International Journal of Bio-Science and Bio-Technology
Vol.7, No.1 (2015), pp.61-70
http://dx.doi.org/10.14257/ijbsbt.2015.7.1.07

Performance Evaluations of Diagnostic Prediction with Neural Networks with Data Filters in Different Types

Hoon Jin[1] and Seungcheon Kim[2,**]

[1]*Dept. of Computer Engineering, Sungkyunkwan University, South Korea*
[2]*Dept. of information Communication Engineering, Hansung University, South Korea*
bioagent@skku.edu[1], kimsc@hansung.ac.kr[2]

Abstract

The advent of the information age and the rapid development of IT skills have led to the construction of massive databases, thus the current research focus is shifting to the efficient utilization of these vast volumes of stored information. Among the data mining algorithms that have been applied to this problem, neural networks can be used with various types, qualities, distributions, or volumes of data and they have high predictive power. Thus, neural networks are known to be the most useful and extensible algorithms, whereas logistic analysis has many constraints. In addition, neural networks obtain better results when the assumptions of linear discriminant analysis cannot be satisfied. The present study evaluated a multilayer perceptron (MLP) and radial-basis function network (RBFN), and their performance levels were compared with logistic regression based on cross-validation using the same data. The experiments showed that MLP delivered better performance than other methods in medical diagnostic applications where numerical data are used. MLP also performed better with the heart disease dataset using finely specified data types compared with the diabetes dataset using simple data types.

Keywords: *Mixed data type; Multilayer perceptron; Neural networks; Different data type; Weka filter*

1. Introduction

Data mining can be applied in diverse environments such as determining the probability of disease onset in hospitals and personal credit scoring in marketing, as well as various other industrial fields. Furthermore, data mining huge volumes of information using statistical or mathematical analyses and pattern recognition techniques can facilitate the discovery of various valuable insights, including novel relationships or biases. The present study compared the performance of two neural network algorithms and a logistic regression model, which are used frequently for data mining in the medical domain, to determine the algorithm that performed better when several evaluating measures related to medical data. This experiment used various clinical datasets from Internet-based public data repositories, but it focused specifically on heart diseases, which are major causes of sudden death in most developed nations. Thus, a heart disease dataset was used in most of the experiments, but a diabetes dataset was also used for comparison with the heart diseases data. The present study evaluates the experimental performance of both algorithms in terms of disease prediction using the same samples.

** Corresponding author is Prof. Seungcheon Kim

ISSN: 2233-7849 IJBSBT

2. Neural Networks and Statistics

As for neural networks, it is as a concept of imitating the human brain's nerve cells, one of modeling techniques to find the embedded patterns taken from the past data through repetitive learning processes. In contrast, statistics is based on mathematics and deals with all aspects of the collection, organization, analysis, interpretation, and presentation of data, including the planning of data collection in terms of the design of surveys and experiments for the study of the collective phenomena [1].

2.1. MLP and RBFN

MLP (Multilayer Perceptron) being used most frequently with the statistical techniques is a traditional neural network [2]. It is composed of input, middle and output layer. The middle layer having one or more intermediate layers is called as hidden layer. Neurons of the input layer are connected to those of the hidden layer and they are connected to the output layer orderly. Also MLP is a feed-forward network not allowing connections within individual layers but supports back-propagation operation between the hidden layer and the output layer [3]. Generally it is known that MLP increases the more numbers of hidden layers, more the characteristics of decision boundaries which perceptron forms are getting sophisticated. When a sigmoid function is applied to nonlinear activation function decision boundaries are appeared in form of gentle curves not straight lines, then since the analysis process of behavior may be a little complicated but the differential is possible, back propagation learning algorithms which can learn the hidden layer start to perform [3]. MLP in learning process using a complete data set composed with known input and output values learns input patterns for the actual network output. A function getting the difference values between actual network outputs and desired network outputs is defined as the objective function and then it can be computed for the weight to minimize the differences [4]. In general, the objective function uses a mean square error or a cross-entropy error [5]. However there are deficiencies in the traditional MLP, since it is slow in learning speed and difficult to understand that the end of trained knowledge represents [6].

RBFN (Radial-Basis Function Network) was proposed as a supplementary method for the deficiency of MLP in learning speed. So it has some advantages; neurons are able to achieve a probability distribution function; learning speed is faster than MLP since no repetition is required to reach the optimum model parameters [7]; some degree of knowledge expressed by the network is more understandable than MLP. In Table 1, the number of hidden layers of RBFN is only one compared to more than one hidden layers MLP can have. Furthermore MLP follows the sigmoid function as an activation function of neurons, but RBFN does a normal Gaussian function.

Table 1. Comparison of MLP and RBFN [8]

	MLP	RBFN
The number of hidden layer	1 or more	1
Activation function	Sigmoid	Gaussian
Activation function parameters	Vector dot product	Euclidean distance
Nonlinear mapping	Global	Regional

In RBFN, all input neurons are connected to neurons of the hidden neurons respectively and those produce the outputs for the input patterns by Gaussian function. Then the output unit puts the cumulative sum which multiplies each output value of the hidden neurons with

the weight of it as the final output of the network. For an example, the final output of RBFN with one output unit can be calculated as the following.

$$y = \sum h_i z_i + h_0$$

, where y represents the result for the output layer, and h_i shows an input vector for the i^{th} hidden layer. Of course z_i means a weight for the i^{th} hidden layer and h_0 is an initial value for the 1st hidden layer.

2.2. Logistic Regression

Logistic regression model, as a special case of generalized linear models which the response variable consists of binary variables with characteristic of categorical types, is used to predict the response variable from several independent variables [9]. In addition, logistic regression model cannot be used to explain the relation and the type of interaction by the model structure but also it can be used to evaluate the impact of the explanatory variables to the response value through parameter estimation. Based on prediction probability such as linear discriminant analysis, it can be used as an analytic method or a classification technique. But compared to the linear discriminant analysis assuming that the explanatory variable follows the same covariance matrix and multivariate normal distribution, because there are few constraints on logistic regression model, it enables to lead to get better results when the assumption of linear discriminant analysis cannot be satisfied [10]. It is assumed that the dependent variable in logistic regression model measured in continuous scale is quantitative. However, there is a case that the dependent variable shows qualitative values with having only two types of variables such as 0 and 1. In this time, the response variable being measured as having binomial types is computed by the following [9].

$$y_i = \pi(x_i) + \epsilon_i$$

$$\pi(x_i) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}, \qquad i = 1,2,\dots,n$$

2.3. Filtering Modthods

Filter in use of data mining algorithms can be used to modify datasets in a systematic fashion as the data preprocessing tools. In Weka, which is a general used integrated tool for data mining, filters are organized into a hierarchical structure. Those in the unsupervised category don't require a class attribute to be set whereas those in the supervised category do. The filters are further divided into ones that operate primarily on the attribute category and ones that operate primarily on the instance category. It is concerned with classes that transform datasets by removing or adding attributes, resampling the dataset, removing examples and so on. Therefore using filters offers useful support for data preprocessing, which is an important step in machine learning [11]. As for BestFirst filter, it is often used to refer to a search with a heuristic that attempts to predict how close the end of a path is to a solution, so that paths that are judged to be closer to a solution are extended first. This specific type of search is called greedy best-first search [12]. The chi-square test is used to examine if there is no association between two attributes, i.e., whether the two variables are independent [13]. It is a nonparametric method often used where the data consist in frequencies or counts, for example the number of students in a class wearing contact lenses as

distinct from quantitative data obtained from measurement of continuous variables such as temperature, height, and so on [14]. Resampling method is commonly used for dealing with the class-imbalance problem. Their advantage over other methods is that they are external and thus, easily transportable. Although such approaches can be very simple to implement, tuning them most effectively is not an easy task [15].

3. Experiments

For the evaluation of the performance of MLP, RBFN and Logistic analysis, we used two medical data sets from open internet repositories [16]. First data set is about heart disease consisting of 13 independent variables helpful to predict the presence of it such as age, sex, types of chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, etc. and one dependent variable showing whether the disease exists or not (see Table 2) [8]. Second data set is about diabetes consisting of 8 independent variables and 1 dependent one. They are represented in the following Table 3 [17].

The total number of former samples is 270 and 13 independent variables are com-posed of 5 continuous variables and 8 categorical variables. That of the latter is 768 and 8 independent variables are all continuous real values. Between the two data sets, we are mainly focusing on finding the difference between the heart disease of which attributes are mixed with categorical and continuous variables and the diabetes of which all attributes are only filled with continuous variables on filter methods. In the attributes of heart disease data, types of chest pain, blood pressure and number of major blood vessels are known as important factors to predict the presence of it ac-cording to the expert's advices. In the case of diabetes disease, because all values are numeric, there need no specific considerations about the data types. Of course the last attribute shows a class value, which means presence or absence of it, like the case of heart disease.

3.1. Previous Study Result

We referred to [1] for the comparative study. In that, after using four algorithms such as NB, DT, MLP and k-NN, they evaluated the results on the basis of 4 criteria, which are accuracy, precision, sensitivity, specificity. [1] used ranking algorithm for feature selection available in WEKA and ordered them by priority on the class. The averages of accuracy, precision, sensitivity and specificity of them are 96.552, 93.698, 0.921 and 0.986 with 12 features, respectively. But when we tried to output prediction results only with default parameters and no filters, the results were very lower than the previous study. It might be presumed that the previous study used one more feature, Globulin, and also used feature selection with ranking algorithm. Our results are showed in the Table 1.

Table 2. The Experimental Heart-statlog Data [8]

Attribute Name	Types	Average	Standard deviation	Comments
Age	C	54.433	9.109	
Sex	D	0, 1		
Chest pain type	D	1, 2, 3, 4		4 values
Resting blood pressure	C	131.344	17.862	
Serum cholesterol	C	249.659	51.686	mg/dl
Blood sugar>120mg/dl	D	0, 1		
Resting electrocardiographic	D	0, 1, 2		values 0,1,2

Maximum heart rate	C	149.678	23.166	
Exercise induced angina	D	0, 1		
Oldpeak	C	1.050	1.145	induced by exercise relative to rest
Slope	D	1, 2, 3		exercise ST segment
Number of major vessels	D	0, 1, 2, 3		colored by fluoroscopy
Thal	D	3, 6, 7		3 = normal; 6 = fixed defect; 7 = reversible defect
Class	D	absence, presence		

(C: continuous data, D: discrete data)

Table 3 represents a brief statistical analysis of the diabetes disease. It is noteworthy that the standard deviation value is bigger than the mean value in the attribute of 2-hour serum insulin.

Table 3. The Brief Statistical Analysis of Diabetes Disease [17]

Attribute Name	Types	Mean	Standard Deviation
Number of times pregnant	C	3.8	3.4
Plasma glucose concentration a 2 hours in an oral glucose tolerance test	C	120.9	32.0
Diastolic blood pressure (mm Hg)	C	69.1	19.4
Triceps skin fold thickness (mm)	C	20.5	16.0
2-Hour serum insulin (mu U/ml)	C	79.8	115.2
Body mass index (weight in kg/(height in m)^2)	C	32.0	7.9
Diabetes pedigree function	C	0.5	0.3
Age (years)	C	33.2	11.8

4. Experiments

Weka package with 3.6.8 version was used with the supervised filters and default options in preprocessing step of both data sets for our test because those are collected considering of supervised learning. Among the various types of filters supported by Weka, BestFirst and ChiSquared filter which belong to the category of attribute selection were used and the Resample filter belonging to the category of instance was used for experiments. Experiments were performed with 10 folds cross-validation to take the averaged value repeated 10 times in all cases and the results of the tests for MLP, RBFN and Logistic analysis are synthetically compared in time to build model, pre-diction accuracy, kappa statistics, RMSE.

5. Results

In Figures 2-5, on the whole, the performances of MLP and Logistic analysis are higher than that of RBFN in case of having the Resample filter in both datasets. It probably means that MLP and Logistic analysis is more adoptable for our data set due to the characteristics of it. The percentage of correctly classified instances is often called an accuracy or sample accuracy. But it has some disadvantages as a performance estimate, therefore we need another measures for representing predicting accuracy effectively.

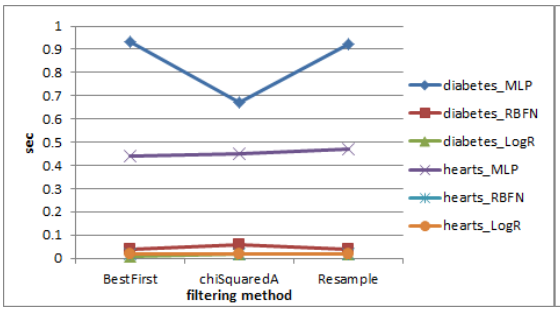

Figure 1. Time to Build Model

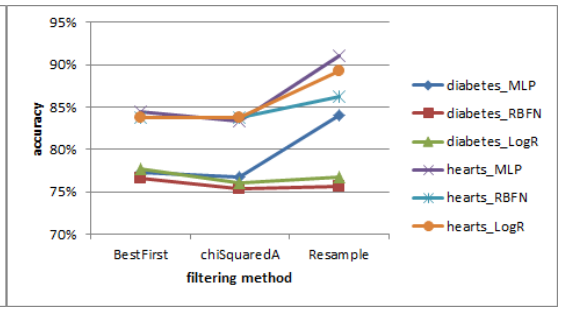

Figure 2. Prediction

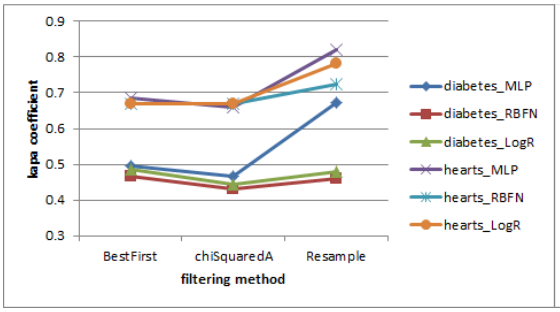

Figure 3. Kappa Statistic

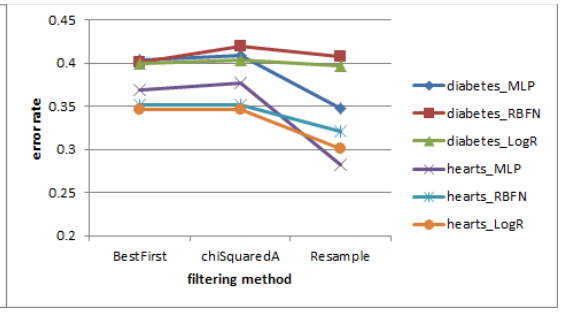

Figure 4. RMSE

Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions [18]. Its objective is to find a set of attribute values that can maximally reduce impurity in Weka [19]. Generally Chi-squared validation is the most common-ly used method in the cross analysis process and it is used to analyze the association between the categorical variables. Class imbalance problems have lately become an important area of study in machine learning and are often solved using intelligent resampling methods to balance the class distribution [20]. It is interesting that MLP tests using the heart disease put better results than others (Figure 3). It is assumed that the accuracy from experiments using the data with finely set-up data types is entirely higher than that with simple data types. Such a tendency is found in Figure 4 showing kappa statistic too. Kappa gives a chance-corrected measure of agreement between the classifications and the true classes [21]. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. So kappa coefficient value shows better prediction accuracy if it is higher. And if there is a value greater than 0, it means that the classifier is doing better than chance. On the contrary side of the above the lower error value is, the bet-ter prediction result RMSE shows like Figure 5. The error rates are used for numeric prediction rather than classification and RMSE is as the abbreviation of root mean squared error, the most intuitive and meaningful accuracy evaluation method to the approximate model. RMSE has an advantage in that if we have a large number of experimental samples, we can evaluate the accuracy of the model more accurately. On the other hand, we tried to find the difference of the performances between two data sets. As mentioned the above, it was intended to identify the difference between the concrete results from the heart disease data with finely set-up data types and the diabetes data with simple set data types.

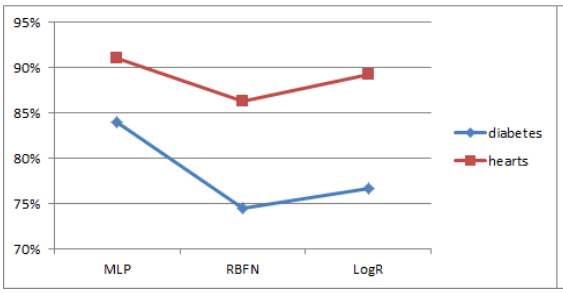

Figure 5. Accuracy per Diseases

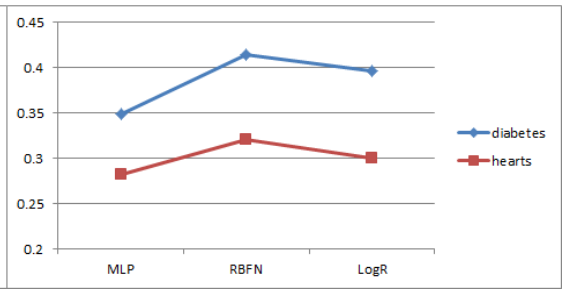

Figure 6. RMSE per Diseases

Figure 5-7 show accuracy, RMSE, building time measured with the Resample filter using two diseases data sets (because the tests with the Resample filter put the better results than those using other filters). Regardless of difference of sample sizes, the heart data shows the better result than the diabetes in accuracy and RMSE. There is no meaningful difference between RBFN and Logistic analysis in Figure 7 for building time per each dataset except MLP. It can be derived that RBFN and Logistic analysis are so fast that we can disregard the size of samples. The result of Logistic analysis over all presents middle leveled performances in our defined measures.

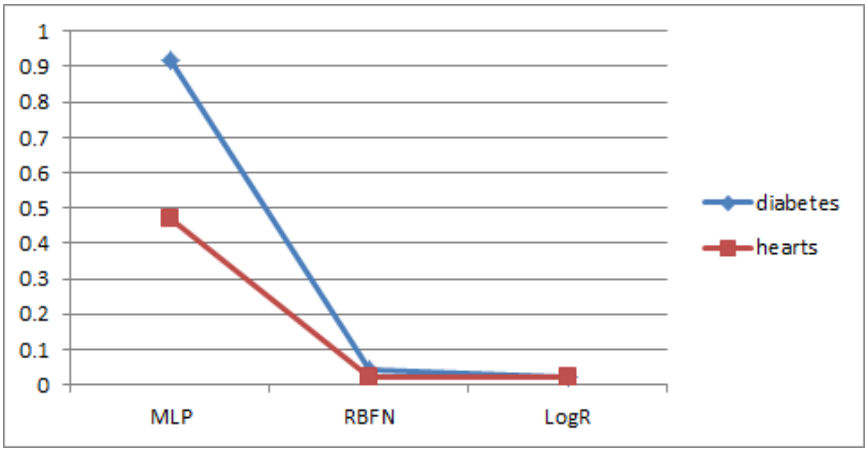

Figure 7. Building Time per Diseases

By controlling the number of hidden layers in preliminary experimental results, we could find the better result with two layers which is composed of 3 and 1 individually, as using with Resample filter.

As a result shown in Figure 8, the overall classification accuracy and the kappa statistic, showed nearly similar leveled results in both of MLP and Logistic analysis. In this Figure, RBFN shows the worst classification accuracy and the highest error rate on the whole. It is probably because RBFN is not efficient than MLP in the aspect of convergence and difficult to make the global optimal model. Additionally because neural network (or neural networks based algorithms) generally needs large number of hidden neurons, the performance of heart disease with 25 input neurons is better than that of diabetes disease with 8. (Compare between Figure 8 and 9).

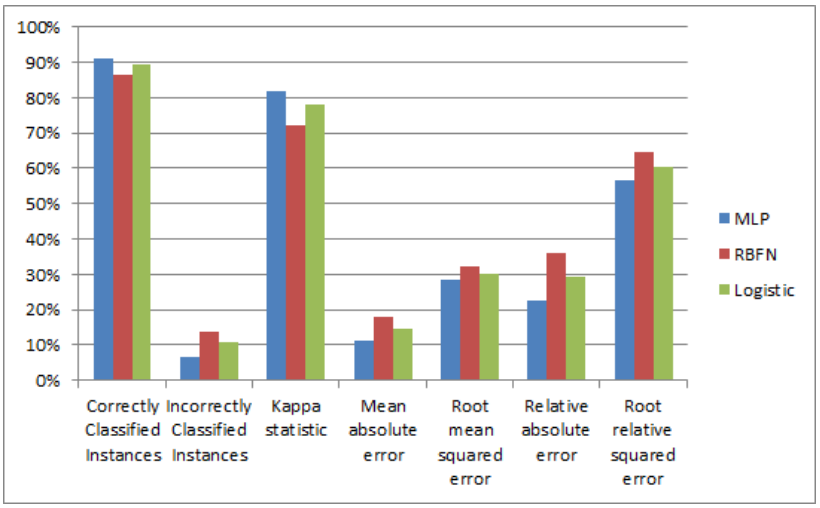

Figure 8. The Performance Results on Each Algorithm using the Heart Disease Data

Figure 9. The Performance Results on Each Algorithm using the Diabetes Data

6. Conclusion

In this paper, we carried out several experiments against three data mining algorithms that were MLP and RBFN, and Logistic analysis using two kinds of data to measure the performance and to understand which was better or adoptable for numerical data analysis in medical domain. For the purpose of the experiment, we mainly used the heart disease data from UCI repository and as a complementary data set used the diabetes data from the freely distributed Weka package [22]. Lastly MLP is, regardless of long time of building model, had the best performance in prediction accuracy, kappa statistics and RMSE than RBFN and Logistic. And also those of the heart disease data with finely set-up data types was better than those of the diabetes with simple set data types. Logistic regression model showed medium performances in all the evaluating measures.

This study includes several defects in that it was not suggested the reason why RBFN had the worst performance than others and why Logistic analysis had good performance to our

data through the entire experiments. But our study, by controlling experimental conditions such as using two different data sets, several filters for pre-processing, various evaluating measures, provides the experimental result that MLP has the better performance than others in medical environment in which numerical data are used. In Future, we have a plan to study to solve the unsolved problems in present study and to prove the effectiveness of the combined model using more than two neural network classifiers for the achievement of the higher leveled recognition rate, such as Boosting and Bagging [23].

References

[1] Y. Dodge, D. Cox, D. Commenges, P. J. Solomon and S. Wilson, "The Oxford dictionary of statistical terms", Oxford University Press, (**2003**).
[2] T. Xie, H. Yu, and B. Wilamowski, "Comparison between traditional neural networks and radial basis function networks", IEEE International Symposum on Industerial Electronics, (**2011**), pp. 1194-1199.
[3] S. Seung, "Multilayer perceptrons and back-propagation learning", Lecture4. 1-6, (2002), Source from: http://hebb.mit.edu/courses/9.641/2002/lectures/lecture04.pdf.
[4] L. Noriega, "Multilayer perceptron tutorial", School of Computing, Staffordshire University, (**2005**).
[5] J. W. Park, R. G. Harley and G. K. Venayagamoorthy, "Comparison of MLP and RBF neural networks using deviation signals for on-line identification of a synchronous generator", In Power Engineering Society Winter Meeting, IEEE, vol. 1, (**2002**), pp. 274-279.
[6] L. S. Thota, and S. B. Changalasetty, "OPTIMUM LEARNING RATE FOR CLASSIFICATION PROBLEM WITH MLP IN DATA MINING", (**1963**).
[7] A. W. Jayawardena, D. A. K. Fernando, and M. C. Zhou, "Comparison of multilayer perceptron and radial basis function networks as tools for flood forecasting", IAHS Publications-Series of Proceedings and Reports-Intern Assoc Hydrological Sciences, vol. 239, (**1997**), pp. 173-182.
[8] Y. G. Jung and H. Jin, "Experimental Comparisons of Neural Networks and Logistic Regression Models for Heart Disease Prediction", International Informational Institute, Information-An International Interdisciplinary Journal, vol. 16, no. 2(B), (**2013**), pp. 1295-1300.
[9] D. W. Hosmer Jr. and S. Lemeshow, "Applied logistic regression", John Wiley & Sons, (**2004**).
[10] H. W. lan and F. Eibe, "Data Mining", Addison Wesley, (**2005**), pp. 315-333.
[11] http://weka.wikispaces.com/Primer. Accessed on 2014.8.13.
[12] H. Wang, T. M. Khoshgoftaar and K. Gao, "A comparative study of filter-based feature ranking techniques", In Information Reuse and Integration (IRI), 2010 IEEE International Conference, (**2010**), pp. 43-48.
[13] http://en.wikipedia.org/wiki/Best-first_search. Accessed on 2014.8.13.
[14] M. F. Zibran, "Chi-Squared test of independence", Department of Computer Science, University of Calgary, Alberta, Canada, (**2012**), [online].
[15] A. Estabrooks, T. Jo and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets", Computational Intelligence, vol. 20, no. 1, (**2004**), pp. 18-36.
[16] A. Frank and A. Asuncion, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]", Irvine, CA: University of California, School of Information and Computer Science, (**2010**).
[17] J. W. Smith, J. E. Everhart, W. C. Dickson, W.C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus", In Proceedings of the symposium on computer applications and medical care, vol. 261, (**1988**), pp. 265.
[18] http://wiki.pentaho.com/display/DATAMINING/BestFirst. Accessed on 2014.8.14.
[19] H. Shi, "Best-first decision tree learning", Master's thesis, University of Waikato, Hamilton, NZ, COMP594, (**2007**).
[20] I. Albisua, O. Arbelaitz, I. Gurrutxaga, A. Lasarguren, J. Muguerza and J. M. Pérez, "The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets", Progress in Artificial Intelligence, vol. 2, no. 1, (**2013**), pp. 45-63.
[21] H. Grain and L. K. Schaper, "Health Information: Digital Health Service Delivery, the Future If Now, IOS Press, (**2013**).
[22] H. Mark, F. Eibe, H. Geoffrey, P. Bernhard, R. Peter and H. W. Ian, "The WEKA Data Mining Software: An Update; SIGKDD Explorations", vol. 11, no. 1, (**2009**).
[23] C. Conversano, R. Siciliano and F. Mola, "Generalized Additive Multi-mixture Model for Data Mining", Computational Statistics and Data Analysis, vol. 38, no. 4, (**2002**), pp. 487-500.

Authors

Hoon Jin, He is a research professor at College of Information & Communication Engineering in Sungkyunkwan University. He learned and studied Artificial Intelligence, in detail, data mining, bioinformatics, agent system, semantic web. He received the Ph.D. degree in Department of Computer Science of Kyonggi University, Suwon, Korea, in 2007. After graduation, He worked at Korea Research Institute of Bioscience and Biotechnology (KRIBB) in Daedeok Science Town of Daejeon, Korea during 3 years as a post-doctoral researcher. Then He worked as a senior researcher at Creative Design Institute aiming for studying the product service systems design in 2010. Recently he has worked as a research professor/senior researcher, at Yonsei Institute of Convergence Technology, Yonsei University for 3 years. His main research interests are in fields of data mining, bioinformatics, ontology, semantic web technologies and now actively researches about biomedical text mining.

Seungcheon Kim, He has received the B.S., M.S. and Ph.D. degrees in Electronic Engineering Department of Yonsei University, Seoul, Korea, in 1994, 1996 and 1999, respectively. He is currently with the Department of Information and Communication Engineering, Hansung University, Seoul, Korea, where he is responsible for teaching and research in wireless data communication networks, and ubiquitous sensor networks. He has worked as a post doctorial research fellow in the School of Electrical and Information Engineering in the University of Sydney, Australia, from 2000 to 2001, where he conducted research about 4G Mobile Wireless Communications. He's also worked as a senior research engineer in the Home Network Group of Digital TV Laboratory and the Digital Tech. Group of DA Laboratory, LG Electronics Inc., from 2001 to 2003, where he designed the Home Network Protocol and developed several Home Networking Devices. He has served as a director of Industrial cooperation research center in Hansung University. He was a visiting scholar in the department of computer science in the University of Oregon, United States, from 2009 to 2010. His research interests include the traffic managements in Wireless and mobile communication networks, architectures of 4G Wireless Networks and the design of Home Networking Protocol and Ubiquitous Network Architecture.