Journal of Universal Computer Science, vol. 24, no. 9 (2018), 1192-1216
submitted: 30/12/17, accepted: 30/6/18, appeared: 28/9/18 © J.UCS

An Effective Risk Factor Detection and Disease Prediction (RFD-DP) Model Applied to Hypertension

Dingkun Li, Yaning Li, Zhou Ye
(Karamay Central Hospital, Karamay, China
{33644251, 153422470}@qq.com, yezhou126@126.com)

Musa Ibrahim, Keun Ho Ryu
(Chungbuk National University, Cheongju, South Korea
{ibrahim, khryu}@dblab.chungbuk.ac.kr)

Seon Phil Jeong
(BNU-HKBU United International College, Zhuhai, China
spjeong@uic.edu.hk)

Abstract: Never before in history is the data growing at such a high volume, variety and velocity. It not only provides multi-sources of information for people to discover useful, important and valuable nuggets of information, but also increases the difficulty in finding such nuggets in almost all fields. Particularly, the field of healthcare is known for its dominical or ontological complexity and variety of clinical data or medical data regarding its variable data standards and data quality and so as the high data dimensionality. In order to effectively use the data at the hand to improve healthcare outcomes and processes, this paper illustrates a model called Risk Factor Detection and Disease Prediction (RFD-DP) model. The model incorporates statistics, data mining and MapReduce techniques on high dimensional clinical data to detect risk factors and generate predicator for a specified disease, hypertension disease. The experimental results indicate that the proposed model outperforms traditional feature selection and classification methods in terms of accuracy, F-score, and AUC. Consequently, the proposed model is promising to be applied to healthcare system.

Keywords: Feature selection, classification, high dimensional data, MapReduce
Categories: J.0, J.3, H.4

1 Introduction

In aging society, the ratio of elderly people has increased in Korea. The percentage of people with chronic disease is also increasing as the population of the elderly increasing. According to a recent report [Kim, 15], 35% of the $ 5.4 billion fund allocated by the Korean government, was spent in medical care due to chronic disease conditions.

Work [Ezzeldin, 17] depicted a bibliometric review via CiteSpace on big data in Healthcare. The detection and prevention of chronic diseases are becoming an important issue in the world and so as Korea. According to the work [Baek, 16], it drew several conclusions: First, the prevalence of chronic diseases increases by 2040 in Korea. The population with hypertension increases 2.04 times; diabetes increases 2.43 times, and cancer increases 3.38 times. Second, health expenditure on chronic

diseases increases as well. Health expenditure on hypertension increases 4.33 times (1,098,753 million won in 2014 to 4,760,811 million won in 2040); diabetes increases 5.34 times (792,444 million won in 2014 to 4,232,714 million won in 2040); and cancer increases 6.09 times.

Data mining techniques have been widely used for the purpose of disease detection based on various data sources. It provides a promising way for disease detection with high accuracy and efficiency. Over many years, a large amount of health-care research work has been completed using DM techniques. In [Kim, 16] and [Amendola, 14], the authors used classification and regression techniques to predict conditions like cardiovascular disease and heart disease. In [Huang, 07] and [Ha, 10], integrated DM techniques are provided for the detection of chronic and physical diseases. Further, a number of other research works, like [Piao, 15], used the advantages of DM to develop new methodologies and frameworks for health-care purposes. However, it becomes difficult to handle continuously growing big data, many tools are emerging to mine big data, such as Hadoop.

In recent years, the Hadoop framework has been widely used for the delivery of health care as a service [Kaur, 14]. Moreover, a wide variety of organizations and researchers have used Hadoop for health-care services and clinical-research projects [Hsu, 12]. Taylor provided a detailed introduction on the use of Hadoop in bioinformatics [Taylor, 10], while Schatz developed an operations support system (OSS) package named CloudBurst that provides an algorithmic parallelization model for which Hadoop MapReduce is used [Schatz, 09]. Indeed, the Hadoop framework has been employed in numerous important works to provide major contributions to the health-care field. The other big-data processing framework, Spark, leverages a synergistic combination of the smartphone and the smartwatch in the monitoring of multidimensional symptoms such as facial tremors, dysfunctional speech, limb dyskinesia, and gait abnormalities [Sharma, 14]. [Li, 18] has developed an big data platform used to analysis the big healthcare data.

Though a lot of work has been done, there are some issues remain for developing healthcare system. The first issue is that it is extremely difficult to get the high-quality relevant data from the observations. One reason is that agencies (such as hospital, family, government, health care center etc.) are not willing to share patients' information according to the privacy policy. Another reason is that medical or healthcare data come from multi-source which consists of redundant, missing, irrelevant and wrong data. Furthermore, this data lacks consistency in respect of its structured, semi-structured and unstructured types. There is no existing system that can handle three kinds of data one for all as far as authors know. Thirdly, most systems lack of the flexible ability to handle both big data and non-big data. High cost is another issue for developing healthcare system due to its expensive equipment such as sensors, alert devices and even call center.

In order to overcome previous problems, the purpose of authors' work is to design an effective model which can be used to 1). detect HTN disease risk factors on high dimensional data set to overcome the problem of "curse of dimensionality". 2). generate the rule-based classifier to predict HTN based on the output of the previous step. We hope the model can be used to improve the accuracy and efficiency of disease detection result.

The remaining paper is organized as follows. Section 2 reviews the disease studied in our work and widely used data mining techniques. Section 3 describes the disease risk factor detection and disease prediction (RFD-DP) model. Section 4 gives a detailed step-by-step implementation depiction of the RFD-DP model. Section 5 gives the experimental result of the proposed method on three data sets. Section 6 makes a summary and discussion of the research and highlights directions for further research.

2 Literature Review

In this chapter, it reviews the disease studies and widely used data mining techniques for this disease detection as well as its risk factor detection.

2.1 Hypertension

Hypertension (HTN) is a condition in which a person's blood pressure is above the normal or optimal limit of 120 mmHg for systolic pressure and 80 mmHg for diastolic pressure. Increased blood pressure in the long term can lead to conditions that could threaten the health of the sufferer. Several conditions can cause disturbances in hypertensive cardiovascular organs such as stroke and heart failure. Hypertension is also called as a silent killer because sufferers sometimes do not realize that he was exposed to conditions of hypertension [Chobanian, 03]. The classification of blood pressure in adult includes 4 classes which have been given in Table 1.

Classification	**Systolic (mmHg)**	**Diastolic (mmHg)**
Normal	<120	And <80
Pre-Hypertension	120~139	or 80~89
HTN Stage 1	140~149	or 90~99
HTN Stage 2	≥150	or ≥100

Table 1: Blood pressure classification

2.2 Feature Selection

Feature selection aims at finding the most relevant features of a problem domain. It is one essential step for data mining which is defined as a multidisciplinary task to find out hidden nugget of information from data [Singh, 14]. Primarily, there are three kinds of feature selection methods, filters, wrappers and embedded methods. The filter methods work fast but its result is not always satisfactory. While the wrapper methods guarantee good results but very slow when applied to wide feature sets which contain thousands or even hundreds of thousands number of features. The third one is embedded methods which reduce the computation time taken up for reclassifying different subsets which is done in wrapper methods.

Following figures compare the difference between filter, wrapper, and embedded methods [Saurav, 16]. The procedure of filter method is shown in Figure 1,

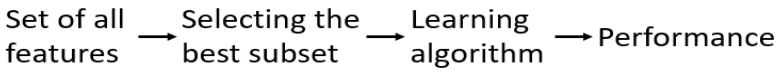

Figure 1: Filter method analysis procedure

Filter methods are generally used as a pre-processing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. The correlation is a subjective term here.

The procedure of the wrapper method is shown in Figure 2.

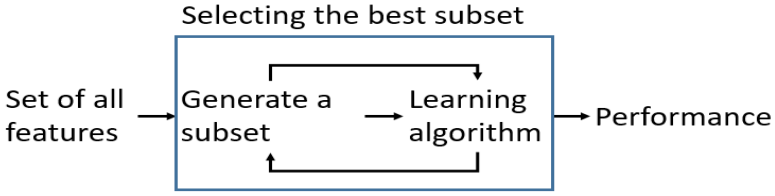

Figure 2: Wrapper method analysis procedure

In wrapper methods, it uses a subset of features and trains a model using them. Based on the inferences that it draws from the previous model, it decides to add or remove features from the subset. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive.

The procedure of the embedded method is shown in Figure 3.

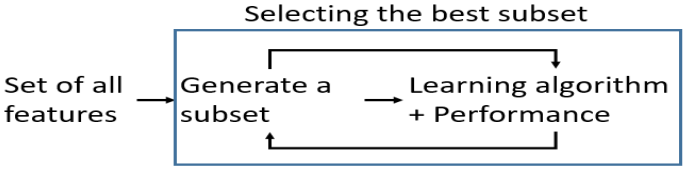

Figure 3: Embedded method analysis procedure

The embedded method combines the qualities of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods.

The benefits of feature selection are reducing data analysis complexity and improve data analysis performance. Besides, there are more benefits such as accuracy improvement, expenditure reduction, etc.

2.3 Classification Methods

Classification is a supervised method defined as a process grouping data objects into one of the predefined classes [Han, 11]. It is widely used in medical research and healthcare field. Trained classifier can be used as the predictor for disease detection. For example, a patient can be classified as "hypertension" and "non-hypertension" based on disease pattern. Several classification methods are used in our work.

Decision Tree (DT):

The tree comprises of the root, non-leaf nodes, and leaf nodes. Each non-leaf node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label [Han, 11].

C5.0 is decision tree based algorithm which is an improved version of C4.5. It includes all functionalities of C4.5 and applies several new technologies, among them the most important application is "boosting" [Freund, 09] technology for improving the accuracy rate of identification on samples.

Another widely used DT method is Chi-square Automatic Interaction Detector (CHAID) [Kass, 80], it creates all possible cross tabulations for each categorical predictor until the best outcome is achieved based on Chi-square test result and no further splitting can be performed it uses multiway splits by default, it needs rather large sample sizes to work effectively. In CHAID analysis, nominal, ordinal, and continuous data can be used.

Logistic Regression (LR):

A powerful statistical method measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. It only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.) as shown in Eq. 1

$$\log[\frac{p(x)}{1-p(x)}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad \text{(Eq. 1)}$$

Where p(x) is the probability of the presence of the characteristic of interest.

Figure 4 compares the Linear Regression with LR.

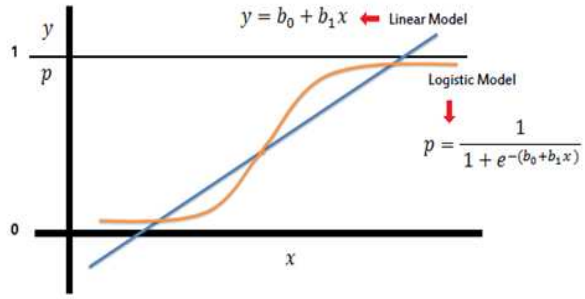

Figure 4: Comparison between linear model and logistic model

K-Nearest Neighbour (K-NN):

When a new tuple whose class is unknown is given, the k-NN classifier compares its proximity with the k-nearest training tuples and assigns the class of K-NN with majority vote or distance weighted vote to the new tuple. Some of the widely used proximity measures for finding nearest neighbours include Euclidean distance, Manhattan distance, Simple Matching coefficient, Jaccard similarity coefficient, Cosine similarity, and correlation coefficient. The similarity function is usually defined in Eq. 2 :

$$D(x,y) = \sqrt[p]{\sum_{i=1}^{n} | x_i - y_i |^p} \quad p \in (1,2,\ldots,n) \qquad \text{(Eq. 2)}$$

The advantages of KNN are that: 1). training speed is very fast, 2). it can be used to lean complex target functions, and 3). do not loose information. The disadvantages are that: 1). it is slow at query time (pre-sorting and indexing training samples into search trees reduces time), 2). it is easily fooled by irrelevant features (attributes). Figure 5 describes one example of distinguishing the class label of the green point, it could belong to red triangle class or blue square class which depends on the proximity measures for finding nearest neighbours. For example, which class does the green object belong to in Figure 5?

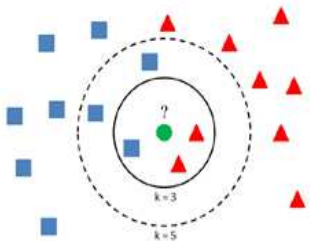

Figure 5: K-NN example

Boosting method:

Boosting is a widely used method developed to improve the performance of learning algorithms that generate multiple classifiers and vote on them [Rahman, 13]. It is an ensemble technique that attempts to create a good classifier from several not so good classifiers, and it is one of the best off-the-shelf classification methods drawing a lot of attention from researches. However, the original algorithm was developed for binary classification problems. In our research, a widely used boost method, called AdaBoost [Rokach, 10] is used to do the experiment. It is an algorithm for constructing a "strong" classifier as linear combination of "simple", and "weak" classifiers.

Naïve Bayesian (NB):

The Naïve Bayesian classifier is based on Bayes theorem and it has been studied widely since the 1950 [Cheng, 01]. The Naïve Bayesian assumption is: attributes that describe data instances are conditionally independent. One of the advantages about the NB classifier is that it can combine any kind of objects (e.g. time series, trees, etc.) to generate classifier, based on the probabilistic model specification. It uses Bayes' theorem and conditional probability to measure the probability of occurrence between classes and attributes and has the advantage of a low computational cost. Calculate the conditional probability of an instance belonging to a class. The likelihood is defined in Eq. 3.

$$p(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)} \quad \text{(Eq. 3)}$$

Where *d* is data, *h* is hypothesis, *P(h)* is prior belief (probability of hypothesis *h* before seeing any data). *P(d|h)* is likelihood (probability of the data 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 NB classification method is one of the most practical learning methods, and used very successfully in medical diagnosis and text classification.

Support Vector Machine (SVM):

Support Vector Machine (SVM) [Hearst, 98] has been used to select features and generate the classifier. 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 6. The key is to find the optimal hyperplane in order to find the maximum margin. The vectors on the dashed line are the support vectors.

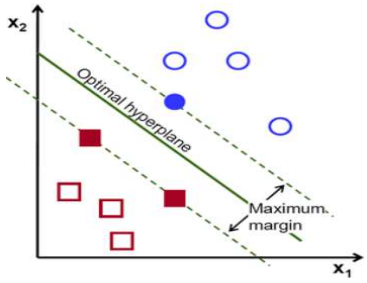

Figure 6: SVM example

Whereas to classify nonlinear data, the original training data is transformed into higher dimension using nonlinear kernel functions such as polynomial, radial, Gaussian, sigmoid etc. is the most powerful classification algorithms in terms of predictive accuracy. The optimal hyperplane is used to classify the data into different classes in two or more dimensionalities.

Given a set of instance-label pairs $(x_i, y_i), x_i \in R^n, y_i \in 1, -1, i = 1, \ldots, l$, SVM solves the following unconstrained optimization problem shown in Eq. 4:

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{l} \zeta(w, b; x_i, y_i) \quad \text{(Eq. 4)}$$

Where $\zeta(w, b; x_i, y_i)$ is a loss function, and C ≥ 0 is a penalty parameter on the training error, b refers to a bias and w denotes an orthogonal vector. Practically, a kernel function which is used to simplify the computation from high dimensionality to low, is defined in Eq. 5:

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \quad \text{(Eq. 5)}$$

SVMs have applications in many disciplines, including customer relationship management (CRM), facial and other image recognition, bioinformatics, text mining concept extraction, intrusion detection, protein structure prediction, and voice and speech recognition.

2.4 MapReduce Framework

MapReduce is a software paradigm for parallel data processing. It consists of two key steps i.e. Map and Reduce. Map is used to divide the large task into smaller ones, while Reduce aggregates result of each task.

Large data dimensionality can badly influence many aspects of data analysis process such as efficiency, accuracy and speed. For problems involving lists of data set with large-scale high dimensional data, the best and reliable way that was proven to be suitable is MapReduce programming technique [Mahmoud, 17]. The procedure can be divided into 2 phases, one is Mapping phase and another one is Reducing phase. Before Mapping phase, the paradigm divides the big data set into a certain number of smaller blocks, then Mapping worker nodes are used to handle the data (key-value pairs) in each block. The output is shuffled, sorted and aggregated by key for Reducing phase. The final output of Reducing worker is the key-value pairs as well; for example, word-count, disease-factors etc.

3 Design and Implementation of Proposed Model

We propose an ensemble framework for high-dimensional feature selection. The selected features are risk factors for HTN. Based on the output of the feature selection step, rule-based classification method CHAID is used to generate the classifier for HTN prediction. Each step of this procedure is depicted in detail in this section. The algorithms are implemented by JAVA, the .jar packages of these algorithms can be submitted to MapReduce cluster for data analysis.

3.1 Key Steps of Proposed Method

As it has been emphasized in the previous chapter, risk factor detection and disease predictor generation are the key steps for the whole model. The detailed framework of these two steps is shown in Figure 7.

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 to achieve a similar feature space for further integration.

Secondly, clinical data is natural 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. However, to our concern, the feature consists of more than 80% of the missing value that should be eliminated because the 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 correlation 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 paper, the 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 classification methods.
More detail will be described in the following sections.

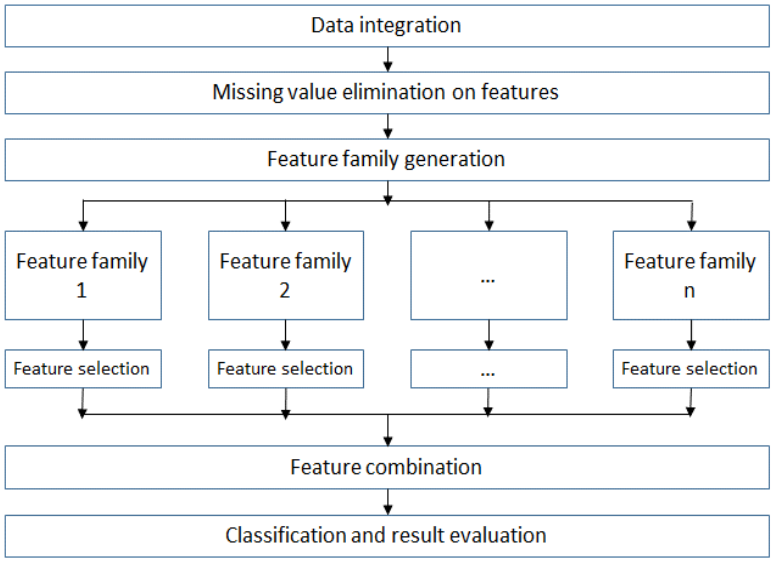

Figure 7: Key steps of proposed model framework

3.2 Feature Family Generation

Feature family is also called feature subset or column family. It is a group of features generated from the whole feature space. Figure 8 depicts one example of the feature families. We can see from the figure that name, sex, height, WC etc. belong to feature family 1 called basic information family, while salt, VB1 etc. belong to feature family 2 called nutrition intake feature family and so on. Ideally, feature families are independent of each other and all the features in the same family are of the same domain, or correlated with each other if they are not selected randomly.

There are three ways to generate the feature families; they are based on the feature domain, correlation or random combination. As the Figure 8 depicted that, features of the same domain are categorized into one family, such as name, sex, height, WC etc. which belong to the basic information family. For other dataset, which has no explicit domain information to categorize the features, some specified criteria such as correlation based methods [Hall, 98] or random feature combination methods are used to generate the feature families.

Feature id	Feature name	Specification	
B_00001	Name	Name of the observation	Feature Family 1 (basic info.)
B_00002	Sex	Gender of the observation	
B_00003	Height	Height of the observation	
B_00004	WC	Waist circumference	
...	...	...	
N_00200	Salt	Salt intake	Feature Family 2 (Nutrition intake)
N_00201	VB1	Vitamin B1 intake	
...	...	...	
L_00300	Alcohol_m	Monthly alcohol intake frequency	Feature Family 3 (lifestyle)
L_00301	Smokinng_m	Monthly smoking frequency	
...	...	...	
D_01108	HE_Dia1	Diabetes I	Feature Family 4 (disease info.)
D_01109	HE_HP	Hypertension	
D_01110	HE_N	Normal	
...	...	...	
			Feature Family n ...
...	...	...	

Figure 8: Examples of feature families

Nevertheless, it is hard to claim whether features within the same family are redundant. For example, for the nutrition intake feature family, nutrition is not redundant because all nutrition is 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 elimination plays an important role for model learning. A Fast Correlation-Based Filter (FCBF) [Yu, 03] approach was proposed to remove the redundant as well as irrelevant features. FCBF algorithm has been integrated into Weka, so we will not provide detail in this paper.

3.3 Feature Selection for Each Feature Family

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

At the beginning of this process, a natural question arises why control of redundancy is useful? Work of [Chakraborty, 15] implies that it can not only increase the effectiveness but also the accuracy of the results. Two filter methods: Information Gain [Thomas, 91], ReliefF [Kononenko, 94] and One wrapper method: Linear-SVM [Hearst, 98] are chosen and modified to be used on our datasets since the performance of these three methods are better than other feature selection algorithms according to our experiment. They are used for redundant feature elimination and candidate feature generation.

(1) Information Gain Based Filter Method

Information Gain (IG) is also known as Kullback-Leibler (KL) [Kullback, 51] divergence or relative entropy that means how much "information" a feature gives us

about the class. Features that perfectly partition should give maximal information, unrelated features should give no information. In order to calculate the IG value, entropy H(X) of the dataset is calculated using Eq. 6.

$$H(X) = -\sum_{l=1}^{n} p(l) \log_2 p(l) \quad \text{(Eq. 6)}$$

Where $p(l)$ is the probability of the observation to be labeled as, which in turns is mathematically defined as Eq. 7.

$$p(l) = \frac{|x_i \in X | y_i = y|}{|X|} \quad \text{(Eq. 7)}$$

The IG of the j^{th} feature of the data set X is calculated by the following Eq. 8.

$$IG(j) = Entropy(X) - Entropy(Y) \quad \text{(Eq. 8)}$$

The algorithm based on information gain has been developed to detect risk factors. Algorithm 1 describes the detailed statement.

Input: data set *S*

Output: Ranked feature list *L*

Procedure:

Entropy(S)
  //for a specified disease,
  $e \leftarrow 0$
  for each class label c_i **do**
    count the instance # $M(c_i)$ of each class label c_i
    calculate the proportion $p(c_i)$ of c_i, $p(c_i) \leftarrow M(c_i)/N$ (where N is # of the whole instances)
    $e \leftarrow e\text{-}p(c_i)*log_2(p(c_i))$ // according to Eq.4.1
  end for
  return *e*
IG-FS(S)
  for each attribute a_i, S_j contains the tuples in *S* with the attribute value(a_i) $\leftarrow v$, **do**
    IG(a_i) $\leftarrow$ Entropy(*S*)- $\sum_{j \in Value(A)} \frac{|S_j|}{|S|} Entropy(S_j)$ //according to Eq.4.3
    add a_i to *L*
  end for
  L $\leftarrow$ sort(*L*) // according to the gain value IG(a_i)
  return *L*

Algorithm 1: IG-FS feature selection algorithm

Function entropy is defined at the beginning to calculate the "information" of each class label. It is invoked in IG-FS() function to achieve information gain value for each attribute. Algorithm returns sorted feature list based on gain value.

(2) ReliefF Based Filter Method

The ReliefF algorithm is an upgraded version of Relief. It is not limited to two class problems, yet it is more robust and it can deal with incomplete and noisy data [Kononenko, 94]. The key idea of ReliefF is that it estimates the quality of attributes according to how well their values distinguish between instances that are near to each other. The algorithm has been defined as below.

Input: data set S

Output: the vector *W*[] of estimations of the qualities of attributes

Procedure:

1 **ReliefF-FS()**

2 set all weight *W*[*A*] ← **0** //*A* is attributes space

3 **for** *i* ← 1 **to** m **do** //m is the repeated times

4 randomly select an instance r_i

5 find *k* nearest hits H_j *//H_j means the k instances which have the same class label as ri*

6 **for** each class *c*≠class(r_i) **do**

7 from class *c* find k nearest misses Mj(*c*) //Mj(c) means the k instances which have the same class label as r_i

for *A* ← l to a **do**

8 $$W[A] <= -W[A] - \sum_{j=1}^{k} \frac{diff(A, r_i, H_j)}{(m.k)} + \sum_{c \neq class(ri)} \frac{p(c)}{1 - p(class(r_i))} \sum_{j=1}^{k} \frac{diff(A, r_i, Mj(C))}{(m.k)}$$

9 //according to Eq. 9 and Eq. 10

10 **return** *W*[*A*]

11 **end**

Algorithm 2: ReliefF-FS feature selection algorithm

At the beginning of the method, the algorithm randomly selects one instance r_i then searches its nearest *k* neighbors form the same class called nearest hits H_j (line 5). It also finds *k* nearest neighbors from each of the different classes called nearest misses *Mj*(*c*). Then it updates the quality estimation W[*A*], finally return W[*A*]. Function diff(*A, I1, I2*) calculates difference between the values of the attributes *A* for two instances I_1, I_2. For nominal attributes, it was originally defined as:

$$diff(A, I_1, I_2) = \begin{cases} 0: & value(A, I_1) = value(A, I_2) \\ 1: & otherwise \end{cases} \quad \text{(Eq. 9)}$$

and for numerical attributes as

$$diff(x_i, d, H_j) = \frac{|\, value(A, I_1) - value(A, I_2) \,|}{\max(A) - m \in (A)} \quad \text{(Eq. 10)}$$

The function diff is also used for calculating the distance between instances to find the nearest neighbors.

(3) Linear-SVM Based Wrapper Method

Linear-SVM is a supervised classifier, generally used in two or multi-class problems. The Linear SVM node uses a linear kernel. A linear kernel function is recommended when linear separation of the data is straightforward. SVMs are particularly suited to analysing data with very large numbers (for example, thousands) of predictor fields. 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 strategy ranks the features according to their influence on the decision hyperplane. Algorithm 3 describes the detailed statements.

Input: data set *S*

Output: ranked feature list *L*

Procedure:

1 **LSVM-FS(*S*)**
2 initial feature list $L \leftarrow \emptyset$, selected subset $S \leftarrow (1, \ldots, d)$;
3 Train a linear SVM with all the training data and variables in *S*
4 **for each** feature f_i in *S* **do**
5 compute the weight vector $w(f_i)$ // $w(f_i)$ could be correlation coefficient
6 compute the ranking scores $c_i \leftarrow (w(f_i))^2$
7 find the feature with the smallest ranking score e, $e \leftarrow \arg\min(c_i)$
8 update $L \leftarrow L(e, L)$
9 update $S \leftarrow S-\{e\}$
10 **return** *L*
11 **end**

Algorithm 3: LSVM-FS feature selection algorithm

It can be seen from the algorithms pseudocode that the SVM classification has been trained (line 3) before the feature selection. The kernel function for SVM is linear kernel function. The selected features are ranked based on the ranking scores such as correlation coefficient. The larger the value is, the more important the feature is.

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

3.4 Feature Integration for Predictor Generation

All features selected from different feature families are integrated based on the primary keys to generate features candidates for classification method. Full outer join (one data integration operation among the four widely used operations which are left join, right join, inner join and full outer join) has been used to merge all selected features together. In our work, full join has been used for feature integration. Features user_id, year, and disease_id are used as the primary key, the joint result is used for the next step.

3.5 Classification and Result Evaluation

Rule-based decision tree classification method CHAID has been applied on feature candidates. The reason for using this method is that it achieves the highest accuracy among other classification methods at this step. Besides, rule-based method provides an easy and effective way for disease interpretation and prediction.

The format of disease rules is same as the IF-THEN rule; for example, IF (edu = elementary, B1 < = 0.86 mg/day, married), THEN (hypertension = yes). The purpose here is the mining of all of the disease-related rules from the training dataset for a further data analysis including disease prediction.

Nevertheless, there is no method that fits for all kinds of the data sets that can achieve highest accuracy and efficiency. One advantage of the proposed method, at this step, the classification method can be replaced by other classification methods if some other datasets are used.

In order to compare the efficiency of the proposed method, several existing methods such as LR, K-NN, AdaBoost, CHAID are used to compare with the proposed method. The result will be given in the experimental chapter.

3.6 Apply RFD-DP Model in MapReduce Environment

The proposed model is suitable to be applied in a parallel environment due to its distributed design architecture. There is another advantage of this model is that it provides flexibility to be applied to both high dimensional data sets and normal datasets.

Figure 9 describes the procedure of the MapReduce framework.

It depicted in Figure 9 that MapReduce consists of mapping phase and reducing phase. During the mapping phase, HBase server t (t=1, 2,..., n) splits the data into different blocks. Map worker is assigned by master to process the data in this block. Its work includes feature family generation and feature selection. During the reducing phase, the output of worker node is shuffled and sorted regarding to year and stored in intermediate data block Dm. Reduce worker is assigned to generate feature candidates of each feature family, integrates these candidates regarding to id to generate the training data set, and executes CHAID classification algorithm to train the classifier.

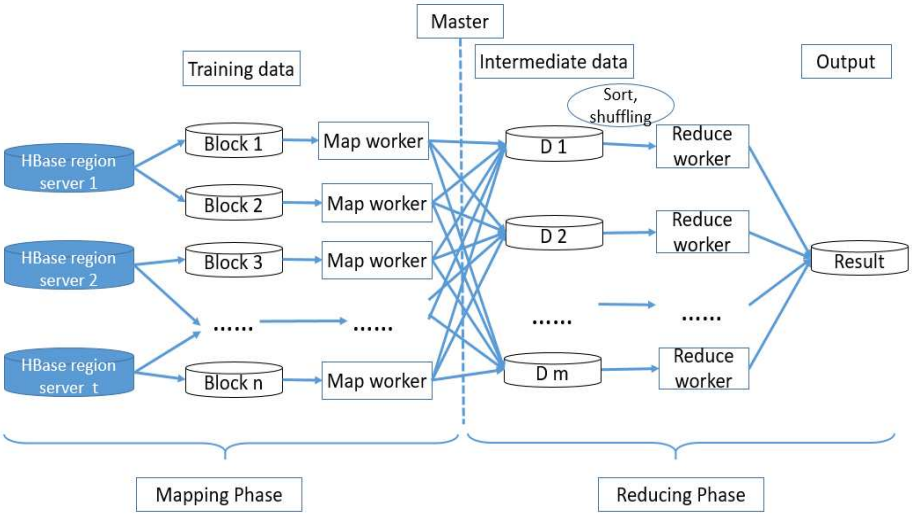

Figure 9: Framework of RFD-DP model running on MapReduce

The pseudo code of the RFD-DP model is given in Algorithm 4.

Input: data set S,

Output: Classifier C

Procedure:

Mapper()

For a specified disease: generate N feature families from S {{id,X_1}, {*id*,X_2}, ... {*id*,X_k}, ... {*id*,X_N}}, $k \in N$, $T_1, T_2, T_3 \leftarrow \emptyset$ // T_1, T_2, T_3 are temp feature sets

for each X_k, **do**

initiate feature set $F_1 \leftarrow \emptyset$

$F_1 \leftarrow$ IG-FS(X_k), $T_1 \leftarrow T_1 \cup F_1$

initiate feature set $F_2 \leftarrow \emptyset$

$F_2 \leftarrow$ ReliefF-FS(X_k), $T_2 \leftarrow T_2 \cup F_2$

initiate feature set $F_3 \leftarrow \emptyset$

$F_3 \leftarrow$ LSVM-FS(X_k), $T_3 \leftarrow T_3 \cup F_3$

end for

return {*id*, T_1, T_2, T_3 }

Reducer()

//generate feature candidate F_C using voting strategy

```
14  F_C ← vote (id, T_1,T_2,T_3)
15  F_L ← Integrate F_C based on the id // F_L is the integrated feature set
16  C_m ← CHAID(F_C) // CHAID() could be replaced by other classification algorithms
17  r_m ← evaluate(C_m) // C_m is rule-based classifier, r_m is the evaluation result of C_m
18  return C_m, r_m
19  vote()
20  initiate feature set F_s ← ∅
21  for each f_i ∈ F_j, where j ∈ {1,2,3} do
22    if ∃f_i ∈ F_k, count (f_i) ++,  k≠j, k ∈ {1,2,3}
23    if count (f_i) >2, return add f_i to F_s
24  end for
25  return F_s
```

Algorithm 4: RFD-DP model algorithm running on MapReduce

As it described from the algorithm that for a specified disease such as HTN, before the MapReduce procedure, the data set S has been divided into several subsets in terms of the feature families, such as (id, X_K). Each Map work node is a filter and wrapper-based feature selection work node. Algorithms IG-FS(),ReliefF-FS() and LSVM-FS() run in this node, the output is the features selected from each feature family.

After that, MapReduce master will start Reducing procedure to integrate selected features into one intermediate data block regarding year. Then on Reduce work node, the method vote() is used to generate feature candidates from the output of the previous step. Then rule-based classification algorithms will run on each Reduce worker node.

The result of each classification algorithm will be compared and the best one is chosen as the classification algorithm on current data sets. In our work, CHAID has been chosen as the rule-based classification algorithm of RFD-DP model.

4 Experiment and Result

4.1 Data Preparation

Data used in this paper come from three data sources, they are KNHANES data [KNHANES, 2017], orlraws10P (Database for faces, 2018) and Prostate [Singh, 02].

The Korea National Health and Nutrition Examination Survey (KNHANES) is a national surveillance system that has been set up since 1998. It collects information on socioeconomic status, health-related behaviors, quality of life, healthcare utilization, anthropometric measures, biochemical and clinical profiles for non-

communicable diseases and dietary intakes. This surveillance 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).

Table 2 summarizes the basic information of the four data sets used in this paper.

Dataset	Instances	Features	Classes	Size
KNHANES(6)	15,587	727	3	184 MB
orlraws10P	100	10,304	10	3.7 MB
Prostate	102	12,600	2	5.6 MB
KNHANES(R3-R6)	733,530	1,194	3	6.5 GB

Table 2: Basic information of selected data sets

One data set called KNHANES(6) used in this paper dates from 2013 to 2015. It consists of 22,948 records and 727 compatible attributes, published at the beginning of 2017 during the 6th publication phase. Another data set called KNHANES(R3-R6) dates from 2009 to 2015 and randomly sampled 10 times in order to generate a distinguishable relative bigger data set compared with other 3 datasets.

According to the data specification provided by KNHANES, attribute HE_HP is the indicator of hypertension. Observation with HE_HP=1 is the normal people without hypertension. Observation with HE_HP=2 is the people who have pre-hypertension. Observation with HE_HP=3 is the people who have hypertension. As it has been predefined, this paper treats people who have pre-hypertension as normal people.

Dataset orlraws10P is an image data set used for face recognition. It consists of 10 classes for different 10 different countenance such as facial expressions (open / closed eyes, smiling / not smiling) and facial details (glasses / no glasses). It is a set of face images taken between April 1992 and April 1994 at the Cambridge University Computer Laboratory. The database was used in the context of a face recognition project carried out in collaboration with the Speech, Vision and Robotics Group of the Cambridge University Engineering Department. The size of each image is 92x112 pixels, with 256 grey levels per pixel. The images are organized in 40 directories (one for each subject), which have names of the form sX, where X indicates the subject number (between 1 and 40). In each of these directories, there are ten different images of that subject, which have names of the form Y.pgm, where Y is the image number for that subject (from 1 to 10).

Prostate cancer is a disease of the prostate, a walnut-sized gland in the male reproductive system. There are two class labels, T means that the observations have prostate disease, N means observations don’t have this disease.

4.2 Experiment Design and Result Comparison

For the purpose of better understanding of the whole experimental procedure, the design of the experiment is given below. It consists of two phases with 3 steps for each phase,

Phase one：

Step 1: Do experiment using the proposed model on KNHANES(6), orlraws10P , Prostate and KNHANES(R3-R6) data sets.

Step 2: Do the same experiment using existing methods, such as Logistic Regression, AdaBoost, K-NN and Decision Tree algorithms based on the same data sets.

Step 3: Compare the performance of the proposed model with existing methods to see if proposed model is preferable.

Phase two：

Step 1: Separate the KNHANES(6) dataset into man and woman data sets.

Step 2: Detect HP risk factors for man and woman separately.

Step 3: Generate HTN predictors for both man and woman separately and evaluate the result

Table 3 to 6 show the comparison result in term of AUC, F-score and Accuracy on four data sets separately after phase one.

Evaluation method	**Proposed**	**LR**	**AdaBoost**	**K-NN**	**CHAID**
AUC	**0.849**	0.828	0.807	0.742	0.822
F-score	**0.845**	0.835	0.821	0.660	0.581
Accuracy	0.774	0.764	0.745	**0.803**	0.760

Table 3: Performance comparison on KNHANES(6) data set

From the Table 3 we can see that the proposed method achieves the highest score regarding AUC and F-score among all the other methods. However, K-NN achieves a better result than the proposed method regarding accuracy. Nevertheless, the problem for accuracy is that it is not suitable to be used as evaluation standards of classifier since the classification is uneven [Ling, 03]. Accuracy works best if false positives and false negatives have a similar cost. Predictive accuracy is a misleading performance measure for highly imbalanced data [Akosa, 17]. AUC holds more value based on this data set.

Evaluation method	**Proposed**	**LR**	**AdaBoost**	**K-NN**	**CHAID**
AUC	**1**	0.990	NA	NA	1
F-score	**1**	0.952	NA	NA	1
Accuracy	**1**	0.909	NA	NA	1

Table 4: Performance comparison on orlraws10P data set

From the Table 4 we can see that the proposed model achieves the best performance in terms of AUC, F-score and accuracy on orlraws10P data set.

Evaluation method	Proposed	LR	AdaBoost	K-NN	CHAID
AUC	**0.990**	0.987	0.981	NA	0.943
F-score	**0.970**	0.914	0.960	NA	0.885
Accuracy	**0.970**	0.911	0.960	NA	.0872

Table 5: Performance comparison on Prostate data set

From the Table 5 we can see that the proposed model achieves the best performance in terms of AUC, F-score and accuracy on Prostate data set.

Evaluation method	Proposed	LR	AdaBoost	K-NN	CHAID
AUC	**0.806**	NA	NA	NA	NA
F-score	**0.736**	NA	NA	NA	NA
Accuracy	**0.735**	NA	NA	NA	NA

Table 6: Performance comparison on KNHANES(R3-R6) data set

From the Table 6 we can see that only the proposed model achieves an experimental result while the other algorithms run failed on this data set.

Hence, from all experimental result, we can draw the conclusion that the proposed model outperforms the existing methods in terms of AUC, F-score and accuracy based on KNHANES(6), orlraws10P, Prostate and KNHANES(R3-R6) data sets. Thus it is going to be applied to KNHANES(6) for risk factor detection and disease predictor generation in Phase two.

The number of observations selected from the raw data set is 15,587. This research cohort is the observations whose HTN information is not missing. There are 2,264 men who have HTN disease among 6,587 records, and 2,561 women who have HTN among 9,000 records.

Table 7 describes the detected risk factors related to hypertension on man data set. 22 features have been selected from 727 features sets, and they are the potential risk factors for men who have hypertension. The significance level describes that for all features, it is less than 0.05 except feature dr_month. Thus, the feature of dr_month should be removed from the result. Risk factor age plays the most important role for HTN followed by BO1. Further analysis can be applied based on the current result.

Related features	B	STD. D	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
							Lower	Upper
age	0.052	0.003	273.758	1	0.000	1.054	1.047	1.060
ainc	0.000	0.000	4.370	1	0.037	1.000	1.000	1.000
mt_nontrt	-0.238	0.108	4.868	1	0.027	0.788	0.638	0.974
BH9_13	-0.002	0.001	6.378	1	0.012	0.998	0.997	1.000
BH1_3	0.020	0.009	4.889	1	0.027	1.020	1.002	1.038
LQ2_ab	0.027	0.012	4.605	1	0.032	1.027	1.002	1.052
BO1	0.424	0.034	151.317	1	0.000	1.529	1.429	1.636
BO2_1	-0.104	0.024	18.409	1	0.000	0.901	0.859	0.945
BD1_11	0.120	0.037	10.636	1	0.001	1.128	1.049	1.213
BD2_31	0.072	0.016	21.283	1	0.000	1.075	1.042	1.109
BD7_4	0.207	0.040	27.385	1	0.000	1.230	1.138	1.330
BD7_6	-0.201	0.054	13.817	1	0.000	0.818	0.736	0.910
BD7_5	-0.379	0.055	47.681	1	0.000	0.685	0.615	0.762
dr_month	0.062	0.141	0.193	1	**0.660**	1.064	0.807	1.402
BA2_2_6	0.000	0.000	6.276	1	0.012	1.000	1.000	1.000
BP7	-0.072	0.018	15.907	1	0.000	0.931	0.899	0.964
BS9_2	0.050	0.020	6.518	1	0.011	1.052	1.012	1.093
O_DMFTP	-0.010	0.005	4.244	1	0.039	0.990	0.980	1.000
T_Q_HR	-0.067	0.016	18.319	1	0.000	0.935	0.907	0.964
T_NQ_PH_T	0.001	0.000	9.055	1	0.003	1.001	1.000	1.001
N_DIET_WHY	-0.033	0.009	13.611	1	0.000	0.967	0.950	0.985
LF_SECUR	0.046	0.023	4.053	1	0.044	1.047	1.001	1.095

Table 7: Risk factors related to hypertension on man data set, HE_HP=3

Table 8 describes the detected risk factors related to hypertension on woman data set.

Related features	B	STD.D	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
							Lower	Upper
age	0.081	0.004	468.719	1	0.000	1.084	1.076	1.092
edu	-0.192	0.035	30.711	1	0.000	0.826	0.772	0.884
occp	0.040	0.016	6.211	1	0.013	1.040	1.008	1.073
tins	-0.005	0.002	4.351	1	0.037	0.995	0.990	1.000
mt_nontrt	-0.201	0.081	6.130	1	0.013	0.818	0.697	0.959
BH9_13	-0.003	0.001	10.122	1	0.001	0.997	0.996	0.999
BO1	0.367	0.030	147.649	1	0.000	1.444	1.361	1.532
BD1_11	0.076	0.025	8.938	1	0.003	1.079	1.026	1.134
BD7_4	0.273	0.081	11.427	1	0.001	1.314	1.122	1.539
BD7_5	-0.367	0.091	16.408	1	0.000	0.693	0.580	0.827
BA2_2_5	0.056	0.013	18.583	1	0.000	1.058	1.031	1.085
BA2_22	-0.213	0.032	44.618	1	0.000	0.808	0.760	0.860
BP7	-0.020	0.010	3.852	1	0.050	0.980	0.961	1.000
BE3_93	0.002	0.001	9.801	1	0.002	1.002	1.001	1.004
O_DMFTP	-0.015	0.005	9.170	1	0.002	0.985	0.975	0.995
OR1	0.059	0.025	5.420	1	0.020	1.061	1.009	1.115
T_Q_HR	-0.134	0.020	44.554	1	0.000	0.874	0.841	0.910
T_NQ_PH_T	0.001	0.000	6.670	1	0.010	1.001	1.000	1.001
T_NQ_OCP_P	0.033	0.014	5.506	1	0.019	1.034	1.005	1.063
N_B2	-0.116	0.043	7.432	1	0.006	0.890	0.819	0.968

Table 8: Risk factors related to hypertension on woman data set, HE_HP=3

Based on the analysis result of the previous step, rule-based classification algorithm CHAID has been used to generate the HTN predicator. The advantage of using a decision tree classifier as the predictor is fairly obvious. 1. It is easy to be interpreted at a glance due to its tree structure, which is very understandable for the user who has no medical background. 2. It is suitable for handling both numerical and nominal features. 3. It is capable of handling missing values in attributes. 4. High-performing regarding searching down a built tree even for massive data sets. Table 9 describes the evaluation result of the predictor for both data sets.

Data set	Sensitivity	Specificity	Precision	Accuracy	F-score		AUC
Man	0.745	0.813	0.609	0.794	0.670		0.859
Woman	0.695	0.841	0.571	0.806	0.627		0.872

Table 9: Evaluation result of the predictor for both data sets

Table 9 describes the performance of the proposed model running on both data sets in term of accuracy, F-score, AUC and etc. Moreover, AUC plays more important role for result comparison.

From the Figure 10, we can see that HTN predictor for woman data set achieves higher performance than man. X axis represents the FP rate, and Y axis represents TP rate,

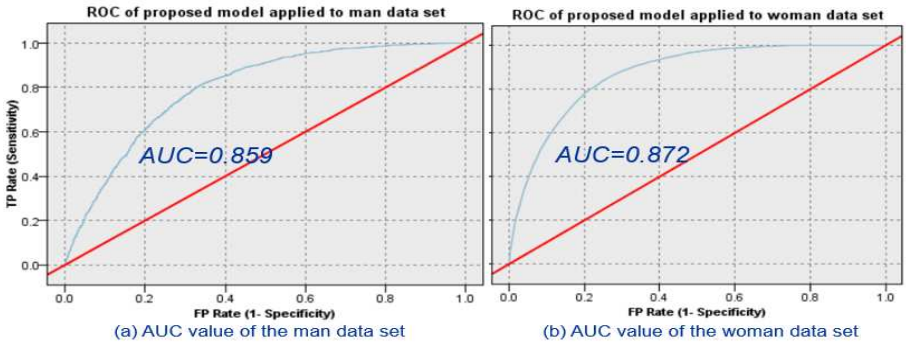

Figure 10: (a) ROC diagram of proposed model applied to man data set and Figure 10: (b) to woman data set

5 Conclusion

In this paper, we have developed an effective feature selection and classification model for high dimensional data set. According to the experimental result, the proposed model achieves the best performance among other well-known classification methods on all datasets. Moreover, the key idea of our model is using a divide-and-conquer strategy to handle high dimensional dataset, it perfectly matches the mechanism of MapReduce. Thus it can be used for high dimensional big data set.

In the future, more advanced models are planning to be compared with our model. For the sake of applying our model to the real practice, continuous improvement of this model is needed along with enhancing its algorithms.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2017R1A2B4010826) and 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 Karamay Central hospital (2017HZ006A) supported by Karamay Office of Science and Technology. Authors Yaning Li, Seon Phil Jeone and Keun Ho Ryu are corresponding athors.

References

[Akosa, 17] Akosa, J. Predictive accuracy: a misleading performance measure for highly imbalanced data, In Proceedings of the SAS Global Forum, 2017.

[Amendola, 14] Amendola, S., Lodato, R., Manzari, S., Occhiuzzi, C., and Marrocco, G. RFID technology for IoT-based personal healthcare in smart spaces, IEEE Internet of Things Journal, 1(2), 144-152, 2014.

[Baek, 16] Baek, M. R., and Jung, K. T. Prediction of changes in health expenditure of chronic diseases between age group of middle and old aged population by using future elderly model, Health Policy and Management, 26(3), 185-194, 2016.

[Chakraborty, 15] Chakraborty, R., and Pal, N. R. Feature selection using a neural framework with controlled redundancy, IEEE Transactions on Neural Networks and Learning Systems, 26(1), 35-50, 2015.

[Cheng, 01] Cheng J, and Greiner R. Learning bayesian belief network classifiers: Algorithms and system, Conference of the Canadian Society for Computational Studies of Intelligence, Springer, Berlin, Heidelberg, 141-151, 2001.

[Chobanian, 03] Chobanian A V, Bakris G L, Black H R, et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure[J]. hypertension, 42(6): 1206-1252, 2003.

[Ezzeldin, 17] Ezzeldin M., Musa I., Li D., Ryu K. Big Data in Healthcare: A bibliometric review via CiteSpace (2012 - 2016), In Proceedings of the 10th Frontiers of Information Technology, Applications and Tools (FITAT 2017), 2017.

[Freund, 09] Freund, Y., and Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, 55(1), 119-139, 1997.

[Ha, 10] Ha, S. H., and Joo, S. H. A hybrid data mining method for the medical classification of chest pain, International Journal of Computer and Information Engineering, 4(1), 33-38, 2010.

[Hall, 98] Hall, M. A. (1998). Correlation-based feature subset selection for machine learning. Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato.

[Han, 11] Han, J., Pei, J., and Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[Hearst, 98] Hearst, M. A. SVM trends and controversies Intelligent Systems and Their Applications, 1998.

[Hsu, 12] Hsu, T. W., Liu, J. S., Hung, S. C., Kuo, K. L., Chang, Y. K., Chen, Y. C., and Tarng, D. C. Renoprotective effect of renin-angiotensin-aldosterone system blockade in patients with predialysis advanced chronic kidney disease, hypertension, and anemia, JAMA Internal Medicine, 174(3), 347-354, 2014.

[Huang, 07] Huang, M. J., Chen, M. Y., and Lee, S. C. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis, Expert Systems with Applications, 32(3), 856-867, 2007.

[Kass, 80] Kass, G. V. An exploratory technique for investigating large quantities of categorical data, Applied Statistics, 119-127, 1980.

[Kaur, 14] Kaur, P. D., and Chana, I. Cloud based intelligent system for delivering health care as a service, Computer Methods and Programs in Biomedicine, 113(1), 346-359, 2014.

[Kim, 15] Kim J. E, Hwang J. W, Park J. S, Park D. S. and Park S. W. Analysis of medical condition of domestic chronic diseases, KHIDI 203, 2015.

[Kim, 16] Kim, H. S., Ishag, M. I. M., Piao, M., Kwon, T., and Ryu, K. H. A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries, Symmetry, 8(6), 47, 2016.

[KNHANES,17] KNHANES (2017). http://knhanes.cdc.go.kr, latest access: Nov, 2017.

[Kononenko, 94] Kononenko, I. Estimating attributes: analysis and extensions of relief, Machine Learning: ECML-94, Springer-Verlag, 171–182, 1994.

[Kullback, 51] Kullback, S., and Leibler, R. A. On information and sufficiency. The annals of mathematical statistics, 22(1), 79-86, 1951.

[Li, 18] Li, D., Park, H. W., Batbaatar, E., Munkhdalai, L., Musa, I., Li, M., and Ryu, K. H. Application of a Mobile Chronic Disease Health-Care System for Hypertension Based on Big Data Platforms. Journal of Sensors, 2018.

[Ling, 03] Ling, C. X., Huang, J., and Zhang, H. AUC: a better measure than accuracy in comparing learning algorithms, Conference of the canadian society for computational studies of intelligenc, Springer, Berlin, Heidelberg, 329-341, 2003.

[Mahmoud, 17] Mahmoud, S, M, Abdulabbas T. E. Multiple MapReduce functions for health care monitoring in a smart environment, E-Health and Bioengineering Conference (EHB), IEEE, 2017, 507-510, 2017.

[Piao, 15] Piao, Y., Piao, M., Jin, C. H., Shon, H. S., Chung, J. M., Hwang, B., and Ryu, K. H. A new ensemble method with feature space partitioning for high-dimensional data classification. Mathematical Problems in Engineering, 2015.

[Rahman, 13] Rahman, A., and Verma, B. Ensemble classifier generation using non-uniform layered clustering and Genetic Algorithm. Knowledge-Based Systems, 43, 30-42, 2013.

[Rokach, 10] Rokach, L. Ensemble-based classifiers, Artificial Intelligence Review, 33(1-2), 1-39, 2010.

[Saurav, 16] Saurav, K. Introduction to Feature Selection methods with an example, https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/, last accessed: May 2018.

[Schatz, 09] Schatz, M. C. (2009). CloudBurst: highly sensitive read mapping with MapReduce, Bioinformatics, 25(11), 1363-1369.

[Singh, 02] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C. and Lander, E. S. (2002). Gene expression correlates of clinical prostate cancer behavior, Cancer Cell, 1(2), 203-209.

[Singh, 14] Singh, B., Kushwaha, N., & Vyas, O. P. A feature subset selection technique for high dimensional data using symmetric uncertainty, Journal of Data Analysis and Information Processing, 2(04), 95, 2014.

[Sharma, 14] Sharma, A. K., Khanna, D., and Balakumar, P. Low-dose dipyridamole treatment partially prevents diabetes mellitus-induced vascular endothelial and renal abnormalities in rats. International journal of cardiology, 172(2), 530-532, 2014.

[Taylor, 10] Taylor, R. C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics, BMC bioinformatics, BioMed Central, 11(12), 2010.

[Thomas, 91] Thomas M. C. and Joy A. T. Elements of information theory, John Wiley &Sons, 1991.

[Yu, 03] Yu, L., and Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution, In Proceedings of the 20th international conference on machine learning (ICML-03), 856-863, 2003.