A Hybrid Feature Selection Method to Classification and Its Application in Hypertension Diagnosis

Hyun Woo Park, Dingkun Li, Yongjun Piao, and Keun Ho Ryu(✉)

Database and Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, Cheongju, South Korea
{hwpark,jerryli,pyz,khryu}@dblab.chungbuk.ac.kr

Abstract. Recently, various studies have shown that meaningful knowledge can be discovered by applying data mining techniques in medical applications, i.e., decision support systems for disease diagnosis. However, there are still several computational challenges due to the high-dimensionality of medical data. Feature selection is an essential pre-processing procedure in data mining to identify relevant feature subset for classification. In this study, we proposed a hybrid feature selection mechanism by combining symmetrical uncertainty and Bayesian network. As a case study, we applied our proposed method to the hypertension diagnosis problem. The results showed that our method can improve the classification performance and outperformed existing feature selection techniques.

Keywords: Classification · Feature selection · Hypertension · KNHANES · Data mining

1 Introduction

In recent years, the proportion of elderly people (over 65 years old) is increasing in Korea. The proportion of elderly people population will increase up to 30% by 2030 according to a recent report by Korea National Statistical Office [1]. Therefore, the number of chronic disease patients is increasing as well according to elderly people increasing. Approximately 80% of men and women 70 years of age have at least one of chronic disease such as hypertension, heart disease and so on [2]. Furthermore, chronic disease is continuous growth of health care cost. According to the report, health care costs for Koreans over age 65, reached 15.4 trillion Korean won in 2011 [3]. In general, chronic disease cannot be prevented by vaccines or cured by medication. Most of the common chronic disease caused by dietary, lifestyle (smoking, drinking), and many other factors. The hypertension is also one of the chronic disease type, the prevalence of hypertension is increasing according to Korean National Health and Nutrient Examination Survey report in 2011 [4]. The hypertension is a major risk factor for heart disease, and many other complications and these complication leading to death. Therefore, the prevention of hypertension has become a major issue in the world.

Most previous studies used the statistical method such as chi-square, and logistic regression for finding risk factors pertaining to chronic diseases [5–7]. Bae et al. [5] investigated the association between hypertension and prevalence of low back pain and

M. Bursa et al. (Eds.): ITBAM 2017, LNCS 10443, pp. 11–19, 2017.
DOI: 10.1007/978-3-319-64265-9_2

osteoarthritis in Koreans using Chi-square and logistic regression. Song et al. [6] examined the associations of total carbohydrate intake, dietary glycaemic load (DGL) and white rice intake with metabolic syndrome risk factors by gender in Korean adolescents using multivariate linear regression. Ha et al. [7] examined whether cardiovascular disease or its risk factors were associated with chronic low back pain (cLBP) in Koreans using logistic regression. The regression models allow for the testing of statistical interactions among independent variables and significance difference in the effects of one or more independent variables.

Recently, various studies [8–13] have shown that it is possible to apply data mining techniques in medical applications [14–17], i.e., decision support systems for disease diagnosis. However, medical data generally contains several irrelevant and redundant features regarding classification target. Those features may lead to low performance and high computational complexity of disease diagnosis. Moreover, most classification methods assume that all features have uniform importance degree during classification [18]. Thus, dimension reduction techniques that discover a reduced set of features are needed to achieve better classification result.

In this study, we proposed a hybrid feature selection method to improve the robustness and the accuracy of the hypertension diagnosis. Symmetrical uncertainty was used to preliminarily remove irrelevant features as a filter and correlation between two features is compared and the lower symmetrical uncertainty is removed. Machine learning algorithms with backward search were used as the wrapper part. The results showed that the proposed method yielded good performance and outperformed other feature selection approaches. The results, although preliminary, are expected to support medical decision making, and consequently reduce the expenditure in medical care.

The reminder of the paper is organized as follows. In Sect. 2, we describe dataset and present hybrid feature selection, classification method. In Sect. 3, shows the framework of experiment and results. The conclusion and future are presented in Sect. 4.

2 Materials and Methods

2.1 Data

In this study, we conducted Korea National Health and Nutrient Examination Survey (KNHNAES). This data set is a national project and it is consisting of four parts. The first part of the survey recorded the socio-demographic characteristics include age, gender, income, education level and so on. The second part of the survey recorded the history of a disease and a third part is recorded health medical examination such as blood pressure (systolic, diastolic). The last part of the survey recorded life pattern and nutritional intake. We conducted KNHANES data from 2007 to 2014. This data set contains lots of missing values and outliers. This data may lead to poor performance, we eliminated this data for generating target population. The basic characteristics of the target population are shows in Table 1.

Table 1. Basic characteristics of target population

	Control (n = 2,938)	Hypertension (n = 2,700)
Age (yr)	45.51 ± 15.46	65.20 ± 10.93
Sex		
Male	1,066	1,151
Female	1,872	1,549
Education		
High school	1,800	2,427
University	1,138	273
Smoking		
Yes	503	574
No	499	1,651
Quick	1,936	475
SBP (mmHg)	112.84 ± 15.00	131.63 ± 16.21
DBP (mmHg)	73.56 ± 9.68	80.53 ± 9.02
BMI (kg/m^2)	22.91 ± 3.11	24.70 ± 3.18
Waist Circumference (cm)	77.79 ± 9.17	85.28 ± 8.93

2.2 Feature Selection

Feature selection is an essential pre-processing procedure in data mining for identifying relevant subset for classification. The high dimensionality of the data may cause a various problem such as increasing the complexity and reducing the accuracy, i.e. curse of dimensionality. The goal of feature selection is to provide faster construction of prediction models with a better performance [8]. Feature selection approaches can be broadly grouped into three categories: filter, wrapper, and hybrid [19]. The main difference of filter and wrapper method is in whether they adopt a machine learning algorithm to guide the feature selection or not. In general, filters employ independent evaluation measures thus are fast but can generate the local-optimal result. In contrast, wrapper methods adopt a searching algorithm to iteratively generate several subsets, evaluate them based on the classification algorithm, and finally choose the subset with best classification performance. Wrappers usually can produce better results than filters but they are computationally expensive. Hybrid methods combined the advantages of filter and wrapper techniques to achieve better learning performance with a similar computational cost of filters [20]. A feature selection procedure can usually be divided into two steps: subset generation and subset evaluation. The most important this process is determined to search strategy and the starting point. Sequential search method, such as Sequential Forward Search (SFS) and Sequential Backward Search (SBS). SFS method start with an empty candidate set and add features until the addition of features does not decrease the criterion. SBS method is removed from a full candidate set until the removal of further features increases the criterion. We illustrate the filter and wrapper approach in Fig. 1.

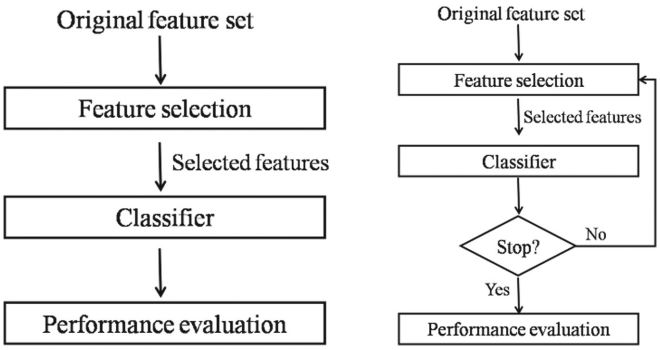

Fig. 1. (a) Filter approach and (b) wrapper approach

In this study, we proposed a hybrid feature selection method to improve the robustness and the accuracy of the hypertension diagnosis. Here, we describe our proposed hybrid feature selection method. The overall framework of generating optimal feature subset is illustrated in Fig. 2. In the proposed method, symmetrical uncertainty was used to preliminarily remove irrelevant features and remove redundant features used Pearson correlation as a filter. Bayesian network [21] with backward search [22] adopted as the wrapper part.

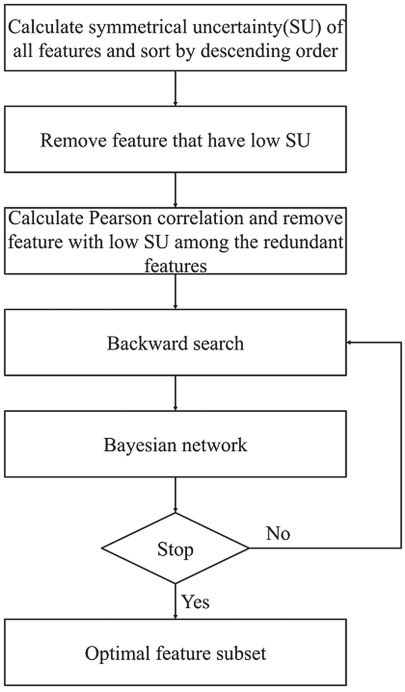

Fig. 2. Framework of proposed feature selection method

For each feature in the original space, the symmetrical uncertainty value is calculated and the features that have smaller symmetrical uncertainty value than the predefined threshold are removed. Then the remaining features are sorted in the descending order of their symmetrical uncertainty value. We compared the correlation between two features and remove the features with low gain ratio among the two features whose correlation is higher than the predefined threshold. Afterward, starting from the whole features generated from the filter part, removing one feature at one time, the subset is evaluated by the Bayesian network until the best feature subset that has the highest accuracy is selected.

The Information gain (IG) is the decrease in the entropy of Y for given information on Y provided by X and is calculated using (3), which is based on (1) and (2). Then the measure of symmetrical uncertainty (SU) is calculated using (4), which can compensate for the problem of a biased IG and normalize its value from 0 to 1.

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y) \tag{1}$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x) \tag{2}$$

$$gain = H(Y) + H(X) - H(X|Y) \tag{3}$$

$$SU = 2.0 \times \left[\frac{gain}{H(Y) + H(X)} \right] \tag{4}$$

Where H(X) is the entropy of feature X, H(C) indicates the entropy of the class, and H(X|C) is the entropy of x after observing class C.

3 Experiment and Results

The framework of the experiment is shown in Fig. 3. Firstly, we need to generate of the target population based on KNHANES dataset. In the next step, we remove missing value and outliers based interquartile ranges and we used proposed feature selection method to remove irrelevant and redundant features. Then, we compared the performance of classification with several commonly used feature selection methods such as Information gain, Gain ratio, RelifF.

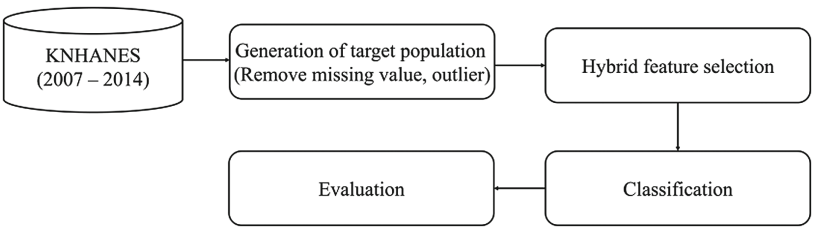

Fig. 3. Framework of experiment

3.1 Generation of Target Population

In this study, we conducted KNHANES from 2007 to 2014. This data contains 65,973 individuals include 9,383 hypertensions patients' data. We eliminated a considerable number of hypertension patients with following exclusion criteria to avoid data bias. The target population data contains 5,698 samples (2,938 hypertensions and 2,700 controls) samples. We describe the procedure of generation of the target population in Fig. 4.

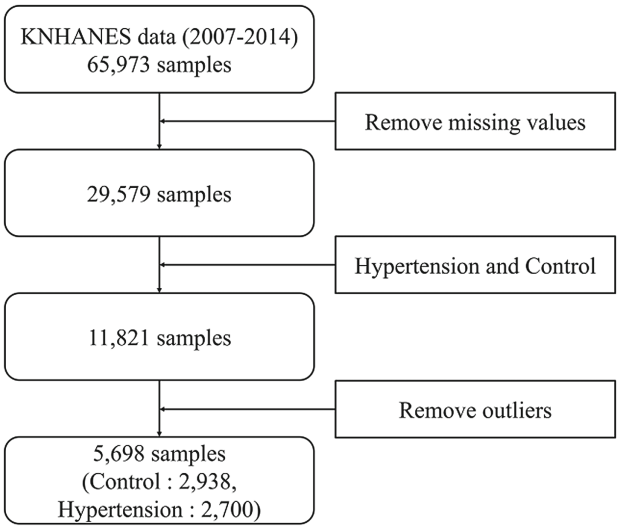

Fig. 4. Procedure of generating target population

3.2 Hybrid Feature Selection

First of all, we eliminated irrelevant features whose SU values are less than 0.01, and sort in descending order. Afterward, we remove redundant features using Pearson correlation between two features. We compared the correlation with the property with highest SU value and remove the variables with low SU value among the features with a correlation greater than 0.5. In the next step we used Bayesian network with SBS search, then we find optimal feature subset.

3.3 Bayesian Network

A Bayesian network is based on probability theory and graphical model. The graph consists of one or more nodes and edges. Each node is the graph represents a random variable, and each arc between every pair of variables. Although the sample size is not large enough for estimating the predictive performance of each model, Bayesian network in general has good prediction abilities.

3.4 Experimental Results

To achieve reliable results, we conducted 10-fold cross-validation during all the experiments. In 10-fold cross-validation, 9 parts were used to train the model and the remaining one was used to test the model. Moreover, we adopted several evaluation measures such as F-measure, sensitivity, specificity, and area under ROC (AUC) to the performance of classification. Table 2 shows the performance of our proposed method. The results were found to be 0.923, 0.923, 0.923, and 0.975 for F-measure, sensitivity, specificity, and AUC, respectively. In Table 2, hypertension indicates the patients with hypertension disease and control indicates those without hypertension and FS refers to apply proposed hybrid feature selection and Non-FS indicates without feature selection. Figure 5. Shows the classification accuracy of 4 feature selection methods. From the Fig. 5. We can easily see that our proposed method outperforms existing feature selection approaches.

Table 2. Classification results of the proposed method

		Hypertension	Control	Average
F-Measure	Non-FS	0.879	0.885	0.882
	Proposed FS	**0.921**	**0.926**	**0.923**
Sensitivity	Non-FS	0.896	0.869	0.882
	Proposed FS	**0.928**	**0.919**	**0.923**
Specificity	Non-FS	0.869	0.896	0.882
	Proposed FS	**0.919**	**0.928**	**0.923**
AUC	Non-FS	0.955	0.955	0.955
	Proposed FS	**0.975**	**0.975**	**0.975**

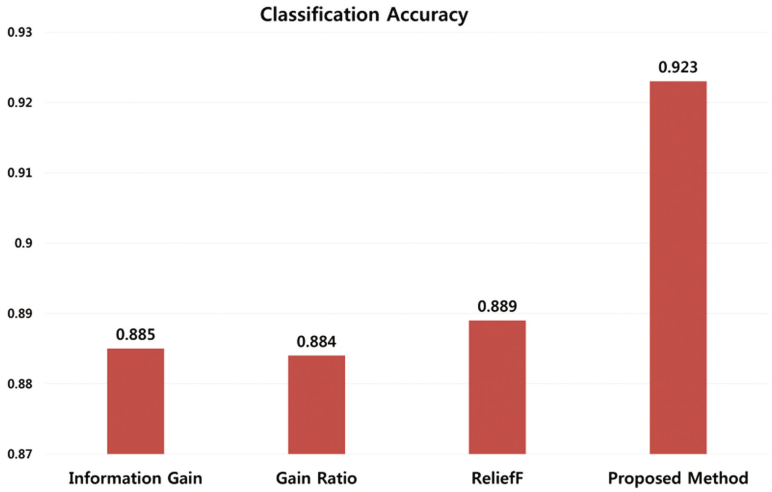

Fig. 5. Performance comparison of the proposed method with other feature selection techniques

4 Conclusion

In this study, we proposed a hybrid feature selection technique based on symmetrical uncertainty and Pearson correlation as a filter approach and Bayesian network as a wrapper approaches for accurately classify hypertensions. To validate the proposed method, we conducted KNHANES 2007–2014 data. We conducted several experiments and compared the performance with existing feature selection approaches. The results showed the proposed method had good performance and outperformed the other feature selection methods.

The hypertension is a major risk factor for heart disease, and many other complications and these complication leading to death. In the future work, we will analyze hypertension and other complications disease such as heart disease, stroke.

Acknowledgment. This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-2013-0-00881) supervised by the IITP (Institute for Information & communication Technology Promotion) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2017R1A2B4010826).

References

1. Korea National Statistical Office: Annual report on the statistical for elderly (2011)
2. Black, D.S., O'Reilly, G.A., Olmstead, R., Breen, E.C., Irwin, M.R.: Mindfulness meditation and improvement in sleep quality and daytime impairment among older adults with sleep disturbances: a randomized clinical trial. JAMA Intern. Med. **175**(4), 494–501 (2015)
3. Jeong, H.S., Song, Y.M.: Contributing factors to the increases in health insurance expenditures for the aged and their forecasts. Korean J. Health Econ. Policy **19**(2), 21–38 (2013)
4. Korea Centers for Disease Control and Prevention: Korea National health & nutrition examination survey (2007–2014)
5. Bae, Y.H., Shin, J.S., Lee, J., Kim, M.R., Park, K.B., Cho, J.H., Ha, I.H.: Association between Hypertension and the prevalence of low back pain and osteoarthritis in Koreans: a cross-sectional study. PloS one, **10**(9) (2015)
6. Song, S., Paik, H.Y., Song, W.O., Song, Y.: Metabolic syndrome risk factors are associated with white rice intake in Korean adolescent girls and boys. Br. J. Nutr. **113**(03), 479–487 (2015)
7. Ha, I.H., Lee, J., Kim, M.R., Kim, H., Shin, J.S.: The association between the history of cardiovascular diseases and chronic low back pain in South Koreans: a cross-sectional study. PloS one **9**(4) (2014)
8. Piao, Y., Piao, M., Park, K., Ryu, K.H.: An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics **28**(24), 3306–3315 (2012)
9. Piao, Y., Piao, M., Ryu, K.H.: Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles. Comput. Biol. Med. **80**, 39–44 (2017)
10. Lee, D.G., Ryu, K.S., Bashir, M., Bae, J.W., Ryu, K.H.: Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. J. Med. Syst. **37**(2), 9896 (2013)

11. Bashir, M.E.A., Shon, H.S., Lee, D.G., Kim, H., Ryu, K.H.: Real-time automated cardiac health monitoring by combination of active learning and adaptive feature selection. TIIS **7**(1), 99–118 (2013)
12. Kim, H., Ishag, M.I.M., Piao, M., Kwon, T., Ryu, K.H.: A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries. Symmetry **8**(6), 4 (2016)
13. Mayer, C., Bachler, M., Holzinger, A., Stein, P.K., Wassertheurer, S.: The effect of threshold values and weighting factors on the association between entropy measures and mortality after myocardial infarction in the Cardiac Arrhythmia Suppression Trial (CAST). Entropy **18**(4), 129, 121–115 (2016)
14. Kaur, H., Wasan, S.K.: Empirical study on applications of data mining techniques in healthcare. J. Comput. Sci. **2**(2), 194–200 (2006)
15. Holzinger, A.: Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Inform. **3**(2), 119–131 (2016)
16. Hund, M., Boehm, D., Sturm, W., Sedlmair, M., Schreck, T., Ullrich, T., Keim, D.A., Majnaric, L., Holzinger, A.: Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the Doctor-in-the-loop. Brain Inform. **3**(4), 233–247 (2016)
17. Park, H.W., Batbaatar, E., Li, D., Ryu, K.H.: Risk factors rule mining in hypertension: Korean National Health and Nutrient Examinations Survey 2007–2014. In: Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–4 (2016)
18. Yang, Y., Liao, Y., Meng, G., Lee, J.: A hybrid feature selection scheme for unsupervised learning and its application in bearing fault diagnosis. Expert Syst. Appl. **38**(9), 11311–11320 (2011)
19. Hsu, H.H., Hsieh, C.W., Lu, M.D.: Hybrid feature selection by combining filters and wrappers. Expert Syst. Appl. **38**(7), 8144–8150 (2011)
20. Xie, J., Wang, C.: Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Expert Syst. Appl. **38**, 5809–5815 (2011)
21. Vapnik, V.: The nature of statistical learning (2013)
22. Han, J., Fu, Y.: Attribute-oriented induction in data mining. Advances in knowledge discovery and data mining, pp. 399–421 (1996)

http://www.springer.com/978-3-319-64264-2

Information Technology in Bio- and Medical Informatics
8th International Conference, ITBAM 2017, Lyon, France, August 28–31, 2017, Proceedings
Bursa, M.; Holzinger, A.; Renda, M.E.; Khuri, S. (Eds.)
2017, X, 135 p. 37 illus., Softcover
ISBN: 978-3-319-64264-2