Projected Clustering Methods for Predicting Heart Disease

Heon Gyu Lee[1]*, Jong Seol Lee[2], Hyun-Sup Kang[1], Keun Ho Ryu[2]

[1] GAION Ltd. Cheongrim-Bldg, 19 Samsung-ro 96 Gil, Kangnam-Gu, Seoul, Republic of Korea.
[2] Chungbuk National Univ. Chungdae-ro 1, Seowon-Gu, Cheongju, ChungbuK, Republic of Korea.

* Corresponding author. Tel.: +82 02-2051-9595; email: hglee@gaion.kr

Manuscript submitted October 20, 2018; accepted November 10, 2018.
doi: 10.17706/ijbbb.2019.9.2.73-81

Abstract: Supervised and unsupervised learning techniques are increasingly applied to improve medical decision-making. Medical-recorded data also have accumulated large amount of information about patients and their medical conditions. Relationship and patterns within this data could provide new medical knowledge. Unfortunately, few methodologies have been developed and applied to discover this hidden knowledge. In this paper, we propose projected clustering method for generating clusters of similar bio-signal patterns from medical data to be analyzed and the various classification methods for reflecting information of heart signal on the classification/prediction model. The experiments show that the optimal cluster is constructed by applying PROCLUS algorithm and it has from 0.881 to 0.9 f1-value index of prediction under test data.

Key words: Heart disease, ECG, de-identification, projected clustering and medical big data.

1. Introduction

Modern medicine generates almost daily, huge amounts of heterogeneous data. For example, medical data may contain bio-signals like ECG, clinical information like temp., cholesterol levels, etc., as well as the physician's interpretation. Those who deal with such data understand that there is a widening gap between data collection and data comprehension. Computerized techniques are needed to help humans address this problem. This paper is devoted to the relatively young and growing field of medical big data analysis. As more and more medical procedures employ bio-signal as a preferred diagnostic tool, there is a need to develop methods for efficient machine learning in big data of bio-signals. Other significant features are security and confidentiality concerns. Moreover, the physician's interpretation of signals and clinical information or other technical data, is written in unstructured language which is very difficult to learning. We propose the application of a standard machine learning technique to the case of bio-signal data. In this paper, cluster analysis and classification techniques are applied in order to predict an accurate heart disease in the medical data. Especially this study uses projected clustering method for generating clusters of similar wind power patterns from data to be analyzed. The framework proposed in this study for the prediction of wind power pattern is as follows.

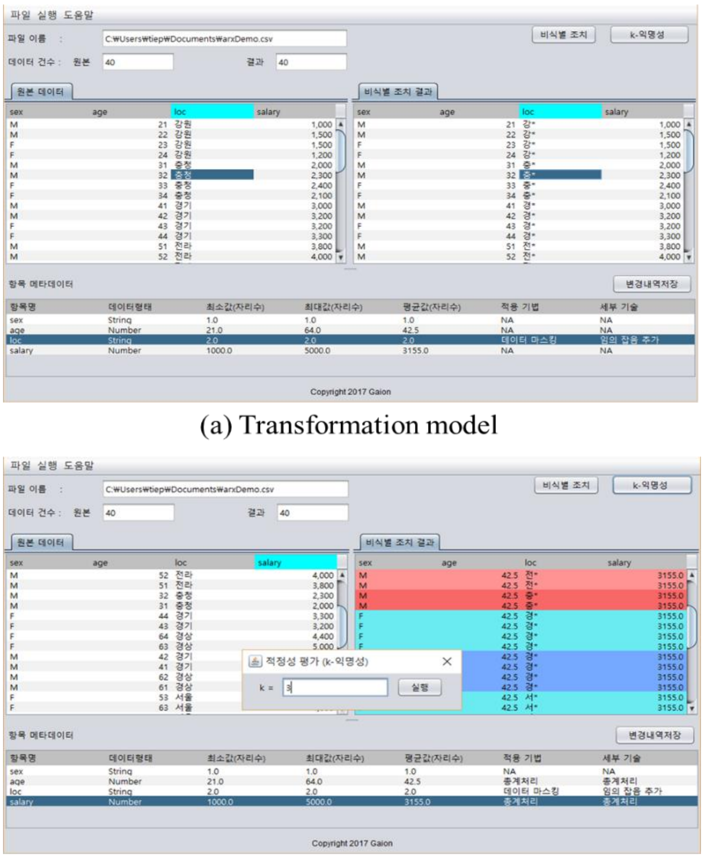
1) Data preprocessing: To preserve privacy, we remove private information. Our method is de-identification for anonymizing sensitive personal data.
2) Projected cluster analysis: to generate a similar group of bio-signal patterns from the preprocessed data
3) Classification: to build the prediction model considering bio-signal features by extending the

projected clustering methods.

2. Data Preprocessing

Privacy preserving data analysis has become an important issue because of the large amount of personal data. Several techniques that are used to preserve privacy are to remove private information. However, private information can still be leaked if abnormal users possess some background knowledge and other data sources. In this study, we proposed a methodology that preserve privacy. Our proposed method is de-identification software for anonymizing sensitive personal data. It supports a wide variety of 1) privacy and risk models, 2) methods for transforming data and 3) methods for analyzing the usefulness of output data (k-anonymity). The software has been used in a variety of contexts, including big data analytics platforms, research projects, human or animal medical data sharing and for training purposes.

To preserve privacy, we develop transformation and privacy model. Fig. 1 is an example of our de-identification software [1].



파일 이름 : C:₩Users₩tiep₩Documents₩arxDemo.csv

데이터 건수 : 원본 40 결과 40

원본 데이터

sex	age	loc	salary
M	21	강원	1,000
M	22	강원	1,500
F	23	강원	1,500
F	24	강원	1,200
M	31	충청	2,000
M	32	충청	2,300
F	33	충청	2,400
F	34	충청	2,100
M	41	경기	3,000
M	42	경기	3,200
F	43	경기	3,200
F	44	경기	3,300
M	51	전라	3,800
M	52	전라	4,000

비식별 조치 결과

sex	age	loc	salary
M	21	강*	1,000
M	22	강*	1,500
F	23	강*	1,500
F	24	강*	1,200
M	31	충*	2,000
M	32	충*	2,300
F	33	충*	2,400
F	34	충*	2,100
M	41	경*	3,000
M	42	경*	3,200
F	43	경*	3,200
F	44	경*	3,300
M	51	전*	3,800
M	52	전*	4,000

항목 메타데이터

항목명	데이터형태	최소값(자리수)	최대값(자리수)	평균값(자리수)	적용 기법	세부 기술
sex	String	1.0	1.0	1.0	NA	NA
age	Number	21.0	64.0	42.5	NA	NA
loc	String	2.0	2.0	2.0	데이터 마스킹	임의 잡음 추가
salary	Number	1000.0	5000.0	3155.0	NA	NA

Copyright 2017 Gaion

(a) Transformation model

파일 이름 : C:₩Users₩tiep₩Documents₩arxDemo.csv

데이터 건수 : 원본 40 결과 40

원본 데이터

sex	age	loc	salary
M	52	전라	4,000
M	51	전라	3,800
M	32	충청	2,300
M	31	충청	2,000
F	44	경기	3,300
F	43	경기	3,200
F	64	경상	4,400
F	63	경상	5,000
M	42	경기	
M	41	경기	
M	62	경상	
M	61	경상	
F	53	서울	
F	63	서울	

적정성 평가 (k-익명성)

k = 3 실행

비식별 조치 결과

sex	age	loc	salary
M	42.5	전*	3155.0
M	42.5	전*	3155.0
M	42.5	충*	3155.0
M	42.5	충*	3155.0
F	42.5	경*	3155.0
F	42.5	경*	3155.0
F	42.5	경*	3155.0
F	42.5	경*	3155.0
	42.5	경*	3155.0
	42.5	경*	3155.0
	42.5	경*	3155.0
	42.5	경*	3155.0
	42.5	서*	3155.0
	42.5	서*	3155.0

항목 메타데이터

항목명	데이터형태	최소값(자리수)	최대값(자리수)	평균값(자리수)	적용 기법	세부 기술
sex	String	1.0	1.0	1.0	NA	NA
age	Number	21.0	64.0	42.5	총계처리	총계처리
loc	String	2.0	2.0	2.0	데이터 마스킹	임의 잡음 추가
salary	Number	1000.0	5000.0	3155.0	총계처리	총계처리

Copyright 2017 Gaion

(b) K-anonymity model

Fig. 1. A screenshot of de-identification software.

2.1. Transformation Model

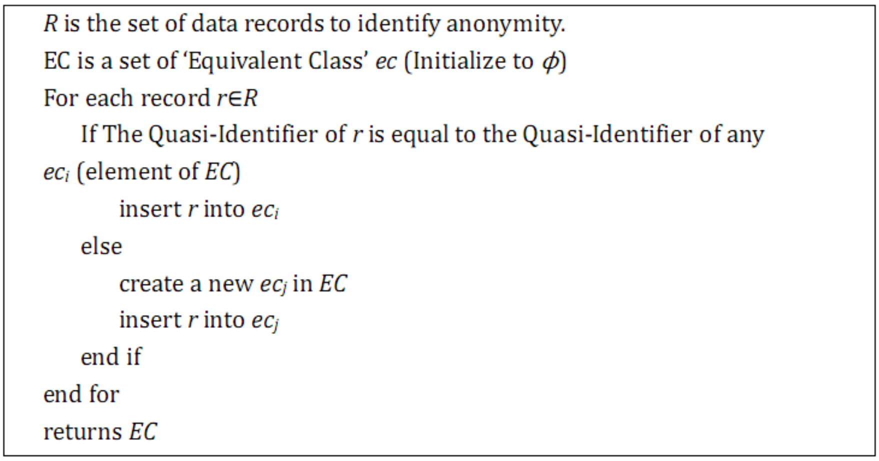
Our software supports a variety of data transformation models, which can also be combined with each other. - Value generalization: User-specified generalization hierarchies form the backbone of data transformation mechanism. Hierarchies can either be used to directly reduce the uniqueness of attribute values or to form clusters that will be transformed using further methods, such as micro-aggregation.

- Random sampling: this model supports multiple methods for drawing a sample from the input dataset. This can be used to relate a dataset to an underlying population table or to reduce privacy risks.

- Record, attribute and cell suppression: As described previously, this model also supports removing individual attributes, attribute values or complete records in the transformation process. This can be controlled by defining appropriate hierarchies (which is supported by specific wizards), by performing local or global transformation and by specifying a limit for the maximal number of records which may be removed.
- Micro-aggregation: Sets of numeric attribute values can be transformed into a common value by user-specified aggregation functions. Prior to aggregation, clustering can be performed based on value generalization hierarchies.
- Categorization: The method provided by the software can be used to create transformation rules that are represented as functions, which can be used to perform on-the-fly categorization of continuous variables during anonymization.

2.2. K-Anonymity

This well-known privacy model aims at protecting datasets from re-identification in the prosecutor model. A dataset is k-anonymous if each record cannot be distinguished from at least k-1 other records regarding the quasi-identifiers. Each group of indistinguishable records forms a so-called equivalence class. The algorithm is shown in Fig. 2



```
$R$ is the set of data records to identify anonymity.
EC is a set of 'Equivalent Class' $ec$ (Initialize to $\phi$)
For each record $r \in R$
    If The Quasi-Identifier of $r$ is equal to the Quasi-Identifier of any
$ec_i$ (element of $EC$)
        insert $r$ into $ec_i$
    else
        create a new $ec_j$ in $EC$
        insert $r$ into $ec_j$
    end if
end for
returns $EC$
```

Fig. 2. K-anonymity algorithm.

3. Projected Clustering

In this study, we use projected clustering approaches for discovering representative feature selection/extraction. Projected clustering is a method for detecting clusters with the highest similarity from the subset of all data dimensions. The biggest difference from the traditional clustering methods is detecting various subsets based on the fact that subsets differ from each other and they include meaningful clusters rather than considering all dimensions given during the clustering process. For example, if attribute is a dimension and bio-signal value is an object in the data, through the projected clustering shown in Fig. 3, time intervals (t7 through t15) can be detected, which are subsets having discriminating power among different clusters.

Projected clustering approaches can be divided into three paradigms based on the detection methods of the subsets [2].

The first paradigm is to divide a data space into grid–cells (cell–based) and form clusters from the cells with sufficient density. The basic concept is defining grid–cell sets first before assigning objects to the proper cells, and calculating the density of each cell. Next, cells with a density of a certain threshold or lower are removed and clusters are built from a series of cells with high density. A popular cell–based method, CLIQUE [3], is a grid–based clustering algorithm that detects clusters of subsets through certain procedures. When multi–dimensional data points of large capacity are given, the data space in general is not uniformly occupied by the data points. The clustering of this approach discriminates sparse and crowded regions in space (or unit), and detects the entire distribution type of the datasets. Clustering of CLIQUE is defined as the biggest group of connected dense units. SCHISM [4] finds subsets by using a support and Chernoff–Hoeffding bound concept and determines the interesting subsets using a depth–first search and backtracking. The second paradigm is density–based projected clustering. It builds clusters by identifying areas with high density divided by areas of low density. Though the overall clustering concept is based on DBSCAN [5], the density calculation here considers only the relevant dimensions. The representative algorithm is FIRES [6], and it applies an efficient filter–refinement method. Above all, the existing base–clusters are created and those that fail to meet the given density conditions are removed in the filtering stage. Next, the base–clusters are merged to create the maximal dimensional projected cluster approximations. Lastly, the final refined clusters are built during the refinement stage. The SUBCLU [7] is a DBSCAN–based greedy algorithm for projected clustering. Unlike grid–based approaches, it can detect clusters with arbitrary shapes. The third paradigm is a clustering–oriented approach. As the data dimension increases in the clustering for high dimensional data, clustering that considers all dimensions can hinder the performance remarkably owing to the presence of sparse data. PROCLUS [8], a famous algorithm, starts from a single dimensional space. Instead, the algorithm of the third paradigm begins by searching the initial estimation regarding clusters in a high dimensional space. Weight is provided for each cluster per each dimension, and the renewed weight is used to create clusters again for the next iteration. STATPC [9] detects relevant subsets based on objects and builds candidate subspaces, which are refined to build local optimal projected clusters. Finally, a greedy search algorithm is used to review all subspaces and build optimal clusters. Table 1 shows the properties of the clustering algorithms used in our study.

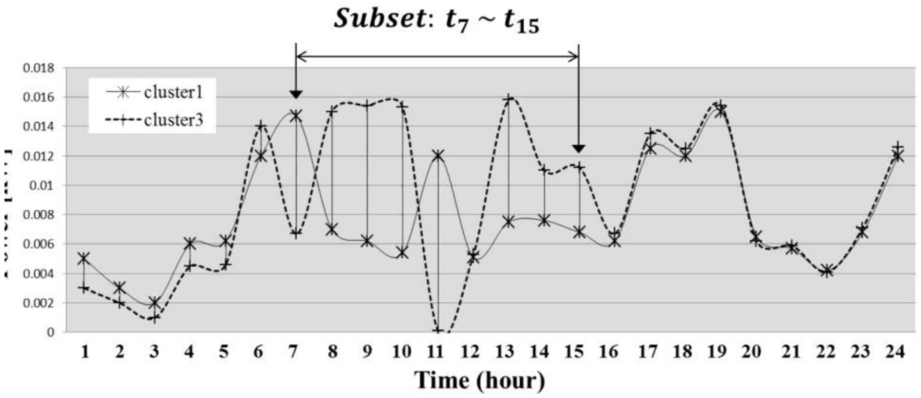


Fig. 3. Different bio-signal patterns of different clusters.

Table 1. Properties of Three Paradigms

Paradigm	Algorithm	Properties
Cell–based	CLIQUE	fixed threshold and grid size, pruning by monotonicity property
	SCHISM	enhanced CLIQUE by variable threshold, using

		heuristics for pruning
Density–based	FIRES	variable density threshold, based on filter–refinement architecture to drop irrelevant base–clusters
	SUBCLU	fixed density threshold, pruning by monotonicity property
Clustering–oriented	PROCLUS	fixed cluster number, iteratively improving result like k–means, partitioning
	STATPC	statistical tests, reducing result size by redundancy elimination

4. Classification Model for Predicting Heart Disease

Feature vectors for the classifier's supervised learning include prior information such as the ST, RRI, PRI, QRS and QTI (see Fig. 4) from ECG signal data, which is a heart disease feature, and class labels are built through clustering. Among all features used for learning.

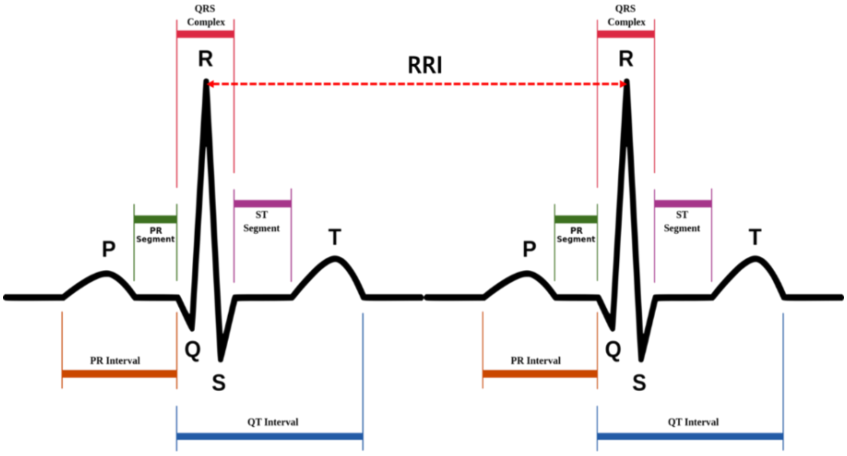


Fig. 4. Bio-signal features for predicting heart disease.

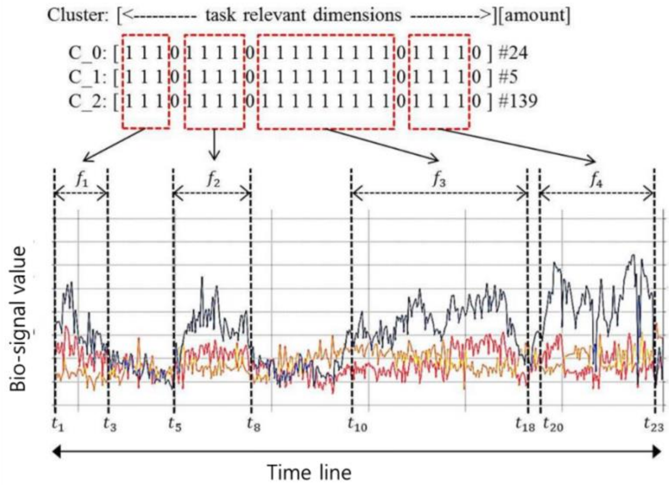


Fig. 5. Example of features for bio-signal.

For instance, Fig. 5 describes the bio-signal dimension involved in building three clusters, C_0, C_1, and C_2. Signal intervals (f_1, f_2, f_3, f_4), i.e., the subsets of all dimensions that are applied to the clustering of the different groups. The classifiers used in the study are the sequential minimal optimization (SMO) algorithm, which shows an excellent performance, and *k*NN and NNs which were evaluated in a related paper [10], [11]. Building representative bio-signal patterns through a clustering analysis of the heart disease patterns can characterize the change in bio-signal patterns depending on the time of objects inside the clusters. The cluster analysis stage describes the cluster analysis results of only SCHISM, FIRES, and PROCLUS. The optimal parameter settings of each algorithm for the data in the test are given in Table 2.

Table 2. Parameter Bracketing for Each Algorithm

Algorithm	Parameter	From	Offset	Op	Steps	To
Cell–based (SCHISM)	TAU	0.1	0.1	+	10	1.0
	XI	1	1	+	24	24
	U	0.05	0	+	1	0.05
	Total number of experiments: 240 (steps: 10*24*1)					
Density–based (FIRES)	BASE_DBSCAN _EPSILON	1.0	0	+	1	1.0
	BASE_DBSCAN _MINPTS	100	0	+	1	100
	GRAPH_K	15	1	+	4	18
	GRAPH_MIN CLU	1	1	+	4	4
	GRAPH_MU	1	1	+	4	4
	GRAPH_SPLIT	0.66	0	+	1	0.66
	POST_DBSCAN _EPSILON	300	0	+	1	300
	POST_DBSCAN _MINPTS	24	0	+	1	24
	PRE_MINIMUM PERCENT	10	0	+	1	10
	Total number of experiments: 64 (steps: 1*1*4*4*4*1*1*1*1)					
Clustering–oriented (PROCLUS)	average Dimensions	1	1	+	24	24
	numberOfClusters	2	1	+	8	9
	Total number of experiments: 192 (steps: 24*8)					

Because the projected clustering algorithm groups similar signal patterns for the all dimensions in the training datasets and classifies which group the test data objects belong to out of the defined clusters, it includes the clustering and classification methods together. Therefore, an evaluation measure such as sum of the squared error or normalized mutual information for a traditional clustering method is inappropriate. The present study used evaluation measures such as the precision, recall, and F1–value to evaluate the three clustering algorithms.

Formal definitions of these measures are given below.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{1}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{2}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

* True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

The bio-medical dataset was applied with five dimensions depending on the parameter set with the range shown in Table 3, and the optimal number of clusters determined was five. Based on the results in Table 4, PROCLUS was used along with the clustering results shown in Table 4.

Table 3. Confusion Matrix of Original vs. Discovered Groups

Original groups	Discovered groups (Cluster:5)				
	C1→ risk	C2 → normal	C3→outlier	C4 →attention	C5→attention
Cluster1 (normal)	0	3810	254	508	0
Cluster2 (attention)	57	228	114	2691	501
Cluster3 (risk)	3426	0	0	0	0

All tests for the three clustering measures use 10–fold cross validation. In addition, the original classes used to evaluate the algorithms in the cluster analysis stage are the patient group IDs (control, attention, risk) of the three regions. Table 4 shows the result of the evaluators based on the clustering methods for datasets.

Table 4. Description of the Summary Results

Algorithm	Precision	Recall	F1	class
Cell–based (SCHISM)	0.748	0.873	0.805	Control
	0.688	0.55	0.611	Attention
	0.647	0.423	0.512	Risk
Density–based (FIRES)	0.762	0.912	0.83	Control
	0.655	0.475	0.551	Attention
	0.824	0.538	0.651	Risk
Clustering–oriented (PROCLUS)	**0.809**	**0.873**	**0.881**	**Control**
	0.769	**0.75**	**0.937**	**Attention**
	0.579	**0.423**	**0.9**	**Risk**

5. Conclusion

This paper use three projected clustering approaches to discover representative bio-signal patterns from data measured from ECG. As subsets of all dimensions required for clustering and an appropriate composition of the clusters were used concurrently, the removal of identification and sensitive information the use of a de-identification method are included. The optimal number of clusters was determined using parameter bracketing provided by PROCLUS algorithm which is a clustering–oriented approach produced the best results.

Currently, heart disease data continue to be accumulated and refined. Therefore, more accurate prediction of the heart disease signal patterns will be possible in future studies.

Acknowledgment

This work was carried out with the support of “Cooperative Research Program for Agriculture Science and Technology Development (Project No. PJ013951)” Rural Development Administration, Republic of Korea.

References

[1] Prasser, F., Eicher, J., Bild, R., Sengler, H., & Kuhn, K. A. (2017). A tool for optimizing de-identified health

data for use in statistical classification. *Proceedings of the 30th IEEE Int'l Symposium on Computer-Based Medical Systems*.

[2] Müller, E., *et al*. (2009). Evaluating clustering in subspace projections of high dimensional data. *VLDB, Vol. 2*, (pp. 1270-1281).

[3] Agrawal, R., *et al*. (1998). Automatic subspace clustering of high dimensional data for data mining applications. *Proceedings of International Conference on Management of Data, Vol. 27*, (pp. 94-105).

[4] Sequeira, K., & Zaki, M. (2004). SCHISM: A new approach for interesting subspace mining. *IEEE Conference on Data Mining,* (pp. 186-193).

[5] Ester, M., *et al*. (1998). Algorithms for characterization and trend detection in spatial databases. *Proceedings of International Conference on Knowledge Discovery and Data Mining*, (pp. 44-50).

[6] Kriegel, H. P., *et al*. (2005). A generic framework for efficient subspace clustering of high–dimensional data. *Proceedings of International Conference on Data Mini*ng, (pp. 250-257).

[7] Kailing, K., Kriegel, H. P., & Kroger, P. (2004). Density–connected subspace clustering for high–dimensional data. *Proceedings of SIAM Conference on Data Mining*, (pp. 246-257).

[8] Aggarwal, C., *et al*. (1999). Fast algorithms for projected clustering. *Proceedings of International Conference on Management of Data*, (pp. 61-72).

[9] Moise, G., & Sander, J. (2008). Finding non–redundant, statistically significant regions in high dimensional data. *Proceedings of International Conference on KDD 2008*, (pp. 533-541).

[10] Akhil, M., Deekshatulua, B. L., & ChandraRahmani, P. (2013). Classification of heart disease using K-nearest neighbor and genetic algorithm. *Proceedings of International Conference on Computational Intelligence: Modeling Techniques and Applications,* (pp. 85-94).

[11] Yu, O., *et al* (2012) Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and clinical parameters. *Journal of Cardiology*, *59*, 190-194.

Heon Gyu Lee received the M.S. and Ph.D. degrees in computer science from Chungbuk National University, Cheongju, Korea, in 2005 and 2009, respectively. In 2016, he joined GAION, Seoul, Rep. of Korea, and works for big data research lab. As a senior vice President. His research interests include data mining, DB, bioinformatics, machine learning and big data analysis.

Jong Seol Lee is with the Database/Bioinformatics Laboratory, School of Electrical and Computer Engineering, Chungbuk National University, South Korea. Currently, he is studying master's course in database bioinformatics laboratory with his major in smart factory in Chungbuk National University (CBNU), South Korea. His research interests include statistics, and smart factory.

Hyun Sup Kang received the bachelor's degrees in social education from Seoul National University, Seoul, Korea, in 1997 and 2000, respectively. In 2007, he established GAION, Seoul, Rep. of Korea. His research interests include database, bioinformatics, machine learning, software engineering, and big data analysis.

Keun Ho Ryu received the PhD degree in computer science and engineering from Yonsei University, South Korea, in 1988, and served in the Reserve Officers' Training Corp (ROTC) of the Korean Army. He is also an honorary doctorate of the National University of Mongolia. He is currently a professor with Chungbuk National University, South Korea, and has been a leader of the Database and Bioinformatics Laboratory, South Korea, since 1986. He has worked at the University of Arizona, U.S.A., as a postdoctoral and a Research Scientist, and also at the Electronics and Telecommunications Research Institute, South Korea, as a Senior Researcher. He is a former Vice-President of the Personalized Tumor Engineering Research Center. Prof. Ryu has served on numerous program committees including roles as a demonstration co-chair of the VLDB, as a panel and tutorial co-chair of the APWeb, and as a FITAT general co-chair. He has published or presented over 1000 referred technical articles in various journals and international conferences, in addition to authoring a number of books. His research interests include temporal databases, spatiotemporal databases, temporal GIS, stream data processing, knowledge-based information retrieval, data mining, biomedical informatics and bioinformatics.