International Journal of Computational Intelligence Systems
Vol. **12(2)**, 2019, *pp.* 649–660
DOI: https://doi.org/10.2991/ijcis.d.190618.001; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

GMDH2: Binary Classification via GMDH-Type Neural Network Algorithms—R Package and Web-Based Tool

Osman Dag[1,2,*], Erdem Karabulut[1], Reha Alpar[1]

[1]*Department of Biostatistics, Faculty of Medicine, Hacettepe University, 06100, Sihhiye/Ankara, Turkey*
[2]*Neoanka Information Technologies Training & Consultancy Services Ltd Co, ODTU Teknokent, 06800, Cankaya/Ankara, Turkey*

ARTICLE INFO

Article History
Received 17 Jan 2019
Accepted 02 Jun 2019

Keywords
Machine learning
Classification
R package
Web-tool

ABSTRACT

Group method of data handling (GMDH)-type neural network algorithms are the self-organizing algorithms for modeling complex systems. GMDH algorithms are used for different objectives; examples include regression, classification, clustering, forecasting, and so on. In this paper, we present **GMDH2** package to perform binary classification via GMDH-type neural network algorithms. The package offers two main algorithms: GMDH algorithm and diverse classifiers ensemble based on GMDH (dce-GMDH) algorithm. GMDH algorithm performs binary classification and returns important variables. dce-GMDH algorithm performs binary classification by assembling classifiers based on GMDH algorithm. The package also provides a well-formatted table of descriptives in different format (R, LaTeX, HTML). Moreover, it produces confusion matrix and related statistics, and scatter plot (2D and 3D) with classification labels of binary classes to assess the prediction performance. Moreover, a user-friendly web-interface of the package is provided especially for non-R users. This web-interface is available at http://www.softmed.hacettepe.edu.tr/GMDH2.

1. INTRODUCTION

Binary classification is a task where binary target labels can be assigned to each observation. Binary classification appears in different areas such as medical studies, economics, agriculture, meteorology, and so on. In literature, the traditional methods used for this purpose are logistic regression [1] and discriminant analysis [2]. There exist certain assumptions of these models such as linearity between logit and independent continuous variables in logistic regression and multivariate normality in discriminant analysis. Moreover, these methods have some drawbacks especially when the number of independent variables is large or/and the variables are highly correlated. Penalized logistic regression models have been proposed to overcome these problems [3]. At times, it is difficult for the researchers to select an appropriate model. Therefore, selecting an appropriate model in an automatic way may be extremely attractive for researchers especially not having enough statistical knowledge and time [4]. For this purpose, there exist many machine learning algorithms of which the most commonly used ones are support vector machines [5], artificial neural network [6], random forest [7], naive Bayes [5] and so on.

The objective of this paper is to present a software for classification. Some of recent studies for the purpose of classification in different fields are the works of Zhang *et al.* [8], Qui [9], Kang *et al.* [10]. In this study, an R package is proposed for the classification of a two-label output through group method of data handling (GMDH) algorithms. First, Ivakhnenko [11] proposed a polynomial to construct high order polynomials. After that, Ivakhnenko [12] presented heuristic self-organization methods—the main working system of GMDH algorithm. Heuristic self-organization method specifies the architecture of GMDH algorithm by following rules such as external criterion. GMDH algorithm is convenient for complex and unstructured systems and also has benefits over high order regression [4].

The development and usage of GMDH algorithms have been increased in the last two decades. Kondo [13] used the heuristic self-organization method in GMDH algorithm. Abdel-Aal [14] applied GMDH algorithm for feature selection and classification of medical data. Kondo and Ueno [15] proposed GMDH algorithm with a feedback loop on medical image recognition of the brain. Sigmoid transfer function was integrated into GMDH algorithm with a feedback loop [16]. Srinivasan [17] utilized GMDH-type neural network to forecast energy demand prediction. El-Alfy and Abdel-Aal [18] used GMDH algorithm for spam detection and e-mail feature analysis. Three transfer functions—sigmoid, radial basis, and polynomial functions—were integrated into feedback GMDH algorithm [19]. Xu *et al.* [20] used GMDH algorithm to forecast the daily power load. Antanasijević *et al.* [21] applied GMDH algorithm on feature selection for the prediction of transition temperatures of bent-core liquid crystals. Dag and Yozgatligil [22] developed an R package, **GMDH**, for short-term forecasting through GMDH algorithms. Xiao *et al.* [23] applied GMDH-based multiple classifiers ensemble for churn prediction in customer relationship management.

*Corresponding author. Email: osman.dag@outlook.com; osman.dag@hacettepe.edu.tr

In this study, we introduce an R package, **GMDH2** [24] which performs binary classification through GMDH-type neural network algorithms. There exist two main algorithms: GMDH algorithm and diverse classifiers ensemble based on GMDH (dce-GMDH) algorithm. GMDH algorithm performs classification for a binary response and returns important variables dominating the system. The dce-GMDH algorithm performs binary classification by assembling classifiers—support vector machines [5], random forest [7], naive Bayes [5], elastic net logistic regression [25], artificial neural network [6]— based on GMDH algorithm. The package also produces a well-formatted table of descriptives for a binary response in different formats (R, LaTeX, HTML). Moreover, it produces confusion matrix, its related statistics and scatter plot (2D and 3D) with classification labels of binary classes to assess the prediction performance in the package version 1.4 and later. The **GMDH2** package is publicly available on the CRAN.

The **GMDH2** package is the first package in R considering GMDH algorithms for the classification purpose. The **GMDH** package is proposed by Dag and Yozgatligil [22]. However, the **GMDH** package contains all structures designed within a time series perspective. The **GMDH2** package is introduced for binary classification. Furthermore, it selects important features via GMDH-type neural network algorithm. Also, it includes dce-GMDH algorithm which takes advantage from other classifiers.

The organization of paper is presented as follows: First, we provide brief details of GMDH and dce-GMDH algorithms. Second, we introduce the **GMDH2** package and demonstrate the applicability of the package on Wisconsin breast cancer data set. Third, the web-interface of the **GMDH2** package is introduced. After that, GMDH and dce-GMDH algorithms are implemented on real data sets. Finally, the paper is concluded with summary and further research.

2. METHODOLOGY

In this section, feature selection and classification through GMDH algorithm is presented. Also, dce-GMDH algorithm for classification is introduced.

2.1. Feature Selection and Classification through GMDH Algorithm

GMDH-type neural network algorithm is a heuristic self-organization method that investigates the relations among the variables. The algorithm defines its structure itself. Ivakhnenko [11] presented the following polynomial—known as the Ivakhnenko polynomial—to construct a high order polynomial:

$$y = a + \sum_{i=1}^{m} b_i \cdot x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij} \cdot x_i \cdot x_j + \ldots \tag{1}$$

where m is the number of variables to be regressed in each neuron and $a, b, c, \ldots$ are weights of variables in the polynomial. Here, y is a response variable, x_i and x_j are the exploratory variables. In this study, only the main effects are included in the model as presented below.

$$y = a + \sum_{i=1}^{m} b_i \cdot x_i \tag{2}$$

The GMDH algorithm, in general, investigates all pairwise combinations of p exploratory variables. Therefore, m is specified as 2 in Equation (2). For this algorithm, there exist three weights to be estimated in each neuron. The weights are estimated via least square estimation.

In model building and evaluation process, the data are divided into three sets: train (60%), validation (20%), and test (20%) sets. Train set is included in model building. Validation set is used for neuron selection. Test set is utilized to estimate the performance of the methods on unseen data.

The GMDH algorithm can be depicted as follows:

i. Each pairwise combination goes into one neuron.

ii. Weights are estimated with train set in each neuron at layer k.

iii. The predicted probabilities of train set are estimated in each neuron at layer k.

iv. The predicted probabilities of validation set are estimated in each neuron at layer k.

v. The external criterion (EC) (i.e., mean square error) is calculated using validation set in each neuron at layer k.

vi. Selection pressure (α) and the maximum number of neurons to be selected need to be specified.

vii. The neurons of which external criteria are smaller than $(\alpha \cdot min(EC) + (1 - \alpha) \cdot max(EC))/2$ are selected. If the number of selected neurons is larger than the specified maximum number of neurons, the neurons—as many as the specified maximum number of neurons—having smaller external criterion compared to the rest of them are selected.

viii. The predicted probabilities of train set obtained from selected neurons become the inputs for the next layer.

ix. This process (i) to (viii) continues until the stopping rule is realized.

x. There are three stopping rules to conclude the algorithm. The first one is an increase in minimum external criterion at consecutive layers. Second, the algorithm stops when the specified maximum number of layers is reached. The third one is that the algorithm stops if only one neuron in a layer is selected.

xi. At the last layer, only one neuron having minimum EC is selected.

GMDH algorithm is a system of layers where the neurons are present. The number of neurons in each layer is determined by the number of inputs. For example, providing that the number of inputs going into a layer is equal to p, the number of neurons in that layer becomes $h = \binom{p}{2}$, since all pairwise combinations of inputs are considered. This does not mean that all layers include h neurons. For instance, the number of inputs in the input layer defines just the number of neurons in first layer. The number of neurons selected in the first layer determines the number of neurons in second layer.

The algorithm organizes the architecture itself. Sample architecture of GMDH algorithm is placed in Figure 1 when there exist three layers and four inputs.

In the GMDH architecture shown in Figure 1, there exist four inputs (X_1, X_2, X_3, X_4). From these input variables, three of them (X_1, X_2, X_4) are dominating the system. X_3 does not have an impact on classification. GMDH algorithm selects these important features having an effect on classification.

2.2. Diverse Classifiers Ensemble Based on GMDH Algorithm

dce-GMDH algorithm is the GMDH algorithm which assemble the well-known classifiers—support vector machines, random forest, naive Bayes, elastic net logistic regression, artificial neural network. These classifiers are available in **e1071** [5], **randomForest** [7], **e1071** [5], **glmnet** [25], **nnet** [6] packages, respectively. Specifically, these classifiers are available in svm (**e1071**), randomForest (**randomForest**), naiveBayes (**e1071**), cv.glmnet (**glmnet**), nnet (**nnet**) functions, respectively. Unlike GMDH algorithm, dce-GMDH algorithm includes base layer. The classifiers are placed at base layer. Predicted probabilities are obtained using all inputs through these classifiers. The predicted probabilities obtained from these classifiers continue their way as inputs of first layer without applying any neuron selection process. The rest of the algorithm is same as GMDH algorithm. The sample architecture of dce-GMDH algorithm is stated in Figure 2.

The dce-GMDH algorithm is a system of layers where the neurons exist. The number of neurons in a base layer is five since the five classifiers are included. The number of neurons in other layers is defined by the number of inputs. The algorithm assembles the most appropriate classifiers by organizing itself.

In the dce-GMDH architecture shown in Figure 2, there exist four inputs (X_1, X_2, X_3, X_4). These four inputs enter each neuron at base layer. There exists a different classifier in each neuron at base layer. Predicted probabilities are obtained by utilizing four inputs through the classifiers. These predicted probabilities obtained from these classifiers continue to first layer without applying any neuron selection process. Since five inputs will enter in the first layer, the number of neurons in that layer becomes $\binom{5}{2} = 10$. According to external criterion, four neurons are selected and six neurons are eliminated from the network. Since four neurons are selected in the first layer, the number of neurons in the second layer becomes $\binom{4}{2} = 6$. This process continues until one of the stopping rules is realized. Also, the algorithm returns which classifiers are assembled.

3. DEMONSTRATION OF GMDH2 PACKAGE

The **GMDH2** package includes several functions especially designed for binary response. In this part, we work with Wisconsin breast cancer data set, collected by Wolberg and Mangasarian [26], available under **mlbench** [27] package in R. This data set includes nine exploratory variables—clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses—and a grouping variable (malignant or benign). After we put missing observations (16 observations) aside, we have a total of 683 observations (239 and 444 observations in each group, respectively).

After installing and loading **GMDH2** package, the functions designed for binary response are available to be used.

```
# load Wisconsin breast cancer data
R> data(BreastCancer, package =
"mlbench")
R> data <- BreastCancer
# obtain complete observations
R> data <- data[complete.cases(data),]
# select the exploratory variables
R> x <- data[,2:10]
# select the grouping variable
R> y <- data[,11]
```

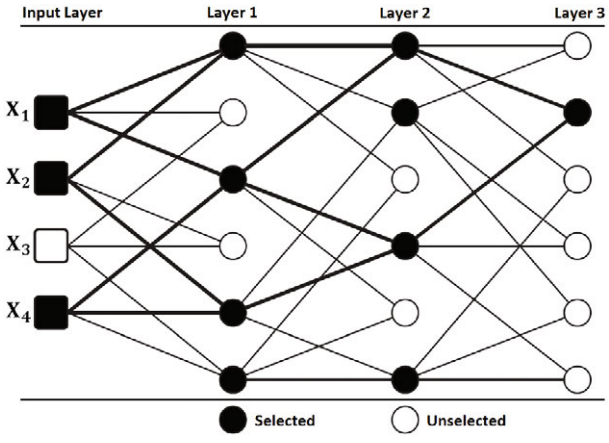


Figure 1 | Architecture of group method of data handling (GMDH) algorithm.

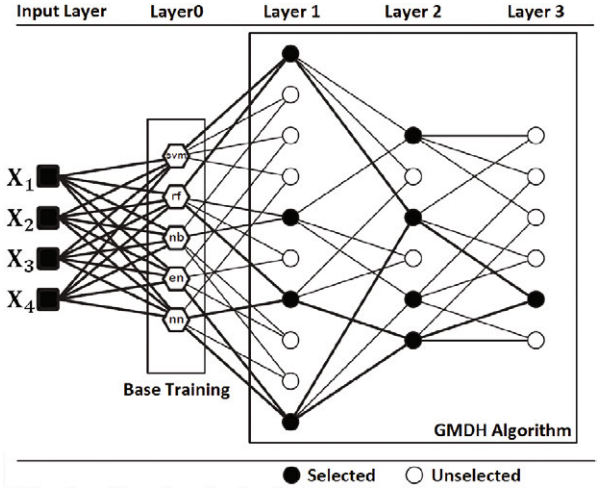


Figure 2 | Architecture of diverse classifiers ensemble based on group method of data handling (dce-GMDH) algorithm.

3.1. Table of the Descriptive Statistics: Table()

Table() produces a table for simple descriptive statistics for a binary response. It returns frequency (percentage) for the variables with class of factor/ordered. Also, this function returns mean ± standard deviation (median, minimum–maximum) or mean ± standard deviation (median, quartile1 - quartile3) for the variables with class of numeric/integer. The option argument is used to return minimum - maximum or quartile1–quartile3 values. When this argument is set to "min-max", this function returns mean ± standard deviation (median, minimum–maximum). When this argument is set to "Q1–Q3," this function returns mean ± standard deviation (median, quartile1–quartile3). The percentages can be specified with the percentages argument as row, column, or total percentages. The ndigits argument is a vector of two numbers utilized to specify the number of digits. The first one is used to specify the number of digits for numeric/integer variables. The second one specifies the number of digits for percentages of factor/ordered variables. Default is set to ndigits = c(2,1). There exists output argument to return the output in a specified format (R, LaTeX, HTML). In this example, we use "LaTeX" output.

```
# obtain a table for simple descriptive
statistics for a binary response
R> Table (x, y, option = "min-max",
percentages = "column", ndigits =
c(2,1), output = "LaTeX")
```

Some portion of the output is given below and its table version is presented in Table 1.

```
\begin{table}[ht]
\centering
\begin{tabular}{rrrrr}
  \hline
  &  & benign &   & malignant \\
  \hline
Observations &   & 444 &   & 239 \\
  Cl.thickness &   &  &   &  \\
  1 &   & 136 (30.6\%) &   &  3 ( 1.3\%) \\
  2 &   &  46 (10.4\%) &   &  4 ( 1.7\%) \\
  3 &   &  92 (20.7\%) &   & 12 ( 5.0\%) \\
  4 &   &  67 (15.1\%) &   & 12 ( 5.0\%) \\
  5 &   &  83 (18.7\%) &   & 45 (18.8\%) \\
  6 &   &  15 ( 3.4\%) &   & 18 ( 7.5\%) \\
  7 &   &   1 ( 0.2\%) &   & 22 ( 9.2\%) \\
  8 &   &   4 ( 0.9\%) &   & 40 (16.7\%) \\
  9 &   &   0 ( 0.0\%) &   & 14 ( 5.9\%) \\
  10 &   &   0 ( 0.0\%) &   & 69 (28.9\%) \\
  Cell.size &   &  &   &  \\
  1 &   & 369 (83.1\%) &   &  4 ( 1.7\%) \\
  2 &   &  37 ( 8.3\%) &   &  8 ( 3.3\%) \\
  3 &   &  27 ( 6.1\%) &   & 25 (10.5\%) \\
  4 &   &   8 ( 1.8\%) &   & 30 (12.6\%) \\
  5 &   &   0 ( 0.0\%) &   & 30 (12.6\%) \\
  6 &   &   0 ( 0.0\%) &   & 25 (10.5\%) \\
  7 &   &   1 ( 0.2\%) &   & 18 ( 7.5\%) \\
  8 &   &   1 ( 0.2\%) &   & 27 (11.3\%) \\
  9 &   &   1 ( 0.2\%) &   &  5 ( 2.1\%) \\
  10 &   &   0 ( 0.0\%) &   & 67 (28.0\%) \\
   \hline
\end{tabular}
\end{table}
```

Table 1 Descriptive statistics.

	benign	malignant
Observations	444	239
Cl.thickness		
1	136 (30.6%)	3 (1.3%)
2	46 (10.4%)	4 (1.7%)
3	92 (20.7%)	12 (5.0%)
4	67 (15.1%)	12 (5.0%)
5	83 (18.7%)	45 (18.8%)
6	15 (3.4%)	18 (7.5%)
7	1 (0.2%)	22 (9.2%)
8	4 (0.9%)	40 (16.7%)
9	0 (0.0%)	14 (5.9%)
10	0 (0.0%)	69 (28.9%)
Cell.size		
1	369 (83.1%)	4 (1.7%)
2	37 (8.3%)	8 (3.3%)
3	27 (6.1%)	25 (10.5%)
4	8 (1.8%)	30 (12.6%)
5	0 (0.0%)	30 (12.6%)
6	0 (0.0%)	25 (10.5%)
7	1 (0.2%)	18 (7.5%)
8	1 (0.2%)	27 (11.3%)
9	1 (0.2%)	5 (2.1%)
10	0 (0.0%)	67 (28.0%)

3.2. Feature Selection and Classification through GMDH Algorithm: GMDH()

In this section, we demonstrate GMDH() function for feature selection and classification. It constructs GMDH algorithm, returns summary statistics of GMDH architecture and important variables. First, we randomly divide data into train, validation and test sets, and then call the GMDH() function. The first and second arguments in this function are a matrix of the exploratory variables and a factor in training set, respectively. The third and fourth arguments are a matrix of the exploratory variables and a factor in validation set, respectively. The alpha argument is the selection pressure. The maxlayers argument is the maximum number of layers specified. The maxneurons argument is the maximum number of neurons allowed in the second and the later layers. The exCriterion argument is the external criterion to be used for neuron selection. The verbose argument is utilized to print the output in R console.

```
# change the class of x to a matrix
R> x <- data.matrix(x)
# the seed number is fixed to 12345 for
reproducibility
R> seed <- 12345
# the number of observations
R> nobs <- length(y)
R> set.seed(seed)
# to split train, validation and test
sets
# to shuffle data
R> indices <- sample(1:nobs)
# the number of observations in each set
R> ntrain <- round(nobs*0.6,0)
R> nvalid <- round(nobs*0.2,0)
R> ntest <- nobs-(ntrain+nvalid)
# obtain the indices of sets
R> train.indices <- sort(indices[1:ntrain])
R> valid.indices <- sort(indices[(ntrain+1)
:(ntrain+nvalid)])
R> test.indices <- sort(indices[(ntrain
+nvalid+1):nobs])
# obtain train, validation and test sets
```

```
R> x.train <- x[train.indices,]
R> y.train <- y[train.indices]
R> x.valid <- x[valid.indices,]
R> y.valid <- y[valid.indices]
R> x.test <- x[test.indices,]
R> y.test <- y[test.indices]
R> set.seed(seed)
# construct model via GMDH algorithm
R> model <- GMDH(x.train, y.train,
x.valid, y.valid, alpha = 0.6,
maxlayers = 10, maxneurons = 15,
exCriterion = "MSE", verbose = TRUE)
 Structure :
 Layer   Neurons  S. neurons    Min MSE
     1        36          15    0.06316
     2       105          15    0.05310
     3       105          15    0.05188
     4       105          15    0.05161
     5       105          15    0.05127
     6       105          15    0.05110
     7       105          15    0.05098
     8       105          11    0.05096
     9        55          15    0.05096
    10       105           1    0.05095
 External criterion : Mean Square Error
 Feature selection  : 8 out of 9
 variables are selected.

    Cl.thickness
       Cell.size
   Marg.adhesion
    Epith.c.size
     Bare.nuclei
     Bl.cromatin
 Normal.nucleoli
         Mitoses
```

Here, the structure includes layer, neurons, s. neurons, and min MSE in the output above. The layer shows the number of layer. The neurons represent the number of neurons in corresponding layer. The s. neurons mean the number of selected neurons. The min MSE respresents the minimum external criterion which is calculated for the neuron gives the minimum external criterion on validation set in the corresponding layer. There exist two options for the external criterion namely, mean square error and mean absolute error.

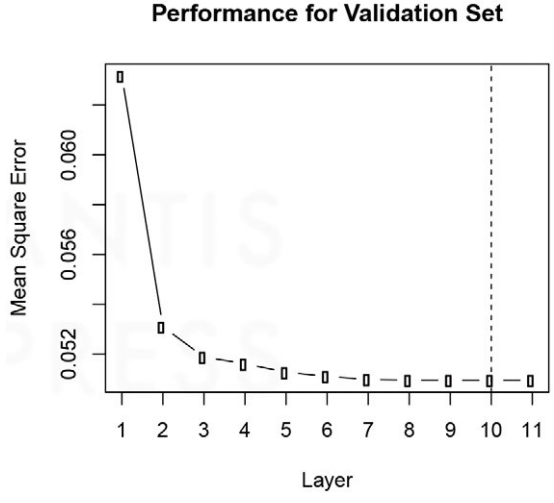


Figure 3 | Minimum external criterion across layers (GMDH algorithm).

Eight variables—clump thickness, uniformity of cell size, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses—are selected by the algorithm. Minimum external criterion can be plotted across layers (presented in Figure 3) by the following code:

```
R> plot(model)
```

Predictions for test set can be made after model building process is completed. Test set has 136 observations, but only 10 of them are reported to save space.

```
R> predict(model, x.test, type = "class")
  [1] benign benign benign benign benign
  [6] benign malignant benign benign
  [10] benign
Levels: benign malignant
R> predict(model, x.test, type =
"probability")
            benign   malignant
  [1,] 1.000000000 0.000000000
  [2,] 0.643870382 0.356129618
  [3,] 0.670641964 0.329358036
  [4,] 0.974398179 0.025601821
  [5,] 0.920988111 0.079011889
  [6,] 0.994693987 0.005306013
  [7,] 0.436033878 0.563966122
  [8,] 0.951034736 0.048965264
  [9,] 1.000000000 0.000000000
 [10,] 0.994693987 0.005306013
```

The GMDH algorithm predicts that the probability of benign for the first and second persons are 100% and 64.4%, respectively. Since the predicted probability of benign is greater than the predicted probability of malignant, these persons are classified as benign.

3.3. Confusion Matrix and Related Statistics: confMat()

The confMat() function produces a confusion matrix for a binary response. It also returns some related statistics. These statistics are accuracy, no information rate, unweighted Kappa statistic, Matthews correlation coefficient, sensitivity, specificity, positive predictive value, negative predictive value, prevalence, balanced accuracy, youden index, detection rate, detection prevalence, precision, recall, and F1 measure. The formulation of these statistics are not stated in this paper, but presented in the manual of GMDH2 package. The positive argument is an optional character string used to specify the positive factor level. The verbose argument is utilized to print the output in R console.

```
# obtain predicted classes for test set
R> y.test_pred <- predict(model, x.test,
type = "class")
# obtain confusion matrix and some
statistics for test set
R> confMat(y.test_pred, y.test, positive
= "malignant")
Confusion Matrix and Statistics
           reference
data        malignant benign
```

```
malignant        51      1
benign            5     79
   Accuracy                :   0.9559
   No Information Rate     :   0.5882
   Kappa                   :   0.9079
   Matthews Corr Coef      :   0.9097
   Sensitivity             :   0.9107
   Specificity             :   0.9875
   Positive Pred Value     :   0.9808
   Negative Pred Value     :   0.9405
   Prevalence              :   0.4118
   Balanced Accuracy       :   0.9491
   Youden Index            :   0.8982
   Detection Rate          :   0.375
   Detection Prevalence    :   0.3824
   Precision               :   0.9808
   Recall                  :   0.9107
   F1                      :   0.9444
   Positive Class          :   malignant
```

Accuracy of GMDH algorithm is estimated to be 0.9559. This algorithm classifies 95.59% of persons in a correct class. Also, sensitivity and specificity are calculated as 0.9107 and 0.9875. The algorithm classifies 91.07% of the persons having breast cancer, 98.75% of the persons not having breast cancer.

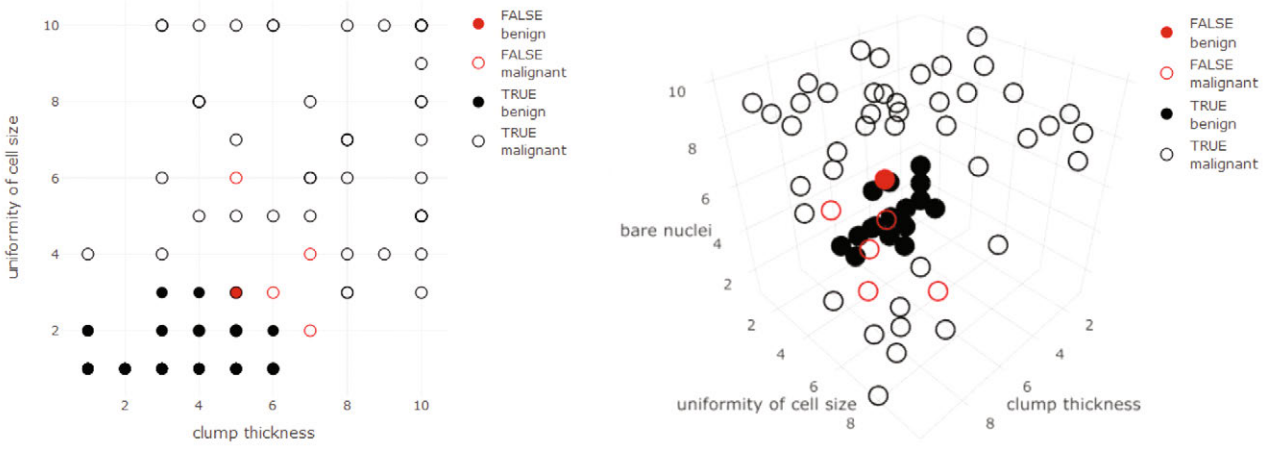
3.4. Scatter Plots with Classification Labels: cplot2d() and cplot3d()

The cplot2d() and cplot3d() functions provide interactive 2-dimensional (Figure 4a) and 3-dimensional (Figure 4b) scatter plots with classification labels, respectively. These functions originally use the plot_ly function from **plotly** [28] package. The first two arguments of cplot2d() are the exploratory variables stated in the *x* and *y* axes of Figure 4a. The first three arguments of cplot3d() are the exploratory variables placed in the *x*, *y*, and *z* axes of Figure 4b. The ypred and yobs arguments are predicted and observed classes. The colors and symbols arguments are used to specify the colors and symbols of true/false classification labels, respectively. The size of symbols can be changed with the size argument. The names of axes can be changed with the arguments xlab, ylab, zlab, and title.

```
# to produce 2D scatter plot with
classification labels for test set
R> cplot2d(x.test[,1], x.test[,2],
y.test_pred, y.test, colors = c("red",
"black"), xlab = "clump thickness",
ylab = "uniformity of cell size")
# to produce 3D scatter plot with
classification labels for test set
R> cplot3d(x.test[,1], x.test[,2],
x.test[,6], y.test_pred, y.test,
colors = c("red", "black"),
xlab = "clump thickness",
ylab = "uniformity of cell size",
zlab = "bare nuclei")
```

3.5. Diverse Classifiers Ensemble Based on GMDH Algorithm: dceGMDH()

In this part, we demonstrate dceGMDH() function for classification. It constructs dce-GMDH algorithm, returns summary statistics of dce-GMDH architecture and assembled classifiers. Like GMDH() function, the first and second arguments are a matrix of the exploratory variables and a factor in training set, respectively. The third and fourth arguments are a matrix of the exploratory variables and a factor in validation set, respectively. The alpha argument is the selection pressure. The maxlayers argument is the specified maximum number of layers. The maxneurons argument is the maximum number of neurons allowed in the second and later layers. The exCriterion argument is the external criterion to be utilized for neuron selection. The verbose argument is utilized to print the output in R console. Also, there are the arguments for options of classifiers. The svm_options argument is a list for options of svm. The randomForest_options argument is a list for options of randomForest. The naiveBayes_options argument is a list for options of naiveBayes. The cv.glmnet_options argument is a list for options of cv.glmnet (the elastic net mixing parameter is fixed to 0.5 as default). The nnet_options argument is a list for options of nnet.



(a) 2-dimensional (b) 3-dimensional

Figure 4 Scatter plots with classification labels.

```
R> set.seed(seed)
# construct model via dce-GMDH algorithm
R> model <- dceGMDH(x.train, y.train,
x.valid, y.valid, alpha = 0.6, maxlayers
= 10, maxneurons = 15, exCriterion =
"MSE", verbose = TRUE)
 Structure :
 Layer   Neurons  S. neurons    Min MSE
     0         5           5    0.04669
     1        10           1    0.04641
 External criterion   : Mean Square Error
 Classifiers ensemble : 2 out of 5
 classifiers are assembled.
       svm
 cv.glmnet
```

In this example, two classifiers—support vector machine and elastic net logistic regression—are assembled by the algorithm. Minimum external criterion can be plotted across layers (presented in Figure 5) by the following line:

```
R> plot(model)
```

Predictions for test set can be made after model building process is completed. Test set has 136 observations; therefore, 10 of them are reported to save space.

```
R> predict(model, x.test, type = "class")
  [1] benign benign malignant benign
  [5] benign benign malignant benign
  [9] benign benign
Levels: benign malignant
R> predict(model, x.test, type =
"probability")
             benign    malignant
  [1,] 0.9571287282 4.287127e-02
  [2,] 0.8317147956 1.682852e-01
  [3,] 0.3400820793 6.599179e-01
  [4,] 1.0000000000 0.000000e+00
  [5,] 0.9876416020 1.235840e-02
  [6,] 1.0000000000 0.000000e+00
  [7,] 0.2762650840 7.237349e-01
  [8,] 1.0000000000 0.000000e+00
  [9,] 1.0000000000 0.000000e+00
 [10,] 1.0000000000 0.000000e+00
```

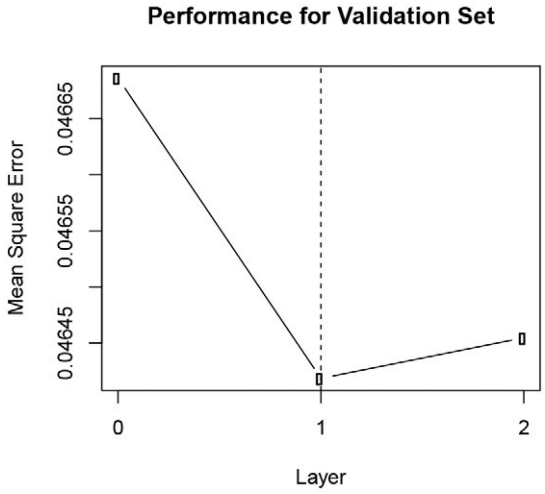


Figure 5 | Minimum external criterion across layers (dce-GMDH algorithm).

The dce-GMDH algorithm predicts that the probability of benign for the first and second persons are 95.7% and 83.2%, respectively. Since the predicted probability of benign is greater than the predicted probability of malignant, these persons are classified as benign.

Confusion matrix and related statistics are obtained through the following codes to investigate the performance measures for the test set:

```
# obtain predicted classes for test set
R> y.test_pred <- predict(model, x.test,
type = "class")
# obtain confusion matrix and some
statistics for test set
R> confMat(y.test_pred, y.test, positive
= "malignant")
Confusion Matrix and Statistics
           reference
data        malignant benign
  malignant        54      1
  benign            2     79
     Accuracy             :   0.9779
     No Information Rate  :   0.5882
     Kappa                :   0.9543
     Matthews Corr Coef   :   0.9545
     Sensitivity          :   0.9643
     Specificity          :   0.9875
     Positive Pred Value  :   0.9818
     Negative Pred Value  :   0.9753
     Prevalence           :   0.4118
     Balanced Accuracy    :   0.9759
     Youden Index         :   0.9518
     Detection Rate       :   0.3971
     Detection Prevalence :   0.4044
     Precision            :   0.9818
     Recall               :   0.9643
     F1                   :   0.973
     Positive Class       :   malignant
```

Accuracy rate of dce-GMDH algorithm is estimated to be 0.9779. This algorithm classifies 97.79% of persons in a correct class.

All in all, using dce-GMDH algorithm increases the classification performance approximately 2% in accuracy compared to GMDH algorithm for this data set.

4. WEB-TOOL DEVELOPMENT

The purpose of this package is to perform binary classification via GMDH-type neural network algorithms. This package presents two main algorithms: GMDH algorithm and dce-GMDH algorithm. GMDH algorithm performs binary classification and returns the variables dominating the system. dce-GMDH algorithm performs binary classification by assembling classifiers depending on GMDH algorithm.

The package provides a well-formatted table of descriptives in different format (R, LaTeX, HTML). Also, it produces confusion matrix, its related statistics and scatter plot (2D and 3D) with classification labels of binary classes to assess the contribution of the variables on the prediction performance. It is sometimes difficult for applied researchers to deal with R codes. Therefore, a web-interface of **GMDH2** package is developed by using **shiny** [29] package. This web-interface is available at url-http://www.softmed.hacettepe.edu.tr/GMDH2.

Researchers can upload their data to the tool through *Data upload* tab (Figure 6a). There is a demo dataset called Wisconsin breast cancer dataset on this tab for the researchers to test the tool. Basic descriptive statistics can be obtained via *Describe data* tab (Figure 6b). These statistics can be obtained in different formats (R, LaTeX, HTML). After describing the data, researchers can specify the algorithm desired through *Algorithms* tab (Figure 6c). In this tab, there exist two main algorithms: GMDH and dce-GMDH algorithms. Researchers can obtain the performance measures of classification through *Results* tab (Figure 6d). Moreover, predicted probabilities and classes can be downloaded via this tab. Researchers can examine the interactive scatter plots with classification labels (Figure 4) via *Visualize* tab (Figure 6e). At last, researchers can upload new data, obtain predicted probabilities and classes through *New data* tab (Figure 6f). Also, these predictions can be downloaded via this tab.

5. REAL DATA APPLICATIONS

5.1. Case Study on Real Datasets

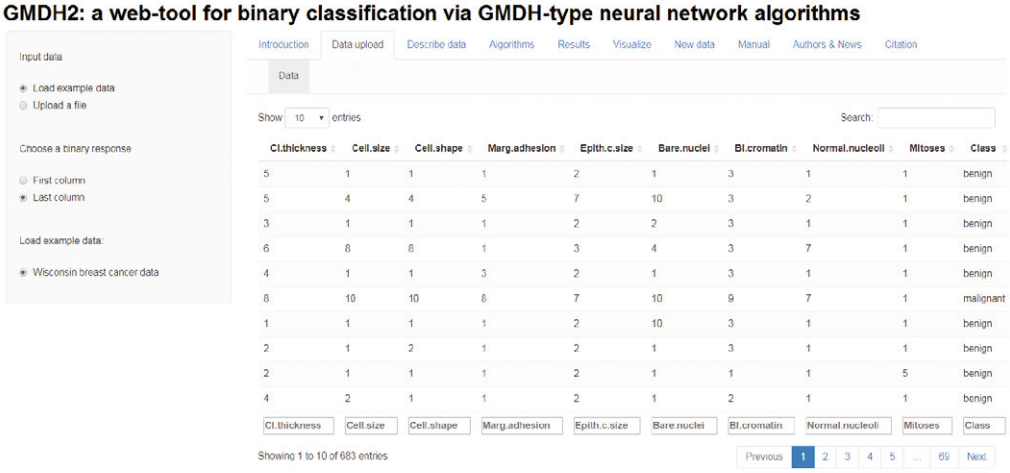
In this section, we will implement **GMDH2** using several real datasets. The first data set is *Wisconsin breast cancer data* collected by Wolberg and Mangasarian [26]. The implementation of the package demonstrated on this data set is introduced in Section 3.

The second data set is *ionosphere data* downloaded from **mlbench** [27] R package. This data set was collected by a system in Goose Bay, Labrador. Free electrons were targeted in the ionosphere. This data set contains 351 radar returns with 34 attributes. Providing that the radar returns illustrate evidence of some type of structure in the ionosphere, they are called "good" radar returns. Otherwise, their signals pass through the ionosphere and they are called "bad" returns. There exist 225 and 126 observations in each group, respectively.

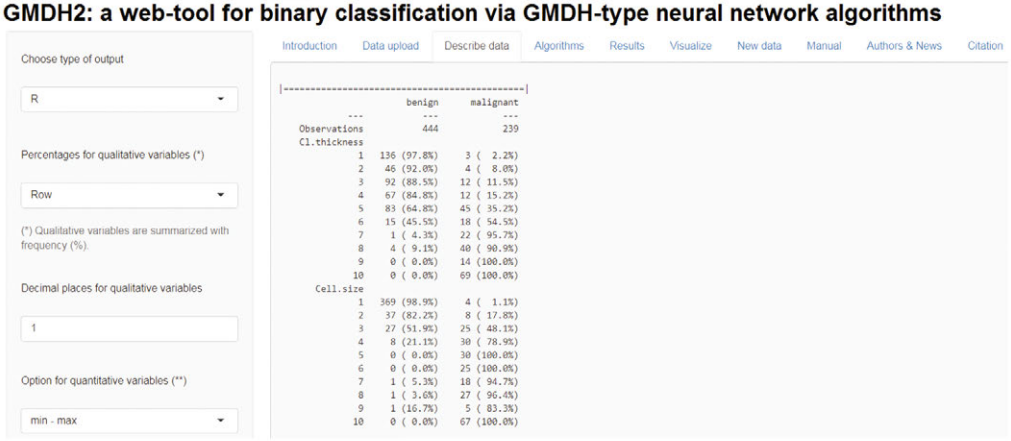
Finally, we utilized *sonar data* downloaded from **mlbench** [27] R package. Sonar signals are classified to be bounced off a metal cylinder or a roughly cylindrical rock. This data set includes 208 sonar signals (111 and 97 observations in each group, respectively) with 60 attributes. Each attribute is in the range 0.0 to 1.0. Each attribute represents the energy within a particular frequency band.

5.2. Implementation of the GMDH2 Package

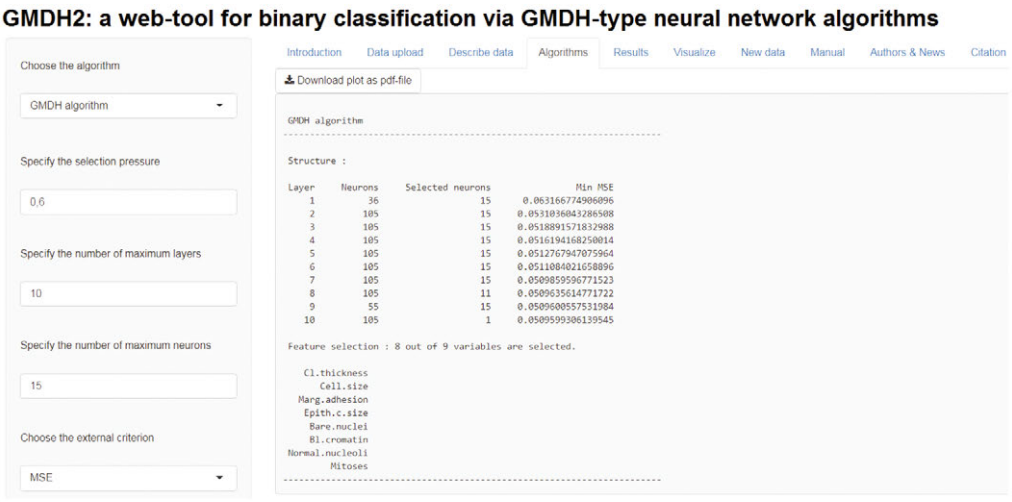
We perform both GMDH and dce-GMDH classifiers on Wisconsin breast cancer data, Ionosphere data, and Sonar data. In the methods, selection pressure is set to 0.6. The maximum number of neurons



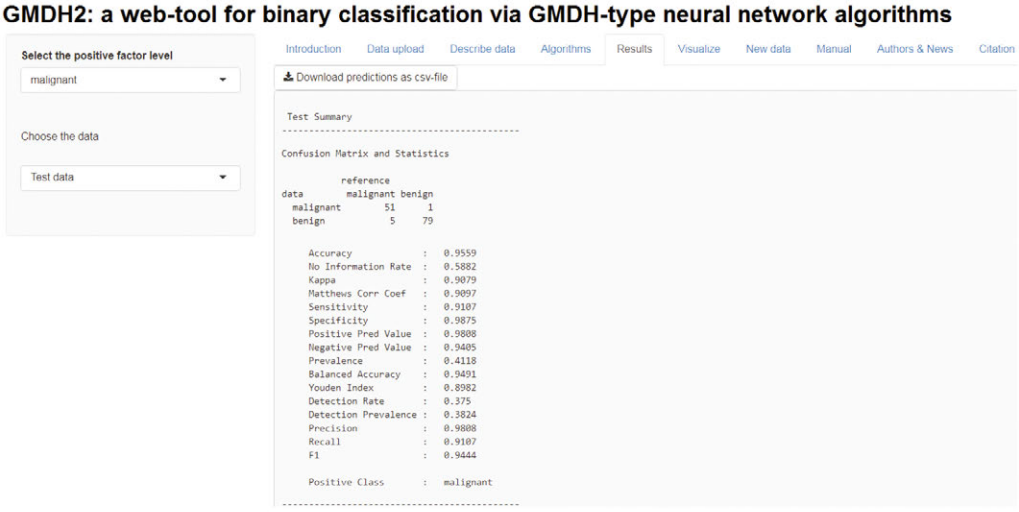
(a) Data upload



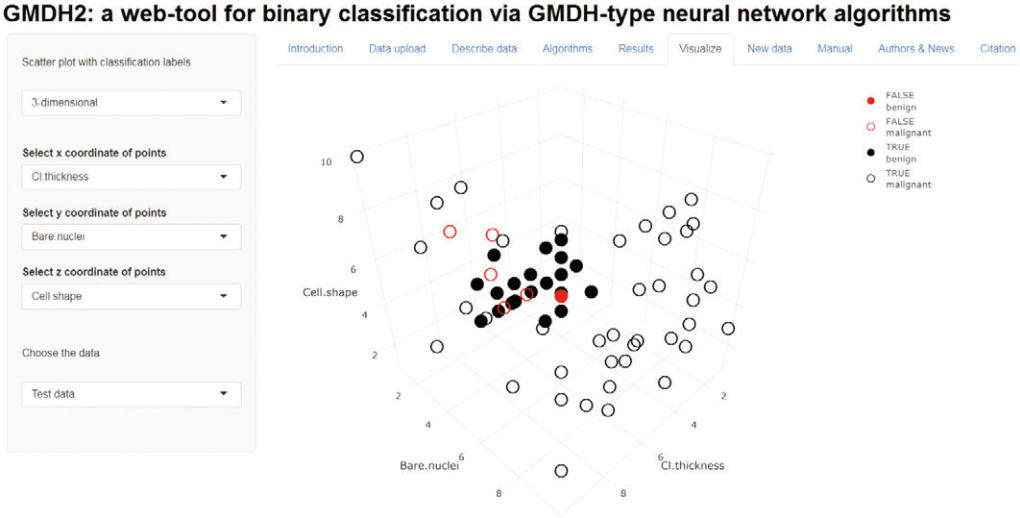
(b) Describe data



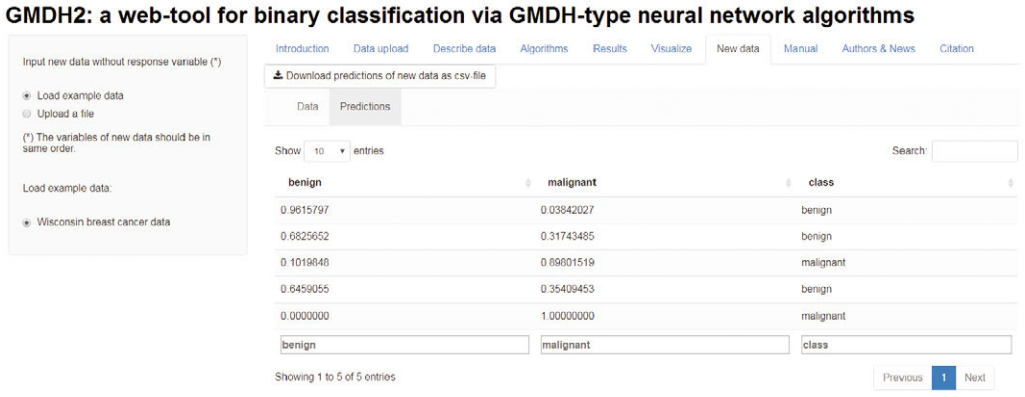
(c) Algorithms



(d) Results



(e) Visualize



(f) New data

Figure 6 Web-tool of GMDH2 package.

is fixed to 15. The maximum number of layers is set to 10. MSE is used as an external criterion in all analyses.

The data set is split into three parts as train, validation and test sets including 60%, 20%, and 20% of all samples, respectively. In this part, the performances of the GMDH classifiers are reported based on the confusion matrices of true and predicted classes for test sets. The seed number is fixed to "12345" for the reproducibility of the results.

5.3. Performance Comparison of GMDH Classifiers

In this section, we discuss the performance of the GMDH classifiers. As it is mentioned in the earlier sections, several measures are considered for the model performances. We reported the results in Table 2 with accuracy, no information rate, Kappa, Matthews correlation coefficient, sensitivity, specificity, positive predictive value, negative predictive value, prevalence, balanced accuracy, youden index, detection rate, detection prevalence, precision, recall, and F1 measure.

According to the measures of accuracy, kappa, Matthews correlation coefficient, positive predictive value (precision), balanced accuracy, youden index and F1, dce-GMDH algorithm is superior to GMDH algorithm in all data sets. With respect to sensitivity (recall) and negative predictive value, dce-GMDH algorithm performs better compared to GMDH algorithm on both breast cancer and sonar data sets, but vice versa is true on ionosphere data set. According to the measure of specificity, dce-GMDH classifier performs as well as GMDH classifier except for ionosphere data, but it performs better than GMDH classifier on ionosphere data.

6. SUMMARY AND FURTHER RESEARCH

Binary classification is a problem in which binary factor labels can be predicted for each observation. Binary classification is used in different disciplines. Examples include medical studies, economics,

Table 2 Classification results for real datasets.

	Breast Cancer		Ionosphere		Sonar	
Number of features	9		34		60	
Number of observations	683		351		208	
Class sizes	239/444		225/126		111/97	
Class ratios	0.538:1		1.786:1		1.144:1	
Train/Validation/Test	410/137/136		211/70/70		125/42/41	
	GMDH	dce-GMDH	GMDH	dce-GMDH	GMDH	dce-GMDH
Accuracy	0.9559	0.9779	0.9000	0.9429	0.7805	0.8049
No information rate	0.5882	0.5882	0.6857	0.6857	0.5854	0.5854
Kappa	0.9079	0.9543	0.7461	0.8641	0.5591	0.6048
Matthews corr coef	0.9097	0.9545	0.7714	0.8661	0.5653	0.6078
Sensitivity	0.9107	0.9643	1.0000	0.9792	0.7500	0.7917
Specificity	0.9875	0.9875	0.6818	0.8636	0.8235	0.8235
Positive pred value	0.9808	0.9818	0.8727	0.9400	0.8571	0.8636
Negative pred value	0.9405	0.9753	1.0000	0.9500	0.7000	0.7368
Prevalence	0.4118	0.4118	0.6857	0.6857	0.5854	0.5854
Balanced accuracy	0.9491	0.9759	0.8409	0.9214	0.7868	0.8076
Youden index	0.8982	0.9518	0.6818	0.8428	0.5735	0.6152
Detection rate	0.3750	0.3971	0.6857	0.6714	0.4390	0.4634
Detection prevalence	0.3824	0.4044	0.7857	0.7143	0.5122	0.5366
Precision	0.9808	0.9818	0.8727	0.9400	0.8571	0.8636
Recall	0.9107	0.9643	1.0000	0.9792	0.7500	0.7917
F1	0.9444	0.9730	0.9320	0.9592	0.8000	0.8261

GMDH, group method of data handling; dce-GMDH, diverse classifiers ensemble based on group method of data handling.

agriculture, meteorology, and so on. In this paper, we present **GMDH2** package to perform binary classification through GMDH-type neural network algorithms.

The **GMDH2** package offers two main algorithms; namely, GMDH and dce-GMDH algorithms. GMDH algorithm makes binary classification and determines which features are important for discrimination of classes. dce-GMDH algorithm assembles the classifiers—support vector machines, random forest, naive Bayes, elastic net logistic regression, artificial neural network—based on GMDH algorithm to perform classification for a binary response. Moreover, the package provides a table of descriptives for a binary factor in different formats (R, LaTeX, HTML). The package also produces confusion matrix, its related statistics, and scatter plot (2D and 3D) with classification labels of binary classes to assess the prediction performance. The package and its web-interface will be updated regularly.

Future studies are planned in the direction of multi-label classification. Moreover, these algorithms can be used for the large number of variables, such as classification of genomics data. With especially GMDH algorithm, selection of important genes can be conducted.

CONFLICT OF INTEREST

The authors declare there is no conflict of interest.

AUTHORS' CONTRIBUTIONS

Osman Dag conceived and designed the study, performed the study, developed the R package and its web-based tool, analyzed the data sets, wrote the paper, prepared figures and table(s).

Erdem Karabulut and Reha Alpar conceived and designed the study, wrote the paper, reviewed drafts of the paper.

Funding Statement

This study is supported by 2211/A Scholarship Program within The Scientific and Technological Research Council of Turkey and by Hacettepe University Scientific Research Projects Coordination Unit with project Number THD-2018-16610.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments and suggestions which helped us to improve the quality of our paper. This study is supported by 2211/A Scholarship Program within The Scientific and Technological Research Council of Turkey and by Hacettepe University Scientific Research Projects Coordination Unit with project Number THD-2018-16610.

REFERENCES

[1] A. Agresti, An Introduction to Categorical Data Analysis, vol. 135, Wiley, New York, 1996.
[2] W.R. Klecka, Discriminant Analysis, Sage, Beverly Hills 1980.
[3] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Methodol. 58 (1996), 267–288.
[4] S.J. Farlow, The GMDH algorithm of Ivakhnenko, Am. Stat. 35 (1981), 210–215.
[5] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: misc functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2015. R package version 1.6-7. https://CRAN.R-project.org/package=e1071
[6] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, fourth ed., Springer, New York, 2002. ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4
[7] A. Liaw, M. Wiener, Classification and regression by randomForest, R News. 2 (2002), 18–22. http://CRAN.R-project.org/doc/Rnews/
[8] Y. Zhang, X. Cui, Y. Liu, B. Yu, Tire defects classification using convolution architecture for fast feature embedding, Int. J. Comput. Intell. Syst. 11 (2018), 1056–1066.
[9] C. Qiu, Bare bones particle swarm optimization with adaptive chaotic jump for feature selection in classification, Int. J. Comput. Intell. Syst. 11 (2018), 1–14.
[10] X. Kang, B. Zhuo, P. Duan, Semi-supervised deep learning for hyperspectral image classification, Remote Sens. Lett. 10 (2019), 353–362.
[11] A. Ivakhnenko, The group method of data handling–a rival of the method of stochastic approximation, Soviet Autom. Control. 13 (1966), 43–55.
[12] A. Ivakhnenko, Heuristic self-organization in problems of engineering cybernetics, Automatica. 6 (1970), 207–219.
[13] T. Kondo, GMDH neural network algorithm using the heuristic self-organization method and its application to the pattern identification problem, in SICE'98. Proceedings of the 37th SICE Annual Conference. International Session Papers, IEEE, Chiba, Japan, 1998, pp. 1143–1148.
[14] R. Abdel-Aal, GMDH-based feature ranking and selection for improved classification of medical data, J. Biomed. Inform. 38 (2005), 456–468.
[15] T. Kondo, J. Ueno, Medical image recognition of the brain by revised GMDH-type neural network algorithm with a feedback loop, Int. J. Innov. Comput. Inf. Control. 2 (2006), 1039–1052.
[16] T. Kondo, J. Ueno, Revised gmdh-type neural network algorithm with a feedback loop identifying sigmoid function neural network, Int. J. Innov. Comput. Inf. Control. 2 (2006), 985–996.
[17] D. Srinivasan, Energy demand prediction using GMDH networks, Neurocomputing. 72 (2008), 625–629.
[18] E.S.M. El-Alfy, R. E. Abdel-Aal, Using GMDH-based networks for improved spam detection and email feature analysis, Appl. Soft Comput. 11 (2011), 477–488.
[19] T. Kondo, J. Ueno, Feedback GMDH-type neural network and its application to medical image analysis of liver cancer, in 42th ISCIE International Symposium on Stochastic Systems Theory and its Applications, Okayama, Japan 2010, pp. 256–263.
[20] H. Xu, Y. Dong, J. Wu, W. Zhao, Application of GMDH to short-term load forecasting, in: Advances in Intelligent Systems, Gary Lee (Ed.), Springer-Verlag, Berlin, Heidelberg, 2012, pp. 27–32.
[21] D. Antanasijević, J. Antanasijević, V. Pocajt, G. Uščumlić, A GMDH-type neural network with multi-filter feature selection for the prediction of transition temperatures of bent-core liquid crystals, RSC Adv. 6 (2016), 99676–99684.

[22] O. Dag, C. Yozgatligil, GMDH: an R package for short term forecasting via GMDH-type neural network algorithms, R J. 8 (2016), 379–386.
[23] J. Xiao, X. Jiang, C. He, G. Teng, Churn prediction in customer relationship management via gmdh-based multiple classifiers ensemble, IEEE Intell. Syst. 31 (2016), 37–44.
[24] O. Dag, E. Karabulut, R. Alpar, GMDH2: binary classification via GMDH-type neural network algorithms, 2018. R package version 1.4. https://CRAN.R-project.org/package=GMDH2
[25] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Softw. 33 (2010), 1.
[26] W.H. Wolberg, O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proc. Natl. Acad. Sci. 87 (1990), 9193–9196.
[27] F. Leisch, E. Dimitriadou, mlbench: machine learning benchmark problems, 2010, p. 1. R package version 21-1. https://CRAN.R-project.org/package=mlbench
[28] C. Sievert, C. Parmer, T. Hocking, S. Chamberlain, K. Ram, M. Corvellec, P. Despouy, Plotly: create interactive web graphics via 'plotly.js', 2017. R package version 4.7.1. https://CRAN.R-project.org/package=plotly
[29] W. Chang, J. Cheng, J. Allaire, Y. Xie, J. McPherson, Shiny: web application framework for R, 2017. R package version 1.0.1. https://CRAN.R-project.org/package=shiny