Journal of AI and Data Mining
Vol 6, No 2, 2018, 277-285
DOI: 10.22044/JADM.2017.4673.1555

Impact of Patients' Gender on Parkinson's disease using Classification Algorithms

M. Abdar[1] and M. Zomorodi-Moghadam[2*]

1. School of Computer Science & Engineering, The University of Aizu, Aizu-Wakamatsu, Japan.
2. Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

Received 24 August 2016; Revised 19 July 2017; Accepted 30 August 2017
**Corresponding author: m_zomorodi@um.ac.ir (M. Zomorodi-Moghadam).*

Abstract
In this work, the accuracy of two machine learning algorithms including the SVM and Bayesian networks were investigated as two important algorithms in the diagnosis of the Parkinson's disease (PD). We used the PD data in the University of California, Irvine (UCI). In order to optimize the SVM algorithm, different kernel functions and C parameters were used, and the results obtained showed that SVM with C parameter (C-SVM) with an average accuracy of 99.18% with the polynomial kernel function in the testing step had a better performance compared to the other kernel functions such as RBF and sigmoid as well as the Bayesian network algorithm. It was also shown that the ten important factors involved in the SVM algorithm were Jitter (Abs), Subject #, RPDE, PPE, Age, Shimmer APQ 11, NHR, Total-UPDRS, Shimmer (dB), and Shimmer respectively. We also proved that the accuracy of our proposed C-SVM and RBF approaches was in direct proportion to the value of the C parameter such that with increase in the amount of C, the accuracy in both kernel functions increased. However, unlike polynomial and RBF, sigmoid had an inverse relation with the amount of C. Indeed, by using these methods, we can find the most effective factors common in both genders (male and female). To the best of our knowledge, there has been no study on PD for identifying the most effective factors common in both genders.

Keywords: *Data Mining, Parkinson's Disease, SVM Algorithm, Bayesian Network Algorithm, C-SVM Algorithm.*

1. Introduction

Early diagnosis of many diseases by physicians can have an important role in preventing the development of a disease, and so the accuracy of the diagnosis is very important. Therefore, using appropriate ways to detect and diagnose diseases with high accuracy can contribute to a better treatment of patients. One of these ways is data mining (DM). Parkinson's disease (PD) [1] is a chronic progressive nervous system disorder that primarily affects movement.

Aging is considered as an important risk factor for PD, and even genetic and environment factors may contribute to PD. The Parkinson's disease was first described by the British scientist Dr. James Parkinson in 1817. He called the disease "shaking palsy" but today it is known as PD after him [2]. PD is a disease of the central nervous system, which mainly occurs in persons who are 40 years old or more with different symptoms such as the gradual stiffening of muscles, and appearance of trembling in various parts of the body. In addition, recent studies have shown that the number of people with PD has increased over the past 60 years [3, 4]. This disorder occurs when a specific area in the brain loses its ability in the production of dopamine (a brain neurotransmitter). According to the studies, PD is considered as the second most common neurodegenerative disorder after the Alzheimer's disease [5].

Although definitive treatment has not been found to eradicate this disease, with the advancement of science, researchers are trying to use a variety of methods to combat against it. Fortunately, with the help of various branches of science, significant progress has been achieved in the control of PD.

One of the emerging techniques that helps the physicians in the early diagnosis and treatment of a disease is DM. According to [6-8], gender is one of the most important factors involved in PD. For this reason, we will concentrate on this factor and its relationship with the other features available in the PD dataset through the use of DM approaches.

1.1. Data mining (DM)

DM and knowledge discovery in databases from large amounts of data have led to the discovery of the hidden knowledge between them. The process of DM includes several steps such as identifying the source data, selecting the data points to be analyzed, extracting the relevant information using some algorithms, and analysis of the results obtained. Several algorithms are used in DM such as SVM, KNN, neural network, C5.0, Apriori, Cox, and K-Means. DM can be used in many scientific fields such as various medical fields [9-12], security [13, 14], marketing [15, 16], web and text mining [17, 18], and various engineering fields [19, 20, 21].

DM has constantly faced several challenges over time, and with increase in the knowledge in this field, a lot of these problems have been solved.

In this work, we used the DM techniques and identified and introduced a useful way to predict the relationship between gender and the important factors in the PD dataset. In this regard, first, we checked two well-known methods in DM in order to predict the patients' gender. We then proposed an improvement in SVM using a regularization parameter (C) on different kernel functions such as Radial Basis Function (RBF) sigmoid, and polynomial. Subsequently, we compared their performance using various metrics such as specificity, sensitivity, precision, FPR, FNR, F1, and accuracy.

As discussed earlier, gender is one the most important features in PD, and for this reason, we concentrated on it. In fact, the previous works have not focused on gender and its relation with other features.

The remainder of this work is organized as follows. In Section 2, we provide a brief review on the related works in the literature. Section 3 illustrates our proposed methods. Our experimental results are presented and discussed in Section 4. In Section 5, we describe our proposed C-SVM algorithm implemented with various kernel functions. Finally, in Section 6, the paper is concluded.

2. Related work

In the recent years, several studies have been done on PD using DM. In this section, some of these works related to PD that use various DM techniques are introduced. In [22], three well-known methods including KNN, random forest, and Ada-Boost algorithms have been implemented on the PD dataset. The results obtained indicate that the KNN algorithm has the best performance with an accuracy of 90.26% when the value of K is equal to 10.

In [23], four DM techniques have been compared with the PD data in the UCI repository dataset. Naïve Bayes classifier, J48, Decision table, and Random tree are those algorithms that have been implemented in [23]. The outcomes showed that Random tree algorithm had a better performance compared to the other algorithms. The accuracy of the Random tree algorithm was 84%.

In another work carried out by Tawseef Ayoub Shaikh [24], the performance of three algorithms used in DM has been investigated. These algorithms include artificial neural network, decision tree algorithm, and Naïve Bayes algorithm. These algorithms were applied to the PD and primary tumor disease datasets. Their results showed that the accuracy of artificial neural network for diagnosis of PD was 90.7692%, which was the best performance among the three algorithms used. Decision trees had an accuracy of 80.5128%, and Naïve Bayes had an accuracy of 69.2308%.

In [25], four methods including neural network, DMNeural, Regression, and Decision tree have been used as the multiple classification methods for diagnosis of PD. The results of this work illustrated that neural network with an accuracy of 92.90% had the best performance compared to the other methods.

3. Method

In this section, we briefly introduce the Support Vector Machine (SVM) and Bayesian network algorithms as two important algorithms in DM. We also discuss about the UCI PD dataset and its factors.

3.1. Support vector machine (SVM)

SVMs are a supervised learning method that can be used for classification and regression. SVM is one of the relatively new methods that have shown good performance for classification over the older methods such as the perceptron neural networks.

This algorithm maps the input into some high dimensional feature space through some non-linear mappings [26]. The input is a vector or pattern of n features. In the most popular form of this algorithm, the data is transferred to a higher-

dimensional space by Phi function. Therefore, to be able to solve the problems with very high dimensions using this method, the Lagrange duality theorem is used for converting the intended minimization problem to its dual form instead of using the complex function Phi [26, 28]. A more detailed description of this algorithm can be found in [26-29].

3.2. Bayesian network

Today many problems are solved with the help of artificial intelligence. One of the main characteristics of these problems is the uncertainty between them. Many techniques in artificial intelligence have been proposed for controlling uncertainties, most of which are based upon the probability theory and the fuzzy theory. One of the useful methods used to control uncertainty in the issues based on the probability theory is the Bayesian network [29-31]. Bayesian network is a directed graph whose nodes contain information about conditional probability values. More precisely, this network includes the following components and features:

a. A collection of random variables constitute the vertices of the graph whose variables can be discrete or continuous.
b. A set of directed edges $X \to Y$, where X is the parent of Y.
c. Each node X_i, has a conditional probability distribution $P(X_i \mid Parents(X_i))$ that shows the effect of the parents' nodes on this node numerically.
d. Graph did not have a direction away, and, in fact, is a directed acyclic graph.

3.3. Dataset

In our modern world, access to different data in different fields is easy, and, at the same time, the volume of data in various fields is increasing. For DM, using reliable data repositories is essential. One of the best sources for obtaining reliable data is the data repository of the University of California, Irvine (UCI). In this paper, we used the PD data available in the UCI data repository [32], which is presented in table 1. The data was related to 42 people with 22 factors for each of them. The total number of data was 5875 records. The total number of data for male patients was 4008, whereas the total number of data for female patients was 1867.

4. Results

The algorithms were executed using IBM SPSS Modeler 14.2 on an Intel core i7 processor with 8GB Ram under the Windows 8.1 operating system. The main goal of this research work was to identify much more effective factors involved in the prediction of gender in PD using the described methods. Reducing the number of factors is important for two reasons. The first reason is to speed up the training phase of the algorithms, and the other one is the increase the prediction accuracy. However, it should be noted that the less important factors should not be overlooked, especially in medical science. Indeed, the smallest signs in medicine are important in order to save the patients' life. Thus in this work, we used all of the existing factors mentioned in table 1. The data was divided into two groups, 70% for training and 30% for testing. In this regard, gender was determined as target in our work. The main purpose for selecting the sex as a target factor is because sex has a major impact on the diagnosis of PD. Thus sex was determined as the target factor, and the other factors were determined as inputs. IBM SPSS Modeler 14.2 was used for implementation of algorithms, and by using the SVM and Bayesian network algorithms, important factors could be identified. In order to compare the performance of these two algorithms, there were 7 important metrics that were calculated according to equations 1 to 7, as follow [33, 34]:

$$Specificity = TNR = TN / TN + FP \quad (1)$$

$$Sensitivity = TPR = TP / TP + FN \quad (2)$$

$$Precision = TP / TP + FP \quad (3)$$

$$FPR = FP / FP + TN = 1 - TNR \quad (4)$$

$$FNR = FN / FN + TP = 1 - TPR \quad (5)$$

$$F_1 = 2TP / \left(2TP + FP + FN\right) \quad (6)$$

$$Accuracy = TP + TN / TP + TN + FP + FN \quad (7)$$

where:

FN = The number of positively labeled data, which falsely has been classified as "Negative".

TN = The number of negatively labeled data, which has been classified as "Correct".

TP = The number of positively labeled data, which has been classified as "Correct".

FP = The number of negatively labeled data, which falsely has been classified as "Positive".

To evaluate the performance of algorithms, the confusion matrix is an appropriate way. For this purpose, in our study, the confusion matrix was utilized [35], which is shown in figure 1.

Actual	Predicted	
	Disease (positive)	No-disease (negative)
Positive	TP	FP
Negative	FN	TN

Figure 1. Confusion matrix in this work.

It should be noted that in our dataset, the total number of male patients was much more than female patients, and for this reason, we considered male as Positive and female as Negative. This approach helped us to find the values for TP, TN, FP, and FN more precisely.

The performances of the SVM and Bayesian networks are as shown in figures 2 and 3. By comparing the results in tables 2 and 3, it can be seen that SVM has a better performance in both the training and testing steps.

Table 1. Dataset from UCI related to PD.

NO	Feature Name and Attribute Information	Range
1	Subject#: Integer that uniquely identifies each subject	[1 - 42]
2	Age: Subject age	[36 - 85]
3	Sex: Subject gender '0' - male, '1' – female	[0-1]
4	Test- time: Time since recruitment into the trial	[-4.2625 - 215.49]
5	Motor-UPDRS: Clinician's motor UPDRS score, linearly interpolated	[5.0377 - 39.511]
6	Total-UPDRS: Clinician's total UPDRS score, linearly interpolated	[7 - 54.992]
7	Jitter (%): measures of variation in fundamental frequency	[0.00083 - 0.09999]
8	Jitter (Abs): measures of variation in fundamental frequency	[0.00000225 - 0.00044559]
9	Jitter:RAP: measures of variation in fundamental frequency	[0.00033 - 0.05754]
10	Jitter:PPQ5: measures of variation in fundamental frequency	[0.00043 - 0.06956]
11	Jitter:DDP: measures of variation in fundamental frequency	[0.00098 - 0.17263]
12	Shimmer: measures of variation in amplitude	[0.00306 - 0.26863]
13	Shimmer (dB): measures of variation in amplitude	[0.026 - 2.107]
14	Shimmer: APQ3: measures of variation in amplitude	[0.00161 - 0.16267]
15	Shimmer: APQ5: measures of variation in amplitude	[0.00194 - 0.16702]
16	Shimmer: APQ11: measures of variation in amplitude	[0.00249 - 0.27546]
17	Shimmer: DDA: measures of variation in amplitude	[0.00484 - 0.48802]
18	NHR: measures of ratio of noise to tonal components in the voice	[0.000286 - 0.74826]
19	HNR: measures of ratio of noise to tonal components in the voice	[1.659 - 37.875]
20	RPDE: A nonl-inear dynamical complexity measure	[0.15102 - 0.96608]
21	DFA: Signal fractal scaling exponent	[0.51404 - 0.8656]
22	PPE: A non-linear measure of fundamental frequency variation	[0.021983 - 0.73173]

Table 2. Comparison of performance of SVM and Bayesian network algorithms through using training dataset for prediction of Sex in PD (%).

Algorithm	Specificity	Sensitivity	Precision	FPR	FNR	F_1	Accuracy
SVM	85.16	**92.45**	92.89	14.84	**7.55**	**92.67**	**90.10**
Bayesian Network	**86.12**	89.13	**94.05**	**13.88**	10.87	91.52	88.27

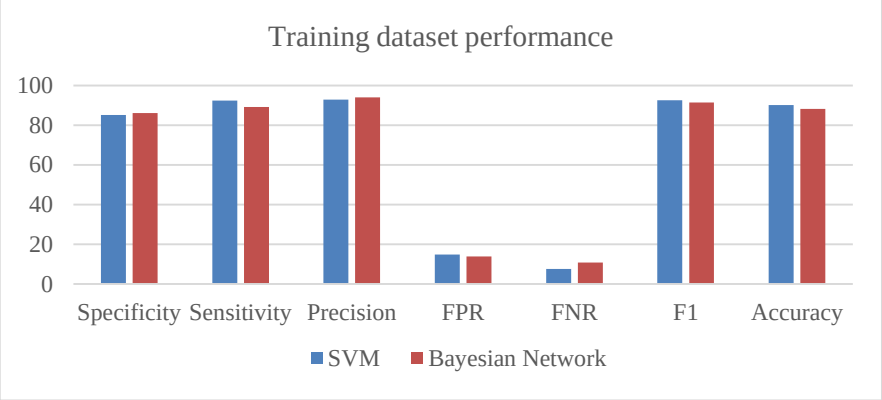

Figure 2. Performance of SVM and Bayesian networks in training dataset.

Table 3. Comparison of performance of SVM and Bayesian network algorithms through using testing dataset for prediction of Sex in PD (%).

Algorithm	Specificity	Sensitivity	Precision	FPR	FNR	F_1	Accuracy
SVM	82.63	**94.89**	92.08	17.37	**5.11**	**93.46**	**90.98**
Bayesian Network	**83.81**	90.43	**93.68**	**16.19**	9.57	92.02	88.62

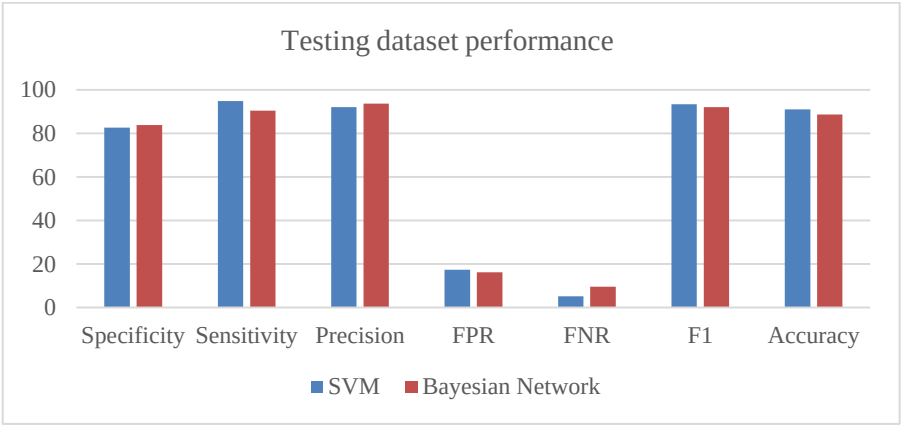

Figure 3. Performance of SVM and Bayesian networks in testing dataset.

Table 4. Conditional probabilities of subject #.

Parents	Probability				
Sex	**< 9.2**	**[9.2 – 17.4)**	**[17.4 – 25.6)**	**[25.6 – 33.8]**	**> 33.8**
1	0.08	0.21	0.13	0.27	0.31
0	0.30	0.17	0.20	0.14	0.19

As it can be seen, the testing accuracy in the SVM and Bayesian network algorithms is more than the training accuracy. These numbers indicate that in PD, both algorithms have similar behaviors.

In figure 4, the most important factors in the SVM algorithm, which are almost half of the whole factors of table 1, are shown. According to figure 4, the most effective and the most important factors for diagnosis of PD are as follow:

Jitter (Abs)

Subject #

RPDE

PPE

Age
Shimmer APQ 11
NHR
Total-UPDRS
Shimmer (dB)
Shimmer

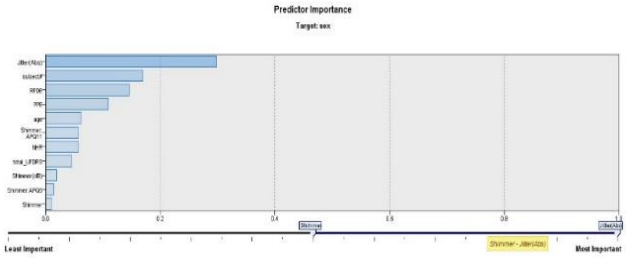

Figure 4. Predictor importance in SVM algorithm.

In figure 5 and table 4, more details about the Bayesian network have been presented.

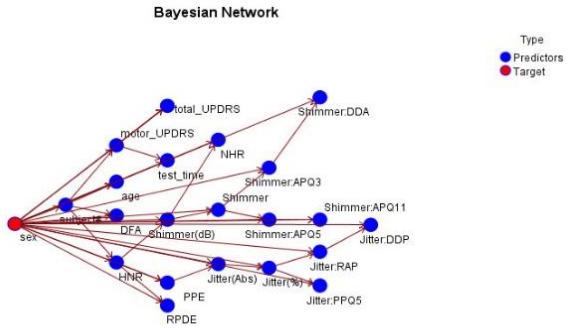

Figure 5. Generated model by Bayesian network for PD.

As it can be seen, when sex = female (1) and subject # is >33.8, the probability of having this disease is greater compared to the other cases and this probability is 0.31, while when sex = male (0) and subject # is <9.2, the probability is greater compared to the other cases, where this probability is 0.30. To further examine the PD dataset and due to better performance through using SVM, this algorithm was utilized in the other two different approaches including polynomial and sigmoid. As a result, by using these approaches, two columns were added to the rest of the data columns, which were \$S-Sex and \$SP-Sex. \$S-Sex are the predicted values for Sex and \$SP-Sex are the scores tendency for prediction. This means that the probability of the predictions for a particular record is correct and is a number between 0 and 1. For instance, when \$SP-Sex is 1, this means that prediction for Sex was done correctly. More details about these approaches are presented in tables 5 and 6, as well as in figures 5 and 6. It should be noted that the performance of simple SVM and SVM with RBF were similar, where for the SVM with RBF approach, the stopping criteria was 1.0E-3. The regularization parameter (C) was 10, regression precision (epsilon) was 0.1, and RBF gamma was 0.1.

The implementation of these approaches reveals that SVM with polynomial had a better performance compared to RBF and sigmoid approaches for finding the relationship between Sex and other factors in PD. To ensure these results, we compared the \$SP-Sex in every three approach. The results obtained using different approaches are presented in tables 5 and 6.

5. Our proposed C-SVM algorithm and experimental comparison on kernel functions

As mentioned earlier, in this work, three different kernel functions were utilized including the Radial Basis Function **(**RBF), sigmoid, and polynomial. In the previous section, we showed that the polynomial approach had a better performance compared to the other kernel functions. In this section, we use the parameter optimization in each kernel function in order to reach a greater accuracy by using them. In this regard, the regularization parameter (C) and SVM were used together as the C-SVM algorithm in all the three kernel functions. The results obtained can be seen in table 7 when the stopping criterion was 1.0E-3 for all kernel functions. According to this table, our proposed C-SVM has a different behavior in terms of different kernel functions and different C values. The sigmoid kernel function had clearly inferior of accuracy rather than RBF and polynomial kernel functions. Table 7 indicates that unlike RBF and polynomial, the accuracy of sigmoid decreased with increase in the number of C. Our results also showed that various regression precisions (epsilons) had almost similar effects on the accuracy. On the other hand, when we changed epsilon with the same value of C, equal accuracy was observed in all the kernel functions. It should be mentioned that we just observed one case in RBF when C = 1 and regression precision (epsilon) = 0.10, which had two accuracies. Also our results indicated that RBF and polynomial had direct relationships with C but we could argue that the amount of C had a much more impact on RBF compared to polynomial. However, in overall, polynomial had the highest accuracy with average of 99.18%, and the average accuracies of RBF and sigmoid were 89.15% and 70.04%, respectively.

By utilizing the parameter optimization approaches, we can ensure the highest accuracy and also guarantee the best performance.

Table 5. Comparison of performance of SVM with RBF, polynomial and sigmoid approaches on PD using training data (%).

Approach	Specificity	Sensitivity	Precision	FPR	FNR	F_1	Accuracy
RBF	82.63	94.89	92.08	17.37	5.11	93.46	90.98
Polynomial	99.54	**99.52**	99.78	0.46	**0.48**	**99.65**	**99.54**
Sigmoid	**100**	68.56	**100**	**0.0**	31.44	81.35	69.10

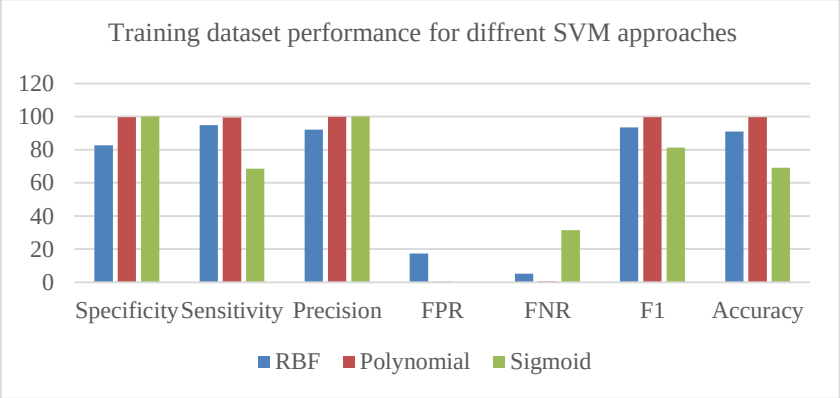

Figure 6. Performance of SVM with various approaches in training dataset.

Table 6. Comparison of performance of SVM with RBF, polynomial, and sigmoid approaches on PD using testing data (%).

Approach	Specificity	Sensitivity	Precision	FPR	FNR	F_1	Accuracy
RBF	82.63	94.89	92.08	17.37	5.11	93.46	90.98
Polynomial	99.24	99.36	99.68	0.76	0.64	99.52	99.33
Sigmoid	100	70.87	100	0.0	29.13	82.95	71.19

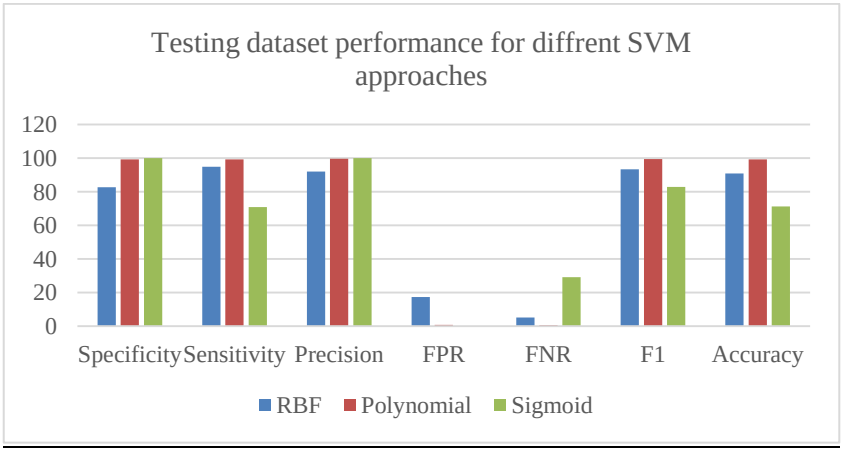

Figure 7. Performance of SVM with various approaches in testing dataset.

6. Conclusion

Parkinson's disease (PD) is a progressive disorder of the nervous system that mainly affects movement. Due to the importance of early diagnosis and treatment, in this paper we provided a useful approach to help for finding the relationship between gender and other features in PD. We used two important algorithms in data mining (DM) including the SVM and Bayesian networks. We assigned Sex as the target and other factors as the inputs. The results obtained indicated that the SVM algorithm had a better performance than the Bayesian network algorithm

for diagnosis of PD. The testing accuracy for the SVM and Bayesian network algorithms were 90.98% and 88.62%, respectively. According to the results obtained, we realized that the SVM algorithm had a remarkable ability to identify the gender of patients who had PD. In addition, we found ten more important factors which are Jitter (Abs), Subject #, RPDE, PPE, Age, Shimmer APQ 11, NHR, Total-UPDRS, Shimmer (dB) and Shimmer respectively. Another result showed that C-SVM with polynomial as a kernel function had a much better performance than the RBF and sigmoid functions. Moreover, our results, with different values of parameter C, indicated that polynomial and RBF have better accuracies when we increased the amount of C, but sigmoid has lower accuracy when we increased the amount of C. According to the outcomes of our experiments, the average accuracy of polynomial function was 99.18%, which is significantly better than RBF and sigmoid with accuracies of 89.15% and 70.04%, respectively. Furthermore, the best accuracy for C-SVM with polynomial was 99.89% when C = 200. Thus we suggest the C-SVM algorithm with the polynomial function to the physicians and researchers to accelerate and improve the diagnosis of PD.

Table 7. Parameter validation with different C values in SVM algorithm on PD.

Regularization parameter (C)	Regression precision (epsilon)	Accuracy (%)		
		RBF	Polynomial	Sigmoid
C = 1	0.10	75.06	**97.42**	70.12
C = 1	0.15, 0.20, 0.25, 0.30, 1.00	75.34	**97.42**	70.12
C = 2	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	79.60	**98.32**	70.12
C = 3	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	83.58	**98.77**	70.07
C = 4	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	85.15	**98.93**	70.07
C = 5	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	86.27	**99.05**	70.07
C = 6	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	86.72	**99.22**	70.07
C = 7	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	87.72	**99.33**	70.07
C = 8	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	88.85	**99.33**	70.07
C = 9	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	89.24	**99.33**	70.07
C = 10	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	89.69	**99.44**	70.07
C = 15	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	91.65	**99.50**	70.01
C = 20	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	92.71	**99.50**	70.01
C = 25	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	93.27	**99.50**	70.01
C = 30	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	94.00	**99.50**	70.01
C = 35	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	94.45	**99.66**	70.01
C = 40	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	94.79	**99.66**	70.01
C = 45	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	94.90	**99.66**	70.01
C = 50	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	95.12	**99.66**	70.01
C = 100	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	96.58	**99.83**	70.01
C = 200	0.10, 0.15, 0.20, 0.25, 0.30, 1.00	97.48	**99.89**	70.01

References

[1] Parkinson's Disease Foundation, http://www.pdf.org/en/about_pd, (2011.09.16).

[2] Parkinson, J. (2002). An essay on the shaking palsy. The Journal of neuropsychiatry and clinical neurosciences, vol. 14, no.2, pp. 223-236.

[3] Van Den Eeden, S. K., et al. (2003). Incidence of Parkinson's disease: variation by age, gender, and race/ethnicity. American journal of epidemiology, vol. 157, no. 11, pp. 1015-1022.

[4] Little, M. A., et al. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE transactions on biomedical engineering, vol. 56, no. 4, pp. 1015-1022.

[5] De Rijk, M. C. (2000). Prevalence of Parkinson's disease in Europe: A collaborative study of. Neurology, vol. 54, no. 5, s214323.

[6] Mayeux, R., et al. (1992). A population-based investigation of Parkinson's disease with and without dementia: relationship to age and gender. Archives of Neurology, vol. 49, no. 5, pp. 492-497.

[7] Picillo, M., et al. (2017). The relevance of gender in Parkinson's disease: a review. Journal of neurology, pp. 1-25.

[8] Georgiev, D., et al. (2017). Gender differences in Parkinson's disease: A clinical perspective. Acta Neurologica Scandinavica, pp. 1-15.

[9] Lavrač, N. (1999). Selected techniques for data mining in medicine. Artificial intelligence in medicine, vol. 16, no. 1, pp. 3-23.

[10] McBride, J. C., et al. (2015). Sugihara causality analysis of scalp EEG for detection of early Alzheimer's disease. NeuroImage: Clinical, vol. 7, pp. 258-265.

[11] Abdar, M., et al. (2017). Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, vol. 67, pp. 239-251.

[12] Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. Expert systems with Applications, vol. 36, no. 2, pp. 3465-3469.

[13] Lee, W., & Stolfo, S. J. (1998). Data Mining Approaches for Intrusion Detection. In Usenix security.

[14] Thuraisingham, B. M. (2006). Data mining for security applications. In Intelligence and security informatics (pp. 1-3). Springer Berlin Heidelberg.

[15] Ling, C. X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions. In KDD, Vol. 98, pp. 73-79.

[16] Suman, M., Anuradha, T., & Veena, K. M. (2011). Direct marketing with the application of data mining. Journal of Information Engineering and Applications, vol. 1, no. 6, pp. 1-4.

[17] Eesa, A. S., Orman, Z., & Brifcani, A. M. A. (2015). A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Systems with Applications, vol. 42, no. 5, pp. 2670-2679.

[18] Pan, S., Morris, T., & Adhikari, U. (2015). Developing a hybrid intrusion detection system using data mining for power systems. IEEE Transactions on Smart Grid, vol. 6, no. 6, pp. 3104-3113.

[19] Cheng, J. C., & Ma, L. J. (2015). A data-driven study of important climate factors on the achievement of LEED-EB credits. Building and environment, vol. 90, pp. 232-244.

[20] Perera, D., et al. (2009). Clustering and sequential pattern mining of online collaborative learning data. IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 6, pp. 759-772.

[21] Shoorabi Sani, S. (2016). A case study for application of fuzzy inference and data mining in structural health monitoring. Journal of AI and Data Mining, vol. 6, no. 1, pp. 105-120.

[22] Sajid, U. K. (2015). Classification of Parkinson's Disease Using Data Mining Techniques. Journal of Parkinson's disease & Alzheimer's disease, vol. 2, pp. 1-4.

[23] Ganesh, H. & Annamary, G. (2014). Comparative study of Data Mining Approaches for Parkinson's Diseases. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 3, pp. 3062-3068.

[24] Shaikh, T. A. (2014). A Prototype of Parkinson's and Primary Tumor Diseases Prediction Using Data Mining Techniques. International Journal of Engineering Science Invention, vol. 3, no. 9, pp. 23-28.

[25] Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications, vol. 37, no. 2, pp. 1568-1572.

[26] Cortes C. & Vapnik, V. (1995). Support-vector networks. Machine learning, vol. 20, no. 3, pp. 273-297.

[27] Bennett, K. P., & Campbell, C. (2000). Support vector machines: hype or hallelujah?. ACM SIGKDD Explorations Newsletter, vol. 2, no. 2, pp. 1-13.

[28] Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

[29] Jensen, F. V. (1996). An introduction to Bayesian networks (Vol. 210). London: UCL press.

[30] Nielsen, T. D., & Jensen, F. V. (2009). Bayesian networks and decision graphs. Springer Science & Business Media.

[31] Cowell, R. G. (Ed.). (2006). Probabilistic networks and expert systems: Exact computational methods for Bayesian networks. Springer Science & Business Media.

[32] Athanasios T, Max L. Parkinsons Telemonitoring Data Set. https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring, [Accessed on 12 September 2015].

[33] Abdar, M. (2015). Using Decision Trees in Data Mining for Predicting Factors Influencing of Heart Disease. Carpathian Journal of Electronic and Computer Engineering, vol. 8, no. 2, pp. 28-33.

[34] Weng, C. H., Huang, T. C. K., & R. P. Han, (2016). Disease prediction with different types of neural network classifiers. Telematics and Informatics, vol. 33, no. 2, pp. 277-292.

[35] Abdar, M., et al. (2015). Data Mining on the Heart Disease with the Use of Different Algorithms. International Journal of Electrical and Computer Engineering, vol. 5, no. 6, pp. 1569-1576.

نشریه هوش مصنوعی و داده کاوی

تأثیر جنسیت بیماران در بیماری پارکینسون با استفاده از الگوریتم‌های دسته‌بندی

مولود آبدار[1] و مریم زمردی مقدم[2،*]

[1] گروه علوم و مهندسی کامپیوتر، دانشگاه آیزو، آیزو-واکاماتسو، ژاپن.

[2] گروه مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، ایران.

ارسال ۲۰۱۶/۰۸/۲۴؛ بازنگری ۲۰۱۷/۰۷/۱۹؛ پذیرش ۲۰۱۷/۰۸/۳۰

چکیده:

در این تحقیق، دقت دو الگوریتم مهم برای تشخیص بیماری پارکینسون شامل الگوریتم‌های شبکه‌های بیزین و SVM از الگوریتم‌های یادگیری ماشین مورد بررسی قرار گرفت. ما از داده‌های پارکینسون موجود در مخزن داده دانشگاه کالیفرنیا، ایروین (UCI)، استفاده کردیم. برای بهینه‌سازی الگوریتم SVM توابع کرنل مختلف و پارامترهای C مختلفی استفاده شدند و نتایج بدست آمده نشان داد که SVM با پارامتر C با متوسط دقت ۹۹,۱۸٪ و با داشتن تابع کرنل چندجمله ای در مرحله تست بازدهی بهتری نسبت به توابع کرنل دیگر مانند RBF و sigmoid و همچنین الگوریتم شبکه بیزین را دارا می‌باشد. همچنین نشان داده شد که ده عامل مهم در الگوریتم SVM شامل Jitter(Abs)،، #Subject، RPDE، PPE، Age، NHR، Shimmer APQ 11، Total-UPDRS، Shimmer (Db) و Shimmer می‌باشد. ما همچنین اثبات کردیم که دقت روش‌های C-SVM و RBF پیشنهادی ما نسبت مستقیم با مقدار پارامتر C داشت. به طوری که با افزایش مقدار C، دقت در هر دو تابع کرنل افزایش یافت. هرچند بر خلاف توابع چند جمله ای و RBF، تابع sigmoid رابطه‌ی معکوس با اندازه ی C دارد. قطعاً با استفاده از این روش‌ها، ما می‌توانیم مؤثرترین عوامل رایج در هر دو جنسیت (مؤنث و مذکر) را پیدا کنیم. تا آنجا که می‌دانیم، مطالعه ای بر روی بیماری پارکینسون برای مشخص کردن مؤثرترین عوامل در هر دو جنسیت صورت نگرفته است.

کلمات کلیدی: داده‌کاوی، بیماری پارکینسون، الگوریتم SVM، الگوریتم شبکه‌ی بیزین، الگوریتم C-SVM.