Neurology, Psychiatry and Brain Research 32 (2019) 77–84

Contents lists available at ScienceDirect

Neurology, Psychiatry and Brain Research

journal homepage: www.elsevier.com/locate/npbr

Dealing with mixed data types in the obsessive-compulsive disorder using ensemble classification

Hesam Hasanpour[a], Ramak Ghavamizadeh Meibodi[a], Keivan Navi[a], Sareh Asadi[b,*]

[a] Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran
[b] Neuroscience Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Keywords:
Ensemble classification
Fluvoxamine
Mixed data types
Predictor
Treatment response

ABSTRACT

Objective: Obsessive-compulsive disorder (OCD) is a psychiatric disorder characterized by recurrent obsessions and/or compulsions. Applying classification algorithms for prediction of treatment response helps to individualize treatment with more effectiveness. OCD data set is heterogeneous including continuous and discrete variables which presents challenges for most of the traditional classifiers to avoid data over-fitting. Here, we aimed to develop an ensemble classifier which is suitable for mixed data types for prediction of treatment response in OCD.

Methods: One hundred fifty-one subjects with OCD aged between 18–65 underwent fluvoxamine pharmacotherapy for 12 weeks and categorized into two groups (responder, non-responder) based on the reduction in their symptom severity following treatment. Decision tree and support vector machines (SVM-tree) were combined to deal with discrete and continuous variables and were used as base classifiers to build an ensemble of classifiers.

Results: Some of the attributes such as sexual obsessions and occupation, factor 1 (aggressive, contamination, sexual, religious, symmetry obsessions), initial obsession score, age at onset and illness duration are the high ranked predictors of treatment response. Comparing accuracy, precision, sensitivity, specificity and f-measure of the new algorithm with traditional classification algorithms such as decision tree, support vector machines (SVM), k-nearest neighbor and random forest showed a stronger performance of the proposed algorithm in the prediction of OCD treatment response.

Conclusion: The proposed strategy introduced an effective classification method to deal with medical datasets with mixed data types which can be of great significance in medical datasets and personalized medicine.

1. Introduction

Obsessive-compulsive disorder (OCD) is a heterogeneous neuropsychiatric condition that affects 1–3% of the population worldwide (Hasler et al., 2005). Selective serotonin reuptake inhibitors (SSRIs) are considered effective and well-established pharmacotherapy for the treatment of this disorder (Dougherty, Rauch, & Jenike, 2004). A significant number of OCD patients (40%–60%) fail to benefit sufficiently from pharmacotherapy (Alarcon, Libb, & Spitler, 1993; Ravizza, Barzega, Bellino, Bogetto, & Maina, 1995). Inability to personalize pharmacotherapy is an important factor that reduces treatment effectiveness (Brandl, Müller, & Richter, 2012), so identification of essential factors in treatment response could be valuable in enhancing clinical outcome. The current study aimed to investigate a new ensemble method to predict the result of OCD pharmacotherapy with fluvoxamine using an OCD dataset from Iranian OCD patients with mixed data types.

The personalized medicine research is entering the era of data mining. The vast amount of data in this field brings both opportunities and new challenges for analysis. Using machine learning and data mining methods researchers are able to examine all of the potential predictors simultaneously. Several machine learning methods including support vector machine (SVM), support vector regression and random forest (RF) were used in clinical research especially in OCD to predict disease severity (Askland et al., 2015; Hoexter et al., 2013).

Medical datasets including OCD patients' dataset usually comprise a mix of continuous and nominal features. In the case of OCD, some features such as age at assessment, age of onset, obsessive and compulsive severities are continuous variables, while features such as sex, marital status, and presence of family history are nominal; so OCD data

* Corresponding author at: Neuroscience Research Center, Shahid Beheshti University of Medical Sciences, P.O. Box 19615-1178, Tehran, Iran.
E-mail address: s.asadi@sbmu.ac.ir (S. Asadi).

https://doi.org/10.1016/j.npbr.2019.04.004
Received 18 December 2018; Received in revised form 8 April 2019; Accepted 25 April 2019
Available online 09 May 2019
0941-9500/

set is a mixed-attribute data set.

Typically, for analysis of mixed-attribute datasets, numerical attributes are transformed resulted in a homogenous categorical data set. There are techniques for converting one type of variable into another such as discretization (Eggermont, Kok, & Kosters, 2004; Ong, Huang, & Tzeng, 2005), although they may introduce noise or lose information (Jabeen & Baig, 2012; Zhang & Jin, 2010). Moreover, the selection of an appropriate discretization algorithm across various methods is a dilemma (Rezvan, Hamadani, & Shalbafzadeh, 2013). There is not a routine predictive model for handling both types of input at the same time without transforming the feature types.

There are reports of investigations considered the problem of mixed data types. Nouaouria and Boukadoum (2014) described a particle swarm optimization (PSO) approach for the challenge of classifying mixed-attribute data sets. They used a novel particle-position update mechanism in their method and a new way for handling mixed-attribute data based on the particle position interpretation. Rezvan et al. (2013) introduced two new case-based reasoning systems with two similarity measures that support mixed categorical and numerical data. They applied their proposed system on Irvine (UCI) data set resulted in good accuracy and interpretability. In another study, an approach based on two-layered genetic programming was presented to classify hybrid datasets. Jabeen and Baig (2012) The proposed method combines the properties of arithmetic expressions (using numerical data) and logical expressions (using categorical data) without data transformation. Hu, Yu, Liu, and Wu (2008) introduced a neighborhood rough set model to deal with heterogeneous data by generalization classical rough set model with neighborhood relations. In another study, Bouguessa et al. proposed a principled approach based on the bivariate beta mixture model to identify outliers in hybrid data. The proposed approach is able to discriminate outliers from inliers automatically and it can be applied to both heterogeneous and homogeneous data without any feature transformation (Bouguessa, 2015). A comprehensive study by Aggarwal demonstrated the different methods for handling mixed data types in outlier detection problem (Aggarwal, 2017). According to his investigation constructing an appropriate distance function is a significant challenge in applying techniques such as nearest neighbor and density base classification. He also stated that although some methods such as probabilistic models have advantages, they also have their limits.

Many studies investigated the possible association of OCD clinical characteristics and treatment response. Factors such as hoarding dimension(Mataix-Cols, Rauch, Manzo, Jenike, & Baer, 1999; Salomoni et al., 2009; Saxena & Maidment, 2004; Stein, Andersen, & Overo, 2007, 2008), contamination and cleaning (Ravizza, Barzega, Bellino, Bogetto, & Maina, 1995; Stein et al., 2007), sexual/religious obsessions, (Alonso et al., 2001; Mataix-Cols, Marks, Greist, Kobak, & Baer, 2002) and poor insight (Erzegovesi et al., 2001; Kishore, Samar, Reddy, Chandrasekhar, & Thennarasu, 2004) have been associated with poor response to OCD treatment. having a partner (Marcks, Weisberg, Dyck, & Keller, 2011; Shavitt et al., 2006) and checking compulsions (Landeros-Weisenberger et al., 2010; Stein et al., 2007) are some predictors with good response to SRI treatment. These studies show that attributes have a different effect on the prediction of treatment response. Selecting appropriate features can enhance the accuracy of classification algorithms.

The aim of the present study is to introduce a new ensemble classification method based on decision tree and SVM that have a good performance on mixed data sets. To our knowledge, ensemble classification for mixed data-type data sets has not yet been applied to predict treatment outcomes in OCD.

2. Methods

2.1. Subjects and treatment procedure

Three hundred and thirty outpatients with Iranian origin meeting DSM-IV criteria for OCD were recruited between 2014 and 2017 from Imam Hossain hospital, Tehran, Iran. Subjects were interviewed by an experienced clinician and met the Diagnostic and Statistical Manual of Mental Disorders (DSM IV-TR) (Association, 2000) criteria for OCD on the Structured Clinical Interview for Axis I Disorders. The exclusion criteria were age under 18 or above 65 years old, having a history of psychotic disorders or mental retardation, reporting severe neurological pathology and history of substance use, diagnosis with other DSM-lV-TR Axis l disorders except for depression, anxiety or tic disorder. A drug-free period of at least 3 weeks was considered as an inclusion criterion for OCD patients.

The socio-demographic data was collected through a questionnaire consist of the patient's full name and address, age at assessment, the age of onset, gender, marital status, educational level, occupation, illness duration and the familial history of any psychiatric disorders specifically OCD. The Persian version of the YBOCS severity scale and checklist (Rajezi Esfahani, Motaghipour, Kamkari, Zahiredin, & Janbozorgi, 2012) were used to assess the severity and types of current obsession and compulsion symptoms.

Patients underwent 12 weeks of treatment with fluvoxamine (150 mg–300 mg). No concomitant pharmacological or behavioral therapy was allowed during the whole treatment period. The treatment procedure was described before (Hasanpour, Meibodi, Navi, & Asadi, 2018). Briefly, the escalating fluvoxamine daily dose was initiated from 25 mg/day, increased up to 100 mg/day in the 3rd week. Fluvoxamine daily dose for the next three weeks was 150 mg/day, and after the sixth week, patients were visited by the psychiatrist and received maximum tolerated dose of the fluvoxamine for the next six weeks. Totally, the patients were given 150 mg/day for 9 weeks. After this period reduction in YBOCS severity score was calculated according to the initial score. Based on this score patients were divided into two groups: group A (responders) was comprised of patients that exhibited > 35% reduction in Y-BOCS severity score after treatment with fluvoxamine; group B (non-responders) was comprised of patients that exhibited < 35% reduction in Y-BOCS scores (Pallanti & Quercioli, 2006). From 330 patients who entered the treatment procedure, only 151 subjects completed pharmacotherapy and their data were included in the subsequent analysis. Others were excluded because of refusal to come for follow-up (N = 108), not taking medication properly (N = 38), discontinuity in pharmacotherapy due to complications, including allergy to the drug (N = 8), meeting exclusion criteria during pharmacotherapy (N = 6) and other reasons (N = 19). The CONSORT diagram in Fig. 1 summarizes the flow of participants through different stages of the trial. The excluded and included patients showed no significant differences in mean baseline scores on the Y-BOCS severity scale and symptom dimensions from the Y-BOCS symptom checklist (data not shown, all $P > 0.1$).

This study was approved by the Research Ethics Committee of Shahid Beheshti University of Medical Sciences (IR.SBMU.PHNS.REC.1396.2).

2.2. Variables

Treatment response (response to treatment, resistant to treatment) was considered as a dichotomous dependent variable. Other variables were considered as predictors including sex, marital status, job status, educational status, ethnicity, initial Y-BOCS score for obsessions, initial Y-BOCS score for compulsions, initial total YBOCS score, Y-BOCS obsession subtypes, and Y-BOCS compulsion subtypes, insight, avoidance, depression and age of onset.

We also applied exploratory factor analysis on the reported scores of

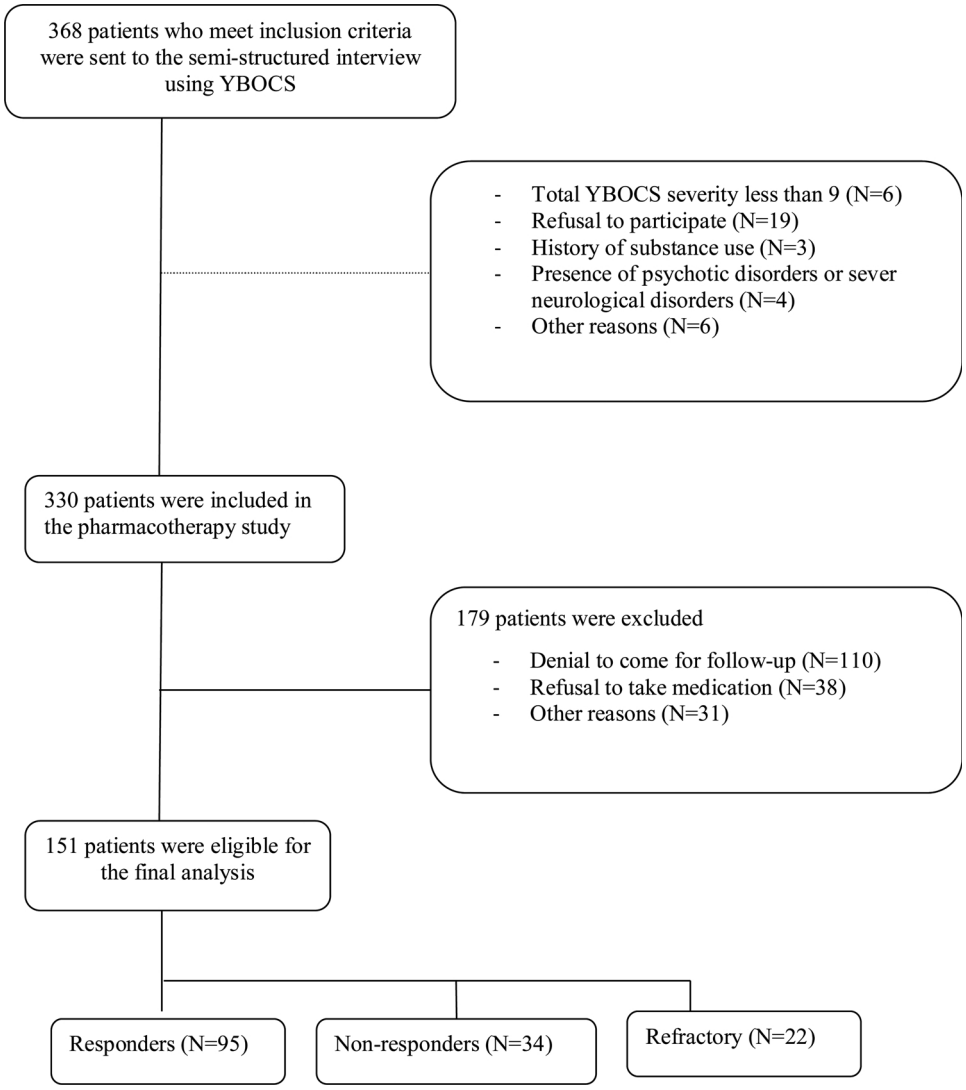

Fig. 1. The consort diagram of the study.
This figure presents a consort diagram of the study showing reasons for exclusion from the study.

Y-BOCS obsession and compulsion categories. Based on our previous work (Hasanpour et al., 2017), we chose four-factor as the appropriate number of factors. After the oblique rotation with Kaiser normalization, items on aggressive, contamination, sexual, religious, symmetry obsessions, as well as repeating and checking compulsions loaded highly on factor 1; symmetry obsession and cleaning, counting and ordering compulsions were considered as factor 2; contamination obsessions and cleaning compulsions were loaded on factor 3; and hoarding obsessions and hoarding compulsions, were considered as factor 4.

2.3. Proposed algorithm

We used ensemble learning for classification of our data set. Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem with better performance. There are a number of training parameters and factors which can be manipulated to create ensemble members: the initial condition, the training data, the architecture of the classifiers, and the training algorithm. In ensemble methods such as Bagging (Breiman, 1996), Boosting (Schapire, 1990) and Random forest (Breiman, 2001) multiple learning algorithms are used to achieve better predictive performance compared to each learning algorithm alone. Combining a set of imperfect classifiers is viewed as a way to enhance the overall recognition capability comparing to the individual classifiers.

In ensemble classification, some classifiers are trained. At the test step, samples are given to all the base classifiers, to determine the class label of each sample, typically through majority vote, based on the output of all base classifiers. Experiments showed that if the individual classifiers have accuracy greater than 0.5 the majority vote gives higher accuracy than a particular classifier (Lam & Suen, 1997; Wanas, 2003). One of the main effects of ensemble averaging is the reduction of the variance of a set of classifiers (Rajapakse & Wang, 2012). For ensemble classification, we need to create diversity in the output of base classifiers. In the case of the strong correlation of classifiers, the addition of base classifiers will not help much to increase the prediction accuracy.

In the current OCD dataset age at assessment, the age of onset, factor scores, obsessive and compulsive severities are continuous variables; sex, marital status, employment status, obsession and compulsion subtypes, the presence of family history and education status are nominal. To analyze this data set with mixed data types we used a method with two steps. At the first step, we removed redundant features and assigned a weight to other ones. To this aim, we assessed the relevance of features and class labels using chi-square test for nominal attributes and independent sample *t*-test for continuous ones. We calculated p-value and utilized 1-(p-value) as the merit of each feature. Finally, we assigned a weight to each feature based on its value. For this

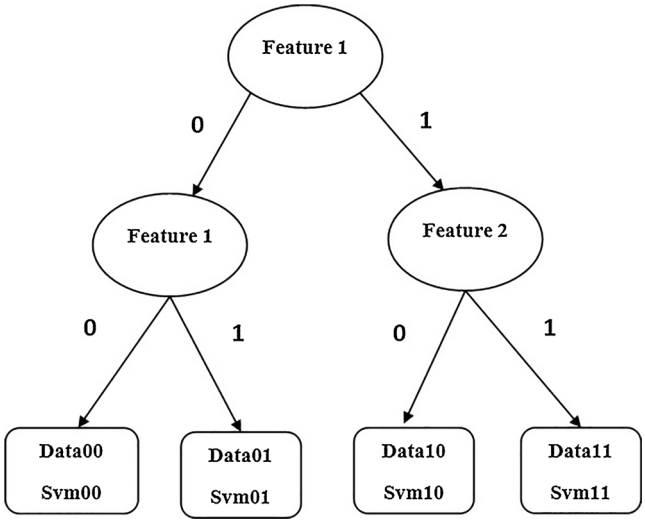

Fig. 2. A base classifier that combined decision tree and support vector machines.
For building this base classifier two nominal and a subset of continuous features are selected randomly using roulette-wheel algorithms. Data00 contains samples with continuous features that their values of feature1 and 2 are 0. SVM00 is a SVM model that built on Data00. Other SVM models including SVM01, SVM10, and SVM11 are built based on Data01, Data10, and Data11, respectively.

purpose, we normalized the merit of each feature by dividing to the sum of the merits of all features. The features whose weight were less than a predefined threshold were considered as redundant features and were removed from further analyses.

At the second step, we combined the decision tree and SVM as the base classifier (SVM-tree) to handle mixed data types. We used an ensemble of these base classifiers for prediction of treatment response in OCD patients. Fig. 2 and Table 1 show the structure of a base classifier and pseudo code of the proposed algorithm.

For constructing a base classifier two discrete features and a number of continuous variables are selected using the roulette-wheel algorithm (De Jong, 1975). Basically, The roulette-wheel algorithm is a genetic operator for selecting potentially useful solutions for recombination. We applied this algorithm on weighted features. According to its approach, candidate features with a higher weight will be more likely to be selected. Feature1 and Feature2 are two nominal features selected from all nominal features of the dataset. We also select a random subset of continuous features for building SVM models (variable S in the pseudo code). This technique is called attribute bagging that is used for improving the accuracy and stability of base classifiers (Bryll, Gutierrez-Osuna, & Quek, 2003; Turner & Oza, 1999). SVM_features variable refers to the S selected continuous features. SVM_data only contains samples with SVM_features. Based on four different values of Feature1 and Feature2, SVM_data is partitioned into four part data00, data02, data10, and data11. For example, data00 contains samples of SVM_data that their values of Feature1 and 2 are 0. SVM00 is an SVM model that built on Data00. According to this nomenclature, we had other SVM models including SVM01, SVM10, and SVM11 based on Data01, Data10, and Data11, respectively. It must be noted that since we used k-fold cross validation (k = 10) for model evaluation, the pseudo code run k times. The low sample size is the main reason for selecting only two nominal feature in building a base classifier. If we select more nominal features then the number of samples that devote to each SVM model is not sufficient for proper learning.

In the next step, we generated a number of base classifiers (SVM-tree). At the time of base classifiers construction, we pruned base classifiers that have accuracy less than a predefined threshold (60% in our study). We split the dataset into training and testing parts. The first part is used for the training the model and the second part is used for testing the model. The accuracy of the model on training subsamples is used for pruning. At the test time, each test samples is given to all base classifiers and its class label is determined using the majority vote. Suppose that we have 20 base classifiers and a two class dataset. The test sample was given to all 20 base classifiers. If the result of 15 base classifiers was class 1 and the other five base classifier opinions were class 2, then according to the majority vote, the test sample was assigned to class 1.

2.4. Performance evaluation

There are a lot of measures used for analyzing the performance of classification models. Accuracy is a primary measure for evaluating a model. The accuracy is the total number of correct classifications out of the total number of samples. This measure is not sufficient for evaluating a model with imbalanced distribution of the class. In this case, other measures such as precision, sensitivity (recall), specificity and F-measure can be used as an alternative (Han, Pei, & Kamber, 2011). Precision (also called positive predictive value) is the fraction of correct positive classifications among all samples that predicted to the positive class. The sensitivity is the proportion of correct positive classifications out of the number of true positives. The specificity is the proportion of correct negative classifications out of the number of true negatives. F-score or F-measure is the harmonic mean of precision and recall. Accuracy (Eq. (1)), precision (Eq. (2)), sensitivity (Eq. (3)), specificity (Eq. (4)) and F-measure (Eq. (5)) are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Sensitivity\,(Recall) = \frac{TP}{TP + FN} \quad (3)$$

$$specificity = \frac{TN}{FP + TN} \quad (4)$$

$$Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

TP (true positive) is the number of positive samples that are correctly identified as positive. FP (false positive) is the number of negative samples that are incorrectly identified as positive. FN (false negative) denotes the number of positive samples that are incorrectly identified as negative. TN (true negative) is the number of negative samples that are correctly identified as negative.

All data preprocessing and analyses were conducted using Matlab version 2014a.

3. Results

The studied sample composed of 151 patients with the following characteristics: more females (64%), more married (66%), more unemployed (58%) and more subjects without academic degrees (76%). Fourteen percent of patients reported no history of mental illness in their family but others (84%) were from the families with a history of psychiatric disorders. Patients' age of onset (mean ± SD) was 22.8 ± 10.4 and age at assessment (mean ± SD) was 34.3 ± 10.1. The mean of Y-BOCS scores at intake (mean ± SD) was 11.2 ± 4.7 for obsession, 9.5 ± 5.4 for compulsion and 20.8 ± 8.7 for the total. The descriptive statistics of all predictors are considered and showed in Tables 2 and 3. To assess the relevance of predictors to class label, independent sample *t*-test was implied for continuous predictors and chi-square test for dichotomous predictors.

As the values of P-values showed, in nominal features, sexual obsessions and occupation were high-ranked predictors for response to

Table 1
Pseudo-code for the proposed method.

```
Assess the relevance of features with class label and assign a weight to each feature
delete features that their weight is less than a predefined threshold (0.01 in our case)
i=1, k=number_of_base classifiers (20 in our case)
numeric_features=features with numeric characteristics
min_numeric_features=number of numeric features/2;
max_numeric_features=number of numeric features*3/4;
nominal_features=features with nominal characteristics
While i<=k
    Feature1=select one attribute from nominal_features using roulette wheel selection
    Feature2=select one attribute from nominal_features using roulette wheel selection
    S= create random number between min_numeric_features and max_numeric_features
    SVM_features= select S attributes from numeric_features using roulette wheel selection
    SVM_data= samples with SVM_features
    Split SVM_data to four part ( based on different values of feature1 and Feature 2) and assign them to Data00, Data01,
Data10 and Data11 ( for example Data00= samples from SVM_data that their Feature1 value is 0 and Feature2 value is 0)
    SVM00= SVM model that built on Data00
    SVM01= SVM model that built on Data01
    SVM10= SVM model that built on Data10
    SVM11= SVM model that built on Data11
    model=a base classifier consist of four SVM model (Figure 1)
    If accuracy(model)>threshold(%60)
        Retain the model
        i=i+1
End while
For each of the test samples %
    For i=1:k
    Select one of the four SVM001, SVM01,SVM10 and SVM11 based on the value of Feature1 and  Feature2 of test sample
        Result(i)=Predict the output using the selected model
    End for i
Assign a label to test sample using majority vote in Result array
End for
```

Table 2
Descriptive statistics of continuous predictors.

	Global Sample		Responder		Non-responder		P-value
predictors	Mean	Std	Mean	Std	Mean	Std	
age	34.3	10.1	34.4	10.6	34.1	9.36	0.85
Age at onset	22.8	10.4	24	10	20.8	10.8	0.06
Illness duration	11.4	10	10.3	9.8	13.2	10	0.08
Initial obsession score	11.2	4.7	10.6	4.7	12.3	4.5	**0.03**
Initial compulsion score	9.5	5.4	9.7	4.9	9.3	6.2	0.67
Initial total score	20.8	8.7	20.3	8.3	21.7	9.4	0.36
Factor1	0	0.86	−0.13	0.96	0.22	0.59	**0.01**
Factor2	0.001	0.81	−0.03	0.84	0.06	0.76	0.46
Factor3	−0.001	0.97	−0.01	0.98	0.02	0.97	0.82
Factor4	−0.002	0.97	0.07	0.95	−0.12	1.01	0.24

treatment while features such as cleaning, checking and hoarding had weak associations with response to treatment; so they excluded from further analysis. In continuous features factor 1 (contamination, sexual, religious) and Initial obsession score, Age at onset and illness duration had a strong association with treatment response but the age at assessment and factor 3 (contamination and cleaning) had weak associations with it.

Table 4 summarizes the average performance of different classification algorithms (multi-layer perceptron (MLP), k-Nearest Neighbor (KNN), support vector machines (SVM), decision tree and random forest (RF) and proposed method) applied on our data set. We used ten-fold stratified cross-validation method for evaluating predictive models which is the best estimation technique, especially in datasets with small sample size. (Kohavi, 1995; Seni & Elder, 2010) It must be noted that our dataset contains 151 patients that belong to two class. 95 patients responded adequately to treatment (positive or responder class) and 56 patient exhibit inadequate responses (negative or non-responder class). The values in the parenthesis are the confidence interval of that measure at a 95% confidence level. Results showed that the MLP algorithm was the weakest classifier for this application with 52% accuracy, 65%

Table 3
Descriptive Statistics of discrete Predictors.

Predictors		Responder	Non-responder	P-value
Sex	Female	65	32	0.16
	Male	30	24	
Marital status	Single	31	20	0.69
	Married	64	36	
Occupation	Unemployed	61	27	0.05
	Employed	34	29	
Educational level	Low	75	40	0.29
	High	20	26	
Family history	Negative	14	10	0.61
	Positive	81	46	
insight	Lack	37	16	0.19
	Presence	58	40	
avoidance	Lack	48	27	0.78
	Presence	47	29	
Contamination obsession	Lack	29	14	0.46
	Presence	66	42	
Sexual obsession	Lack	72	35	0.08
	Presence	23	21	
Aggressive obsession	Lack	20	13	0.75
	Presence	75	43	
Hoarding obsession	Lack	58	37	0.53
	Presence	37	19	
Religious obsession	Lack	37	18	0.4
	Presence	58	38	
Symmetry obsession	Lack	38	29	0.16
	Presence	57	27	
Mis obsession	Lack	11	8	0.62
	Presence	88	48	
Somatic obsession	Lack	32	16	0.51
	Presence	63	40	
Cleaning compulsion	Lack	36	20	0.78
	Presence	59	36	
Checking compulsion	Lack	32	18	0.84
	Presence	63	38	
Repeating compulsion	Lack	47	21	0.15
	Presence	48	35	
Counting compulsion	Lack	69	43	0.57
	Presence	26	13	
Ordering compulsion	Lack	42	30	0.26
	Presence	53	26	
Hoarding compulsion	Lack	68	40	0.98
	Presence	27	16	
Mis compulsion	Lack	26	18	0.53
	Presence	69	38	

Table 5
Confusion matrix of different algorithms.

Algorithm			Actual class		
			Responder	Non-responder	Total
MLP	Prediction class	Responder	50	27	77
		Non-responder	45	29	74
KNN	Prediction class	Responder	75	41	116
		Non-responder	20	15	35
Decision tree	Prediction class	Responder	91	43	134
		Non-responder	4	13	17
SVM	Prediction class	Responder	73	42	115
		Non-responder	22	14	36
Random forest	Prediction class	Responder	80	44	124
		Non-responder	15	12	27
Proposed method	Prediction class	Responder	73	23	96
		Non-responder	22	33	55

precision, 53% sensitivity, 52% specificity and 58% f-measure. In addition, nearly all of the algorithms except the proposed method are weak at the prediction of non-responder patients. Accuracy, precision, sensitivity, specificity, and f-measure of the proposed method were 70%, 76%, 77%, 59%, and 76%, respectively, which were higher than other classification algorithms.

Table 5 shows the confusion matrix for different algorithms. The confusion matrix is a table with two rows and two columns that reports the number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). As results show, some algorithms such as decision tree are good at the prediction of responder class but weak at non-responder class. The value of TP (true positive) for decision tree is 91. It means that decision tree algorithms correctly assign 91 out of 95 patients to responder class. This algorithm incorrectly assigned 43 of non-responder patients to responder class (FP or false positive). The value of FN (false negative) and TN (true negative) for decision tree algorithm is 4 and 13, respectively. In other word, four patients of responder class incorrectly assign to non-responder class and 13 patients of non-responder class correctly assign to non-responder class. As a result of the decision tree shows, this algorithm assigns 134 patients to responder class and the remaining ones (17 patients) to non-responder class. The values of TP, FP, FN and TN for Our proposed method are 73, 23, 22 and 33 respectively. Proposed algorithm classified 73 out of 95 patient to responder class and 33 of 55 patient to non-responder class.

4. Discussion

The main aim of the current study was introducing a classification model for the analysis of datasets with mixed data types. We combined the decision tree and SVM (SVM-tree) to handle both discrete and continuous data. We used a number of SVM-trees as base classifiers and built an ensemble of base classifiers. For constructing a base classifier, we divided features into two categories: nominal and continuous features. For each category, we sort the features based on the association

Table 4
Accuracy, sensitivity and specificity of different classification algorithms applied on the current OCD data set based on 20 repetitions of 10-fold cross-validation.

	Accuracy % (Upper-Lower)	Precision % (Upper-Lower)	Sensitivity % (Upper-Lower)	Specificity % (Upper-Lower)	F-measure % (Upper-Lower)
MLP	52 (44-59.8)	65 (57-72)	53 (45-60)	52 (44-59.8)	58 (50-65)
Decision tree	69 (61.2-75.8)	68 (60-74.9)	96 (91.5-98)	23 (17-30)	79 (71.8-84.7)
KNN	60 (52-67.4)	65 (57-72)	79 (71.8-84.7)	28 (21.4-35.6)	71 (63.3-77.6)
SVM	58 (50-65.5)	63 (55-70)	77 (69.6-82.9)	25 (18.7-32.4)	69 (61.2-75.8)
Random Forest	61 (53-68.4)	64(56-71.2)	84 (77.3-88.9)	21 (15.2-28.1)	73 (65.7-79.4)
Proposed method	70 (62.2-76.7)	76 (68.6-82.1)	77 (69.6-82.9)	59 (51-66.5)	76 (68.6-82.1)

of features to the treatment response and excluded less associated features from future analysis. As a result, the proposed classification model has the ability to handle both continuous and nominal data types without any transformation and is based on ensemble learning. Results showed that some attributes such as sexual obsessions, occupation, factor 1 (contamination, sexual, religious), Initial obsession score, Age at onset and illness duration are the most essential features for prediction of treatment response, and some attributes such as age at assessment, cleaning, checking, hoarding and factor 3 (contamination and cleaning) are the least important predictors for response to treatment in OCD patients.

We used an ensemble of classifiers which has been shown that this strategy can improve class prediction even though individual classifiers may be rather weak and error-prone in making decisions (Ahn et al., 2006; Hastie, Tibshirani, & Friedman, 2002; Rokach, 2010). Successful applications of ensemble methods were reported in bioinformatics (Tan, Gilbert, & Deville, 2003) and medicine (Mangiameli, West, & Rampal, 2004; Özçift, 2011). The outcomes of applying the proposed method on the current data set showed it can predict fluvoxamine pharmacotherapy in OCD patients stronger than the other classification methods.

Many studies have been published investigating the role of genetic factors in response to SSRIs pharmacotherapy in OCD (Brandl et al., 2012; Qin et al., 2016; Ritter & Qin, 2018), but there are other studies investigating the importance of clinical characteristics in OCD treatment response. Bloch et al discussed the association of hoarding symptoms with treatment response in OCD (Bloch et al., 2014). In another study, clinical predictors of worse prognosis were defined as the earlier age at onset, longer duration of illness, presence of at least one comorbid psychiatric disorder, comorbidity with a mood disorder, higher baseline Beck-Depression scores, positive family history of tics, and positive family history of anxiety disorders (Jakubovski et al., 2013). Other predictors of poor treatment response were defined as living without a partner (Marcks et al., 2011), higher OCD severity at intake (Catapano et al., 2006; Fineberg et al., 2013), sexual subtype (Mataix-Cols et al., 2002), poor insight into obsessions, symmetry/hoarding and contamination/washing dimension and the presence of specific personality disorders (Hazari, Narayanaswamy, & Arumugham, 2016).

One of the main limitations of the current study which may affect the accuracy of the proposed method is the sample size. Since we build four SVM model in each base classifier, the number of samples that devote to each SVM model may be not adequate for proper training. Therefore bigger sample size may improve the efficiency of the proposed method. In conclusion, a new ensemble of SVM-tree classifiers was proposed for the analysis of mixed feature variables of OCD dataset. This algorithm allows dealing with heterogeneous datasets, without any transformation with good accuracy, sensitivity, and specificity. The proposed algorithm could be applied to treat similar datasets, especially in medicine with clinical importance.

Funding

This research was supported by grants Shahid Beheshti University of Medical Sciences, Iran, grant No 14553-6.

Availability of data and material

The dataset that was used during the current study will be available from the corresponding author on reasonable request.

Ethics approval and consent to participate

This research was performed in accordance with the latest version of Declaration of Helsinki and with the approval of the research ethics committee of "Neuroscience Research Center, Shahid Beheshti University of Medical Sciences" (IR.SBMU.PHNS.REC.1397.2). Written consents for the participate statement were obtained from participants after being informed of the nature of the research.

Disclosure statement

The authors declare that they have nothing to disclose.

Author contributions

HH contributed in conception and design of the study, literature search, acquisition and analysis of data, and drafting the manuscript; R G M contributed in conception and design of the study, manuscript editing, and review. K N in conception and design of the study, manuscript editing and review. S A contributed in conception and design of the study, literature search, acquisition of clinical data, and manuscript editing and review.

Acknowledgments

This research was supported by a grant from Neuroscience Research Center, Shahid Beheshti University of Medical Sciences, Iran. We are also grateful to Imam Hossain staff for their cooperation with our research team.

References

Aggarwal, C. C. (2017). *An introduction to outlier analysis. Outlier analysis*. Springer1–34.
Ahn, H., Moon, H., Fazzari, M. J., Lim, N., Chen, J. J., & Kodell, R. L. (2006). Classification by ensembles from random partitions. *Journal of Computational Statistics and Data Analysis, 2007*, 51.
Alarcon, R. D., Libb, J. W., & Spitler, D. (1993). A predictive study of obsessive-compulsive disorder response to clomipramine. *Journal of clinical psychopharmacology, 13*(3), 210–213.
Alonso, P., Menchon, J. M., Pifarre, J., Mataix-Cols, D., Torres, L., Salgado, P., & Vallejo, J. (2001). Long-term follow-up and predictors of clinical outcome in obsessive-compulsive patients treated with serotonin reuptake inhibitors and behavioral therapy. *The Journal of Clinical Psychiatry, 62*(7), 535–540.
Askland, K. D., Garnaat, S., Sibrava, N. J., Boisseau, C. L., Strong, D., Mancebo, M., ... Eisen, J. (2015). Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy. *International Journal of Methods in Psychiatric Research, 24*(2) 156–16.
Association, A. P. (2000). *Diagnostic and statistical manual of mental disorders-IV-TR*. Washington, DC: American Psychiatric Association.
Bloch, M., Bartley, C., Zipperer, L., Jakubovski, E., Landeros-Weisenberger, A., Pittenger, C., ... Leckman, J. (2014). Meta-analysis: Hoarding symptoms associated with poor treatment outcome in obsessive–compulsive disorder. *Molecular Psychiatry, 19*, 1025–1030.
Bouguessa, M. (2015). A practical outlier detection approach for mixed-attribute data. *Expert Systems with Applications, 42*(22), 8637–8649.
Brandl, E. J., Müller, D. J., & Richter, M. A. (2012). Pharmacogenetics of obsessive-compulsive disorders. *Pharmacogenomics, 13*(1), 71–81.
Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Bryll, R., Gutierrez-Osuna, R., & Quek, F. (2003). Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognition, 36*(6), 1291–1302.
Catapano, F., Perris, F., Masella, M., Rossano, F., Cigliano, M., Magliano, L., ... Maj, M. (2006). Obsessive–compulsive disorder: A 3-year prospective follow-up study of patients treated with serotonin reuptake inhibitors: OCD follow-up study. *Journal of Psychiatric Research, 40*(6), 502–510.
De Jong, K. A. (1975). *Analysis of the behavior of a class of genetic adaptive systems*.
Dougherty, D. D., Rauch, S. L., & Jenike, M. A. (2004). Pharmacotherapy for obsessive-compulsive disorder. *Journal of Clinical Psychology, 60*(11), 1195–1202.
Eggermont, J., Kok, J. N., & Kosters, W. A. (2004). *Genetic programming for data classification: Partitioning the search space. Paper presented at the proceedings of the 2004 ACM symposium on applied computing*.
Erzegovesi, S., Cavallini, M. C., Cavedini, P., Diaferia, G., Locatelli, M., & Bellodi, L. (2001). Clinical predictors of drug response in obsessive-compulsive disorder. *Journal of Clinical Psychopharmacology, 21*(5), 488–492.
Fineberg, N. A., Hengartner, M. P., Bergbaum, C., Gale, T., Rössler, W., & Angst, J. (2013). Remission of obsessive-compulsive disorders and syndromes; evidence from a prospective community cohort study over 30 years. *International Journal of Psychiatry in Clinical Practice, 17*(3), 179–187.
Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3th ed.). Elsevier.
Hasanpour, H., Asadi, S., Meibodi, R. G., Daraeian, A., Ahmadiani, A., Shams, J., ... Navi,

K. (2017). A critical appraisal of heterogeneity in Obsessive-Compulsive Disorder using symptom-based clustering analysis. *Asian Journal of Psychiatry, 28*, 89–96.
Hasanpour, H., Meibodi, R. G., Navi, K., & Asadi, S. (2018). Novel ensemble method for the prediction of response to fluvoxamine treatment of obsessive–compulsive disorder. *Neuropsychiatric Disease and Treatment, 14*, 2027.
Hasler, G., LaSalle-Ricci, V. H., Ronquillo, J. G., Crawley, S. A., Cochran, L. W., Kazuba, D., ... Murphy, D. L. (2005). Obsessive–compulsive disorder symptom dimensions show specific relationships to psychiatric comorbidity. *Psychiatry Research, 135*(2), 121–132.
Hastie, T., Tibshirani, R., & Friedman, J. (2002). The elements of statistical learning: Data mining, inference, and prediction. *Biometrics, 5*, 453–480.
Hazari, N., Narayanaswamy, J. C., & Arumugham, S. S. (2016). Predictors of response to serotonin reuptake inhibitors in obsessive-compulsive disorder. *Expert Review of Neurotherapeutics, 16*(10), 1175–1191.
Hoexter, M. Q., Miguel, E. C., Diniz, J. B., Shavitt, R. G., Busatto, G. F., & Sato, J. R. (2013). Predicting obsessive–compulsive disorder severity combining neuroimaging and machine learning methods. *Journal of Affective Disorders, 150*(3), 1213–1216.
Hu, Q., Yu, D., Liu, J., & Wu, C. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Information Sciences, 178*(18), 3577–3594.
Jabeen, H., & Baig, A. R. (2012). Two layered genetic programming for mixed-attribute data classification. *Applied Soft Computing, 12*(1), 416–422.
Jakubovski, E., Diniz, J. B., Valerio, C., Fossaluza, V., Belotto-Silva, C., Gorenstein, C., ... Shavitt, R. G. (2013). Clinical predictors of long-term outcome in obsessive-compulsive disorder. *Depression and Anxiety, 30*(8), 763–772.
Kishore, V. R., Samar, R., Reddy, Y. J., Chandrasekhar, C., & Thennarasu, K. (2004). Clinical characteristics and treatment response in poor and good insight obsessive–compulsive disorder. *European Psychiatry, 19*(4), 202–208.
Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at the Ijcai*.
Lam, L., & Suen, S. (1997). Application of majority voting to pattern recognition: An analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 27*(5), 553–568.
Landeros-Weisenberger, A., Bloch, M. H., Kelmendi, B., Wegner, R., Nudel, J., Dombrowski, P., ... Leckman, J. F. (2010). Dimensional predictors of response to SRI pharmacotherapy in obsessive–compulsive disorder. *Journal of Affective Disorders, 121*(1), 175–179.
Mangiameli, P., West, D., & Rampal, R. (2004). Model selection for medical diagnosis decision support systems. *Decision Support Systems, 36*(3), 247–259.
Marcks, B. A., Weisberg, R. B., Dyck, I., & Keller, M. B. (2011). Longitudinal course of obsessive-compulsive disorder in patients with anxiety disorders: A 15-year prospective follow-up study. *Comprehensive Psychiatry, 52*(6), 670–677.
Mataix-Cols, D., Marks, I. M., Greist, J. H., Kobak, K. A., & Baer, L. (2002). Obsessive-compulsive symptom dimensions as predictors of compliance with and response to behaviour therapy: Results from a controlled trial. *Psychotherapy and Psychosomatics, 71*(5), 255–262.
Mataix-Cols, D., Rauch, S. L., Manzo, P. A., Jenike, M. A., & Baer, L. (1999). Use of factor-analyzed symptom dimensions to predict outcome with serotonin reuptake inhibitors and placebo in the treatment of obsessive-compulsive disorder. *American Journal of Psychiatry, 156*(9), 1409–1416.
Nouaouria, N., & Boukadoum, M. (2014). Improved global-best particle swarm optimization algorithm with mixed-attribute data classification capability. *Applied Soft Computing, 21*, 554–567.
Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications, 29*(1), 41–47.
Özçift, A. (2011). Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Computers in Biology and Medicine, 41*(5), 265–271.
Pallanti, S., & Quercioli, L. (2006). Treatment-refractory obsessive-compulsive disorder: Methodological issues, operational definitions and therapeutic lines. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 30*(3), 400–412.
Qin, H., Samuels, J., Wang, Y., Zhu, Y., Grados, M., Riddle, M. A., ... McCracken, J. (2016). Whole-genome association analysis of treatment response in obsessive-compulsive disorder. *Molecular Psychiatry, 21*(2), 270.
Rajapakse, J. C., & Wang, L. (2012). *Neural information processing: Research and development, vol. 152*. Springer.
Rajezi Esfahani, S., Motaghipour, Y., Kamkari, K., Zahiredin, A., & Janbozorgi, M. (2012). Reliability and validity of the Persian version of the Yale-Brown Obsessive-Compulsive scale (Y-BOCS). *Iranian Journal of Psychiatry and Clinical Psychology, 17*(4), 297–303.
Ravizza, L., Barzega, G., Bellino, S., Bogetto, F., & Maina, G. (1995a). Predictors of drug treatment response in obsessive-compulsive disorder. *The Journal of Clinical Psychiatry, 56*(8), 368–373.
Ravizza, L., Barzega, G., Bellino, S., Bogetto, F., & Maina, G. (1995b). Predictors of drug treatment response in obsessive-compulsive disorder. *The Journal of Clinical Psychiatry, 56*(9), 321–326.
Rezvan, M. T., Hamadani, A. Z., & Shalbafzadeh, A. (2013). Case-based reasoning for classification in the mixed data sets employing the compound distance methods. *Engineering Applications of Artificial Intelligence, 26*(9), 2001–2009.
Ritter, M., & Qin, H. (2018). Whole-genome association analysis of treatment response from obsessive-compulsive disorder. *Applied Computational Genomics,* 45–57 Springer.
Rokach, L. (2010). *Pattern classification using ensemble methods*, *Vol. 75*. World Scientific.
Salomoni, G., Grassi, M., Mosini, P., Riva, P., Cavedini, P., & Bellodi, L. (2009). Artificial neural network model for the prediction of obsessive-compulsive disorder treatment response. *Journal of Clinical Psychopharmacology, 29*(4), 343–349.
Saxena, S., & Maidment, K. M. (2004). Treatment of compulsive hoarding. *Journal of Clinical Psychology, 60*(11), 1143–1154.
Schapire, R. E. (1990). The strength of weak learnability. *Machine learning, 5*(2), 197–227.
Seni, G., & Elder, J. F. (2010). Ensemble methods in data mining: Improving accuracy through combining predictions. *Synthesis Lectures on Data Mining and Knowledge Discovery, 2*(1), 1–126.
Shavitt, R. G., Belotto, C., Curi, M., Hounie, A. G., Rosário-Campos, M. C., Diniz, J. B., ... Miguel, E. C. (2006). Clinical features associated with treatment response in obsessive-compulsive disorder. *Comprehensive Psychiatry, 47*(4), 276–281.
Stein, D. J., Andersen, E. W., & Overo, K. F. (2007). Response of symptom dimensions in obsessive-compulsive disorder to treatment with citalopram or placebo. *Revista Brasileira de Psiquiatria, 29*(4), 303–307.
Stein, D. J., Carey, P. D., Lochner, C., Seedat, S., Fineberg, N., & Andersen, E. W. (2008). Escitalopram in obsessive-compulsive disorder: Response of symptom dimensions to pharmacotherapy. *CNS Spectrums, 13*(06), 492–498.
Tan, A. C., Gilbert, D., & Deville, Y. (2003). Multi-class protein fold classification using a new ensemble machine learning approach. *Genome Informatics, 14*, 206–217.
Turner, K., & Oza, N. C. (1999). *Decimated input ensembles for improved generalization. Paper presented at the international joint conference on neural networks, 1999. IJCNN'99*.
Wanas, N. (2003). *Feature-based architectures for decision fusion.* University of Waterloo [Department of Systems Engineering].
Zhang, K., & Jin, H. (2010). *An effective pattern based outlier detection approach for mixed attribute data. Paper presented at the Australasian conference on artificial intelligence*.