An International Journal
JSCDSS
http://www.jscdss.com
Vol.4 No.5 October 2017: 8-16
Article history:
Accepted 3 September 2017
Published online 3 September 2017

Journal of Soft Computing and Decision Support Systems

E-ISSN: 2289-8603

Knowledge Discovery and Diseases Prediction: A Comparative Study of Machine Learning Techniques

Mehrbakhsh Nilashi [a,*], Hossein Ahmadi [b], Leila Shahmoradi [b,*], Abbas Mardani [c], Othman Ibrahim [a], Elaheh Yadegaridehkordi [d]

[a] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

[b] Health Information Management Department, 5th Floor, School of Allied Medical Sciences, Tehran University of Medical Sciences, No #17, Farredanesh Alley, Ghods St, Enghelab Ave, Tehran, Iran

*** Corresponding author email address**: nilashidotnet@hotmail.com

Abstract

The use of medical datasets has attracted the attention of researchers worldwide. Data mining techniques have been widely used in developing decision support systems for disease classification through a set of medical datasets. In this paper, we propose a predictive method for diseases prediction using machine learning techniques. The proposed method is developed through clustering, noise removal, and supervised data mining techniques. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Neural Network (NN), Adaptive Network-Based Fuzzy Inference System (ANFIS), Support Vector Regression (SVR) and Classification and Regression Trees (CART) are used for diseases prediction task. We also use the Principal Component Analysis (PCA) for dimensionality reduction and to address multi-collinearity problems in the experimental datasets. We test our proposed method on several public medical datasets. Experimental results on Wisconsin Diagnostic Breast Cancer, StatLog, Cleveland and Parkinson's telemonitoring datasets show that proposed method remarkably improves diseases prediction accuracy. The predictive method can assist medical practitioners in the healthcare practice as an analytical method.

Keywords: Malignant mesothelioma, Clustering, Incremental PCA, Incremental SVM, Machine Learning.

1. Introduction

The potential of data mining in solving the problems of diseases classification was identified by World Health Organization (WHO) (Gulbinat, 1997). The use of data mining techniques in knowledge discovery for diseases classification has been one of the interesting and important topics addressed by the researchers (Bellazzi et al., 2008; Cauchi et al., 2015; Chen et al., 2016). Data mining is the unified name for all tools that can be used when searching for relationships and trends in large amounts of data (Han et al., 2011). Data mining is a process of discovering useful knowledge from database to build a structure (i.e., model or pattern) that can meaningfully interpret the data. Data mining techniques are pattern recognition techniques can be used to assist physicians in diagnosing and predicting diseases so they can provide the necessary treatment and prevent the impact, including the possibility of death.

Supervised learning techniques construct prediction models for classifying future events in a way consistent with historical information (Kotsiantis et al., 2007; Caruana and Niculescu-Mizil, 2006). Supervised data mining techniques have frequently appeared in various practical areas especially in health care (Zadeh et al., 2010). Classification and prediction problems have a vital role in medical decision making (Pendharkar et al., 1999). Accordingly, due to diseases diagnosis importance to mankind, several studies have been conducted on modeling procedures for their classification (Fida et al., 2011; Lahsasna et al., 2012; Palaniappan and Awang, 2008; Cauchi et al., 2015; Chen et al., 2016; Nilashi et al., 2017a; Nilashi et al., 2017b).

Using clinical data, data-driven predictive models are increasingly being implemented. The main aim of using the data mining techniques is to find best ways to construct generalizable models using clinical data. There is a vast sea of different techniques and algorithms used in data mining especially for supervised machine learning techniques; therefore, selecting the appropriate techniques has been a challenge among researchers in developing the medical diagnosis systems. In addition, these data mining methods can be used to classify the diseases through a set of real-world datasets (see Table 1). According to Frank et al., (2004), "no single algorithm is superior on all data mining problems. The algorithm needs to match the structure of the problem to obtain useful information or an accurate model." Accordingly, in order to improve accuracy of diseases prediction, we propose a new method using clustering, noise removal, and supervised data mining techniques. We then evaluate the method on public medical datasets and report the results.

In this paper, we incorporate the data mining techniques and propose a new predictive method using Principal Component Analysis (PCA), Gaussian mixture model with Expectation Maximization (EM) and supervised data mining techniques. We then evaluate the proposed method on real-world datasets. These datasets are taken from Data Mining Repository of the University of California, Irvine (UCI) (Newman et al., 1998). Thus, in comparison with research efforts found in the literature, our work has the following differences. In this research:

- a predictive method is proposed using EM, PCA, supervised data mining techniques for increasing the diseases prediction accuracy.
- Expectation Maximization (EM) (Mitra et al., 2003) is used for data clustering.
- PCA is used for dimensionality reduction and dealing with the multi-collinearity problem in the experimental data (Nilashi et al., 2016c; NIlashi et al., 2016d).
- Support Vector Machine (SVM), K-Nearest Neighbor (KNN) (Nilashi et al., 2014), Neural Network (NN), Adaptive Network-Based Fuzzy Inference System (ANFIS) (Nilashi et al., 2015; Nilashi et al., 2016e), Support Vector Regression (SVR) (Nilashi et al., 2014), and Classification and Regression Trees (CART) (Briollais et al., 2007) are used for diseases prediction task.

Our study at hand is organized as follows: Section 2 provides the research methodology. Section 3 presents the methods evaluations and discussion. In Section 4 we present the research implications and finally, conclusions and future work is provided in the Section 5.

Table 1
List of previous work based on diseases and techniques used for their classification

Disease	Author	Methods														
		SVM	KNN	NN	ANFIS	Fuzzy Logic	K-Means	GP	EM	PCA	Random Forest	LDA	DT	Association Rule	PSO	NB
Diabetes	Polat et al. (2008)	*														
	Aslam et al. (2013)	*	*					*								
	Kahramanli and Allahverdi (2008)			*		*										
	Erkaymaz and Ozer (2016)			*												
	Ganji and Abadeh (2011)					*										
	Dogantekin et al. (2010)				*							*				
	Temurtas et al. (2009)			*												
	Çalişir and Doğantekin (2011)	*										*				
	Nilashi et al. (2016a)	*							*	*						
	Hayashi and Yukita (2016)												*			
Breast Cancer	Şahan et al. (2007)					*										
	Polat and Güneş (2007)	*														
	Übeyli (2007)	*														
	Marcano-Cedeño (2011)			*												
	Zheng et al. (2014)	*														
	Chen et al. (2014)	*														
	Chen (2014)													*		
	Bhardwaj and Tiwari (2015)			*												
	Onan (2015)		*			*										
	Karabatak (2015)															*
	Abdel-Zaher and Eldeib (2016)			*											*	
	Sheikhpour et al. (2016)														*	

Table 1
List of previous work based on diseases and techniques used for their classification (Cont.)

Disease	Author	Methods														
		SVM	KNN	NN	ANFIS	Fuzzy Logic	K-Means	GP	EM	PCA	Random Forest	LDA	DT	Association Rule	PSO	NB
Parkinson	Guo et al. (2010)							*	*							
	Das (2010)			*												
	Bhattacharya and Bhatia (2010)	*														
	Åström and Koker (2011)					*										
	Li et al. (2011)					*										
	Ozcift (2012)	*														
	Polat (2012)		*			*	*									
	Eskidere et al. (2012)	*		*												
	Chen et al. (2013)	*	*			*				*						
	Babu and Suresh (2013)			*												
	Peterek et al. (2013)										*					
	Hariharan et al. (2014)			*					*	*		*				
	Froelich et al. (2015)												*			
	Khan et al. (2016)			*				*								
	Buza and Varga (2016)			*												
	Naranjo et al. (2016)															*
	Al-Fatlawi et al. (2016)			*												
	Jain and Shetty (2016)		*	*										*		
	Behroozi and Sami (2016)	*	*													*
	Avci et al. (2016)							*								
	Nilashi et al. (2016a)	*			*				*	*						
Heart	Bhatia et al. (2008)	*														
	Allahverdi (2008)			*		*		*								
	Das et al. (2009)			*												
	Adeli et al. (2010)					*										
	Soni et al. (2011)													*		
	Gudadhe et al. (2010)	*		*												
	Ghumbre et al. (2011)	*		*												
	Anooj (2012)					*										
	Rout (2012)					*										
	Nahar et al. (2013)													*		
	Shilaskar and Ghatol (2013)	*														
	Shao et al. (2014)			*												
	Long et al. (2015)					*										
	Nguyen et al. (2015a)					*		*								
	Nguyen et al. (2015b)					*										
	Kausar et al. (2016)	*					*			*						

2. Methodology of research

Focusing on the diseases prediction problem, the present study uses PCA, EM, and supervised data mining techniques. The general framework of proposed method is shown in Fig. 1.

In this framework, clustering techniques can be used as an unsupervised classification method to cluster the data of experimental datasets into similar groups. We suggest that before the classification task patients' data be clustered in similar groups. This way will improve the readability and handling the data. In addition, for big datasets the clustering of data will improve the complexity issue of data

processing. As a dimensionality reduction technique, the clustering techniques need to be incorpporated into the diseases classification methods to reflect its broad appeal and usefulness as one of the steps in exploratory health data analysis. Several clustering techniques have been proposed and the selection of these techniques also is important. Note, as we suggest clustering approach in the framework, linearly dependent between the variables may effect on the clustering results. To overcome this issue, we need to pre-process the data to remove correlated features or reduce the correlations between the variables. Accordingly, in the preprocessing step, we can examine the datasets for correlation coefficients between the input variables. In this framework, we also suggest for dimensionality reduction because the greatest source of difficulties in using classification methods is the existence of multi-collinearity in many sets of data. Multi-collinearity can significantly effect on the classification results. For multivariate analysis and solving this issue, it has been suggested to use as a dimensionality reduction technique for data compression to retain the essential information. The central idea of using dimensionality reduction techniques is to reduce the number of dimensions of the data while preserving as much as possible of the variations in the original dataset. In the final step of our proposed framework, we propose to rely on supervised classification techniques to learn the classification models.

The datasets for the evaluation of method are taken from Data Mining Repository of the University of California, Irvine (UCI) (Newman et al., 1998). The datasets are WDBC, StatLog Heart Disease, Cleveland Heart Disease, and Parkinson's Telemonitoring.

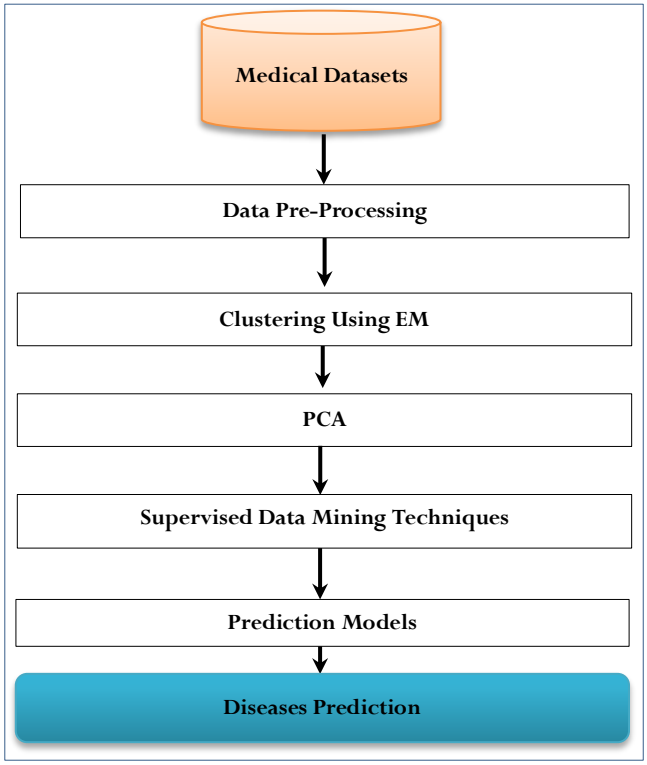

Fig. 1. Proposed model for the diseases diagnosis

To experimentally show the effectiveness of clustering, prediction techniques (SVR and ANFIS) and PCA, we perform the experiments and compare with the techniques ANFIS, Neural Network (NN) and SVR without using clustering and PCA. It should be noted that, for ANFIS models, we selected the best configurations in terms of MFs type, type of trainings and number of training. The Gaussian MF type showed the best performance in relation to the Triangular one. In addition, we selected hybrid learning (training) algorithm in ANFIS. This type of learning algorithm combines the least squares estimator and the gradient descent method. Using the hybrid method, the ANFIS models generated rules by enumerating all possible combinations of MFs of all original inputs and PCs. Compared with the ANFIS for prediction of outputs, the models that used ANFIS with incorporating PCA obtained lower computation time in all models as the computation time for ANFIS is moderately large when the number of inputs is increased (curse of dimensionality) (Brown et al., 1995). Hence, this problem was solved with incorporating the PCA before applying ANFIS. This incorporation of PCA caused the reduction in number of inputs and accordingly hidden layers, number of MFs and rules. Evidently, the training time of prediction models was significantly reduced.

For error estimation in the clusters of EM, after 200 epochs, the averages MAE and R^2 were calculated as presented in Table 2. The MAE and R^2 were calculated based on output (Motor-UPDRS and Total-UPDRS) prediction. It should be noted that we used 10-fold cross validation and average test accuracy for each cluster.

Table 2

MAE and R^2 for predicting Motor-UPDRS and Total-UPDRS in Parkinson's telemonitoring dataset

Method	Output	MAE	R^2
PCA-NN	Motor-UPDRS	0.861	0.721
	Total-UPDRS	0.841	0.745
PCA-ANFIS	Motor-UPDRS	0.662	0.791
	Total-UPDRS	0.634	0.812
PCA-SVR	Motor-UPDRS	0.611	0.825
	Total-UPDRS	0.599	0.831
EM-PCA-ANFIS	Motor-UPDRS	0.585	0.887
	Total-UPDRS	0.532	0.923
EM-PCA-SVR	**Motor-UPDRS**	**0.4721**	**0.977**
	Total-UPDRS	**0.4431**	**0.991**
PCA-CART	Motor-UPDRS	0.669	0.781
	Total-UPDRS	0.641	0.803

The results demonstrate that the accuracy of SVR using RBF kernel are the best on Total-UPDRS and Motor-UPDRS in relation to other methods. Comparison of performance in predicting Motor-UPDRS and Total-UPDRS for PCA-SVR, PCA-NN, PCA-ANFIS and PCA-CART on experimental dataset show that the proposed PCA-ANFIS method is more accurate. However, when compared with EM-PCA-SVR, it can found that prediction errors for PCA-SVR models of EM clusters are lower than other methods with high values of coefficient of determination. Hence, in relation to the PCA-ANFIS, our method using EM, PCA and SVR helps to improve the

JSCDS
E-ISSN: 2289-8603

prediction accuracy of Motor-UPDRS and Total-UPDRS by more than 6% and 9% for Motor-UPDRS and Total-UPDRS, respectively. Moreover, it can be found that the accuracy of methods which uses prediction techniques with EM and PCA is higher than those methods that only use PCA. These show the effectiveness of incorporating the clustering and PCA techniques for the prediction accuracy of PD progression. In addition, the superiority of EM-PCA-ANFIS and EM-PCA-SVR can be explained by the fact that these methods have used clustering and noise removal methods before the prediction of Motor-UPDRS and Total-UPDRS while the other methods solely rely on prediction methods with PCA.

To experimentally show the effectiveness of noise removal, clustering and classification methods (SVM and KNN), we performed the experiments on the proposed methods.We applied SVM with different types of kernels (Polynomial, RBF, Sigmoid and Linear) on experimental dataset clustered by EM algorithm. We use Area under the receiver operating characteristic curve (AUC) which has been defined as a graphical display that provides the measure of the prediction/classification accuracy of the model by two measures of accuracy, the specificity and sensitivity. Specificity is a measure of accuracy for predicting nonevents that is equal to the true negative/total actual negative of a classifier for a range of cutoffs. Sensitivity is a measure of accuracy for predicting events that is equal to the true positive/total actual positive.

Table 3 presents the accuracy results of applying classification methods for SVM (RBF, Polynomial, Sigmoid and Linear kernels) for Cleveland, StatLog, and Wisconsin Diagnostic Breast Cancer datasets. It can be seen that for Cleveland, the accuracy is calculated about 0.9943% using SVM through the RBF kernel. And, for StatLog the accuracy is calculated about 0.9821% using SVM through the RBF kernel. The results showed that the difference of accuracy obtained by kernels is not significant but the SVM using RBF kernel outperforms the others kernels. The results also indicated that the linear kernel give the best performance in case of computation time. In addition, the worst classification accuracy is obtained using polynomial kernel and computation time of sigmoid kernel is higher than other kernels. Furthermore, from the results in Table 3 it can be found that using all types of kernels, EM-PCA-SVM outperforms EM-PCA-KNN method.

Table 3
Prediction accuracy of diseases

Method	Kernel	StatLog Accuracy	Cleveland Accuracy	Wisconsin Diagnostic Breast Cancer Accuracy
EM-PCA-SVM	Polynomial	0.9134	0.9569	0.9623
	RBF	0.9821	0.9943	0.9967
	Sigmoid	0.9567	0.9710	0.9812
	Linear	0.9255	0.9611	0.9716
PCA-SVM	Polynomial	0.8134	0.8531	0.8623
	RBF	0.8821	0.8943	0.8982
	Sigmoid	0.8560	0.8710	0.8840
	Linear	0.8313	0.8567	0.8735
EM-PCA-KNN	-	0.8856	0.8956	0.9144
PCA-KNN	-	0.7533	0.7924	0.8122

3. Conclusion

In this paper, we propose a new predictive method for diseases prediction using data mining learning techniques. We applied EM clustering algorithm to cluster the experimental disease datasets and supervised data mining techniques for disease prediction. In addition, PCA was used for dimensionality reduction and to address multi-collinearity in the datasets. We evaluated the effectiveness of the proposed method by performing several experiments on real-word diseases datasets retrieved from UCI. Using measures of accuracy, ROC, R2 and MAE, we obtained high accuracy for prediction of disease types in all datasets. The results also demonstrated that EM-PCA-SVM and EM-PCA-SVR outperform other supervised data mining techniques. All of the approaches used in this study, may also be applicable to other prediction problems within the medical domain. However, there is still plenty of work in conducting researches on combination of PCA, EM with supervised data mining techniques for disease diagnosis in order to exploit all their potential and usefulness.

References

Abdel-Zaher, A. M., & Eldeib, A. M. (2016). Breast cancer classification using deep belief networks. Expert Systems with Applications, 46, 139-144.

Adeli, A., Mehdi.Neshat, 2010."A Fuzzy Expert System for Heart Disease Diagnosis". Proceedings of the International Multi Conference of Engineers and Computer Scientists, Vol I, ISSN 2078-0966.

Akaike H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6), 716-723.

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications, 36(2), 3240-3247.

Al-Fatlawi, A. H., Jabardi, M. H., & Ling, S. H. (2016, November). Efficient diagnosis system for Parkinson's disease using deep belief network. In Evolutionary Computation (CEC), 2016 IEEE Congress on (pp. 1324-1330). IEEE.

Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University-Computer and

Information Sciences, 24(1), 27-40.
Åström F., & Koker R. (2011). A parallel neural network approach to prediction of Parkinson's Disease. Expert systems with applications, 38(10), 12470-12474.
Avci, D., & Dogantekin, A. (2016). An Expert Diagnosis System for Parkinson Disease Based on Genetic Algorithm-Wavelet Kernel-Extreme Learning Machine. Parkinson's Disease, 2016.
Babu G. S., & Suresh S. (2013). Parkinson's disease prediction using gene expression–A projection based learning meta-cognitive neural classifier approach. Expert Systems with Applications, 40(5), 1519-1529.
Bache, K. and Lichman, M., UCI machine learning repository, 2013.
Behroozi, M., & Sami, A. (2016). A Multiple-Classifier Framework for Parkinson's Disease Detection Based on Various Vocal Tests. International journal of telemedicine and applications, 2016.
Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using genetically optimized neural network model. Expert Systems with Applications, 42(10), 4611-4620.
Bhatia, S., Prakash, P., & Pillai, G. N. (2008, October). SVM based decision support system for heart disease classification with integer-coded genetic algorithm to select critical features. In Proceedings of the World Congress on Engineering and Computer Science, WCECS (pp. 22-24).
Bhattacharya I., & Bhatia M. P. S. (2010, September). SVM classification to distinguish Parkinson disease patients. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India (p. 14). ACM.
Burges C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.
Buza, K., & Varga, N. Á. (2016). ParkinsoNET: Estimation of UPDRS Score Using Hubness-Aware Feedforward Neural Networks. Applied Artificial Intelligence, 30(6), 541-555.
Cattell R. B. (1966). The scree test for the number of factors. Multivariate behavioral research, 1(2), 245-276.
Cauchi, M., Fowler, D. P., Walton, C., Turner, C., Waring, R. H., Ramsden, D. B., ... & Bessant, C. (2015). Comparison of GC-MS, HPLC-MS and SIFT-MS in conjunction with multivariate classification for the diagnosis of Crohn's disease in urine. Analytical Methods, 7(19), 8379-8385.
Cauwenberghs G., & Poggio T. (2001). Incremental and decremental support vector machine learning. Advances in neural information processing systems, 409-415.
Chen H. L., Huang C. C., Yu, X. G., Xu, X., Sun, X., Wang, G., & Wang, S. J. (2013). An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert systems with applications, 40(1), 263-271.
Chen M. S., Han J., & Yu P. S. (1996). Data mining: an overview from a database perspective. Knowledge and data Engineering, IEEE Transactions on, 8(6), 866-883.
Chen, C. H. (2014). A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. Applied Soft Computing, 20, 4-14.
Chen, H. L., Yang, B., Liu, J., & Liu, D. Y. (2011). A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications, 38(7), 9014-9022.
Chen, P., Shen, A., Zhou, X., & Hu, J. (2011). Bio-Raman spectroscopy: a potential clinical analytical method assisting in disease diagnosis. Analytical Methods, 3(6), 1257-1269.
Daneault J. F., Carignan B., Sadikot, A. F., & Duval, C. (2013). Are quantitative and clinical measures of bradykinesia related in advanced Parkinson's disease?. Journal of neuroscience methods, 219(2), 220-223.
Dangare, C. S., & Apte, S. S. (2012). A data mining approach for prediction of heart disease using neural networks. International Journal of Computer Engineering and Technology (IJCET), 3(3).
Das R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications, 37(2), 1568-1572.
Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert systems with applications, 36(4), 7675-7680.
Duijm, L. E. M., Groenewoud, J. H., Jansen, F. H., Fracheboud, J., van Beek, M., & de Koning, H. J. (2004). Mammography screening in the Netherlands: delay in the diagnosis of breast cancer after breast cancer screening. British journal of cancer, 91(10), 1795-1799.
Ene M. (2008). Neural network-based approach to discriminate healthy people from those with Parkinson's disease. Annals of the University of Craiova-Mathematics and Computer Science Series, 35, 112-116.
Ephzibah, E. P. (2010). Cost effective approach on feature selection using genetic algorithms and LS-SVM classifier. IJCA Special Issue on Evolutionary Computation for Optimization Techniques, ECOT.
Eskidere, Ö., Erta F., & Hanilçi C. (2012). A comparison of regression methods for remote tracking of Parkinson's disease progression. Expert Systems with Applications, 39(5), 5523-5528.
Farnikova K., Krobot A., & Kanovsky P. (2012). Musculoskeletal problems as an initial manifestation of Parkinson's disease: A retrospective study. Journal of the neurological sciences, 319(1), 102-104.
Fida, B., Nazir, M., Naveed, N., & Akram, S. (2011, December). Heart disease classification ensemble optimization using genetic algorithm. In Multitopic Conference (INMIC), 2011 IEEE 14th International (pp. 19-24). IEEE.
Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. Bioinformatics, 20(15), 2479-2481.

E-ISSN: 2289-8603

Froelich W., Wrobel K., & Porwik P. (2015). Diagnosis of Parkinson's Disease Using Speech Samples and Threshold-Based Classification. Journal of Medical Imaging and Health Informatics, 5(6), 1358-1363.

Ghumbre, S., Patil, C., & Ghatol, A. (2011, December). Heart disease diagnosis using support vector machine. In International conference on computer science and information technology (ICCSIT') Pattaya.

Gudadhe, M., Wankhade, K., & Dongre, S. (2010, September). Decision support system for heart disease based on support vector machine and artificial neural network. In Computer and Communication Technology (ICCCT), 2010 International Conference on (pp. 741-745). IEEE.

Guo P. F., Bhattacharya P., & Kharma N. (2010). Advances in detecting Parkinson's disease. In Medical Biometrics (pp. 306-314). Springer Berlin Heidelberg.

Hall, P. M., Marshall, A. D., & Martin, R. R. (1998, September). Incremental Eigenanalysis for Classification. In BMVC (Vol. 98, pp. 286-295).

Han J., & Kamber M. (2001). Data mining: concepts and techniques. Morgan Kaufmann, San Francisco, Calif, USA, 2nd edition, 2011.

Hariharan M., Polat K., & Sindhu R. (2014). A new hybrid intelligent system for accurate detection of Parkinson's disease. Computer methods and programs in biomedicine, 113(3), 904-913.

Hayashi, Y., & Yukita, S. (2016). Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked, 2, 92-104

He, Y., Tang, Y., Zhang, Y. Q., & Sunderraman, R. (2006). Adaptive Fuzzy Association Rule mining for effective decision support in biomedical applications. International journal of data mining and bioinformatics, 1(1), 3-18.

Hwang J. P., Park S., & Kim E. (2011). Dual margin approach on a Lagrangian support vector machine. International Journal of Computer Mathematics, 88(4), 695-708.

Jain, S., & Shetty, S. (2016, April). Improving accuracy in noninvasive telemonitoring of progression of Parkinson'S Disease using two-step predictive model. In 2016 Third International Conference on Electrical, Electronics, Computer Engineering and their Applications (EECEA) (pp. 104-109). IEEE.

Jung, Y. G., Kang, M. S., & Heo, J. (2014). Clustering performance comparison using K-means and expectation maximization algorithms. Biotechnology & Biotechnological Equipment, 28(sup1), S44-S48.

Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert Systems with Applications, 35(1), 82-89.

Karabatak, M. (2015). A new classifier for breast cancer detection based on Naïve Bayesian. Measurement, 72, 32-36.

Kaufman, L. R., & Rousseeuw, P. PJ (1990) Finding groups in data: An introduction to cluster analysis. Hoboken NJ John Wiley & Sons Inc.

Kaur A., & Kaur K. (2013). Statistical Comparison Of Modelling Methods For Software Maintainability Prediction. International Journal of Software Engineering and Knowledge Engineering, 23(06), 743-774.

Kaur, P., Goyal, M., & Lu, J. (2014, July). A price prediction model for online auctions using fuzzy reasoning techniques. In Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on (pp. 1311-1318). IEEE.

Kausar, N., Abdullah, A., Samir, B. B., Palaniappan, S., AlGhamdi, B. S., & Dey, N. (2016). Ensemble Clustering Algorithm with Supervised Classification of Clinical Data for Early Diagnosis of Coronary Artery Disease. Journal of Medical Imaging and Health Informatics, 6(1), 78-87. Singapore.

Khan, M. M., Chalup, S. K., & Mendes, A. (2016, February). Parkinson's Disease Data Classification Using Evolvable Wavelet Neural Networks. In Australasian Conference on Artificial Life and Computational Intelligence (pp. 113-124). Springer International Publishing.

Kharazmi, E., Försti, A., Sundquist, K., & Hemminki, K. (2016). Survival in familial and non-familial breast cancer by age and stage at diagnosis. European Journal of Cancer, 52, 10-18.

Kohavi R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).

Lahsasna, A., Ainon, R. N., Zainuddin, R., & Bulgiba, A. (2012). Design of a fuzzy-based decision support system for coronary heart disease diagnosis. Journal of medical systems, 36(5), 3293-3306.

Li D. C., Liu C. W., & Hu S. C. (2011). A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. Artificial Intelligence in Medicine, 52(1), 45-52.

Long, N. C., Meesad, P., & Unger, H. (2015). A highly accurate firefly based algorithm for heart disease prediction. Expert Systems with Applications, 42(21), 8221-8231.

Loukas C., & Brown P. (2012). A PC-based system for predicting movement from deep brain signals in Parkinson's disease. Computer methods and programs in biomedicine, 107(1), 36-44.

Luengo-Fernandez, R., Leal, J., & Gray, A. M. (2012). UK research expenditure on dementia, heart disease, stroke and cancer: are levels of spending related to disease burden?. European journal of Neurology, 19(1), 149-154.

Mandal I., & Sairam N. (2013). Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system. International journal of medical informatics, 82(5), 359-377.

Marcano-Cedeño, A., Quintanilla-Domínguez, J., & Andina, D. (2011). WBCD breast cancer database

E-ISSN: 2289-8603

classification applying artificial metaplasticity neural network. Expert Systems with Applications, 38(8), 9573-9579.

McCarthy, A. M., Yang, J., & Armstrong, K. (2015). Increasing disparities in breast cancer mortality from 1979 to 2010 for US black women aged 20 to 49 years. American journal of public health, (0), e1-e3.

Mendis, S., Puska, P., & Norrving, B. (2011). Global Atlas on Cardiovascular Disease Prevention and Control. World Health Organization in Collaboration with World Heart Federation, World Stroke Organization.

Mitra, P., Pal, S. K., & Siddiqi, M. A. (2003). Non-convex clustering using expectation maximization algorithm with rough set initialization. Pattern Recognition Letters, 24(6), 863-873.

Nahar, J., Imam, T., Tickle, K. S., & Chen, Y. P. P. (2013). Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Applications, 40(4), 1086-1093.

Naranjo, L., Pérez, C. J., & Martín, J. (2016). Addressing voice recording replications for tracking Parkinson's disease progression. Medical & biological engineering & computing, 1-9.

Nathiya G., Punitha S. C., & Punithavalli M. (2010). An analytical study on behavior of clusters using k means, em and k- means algorithm. International Journal of Computer Science and Information Security, 7(3), 155-190.

Newman, D. J., Hettich, S., Blake, C. L., Merz, C. J., & Aha, D. W. (1998). UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA. In 1998 of Conference, http://archive. ics. uci. edu/ml/datasets. html.

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015a). Classification of healthcare data using genetic fuzzy logic system and wavelets. Expert Systems with Applications, 42(4), 2184-2197.

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. (2015b). Medical data classification using interval type-2 fuzzy logic system and wavelets. Applied Soft Computing, 30, 812-822.

Nilashi, M., bin Ibrahim, O., & Ithnin, N. (2014). Multi-criteria collaborative filtering with high accuracy using higher order singular value decomposition and Neuro-Fuzzy system. Knowledge-Based Systems, 60, 82-101.

Nilashi, M., bin Ibrahim, O., & Ithnin, N. (2014). Multi-criteria collaborative filtering with high accuracy using higher order singular value decomposition and Neuro-Fuzzy system. Knowledge-Based Systems, 60, 82-101.

Nilashi, M., Ibrahim, O. B., Ithnin, N., & Zakaria, R. (2015). A multi-criteria recommendation system using dimensionality reduction and Neuro-Fuzzy techniques. Soft Computing, 19(11), 3173-3207.

Nilashi, M., Ibrahim, O. B., Mardani, A., Ahani, A., & Jusoh, A. (2016a). A soft computing approach for diabetes disease classification. Health Informatics Journal, 1, 15.

Nilashi, M., Ibrahim, O., & Ahani, A. (2016b). Accuracy Improvement for Predicting Parkinson's Disease Progression. Scientific Reports, 6.

Nilashi, M., Esfahani, M. D., Roudbaraki, M. Z., Ramayah, T., & Ibrahim, O. (2016c). A multi-criteria collaborative filtering recommender system using clustering and regression techniques. Journal of Soft Computing and Decision Support Systems, 3(5), 24-30.

Nilashi, M., Bin Ibrahim, O., Mardani, A., Ahani, A., & Jusoh, A. (2016d). A soft computing approach for diabetes disease classification. Health Informatics Journal, 1460458216675500.

Akbari, E., Buntat, Z., Shahraki, E., Zeinalinezhad, A., & Nilashi, M. (2016e). ANFIS modeling for bacteria detection based on GNR biosensor. Journal of Chemical Technology and Biotechnology, 91(6), 1728-1736.

Nilashi, M., Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017a). A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics, 34(4), 133-144.

Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017b). An Analytical Method for Diseases Prediction Using Machine Learning Techniques. Computers & Chemical Engineering, 106, 212-223.

Nilashi, M., Bagherifard, K.., Rahmani, M., & Rafe, V. (2017c). A Recommender System for Tourism Industry Using Cluster Ensemble and Prediction Machine Learning Techniques. Computers & Industrial Engineering, 109, 357-368.

Nilashi, M., Dalvi, M., Ibrahim, O., Fard, K. B., Mardani, A., & Zakuan, N. (2017d). A Soft Computing Method for the Prediction of Energy Performance of Residential Buildings. Measurement, 109, 268-280.

Nilashi, M., Ahmadi, H., Shahmoradi, L., Salahshour, M., & Ibrahim, O. (2017d). A Soft Computing Method for Mesothelioma Disease Classification. Journal of Soft Computing and Decision Support Systems, 4(1), 16-18.

Onan, A. (2015). A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Systems with Applications.

Ordonez, C., & Omiecinski, E. (2002, November). FREM: fast and robust EM clustering for large data sets. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 590-599). ACM.

Ozcift, A. (2012). SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease. Journal of medical systems, 36(4), 2141-2147.

Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on (pp. 108-115). IEEE.

Patil, S. B., & Kumaraswamy, Y. S. (2009). Intelligent and effective heart attack prediction system using data

E-ISSN: 2289-8603

mining and artificial neural network. European Journal of Scientific Research, 31(4), 642-656.

Pelleg, D., & Moore, A. W. (2000, June). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In ICML (pp. 727-734).

Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N., & Benner, M. (1999). Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications, 17(3), 223-232.

Peterek T., Dohnalek P., Gajdos, P., & Smondrk M. (2013, December). Performance evaluation of Random Forest regression model in tracking Parkinson's disease progress. In Hybrid Intelligent Systems (HIS), 2013 13th International Conference on (pp. 83-87). IEEE.

Poggio, T., & Cauwenberghs, G. (2001). Incremental and decremental support vector machine learning. Advances in neural information processing systems, 13, 409.

Polat K. (2012). Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering. International Journal of Systems Science, 43(4), 597-609.

Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. Digital Signal Processing, 17(4), 694-701.

Postuma R. B., & Montplaisir J. (2009). Predicting Parkinson's disease–why, when, and how?. Parkinsonism & related disorders, 15, S105-S109.

Que, J., Jiang, X., & Ohno-Machado, L. (2012). A collaborative framework for distributed privacy-preserving support vector machine learning. In AMIA Annual Symposium Proceedings (Vol. 2012, p. 1350). American Medical Informatics Association.

Romenets S. R., Gagnon J. F., Latreille V., Panniset M., Chouinard S., Montplaisir J., & Postuma R. B. (2012). Rapid eye movement sleep behavior disorder and subtypes of Parkinson's disease. Movement Disorders, 27(8), 996-1003.

Şahan, S., Polat, K., Kodaz, H., & Güneş, S. (2007). A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine, 37(3), 415-423.

Schölkopf B., & Smola A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press.

Shao, Y. E., Hou, C. D., & Chiu, C. C. (2014). Hybrid intelligent modeling schemes for heart disease classification. Applied Soft Computing, 14, 47-52.

Sheikhpour, R., Sarram, M. A., & Sheikhpour, R. (2015). Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. Applied Soft Computing.

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. Expert Systems with Applications, 40(10), 4146-4153.

Shradhanjali, 2012. "Fuzzy Petry Net Application: Heart Disease Diagnosis", International Conference on Computing and Control Engineering.

Soni, J., Ansari, U., Dipesh Sharma, 2011, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative classifiers". International Journal on Computer Science and Engineering, vol 3, No. 6, pp.2385- 2392.

Tabar, L., Tot, T., & Dean, P. B. (2004). Breast Cancer: The Art and Science of Early Detection with Mammography-Perception. Interpretation, Histopatholigic Correlation, 1st edn Georg Thieme Verlag.

Tsanas A., Little M., McSharry P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. Biomedical Engineering, IEEE Transactions on, 57(4), 884-893.

Übeyli, E. D. (2007). Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications, 33(4), 1054-1062.

Vapnik V., Golowich S. E., & Smola A. (1996). Support vector method for function approximation, regression estimation, and signal processing. In Advances in Neural Information Processing Systems 9.

World Health Organization. World health organization cause-specific mortality estimates for 2000-2012. June 2014.

World Health Organization. World health organization estimates of causespecific disability-adjusted life year for 2000-2012. June 2014.

World Health Organization. World health organization projections of causes of death, 2015 and 2030. July 2013.

Wu, D., Zhang, G., & Lu, J. (2013, October). A fuzzy tree similarity measure and its application in telecom product recommendation. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on (pp. 3483-3488). IEEE.

Yapa S. S. (1992). Detection of subclinical ascorbate deficiency in early Parkinson's disease. Public health, 106(5), 393-395.

Zadeh, L. A. (1965). Fuzzy sets. Information and control, 8(3), 338-353.

Zhang, Z., Lin, H., Liu, K., Wu, D., Zhang, G., & Lu, J. (2013). A hybrid fuzzy-based personalized recommender system for telecom products/services. Information Sciences, 235, 117-129.

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.

E-ISSN: 2289-8603