JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 4, pp. 739-747, Apr. 2018

KNHNAES (2013~2015) 에 기반한 대형 특징 공간 데이터집 혼합형 효율적인 특징 선택 모델

권 태 일[1,2] · 이 정 곤[2*] · 박 현 우[2] · 류 광 선[2] · 김 의 탁[2] · 박 명 호[3]
[1]빅썬시스템즈(주)
[2]충북대학교 전기전자정보컴퓨터학부 데이터베이스/바이오인포매틱스연구실
[3]충북대학교 스마트 팩토리 사업단

A Hybrid Efficient Feature Selection Model for High Dimensional Data Set based on KNHNAES (2013~2015)

Tae il Kwon[1,2]·Dingkun Li[2*]·Hyun Woo Park[2]·Kwang Sun Ryu[2]·Eui Tak Kim[2]·Minghao Piao[3]

[1]BigSun Systems Co. LTd., Seoul, South Korea

[2]Database/Bioinformatics Lab, School of Electrical & Computer Engineering, Chungbuk National University, Cheongju, South Korea

[3]Agency of Smart Factory, Chungbuk National University, Cheongju, South Korea

[요 약]

고차원 데이터에서는 데이터마이닝 기법 중에서 특징 선택은 매우 중요한 과정이 되었다. 그러나 전통적인 단일 특징 선택방법은 더 이상 효율적인 특징선택 기법으로 적합하지 않을 수 있다. 본 논문에서 우리는 고차원 데이터에 대한 효율적인 특징선택을 위하여 혼합형 특징선택 기법을 제안하였다. 본 논문에서는 KNHANES 데이터에 제안한 혼합형 특징선택기법을 적용하여 분류한 결과 기존의 분류기법을 적용한 모델보다 5% 이상의 정확도가 향상되었다.

[Abstract]

With a large feature space data, feature selection has become an extremely important procedure in the Data Mining process. But the traditional feature selection methods with single process may no longer fit for this procedure. In this paper, we proposed a hybrid efficient feature selection model for high dimensional data. We have applied our model on KNHNAES data set, the result shows that our model outperforms many existing methods in terms of accuracy over than at least 5%.

색인어 : 특징 선택, 고차원 데이터, 하이브리드 특징 선택 모델, 데이터 마이닝, 병렬 컴퓨팅

Key word: feature selection, high dimensional data, hybrid feature selection model, data mining, parallel computing.

http://dx.doi.org/10.9728/dcs.2018.19.4.739

Received 05 April 2018; **Revised** 18 April 2018
Accepted 25 April 2018

***Corresponding Author; Dingkun Li**

Tel: +82-10-8969-8680
E-mail: jerryli@dblab.chungbuk.ac.kr

Ⅰ. Introduction

Never before in history is the data growing at such a high speed as well as variety and quantity. Vast records (text, transactions, images, etc.) are created in every second and the properties, also can be defined as features or attributes, used to describe these records are increasing as well. For example, more attributes are included to describe the single patient health condition for the purpose of all around healthcare service. That's why generally large scale medical databases are having large number of attributes or dimensions. This large data dimensionality can badly influence many aspects of analysis process. The increase in the data dimensionality may cause several issues with respect to scalability and learning performance in these classification algorithms. This is so called "curse of dimensionality" problem [1].

Feature selection is an important technique for handling high dimension data. It is one essential step for data mining which is defined as a multidisciplinary task to find out hidden nugget of information from data [2]. The main goal of feature selection is to find subset from the large feature space that this subset can be used to identify the specified object, such as disease prediction. There is a hidden rule behind this subset, although the dimensions is reduced, the discriminative capability should not be reduced. Generally, the benefits for feature selection are reducing data analysis complexity and improve data analysis performance. Nevertheless, there are more than that such as accuracy improvement, expenditure reduction, etc.

Generally, feature selection methods can be categorized into three types: filter, wrapper and embedded methods [3]. Data mining algorithms are not included in filter methods. They are strongly relies on underlying characteristics of the data variable depends on certain criteria. For example feature selection using, information gain, fisher score [4] etc. Wrapper model consists a learning process, the induced algorithms are used as "black box". It uses forward or backward or embedded strategy, that gives more discriminative power with that particular learning process; therefore, it consume more time compare to filter [3]. The filters work fast but its result is not always satisfactory. While the wrappers guarantee good results but very slow when applied to wide feature sets which contain hundreds or even thousands of features [5].

In the past, various feature selection techniques have been proposed such as chi-square test, mutual information, Pearson correlation coefficients, and Relief etc.[6] Although these methods are fast, they lack robustness when interactions among features exist. That means they assume that the features are independent with each other. In literature, a large number of feature selection algorithms have been already proposed and they were applied to different fields: bioinformatics [7][8][32][33], healthcare [9], image processing [10], etc. However there is no compatible algorithm or framework can be used in all fields. But the traditional feature selection methods with single process may no longer fit for a large feature space data.

In healthcare field, clinical databases often consist of a large number of features. For clinical data analysis, some features are not useful, some are redundant, and some are key factors. Our research purpose is to find practical, efficient and well fitted methods for clinical feature selection among large feature space based on Korea National Health and Nutrient Examination Survey (KNHANES) dataset [11]. The target disease is hypertension (HP) in our work.

Hypertension is a condition in which a person's blood pressure is above normal or optimal limit of 120 mmHg for systolic pressure and 80 mmHg for diastolic pressure. The categories of HP with normal people information in shown in Table 1 [27]. In our work we treat both pre-HP and normal people as normal in KNHANES data set.

표 1. KNHANES 데이터 세트의 HP 카테고리
Table 1. HP Categories of KNHANES data set

Attribute Name	Classification	Systolic(mmHg)	Diastolic(mmHg)
HE_HP=1	normal	<120	and <80
HE_HP=2	pre-HP	120~139	or 80~89
HE_HP=3	HP	>140	or ≥100

The contributions of our work are: 1). developed an efficient paralleled model for feature selection and DM model generation on large feature space data set. 2). Provide a framework which can be extended to big data analysis framework. We have compared our model with existing algorithms such as Logistical Regression classifier, Naive-bayes classifier, KNN and C5.0, the results show that our model outperforms the above algorithms.

This paper is organized as follows: A brief introduction about data mining, feature selection and the purpose of our work is give in Section Ⅰ. Section Ⅱ depicts previous techniques and work related to our work. An overview of the proposed system with its implementation is introduced in Section Ⅲ. Experimental results with proposed method evaluation is given in Section Ⅳ. Section Ⅴ concludes this work and introduces future work.

Ⅱ. Previous Techniques

This section describes some algorithms which are used to compare with our proposed methods.

2-1 Logistic Regression

The Logistic Regression (LRc) has been developed by statistician David in 1958 [20]. There are two models, one is binary logistic regression, where the output can have only two values, another is multinomial logistic regression, which has more than two outcome categories. It is widely used in clinical field. The LR model solves these problems:

$$\ln[\frac{P}{1-P}] = \alpha + \beta X + e \quad (1)$$

Where p is probability that the event Y ouuurs, p(y=1), p/(1-p) is the "odds ration".

2-2 Naive-bayes Classfier

The Naive-bayes classifier (NBc) is based on Bayes theorem and has been studied widely since the 1950 [21]. The Naive Bayes assumption is: attributes that describe data instances are conditionally independent. One of the advantages about NBc is that, it can combine any kind of objects (e.g. time series, trees, etc.) to generate classifier, based on probabilistic model specification.

$$p(h|d) = \frac{P(d|h)P(h)}{P(d)} \quad (2)$$

Where *d* is data, *h* is hypothesis, *P(h)* is prior belief (probability of hypothesis *h* before seeing any data). *P(d|h)* is like hood (probability of the dat if the hypothesis *h* is true). $P(d)=\sum P(d|h)P(h)$ is data evidence (marginal probability of the data). *P(h|d)* is posterior (probability of hypothesis *h* after having seen the data *d*).

The NBc is one of the most practical learning methods, and used very successfully in medical diagnosis and text classification.

2-3 K Nearest Neighbors

KNN, short for K Nearest Neighbors, is a non-parametric method used for classification. An instance is classified by a certain kind of similarity method of its neighbors with it being assigned to the class most common among its k nearest neighbors. The similarity function is usually defined as：

$$D(x,y) = \sqrt[q]{\sum_{i=1}^{n} |x_i - y_i|^p} \quad (3)$$

Where $p \in \{1,2,...n\}$.

The advantages of KNN is that: training is very fast, it can be used to lean complex targe functions, and do not loose information. The disadvantages is that: it is slow at query time (pre-sorting and indexing training samples into search trees reduces time), it is easily fooled by irrelevant features (attributes).

2-4 C5.0

The C5.0 is a rule based classification technique, the output consists of IF...THEN rules which are basic of decision tree [22]. C5.0 follows the strategy of C4.5 [23], but it is more advanced in terms of memory efficiency and error rate with acknowledge on noise and missing data.

2-5 SVM

Support Vector Machine (SVM) [24] has been used to select features and generate the classifier. For feature selection, this method is a backward sequential selection approach. One starts with all the features and removes one feature at a time until only r features are left. The operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. The basic concept is described using Figure 1.

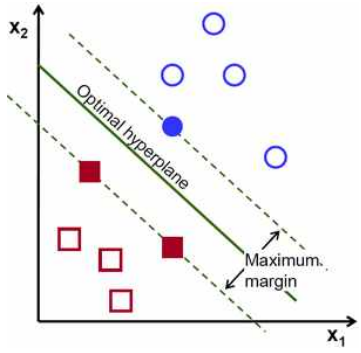

그림 1. 최대 마진, 점선의 벡터는 서포트 벡터
Fig. 1. Maximum Margin, the vectors on the dashed line are the support vectors

The strategy ranks the features according to their influence on the decision hyperplane. The optimal hyperplane is used to classify the data into different classes in two or more dimensionalities.

Ⅲ. Proposed Method

We propose an ensemble framework for high-dimensional feature selection based on KNHANES data set. The each step detail of this framework is depicted in this section.

3-1 Framework of Proposed Method

The framework of the proposed method is shown in Figure 2.

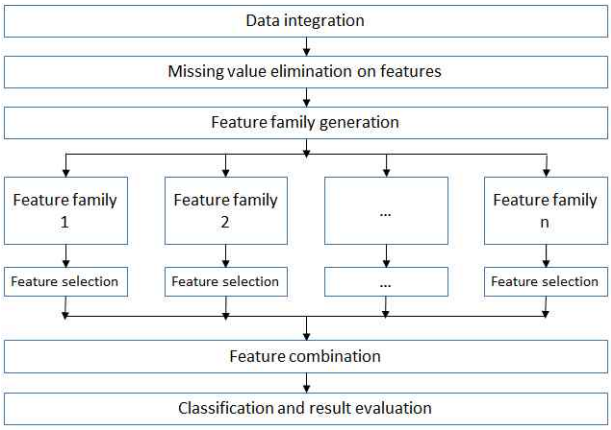

그림 2. 연구 모델
Fig. 2. Research Model

Firstly, the method integrates data from multi-source in term of year and region. Data sets of all data source should consist of similar feature space so that they can be integrated without conflict. Or these data sets should be processed for having similar feature space for integration.

Secondly, clinical data is nature to have a big amount of features and big amount of missing value. There are many ways to handle missing value such as using mean, median or user specified value to take place of the missing value. But to our concern, the feature consists more than 80% missing value should be eliminated because too much missing value will reduce the importance of the feature. If a feature is blank or only has one value, it holds no meaning.

Thirdly, feature families (sub feature sets) are generated based on the domain or corelation between each other, or random combination of them.

Fourthly, feature selection methods are used to select features according to a certain kind of criteria such as voting or weight. In this work, voting strategy has been used to generate the feature candidates. Next, only important features are selected and combined for further model training.

Finally, the prediction model will be generated and the result will be compared with existing widely used model.

More detail will be described in the following sections.

3-2 Feature Subset Generation

Feature subset is also called feature family. Ideally feature families are independent with each other and all the features in the same family are within the same domain, or correlated with each other if they are not selected randomly. But we can not say they are independent or not.

Data set KNHANES is a well feature-defined data set. These features can be assigned to different family easily based on the domain which is distinguished by prefix of the feature name. For example, feature name starts with "N_" means this feature indicate nutrition intake of observation. Feature name starts with "D_" means this feature is related to a certain kind of disease, and so forth. 19 feature families have been achieved by using this prefix method. It also guarantees the independence of each family.

Nevertheless, it is hard to claim that features within the same family are redundant or not. For example, for the nutrition intake feature family, nutritions are not redundant because all nutritions are different, but for lifestyle feature family, weekly smoking times and monthly smoking times are redundant because monthly value can be generated from weekly data. The redundant feature removement plays an important role for model learning.

3-3 Feature Selection for Each Feature Family

Feature selection is an assembled process. The purpose for this step is to generate the candidate features for next step.

At the beginning of this process, a natural question arises why control of redundancy is useful? Work [12] implies that it can not only increase the effectiveness but also accuracy of the results. We choose 3 filter methods Information Gain [28], Symmetrical Uncertainty [29] and CfsSubsetSelect [30], these methods are used for redundant feature elimination and candidate feature generation.

After that several important features can be selected from each feature family as the candidates. A voting strategy has been used to generate the candidates in each feature family. The idea is that: the feature has more than 2 votes among 3 methods is selected as the candidate for next step.

3-4 Feature Combination

Combination of these features are based on the ID and year features. The authors assign each feature family the same weight to void the data bias problem [16].

3-5 Classification and Result Evaluation

Over the past few years, SVM has been widely used for

classification. In addition, SVM can be used for feature selection as well. Features that do not contribute to classification are eliminated in each round until no further improvement in the classification can be achieved [31]. The selected features is based on its weight obtained from training SVMs with candidate feature set.

The generated classifier with its selected features are evaluated and compared with Logistical Regression classifier, Naive-bayes classifier, KNN and C5.0 classification methods. The experimental result are given in next section.

Ⅳ. Experimental Results

We present the detail of data sets used for feature selection and model learning. The selected feature are used as input features for further classifier generation.

Other classification methods have been used to compare the performance with our method. The detail will be described in the following sections.

4-1 Dataset Preparation

The Korea National Health and Nutrition Examination Survey (KNHANES) is a national surveillance system that has been set up since 1998 [15]. It collection information on socioeconomic status, health-related behaviours, quality of life, healthcare utilization, anthropometric measures, biochemical and clinical profiles for non-communicable diseases and dietary intakes. This serveillance system has been conducted by the Korea Center for Disease Control and Prevention (KCDC).

The report and microdata of KNHANES release annually. All resources are available through the official website (http://knhanes.cdc.go.kr). The data the authors downloaded starts from 2013 to 2015. It consists of 22,948 records and 727 compatible attributes. The general statistical information about the selected attributes is shown in Table 2. The valid count of hypertension record is 15,587.

4-2 Experiment Design

In order to compare the experimental result with existing methods, the authors designed experiment in the following steps: 1). Without lose of generality, the authors run proposed method 3 times on data set to get the average performance result. 2). The authors do the same run by existing algorithms, 3). Compare the average performance result with these classifier methods.

표 2. 선택한 속성에 대한 기분 정보
Table 2. The General Statistical Information about the Selected Attributes

Feature Name	Description	Min	Max	Mean	SD
Sex	Sex(1:man, 2:woman)	1	2	1.55	0.50
age	age of the observation	1	80	41.93	22.77
marri_1	Marital Status(1:married, 2:unmarried, 9:unknow)	1	9	1.35	0.55
incm	income(1:low; 2:below avg.; 3:above avg.; 4:high)	1	4	2.50	1.18
edu	education level(1:elementary; 2:middle; 3:high; 4:university)	1	4	2.38	1.21
HE_HP	Hypertension info. (1:normal; 2:pre-HP; 3:HP)	1	3	1.85	0.86

SD: Standard Deviation.

4-3 Performance Evaluation and Comparition

The method we used to evaluate is the method widely used in machine learning is called confusion matrix [25]. It is a table with two rows and two columns that reports the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). TP rate and FP rate refer to the proportion of actual positive instances correctly predicted as positive and the proportion of actual negative instances wrongly predicted as positive. Sensitivity is defined as TP/(TP+FN), Specificity is TN/(FP+TN), Precision is TP/(TP+FP) and Accuracy is (TP+TN)/(P+N), F-measure is 2TP/(2TP+FP+FN), which is a measure of a test's accuracy. F-measure is more robust compared with accuracy. Roc stands for receiver operating characteristic curve, is TP rate again FP rate at various threshold settings. AUC is the area under the Roc curve. [17,18,19, 26]. The machine learning community most often uses the ROC AUC statistic for model comparison.

We performed the proposed method by 10-fold validation on each feature family, which was randomly partitioned into 10 parts. 9 parts were used as the training set, and the last one was used as testing dataset.

After feature combination, to achieve a statistically reliable result, the authors run the proposed method 3 time, the result is given in Table 3.

표 3. 방법 성능
Table 3. Performance of proposed method.

Run	Sensitivity	Specificity	Precision	Accuracy	F-score	AUC
1st	0.804	0.678	0.890	0.7739	0.845	0.849
2nd	0.804	0.678	0.890	0.7739	0.845	0.849
3rd	0.803	0.680	0.891	0.7739	0.844	0.848
Ave.	0.804	0.676	0.890	0.7739	0.8445	0.849

Each row of Table 3 indicates the performance of each run and the last row shows the weighted average of 3 runs.

The author do the same run by using existing algorithms: Logistic Regression classifier, Naive-bayes classifier, KNN, C5.0 algorithm 3 times separately. The AUC comparition result is give in Table 4.

표 4. 제안 된 방법과 다른 방법 사이의 AUC 비교
Table 4. Comparison of AUC Between Proposed Method and other Methods

Run	Proposed	LR	NB	KNN	C5.0
1st	0.849	0.613	0.513	0.508	0.795
2nd	0.849	0.600	0.580	0.610	0.780
3rd	0.848	0.692	0.624	0.635	0.756

For the purpose of describing AUC comparition result visually according to Table 4, we plotted the result on box plot diagram in Figure 3.

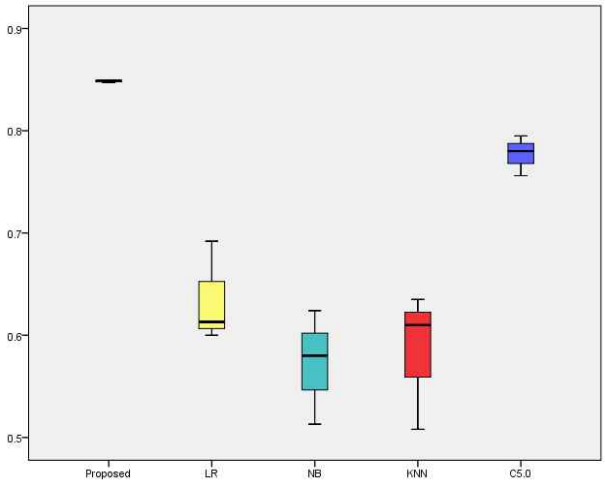

그림 3. AUC 결과의 박스 플롯
Fig. 3. Box plot of AUC result

We can draw conclusion from the Figure 3 that the proposed model shows the best average prediction result in terms of AUC. The narrow box of the proposed method also shows the stability of our method.

Ⅴ. Conclusion and future work

In this paper, we presented a hybrid efficient feature selection model on large feature space. The data set used for our experiment is KNHANES which is a widely used and testified. Filter based feature selection methods combined with wrapper method consist of our hybrid feature selection model. We demonstrated that dividing the large feature space into sub feature space, called feature family, for hybrid method can achieve better performance on KNHANES data set.

In the future, we will continuously improve the model accuracy by improving feature selection method on feature family and classification method. More data sets will be tested using this model if the feature family generation method is extended based on domain, characters of large feature space or even by using random feature choosing method. We strongly consider that this model can be used for extremely large-feature-space data since it is nature to be deployed under parallel framework such as MapReduce.

Acknowledgement

This work was supported by the Big Sun Systems Korea (2016090742), by the KIAT(Korea Institute for Advancement of Technology) grant funded by the Korea Government(MOTIE : Ministry of Trade Industry and Energy) (No. N0002429), and also by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2018-2013-1-00881) supervised by the IITP(Institute for Information & communication Technology Promotion).

References

[1] K. Eamonn, A. Mueen, "Curse of dimensionality", Encyclopedia of Machine Learning and Data Mining, Springer, pp.314-315, 2017.

[2] S. Bharat, N. Kushwaha, O. P. Vyas, "A feature subset selection technique for high dimensional data using symmetric uncertainty." Journal of Data Analysis and Information Processing,Vol. 2 No. 04, pp. 95, 2014.

[3] G. Isabelle, A. Elisseeff, "An introduction to variable and feature selection." Journal of machine learning research, Vol. 3, pp. 1157-1182, Mar, 2003.

[4] Q. Gu, Z. Li, J. Han, "Generalized fisher score for feature selection." arXiv preprint arXiv, 1202.3725, 2012.

[5] H. H. Hsu, C. W. Hsieh, M. D. Lu, “Hybrid feature selection by combining filters and wrappers.” Expert Systems with Applications, Vol. 38, No. 7, pp. 8144-8150, 2011.

[6] Y. Lei, H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution." Proceedings of the 20th international conference on machine learning (ICML-03). 2003.

[7] Z. M. Hira, D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data." Advances in bioinformatics 2015, 2015.

[8] L. Wang, Y. Wang, Q. Chang, "Feature selection methods for big data bioinformatics: A survey from the search perspective." Methods, Vol. 111, pp. 21-31, 2016.

[9] N. A. Capela, E. D. Lemaire, N. Baddour, "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients." PloS one, Vol. 10, No. 4, 2015.

[10] E. Guldogan, M. Gabbouj, "Feature Selection for Content-Based Image Retrieval", Signal, Image and Video Processing, Vol. 2, pp. 241-250, 2008.

[11] KNHANES, Available: https://knhanes.cdc.go.kr/knhanes/sub03/sub03_02_02.do

[12] R. Chakraborty, R. P. Nikhil, "Feature selection using a neural framework with controlled redundancy", IEEE transactions on neural networks and learning systems Vol. 26, No. 1, pp. 35-50, 2015.

[13] L. Yu, H. Liu, "Feature selection for high-dimensional data: a fast correlation-based filter solution", Proceedings of the 12th International Conference on Machine Learning, Washington, DC, USA, 2003.

[14] K. I. Kim, M. I. M. Ishag, M. Kim, J. S. Kim, and K. H. Ryu, "Proposal of a Resource-Monitoring Improvement System Using Amazon Web Service API." In Advances in Computer Science and Ubiquitous Computing, pp. 1103-1107, 2016

[15] S. Kweon, et al., "Data resource profile: the Korea national health and nutrition examination survey (KNHANES)", International journal of epidemiology, Vol. 43, No. 1, pp. 69-77, 2014.

[16] C. B. Begg, A. B. Jesse, "Publication bias: a problem in interpreting medical data", Journal of the Royal Statistical Society. Series A (Statistics in Society), pp. 419-463, 1988.

[17] M. Piao, H. S. Shon, J. Y. Lee, and K. H. Ryu, "Subspace projection method based clustering analysis in load profiling", IEEE Transactions on Power Systems, vol. 29, no. 6, pp. 2628–2635, 2014.

[18] M. E. A. Bashir, D. G. Lee, M. Li et al., "Trigger learning and ECG parameter customization for remote cardiac clinical care information system", IEEE Transactions on Information Technology in Biomedicine, vol. 16, no. 4, pp. 561–571, 2012.

[19] Y. Lee, Y. J. Jung, K. W. Nam, S. Nittel, K. Beard, and K. H. Ryu, "Geosensor data representation using layered slope grids", Sensors, vol. 12, no. 12, pp. 17074–17093, 2012.

[20] D. R. Cox, "The regression analysis of binary sequences (with discussion)", J Roy Stat Soc B. Vol. 20, pp. 215–242, 1958.

[21] S. Russell, P. Norvig, Artificial Intelligence: "A Modern Approach (2nd ed.)." Prentice Hall, 1995

[22] R. Pandya, P. Jayati, "C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning", International Journal of Computer Applications Vol. 117, No. 16, 2015.

[23] J. R. Quinlan, "C4. 5: programs for machine learning." Elsevier, 2014.

[24] C. Cortes, V. Vapnik, "Support-vector networks", Machine learning, Vol. 20, No. 3, pp. 273-297, 1995.

[25] T. Fawcett, "An Introduction to ROC Analysis", Pattern Recognition Letters, Vol. 27, No. 8, pp. 861–874, 2006.

[26] http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

[27] A. V. Chobanian, G. L. Bakris, H. R. Black, et al., "Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure", Hypertension, Vol. 42, No. 6, pp. 1206-1252, 2003

[28] T. M. Cover, J. A. Thomas, Elements of Information Theory (Wiley ed.), 1991.

[29] S. S. Kannan, N. Ramraj, "A Novel Hybrid Feature Selection via Symmetrical Uncertainty Ranking Based Local Memetic Search Algorithm", Knowledge-Based Systems, Vol. 23, pp. 580-585, 2010.

[30] M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning", Hamilton, New Zealand, 1998.

[31] S. Maldonado, R. Weber, and J. Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines", Information Sciences, vol. 181, no. 1, pp. 115–128, 2011.

[32] H. Kim, M. I. M.Ishag, M. Piao, T. Kwon, and K. H. Ryu, "A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries", Symmetry, vol. 8, no.6, 47, 2016.

[33] P, Li, Y. Piao, H. S. Shon, K. H. Ryu, "Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data", BMC bioinformatics, , 16(1): 347, 2015.

권태일(Tae-il Kwon)

1987년: 숭실대학교 숭실대학교 (공학사)
1991년: 한양대학교 산업대학원 전자공학과 (공학석사)
2013년 ~ 현재: 충북대학교 전자정보대학원 (박사수료)

1998년~2002년 교육 과학 기술부 교육정보화 위원
2004년~2006년 교육 과학 기술부 이러닝 위원
2000년~2005년 한국 정보처리학회 이사
2005년~2006년 한국 정보과학회 감사
2006년~2008년 한국 IT 서비스학회 이사
2007년~2008년 한국 정보처리학회 이사
2008년~2011년 한국 정보처리학회 감사
2011년~현재 빅썬시스템즈 대표이사
※관심분야 : 데이터마이닝(Data Mining), 바이오메디칼 및 바이오인포매틱스 (Biomedical and Bioinformatics),마케팅 (Marketing) 등

이정곤(Dingkun Li)

2009년: 중국 하얼빈공업대학교 (공학사)
2014년: 충북대학교 전자정보대학원 (공학석사)
2014년 ~ 현재: 충북대학교 전자정보대학원 (박사수료)

2009년~2011년 중국 Kingdee Co. LTd. 소프트웨어 엔지니어
2011년~2013년 중국 United International College, 강사
※관심분야 : 데이터마이닝(Data Mining), 헬스케어(Healthcare), 빅 데이터 (Big Data)

박현우(Hyun Woo Park)

2011년: 공주대학교 산업시스템공학과 (공학사)
2011년 ~ 현재: 충북대학교 전자정보대학원 (석박사통합수료)

※관심분야 : 데이터마이닝(Data Mining), 헬스케어(Healthcare), 빅 데이터 (Big Data), 바이오메디칼 및 바이오인포매틱스 (Biomedical and Bioinformatics) 등

류광선(Kwang Sun Ryu)

2017: 충북대학교 컴퓨터과학과 (공학박사)

2017년 ~ 현재: 충북대학교 줄기세포 연구소 박사 후 연구원
※관심분야 : 데이터마이닝(Data-Mining), 메디컬인포메틱스 (Medical Informatics), 패턴마이닝 (pattern-mining)

김의탁(Eui Tak Kim)

1997년 2월 : 대전대학교 컴퓨터공학과(공학사)
1999년 8월 : 대전대학교 컴퓨터공학과(공학석사)
2018년 현재 : 충북대학교 전자계산학과(공학박사)

2001년 ~ 2005년 : ㈜아이언마스크 기술이사
2005년 ~ 2010년 : 티에스온넷(주) 정보보호연구소 부장
2010년 ~ 현재 : ㈜하우리 기술연구소 연구소장
※관심분야 : 접근통제시스템, 악성코드 분석 및 탐지, 데이터 마이닝, 클라우드 컴퓨팅, 핀테크 보안, 인공지능

박명호(Minghao Piao)

2007: 중국 연변과학기술대학교(공학사)
2009: 충북대학교 바이오인포메틱스 (공학석사)
2014: 충북대학교 컴퓨터과학 (공학박사)

2015~2017: 동국대학교 경주캠퍼스 컴퓨터공학과 연구교수
2017~현재: 충북대학교 소프트웨어학과 초빙교수
※관심분야 : 데이터마이닝, 바이오인포메틱스, 빅데이터