conditions that require patients to adhere to a lifelong therapeutic regimen such as medication, exercise, diet, fluid restriction and daily weight measurements to achieve optimal outcomes [11]. Especially, patients with HF require multiple medications to decrease morbidity and mortality and to attain maximal therapeutic benefits [10,12]. Estimates of medication adherence rates in patients with HF range from 7% to 90%, depending on the definition and how adherence is measured [10]. Adherence rates decrease as patients stay longer with the condition. Medication nonadherence is thought to be the most common cause of HF exacerbation and subsequent hospital readmission in patients with HF. Because of the importance of medication adherence in managing HF, a full understanding of the factors associated with medication adherence in patients with HF is needed so that effective interventions to improve medication adherence can be developed [13].

In Western countries, studies related to application of SVM in cardiovascular patients have been continuously conducted [14,15]. On the other hand, in Korea, classification algorithm or pattern analysis was mainly focused on Bayesian and artificial neural networks in broad healthcare domain [16,17]. Up to date, SVM have not yet been studied for prediction of medication adherence in HF patients in Korea nor in other countries. Therefore, the aim of this study is to investigate the use of a SVM based classification model for determining the predictors of medication adherence in HF patients.

II. Methods

1. Data and Data Preprocessing
For the model building, Ninety four consecutive HF patients who visited at outpatient clinics from January to April 2010 participated in the study. As eighteen patients with 1 or more missing data points were excluded, 76 (21 men, 55 women) were used for SVM models. The mean age of the HF patients was 74.8 years (SD, 5.9). Table 1 shows the demographic and

Table 1. Demographic and clinical characteristics of patients with heart failure (n = 76)

Characteristics	Category	No. (%)	Mean ± SD
Gender	Men	21 (27.6)	
	Women	55 (72.4)	
Age (yr)			74.8 ± 5.9
Spouse	No	39 (51.3)	
	Yes	37 (48.7)	
Education	Illiteracy	38 (50.0)	
	Elementary school	23 (30.3)	
	Middle school	6 (7.9)	
	High school	7 (9.2)	
	Above college	2 (2.6)	
Monthly income	None	53 (69.7)	
(1,000 won)	1-490	14 (18.4)	
	500-990	4 (5.3)	
	1,000-1,499	2 (2.6)	
	1,500-1,999	2 (2.6)	
	≥2,000	1 (1.3)	
MMSE-K			21.5 ± 4.1
Duration of heart failure diagnosis (mo)			32.1 ± 38.0
Ejection fraction (%)			57.8 ± 14.6
NYHA	I	20 (26.3)	
functional	II	39 (51.3)	
class	III/IV	17 (22.4)	
Daily frequency	1	26 (34.2)	
of medication	2	41 (53.9)	
(per day)	3	9 (11.8)	
	≥4	0 (0)	
Medication knowledge			12.8 ± 3.3

MMSE-K: Mini-Mental Status Exam-Korean, NYHA: New York Heart Association.

clinical characteristics of the patients with HF.

The study was approved by the ethics committee of the Soonchunhyang University Hospital prior to the start of the data collection. The committee waived informed consent because of the non-interventional study design.

2. Measurements
The questionnaire was designed to yield information about

demographic characteristics such as age, gender, education, spouse and monthly income. To assess ability of cognition, we used Mini Mental Status Examination–Korean version (MMSE-K) by interviewing. Clinical data were obtained from medical records. New York Heart Association (NYHA) functional class was used as an index of functional impairment. The NYHA functional class identifies patients in one of four categories based on physical symptoms and activity restriction: Class I is no symptoms with ordinary activity and no limitations on activity; Class II is slight to moderate symptoms with normal activity and slight limitation of activity; Class III is moderate symptoms with less than normal activity and marked limitation of activity; Class IV is inability to carry out any physical activity without discomfort and symptoms may occur ar rest. Ejection fraction (EF) was used as an index of cardiac function. EF was determined by left ventricular angiography or echocardiography.

We developed five self-report questions to assess patients' medication knowledge. These items asked how well the patients knew the names, purposes, recommended doses, frequencies, and side effects of their medications. The 5 items with a 5-point Likert type scale was used. Internal consistency measured by Cronbach's alpha was 0.83. A simple single item to measure medication adherence was used. The patient was categorized as either compliant or non-compliant.

3. Variable Selection

A total of 11 variables were used in the model building and analysis. The variables were selected because they either had shown to have an impact on medication adherence in HF patients in previous research [8-10,12,13] or were of potential clinical importance as indicated by a panel of experts. The 11 variables were gender, age, spouse, education, monthly income, and duration of HF diagnosis, daily frequency of medication, EF, MMSE-K, medication knowledge, and NYHA functional class.

4. Analytical Tool – SVM Basics

Among various data mining methods, SVM is well known for its discriminative power for classification, especially in the cases where sample sizes are small and a large number of features (variables) are involved (i.e., high-dimensional space). Kim et al. [18] demonstrated an outstanding performance of SVM on the classification for prognosis prediction of Class III malocclusion in comparison to a conventional statistical method, stepwise Guyon et al. [19] also showed that SVM performs much better than correlation based techniques in a gene selection problem for cancer classification with the feature selection method, recursive feature elimina-

tion (RFE). In addition, in comparing SVM with logistic regression in predicting diabetes and pre-diabetes, Yu et al. [3] demonstrated that SVM can be a promising tool.

To make computers automatically learn to identify complex pattern and enable intelligent decision making, machine learning techniques have been developed. Machine learning techniques are broadly classified into two types: supervised learning and unsupervised learning. In supervised learning, generally used for classification, an algorithm is provided. On the other hand, in unsupervised learning, generally used for clustering, no learning algorithms are provided.

SVM is one of the most well-known supervised machine learning algorithms for classification. For a given set of training data, each marked as belonging to one of two categories, SVM training algorithm develops a model by finding a hyperplane, which classifies the given data as correctly as possible by maximizing the distance between two data clusters [1].

5. Development of the SVM Based Models

In practice, however, it is frequently not possible to clearly separate the given data set because some of the data points in the two classes might fall into gray area that is not easy to separate linearly. As one of the solutions for this problem, data are mapped to a higher dimension such that the two classes could be separable in the higher dimension (called kernel function). Consider a training set of vector-label pairs (x_i, y_i), i = 1, …, l where $x_i \in R^n$ and $y_i \in \{1, -1\}$ where xi is a vector in an n-dimensional space. If the training set is linearly separable, there exist infinitely many hyperplanes f(x) = $w_i x_i$ + b which correctly classify all vectors in the training set, i.e. sign(f(x_i)) = y_i. For n-dimensional space, the hyperplane will be n-1 dimensional. The objective of SVM is to choose the optimal hyper-plane that clearly separates vectors in the training set into two groups, +1 or -1, and maximizes the distance (margin) between the hyperplane and the support vectors. In summary, the purpose of SVM can be regarded as the solution of the following optimization problem:

$$\min_{w,b,\varepsilon} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \varepsilon_i$$ subject to y_i ($w^T \emptyset(x_i) + b$) ≥ 1-ε_i, where $\varepsilon_i \geq 0$, where the training data are mapped to a higher dimensional space by the function $\emptyset$, and C is a penalty parameter on the training error. Therefore, for any testing vector x, the decision function (predictor) is $f(x) = sign(w^T \emptyset(x) + b)$.

In addition, as mentioned above, most training sets used in a variety of domains are not linearly separated so that it is hard to derive the optimal hyperplane correctly classifying vectors in two classes. To solve this non-linearity problem, several solutions called kernel functions have been proposed

and adopted for SVM. A kernel function is written $K(x_i, x_j) \equiv \emptyset(x_i)^T \emptyset(x_j)$ and the most widely used four kernel functions are

Linear: $K(x_i, x_j) = x_i^T x_j$

Polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$

Radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$

Sigmoid: $K(x_i, x_j) = \tan h(\gamma x_i^T x_j + r)$.

where γ, r are kernel parameters.

6. SVM Simulation Method

We used LIBSVM [20], a freely available SVM software library, to generate the SVM models. To generate the data set for model training, we randomly selected non-cases to match the number of cases in the training data set. To meet the format requirement of SVM, the collected data are transformed as follows. Values of selected input features (equivalent to 'input variables' in regression analysis) were normalized to values from -1 to +1. Values of categorical variables such as Education are arbitrarily assigned to numbers between -1 and +1. For example, -1, -0.5, 0, 0.5, 1 represents illiteracy, elementary school graduates, middle school graduates, high school graduates, college graduates, respectively. Values of continuous variables were transformed into values between -1 and +1 by dividing them by an appropriate number. In the training data set, the last column of the input data was set to the known outcome, i.e., 1 for positive, -1 for negative. Different kernel functions, including linear, polynomial, sigmoid, and radial basis functions (RBF), were tested and selected for the models on the basis of performance.

7. Performance Evaluation

To evaluate the robustness of the estimates from the SVM models, the leave-one-out cross-validation (LOOCV) was performed. In LOOCV, one test sample is extracted from a total of n samples. This test sample is then used for calculating the classification accuracy of the remaining n-1 training samples, and this process is repeated n times. That is, LOOCV is a special case of k-fold cross-validation where k is the same as the number of samples in training data set [3,18,21]. To generate summary performance estimates, we used the statistics including sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the cross-validations.

Five widely used statistics were adopted to evaluate the performance of a model: sensitivity, specificity, PPV, NPV, and accuracy.

$$Sensitivity = \frac{TP}{(TP + FN)}, \; Specificity = \frac{TN}{(TN + FP)}$$

$$PPV = \frac{TP}{(TP + FP)}, \; NPV = \frac{TN}{(TN + FN)},$$

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

where TP, FP, TN, and FN refer to the number of true positives, false positives, true negatives, and false negatives statuses, respectively.

III. Results

1. Feature Selection

To build an optimal solution which identifies predictors of medication adherence in HF patients using our questionnaire survey data, we applied SVM for all possible feature-combinations from the dataset. We generated all possible combinations(2^{11} - 1 = 2,047) of 11 features (gender, age, spouse, education, monthly income, duration of HF diagnosis, daily frequency of medication, ejection fraction and MMSE-K, medication knowledge, and NYHA functional class) and found out combinations that represent the highest classification accuracy.

Feature selection technique is one of the key issues in data mining for reducing classifier-building time as well as increasing the performance of classifiers. Fortunately, our sample data consists of 11 dimensional vectors which have relatively small dimensionality, therefore it allowed us to build classifiers for 2,047 (2^{11} - 1) combinations of features and to evaluate their performance in a reasonable time.

2. Cross-Validation Test

Consequently, using the four kernel functions, four models for partitioning people into two categories - medication adherence of HF patients and the others - were developed. The performance of each model was evaluated using LOOCV which randomly partitions the dataset into 76 equal size subsets having a sample as its element and uses each subset as a test dataset to verify the performance of each model which is derived by 75 remaining data subsets.

Table 2 depicts the comparison of the best accuracy of the four models (Linear, Polynomial, RBF, and Sigmoid). Since SVM with RBF showed better performance than with other kernel functions in the accuracy, hereafter we only report the performance of models with RBF.

The model showing the best accuracy consists of five features (gender, spouse, daily frequency of medication, medication knowledge, NYHA functional class) and seven features (age, education, monthly income, EF, MMSE-K, medication knowledge, NYHA functional class).

Table 3 shows the performance of the best model in detail

with statistics of sensitivity, specificity, PPV, and NPV.

Table 4 shows that the two sets of medication adherence-related variables with the best classification performance: a set of five variables (gender, daily frequency of medication, medication knowledge, NYHA functional class, spouse) and seven variables (age, education, monthly income, EF, MMSE-K, medication knowledge, NYHA functional class). The best detection accuracy was 77.63%. The 2 variables, medication knowledge and NYHA functional class, are shared by the two models. Swets suggests the following guidelines for interpretation: 0.5-0.7, rather low accuracy; 0.7-0.9, moderate accuracies useful for some purposes; and >0.9, rather high accuracy [22]. Therefore, the classification accuracy of these five-variable and seven-variable models can be described as the level of fair performance.

IV. Discussion

A full understanding of the factors associated with medication adherence in patients with HF is needed so that effective interventions to improve medication adherence can be developed [3]. But, most studies of medication adherence in patients with HF have had a number of limitations that reduce the usefulness of their findings. One of these limitations is the failure to use multivariate analysis methods to study medication adherence. This study addressed this problem by using SVM in identifying predictors of medication adherence in HF patients.

Our application of SVM found models that achieved fair classification performance, with a leave-one-out cross validated accuracy of around 77.6%. The LOOCV performance is a realistic indicator of performance of a classifier on unseen data and is a widely used statistical technique. The LOOCV is used during the training of a classifier to prevent overfitting of a classifier on the training set. But the LOOCV is rarely adopted in large-scale applications since it is computationally expensive [23]. In further research, if there are many samples, we may have to consider other ways such as 10-fold cross validation, etc.

Some of the selected variables in HF patients showed unexpected tendencies that are contrary to common medical thoughts. For example, a longer duration of HF diagnosis might be associated with medication non-adherence. However, this variable was absent in the two models that showed the best accuracy. In future work, we will investigate the performance using the SVM algorithm as well as other features selection methods to identify a group of significant factors.

In general, factors known to influence medication adherence in HF include medication knowledge and NYHA functional class [24,25]. Each of the models showed a strong preference for these factors. Especially, medication knowledge as a single variable was a strong predictor. Medication knowledge is well known as risk factor of medication nonadherence in HF [24,25]. The results of this study suggest that patients with higher levels of knowledge utilized the health care system more often, perhaps have greater awareness, hence a stronger tendency of self care.

Although our sample size was large enough to demonstrate fair accuracy, a larger sample size and more heterogeneous sample may be needed to more thoroughly investigate predictors of medication adherence and to obtain results that can be generalized to large populations.

Our study has several limitations. The primary limitation of this study is its small sample size, which made it very difficult for any of the endpoints to achieve statistical significance. The second limitation was that we did not directly

Table 2. Comparison of the best accuracy for four kernel functions

	Linear	Polynomial	RBF	Sigmoid
Best accuracy	73.438	67.188	77.632	64.063

RBF: radial basis function.

Table 3. Detailed sensitivity, specificity, PPV, NPV and accuracy for the best model of RBF kernel function

Sensitivity	Specificity	PPV	NPV	Accuracy
0.776	0.776	0.778	0.816	0.776

PPV: positive predictive value, NPV: negative predictive value, RBF: radial basis function.

Table 4. Predictors of medication adherence in heart failure patients

No.of features	Model	Accuracy (%)
5	Gender, daily frequency of medication, medication knowledge, NYHA functional class, spouse	77.632
7	Age, education, monthly income, EF, MMSE-K, medication knowledge, NYHA functional class	77.632

NYHA: New York Heart Association, EF: ejection fraction, MMSE-K: Mini-Mental Status Exam-Korean.

measure medication adherence. Third, the cross-sectional design does not allow any inference to be drawn with respect to the causal relationship among variables. Finally, our data were mainly based on patient information, and so, Berkson's bias [26] may contribute to inaccurate or excessive approval. SVM has been widely used in various areas, such as recognition, reliability evaluation, bioinformatics and medicines, for survival time classification and assessment of the severity of many acute and chronic diseases [27]. However, to our knowledge, this is the first study to investigate predictors of medication adherence and non-adherence in Korean patients with HF by means of SVM.

This study only illustrates a potential use of the SVM technique. Further studies should be conducted where the discriminative power of our SVM is compared with that of commonly used logistic regression, Bayesian network and neural network models.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

References

1. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995; 20: 273-297.

2. Verplancke T, Van Looy S, Benoit D, Vansteelandt S, Depuydt P, De Turck F, Decruyenaere J. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. BMC Med Inform Decis Mak 2008; 8: 56.

3. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak 2010; 10: 16.

4. Maglogiannis I, Loukis E, Zafiropoulos E, Stasis A. Support vectors machine-based identification of heart valve diseases using heart sounds. Comput Methods Programs Biomed 2009; 95: 47-61.

5. Thurston RC, Matthews KA, Hernandez J, De La Torre F. Improving the performance of physiologic hot flash measures with support vector machines. Psychophysiol-

ogy 2009; 46: 285-292.

6. Chu A, Ahn H, Halwan B, Kalmin B, Artifon EL, Barkun A, Lagoudakis MG, Kumar A. A decision support system to facilitate management of patients with acute gastrointestinal bleeding. Artif Intell Med 2008; 42: 247-259.

7. Rice SB, Nenadic G, Stapley BJ. Mining protein function from text using term-based support vector machines. BMC Bioinformatics 2005; 6 Suppl 1: S22.

8. Hauptman PJ. Medication adherence in heart failure. Heart Fail Rev 2008; 13: 99-106.

9. Komajda M, Lapuerta P, Hermans N, Gonzalez-Juanatey JR, van Veldhuisen DJ, Erdmann E, Tavazzi L, Poole-Wilson P, Le Pen C. Adherence to guidelines is a predictor of outcome in chronic heart failure: the MAHLER survey. Eur Heart J 2005; 26: 1653-1659.

10. Toh CT, Jackson B, Gascard DJ, Manning AR, Tuck EJ. Barriers to medication adherence in chronic heart failure patients during home visits. J Pharm Pract Res 2010; 40: 27-30.

11. Riegel B, Moser DK, Anker SD, Appel LJ, Dunbar SB, Grady KL, Gurvitz MZ, Havranek EP, Lee CS, Lindenfeld J, Peterson PN, Pressler SJ, Schocken DD, Whellan DJ; American Heart Association Council on Cardiovascular Nursing; American Heart Association Council on Clinical Cardiology; American Heart Association Council on Nutrition, Physical Activity, and Metabolism; American Heart Association Interdisciplinary Council on Quality of Care and Outcomes Research. State of the science: promoting self-care in persons with heart failure: a scientific statement from the American Heart Association. Circulation 2009; 120: 1141-1163.

12. van der Wal MH, Jaarsma T, Moser DK, Veeger NJ, van Gilst WH, van Veldhuisen DJ. Compliance in heart failure patients: the importance of knowledge and beliefs. Eur Heart J 2006; 27: 434-440.

13. Hope CJ, Wu J, Tu W, Young J, Murray MD. Association of medication adherence, knowledge, and skills with emergency department visits by adults 50 years or older with congestive heart failure. Am J Health Syst Pharm 2004; 61: 2043-2049.

14. Rud S, Yang JS. A support vector machine (SVM) classification approach to heart murmur detection. Lect Notes Comp Sci 2010; 6004: 52-59.

15. Balasubramanian VN, Gouripeddi R, Panchanathan S, Vermillion J, Bhaskaran A, Siegel RM. Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure. In: Proceedings of Computer in Cardiology 2009;

2009 Sept 13-16; Park City, UT, USA. p5-8.

16. Choi K, Chung S, Rhee H, Suh Y. Classification and sequential pattern analysis for improving managerial efficiency and providing better medical service in public healthcare centers. Healthc Inform Res 2010; 16: 67-76.

17. Choi JP, Han TH, Park RW. A hybrid Bayesian network model for predicting breast cancer prognosis. J Korean Soc Med Inform 2009; 15: 49-57.

18. Kim BM, Kang BY, Kim HG, Baek SH. Prognosis prediction for class III malocclusion treatment by feature wrapping method. Angle Orthod 2009; 79: 683-691.

19. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002; 46: 389-422.

20. Chang CC, Lin CJ. LIBSVM: a library for support vector machines [Internet]. [cited at 2010 Oct 29]. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

21. Rennie JDM. On the value of leave-one-out cross validation bounds [data file]. [cited at 2010 Oct 29]. Available from: http://www.ai.mit.edu/~jrennie/writing/loocv.ps.gz.

22. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988; 240: 1285-1293.

23. Kim H, Drake BL, Park H. Adaptive nonlinear discriminant analysis by regularized minimum squared errors. IEEE Trans Knowl Data Eng 2006; 18: 603-612.

24. Vlasnik JJ, Aliotta SL, DeLor B. Medication adherence: factors influencing compliance with prescribed medication plans. Case Manager 2005; 16: 47-51.

25. Morrow DG, Weiner M, Young J, Steinley D, Deer M, Murray MD. Improving medication knowledge among older adults with heart failure: a patient-centered approach to instruction design. Gerontologist 2005; 45: 545-552.

26. Berkson J. Limitations of the application of fourfold table analysis to hospital data. Biometrics 1946; 2: 47-53.

27. Cho BH, Yu H, Kim KW, Kim TH, Kim YI, Kim SI. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artif Intell Med 2008; 42: 37-53.