Gaussian Process Regression Method for Classification for High-Dimensional Data with Limited Samples

Nian Zhang
Dept. of Electrical and Computer Eng.
Univ. of the District of Columbia
Washington, D.C. 20008 USA
nzhang@udc.edu

Jiang Xiong and Jing Zhong
College of Computer Science and Eng.
Chongqing Three Gorges University
Chongqing, 404000, China
xjcq123@sohu.com, zhongandy@sohu.com

Keenan Leatham
Dept. of Electrical and Comp. Eng.
Univ. of the District of Columbia
Washington, D.C. 20008 USA
keenan.leatham@udc.edu

***Abstract*—We present a Gaussian process regression (GPR) algorithm with variable models to adapt to numerous pattern recognition data for classification. The algorithms of the Gaussian process regression (GPR) models including the rational quadratic GPR, squared exponential GPR, matern 5/2 GPR, and exponential GPR are described. The response plot, predicted vs. actual plot, and residuals plot of these GPR models are demonstrated. In addition, a comprehensive comparison of classification performance among rational quadratic GPR, squared exponential GPR, matern 5/2 GPR, and exponential GPR is presented in terms of various model statistics. Furthermore, the classification error rates of these four GPR based models are in comparison to the extended nearest neighbor (ENN), classic k-nearest Neighbor (KNN), naive Bayes, linear discriminant analysis (LDA), and the classic multilayer perceptron (MLP) neural network. The excellent experimental results demonstrated that the Gaussian process regression models provide a very promising feature selection solution to numerous pattern recognition problems. The algorithm is able to learn from the global distribution, therefore improving pattern recognition performance.**

Keywords— Gaussian process regression (GPR);high-dimensional data; classification

I. INTRODUCTION

Recent advances in modern technologies, such as photo-thermal infrared (IR) imaging spectroscopy technology in the application of remote explosive detection, 4D CT-scans technology, and DNA microarrays have produced numerous massive and imbalanced data. The needs of classification ubiquitously exist in real-world data-intensive applications, ranging from civilian applications such as cancer diagnoses and outlier detection in stock market time series, to homeland security or defense related applications such as remote explosive detection, illegal drug detection, and abnormal behavior recognition.

In the situation when the dimensionality of data is high but with few data, feature selection usually becomes imperative to the learning algorithms because high-dimensional data tends to negatively affect the efficiency of most learning algorithms. Feature selection is an efficient dimensionality reduction technique that selects an optimal subset of the original features that provide the best predictive power in modeling the data. They are the most distinct features that can be used to differentiate samples into different classes.

There are a large number of state-of-the-art feature selection methods. A simultaneous spectral-spatial feature selection and extraction algorithm was proposed for hyperspectral images spectral-spatial feature representation and classification. However, it lacks of kernel version and thus its performance on complex datasets is unknown [1]. A regularized regression based feature selection classifier was modified into a cost-sensitive classifier by generating and assigning different costs to each class. Features will be selected according to the classifier with optimal F-measure in order to solve the class imbalance problem [2]. A feature selection algorithm using AdaBoost was presented to deal with Haar-like features for vehicle detection. The normalized feature set is used to cross validate the RBF-SVM classifier to select the optimal parameters [3]. A support vector machine (SVM) based classifier is designed to identify abnormal residual functional capacities in athletes suffering from concussion. The total accuracy of the classifier using 10 prominent features on a multichannel EEG data set was 77.1% [4]. However, these methods require a lot of training data to estimate the underlying function and their accuracy need to be improved. Therefore, it is imperative to develop a new algorithm to adapt to the high-dimensional but relatively small samples for classification. We will use the banknote authentication data set and other 19 other data sets as a demonstration.

The rest of the paper is organized as follows. In Section II, the Gaussian process regression (GPR) models including the rational quadratic GPR, squared exponential GPR, matern 5/2 GPR, and exponential GPR are described. In Section III, the banknote authentication data set is introduced. In Section IV, the response plot, predicted vs. actual plot, and residuals plot of these GPR models are demonstrated. In addition, a comprehensive comparison of classification performance among rational quadratic GPR, squared exponential GPR, matern 5/2 GPR, and exponential GPR is presented in terms of various model statistics. Furthermore, the classification error rates of these four GPR based models are in comparison to the ENN [5], classic KNN, naive Bayes, linear discriminant analysis (LDA), and the classic multilayer perceptron (MLP) neural network. In Section V, the paper is concluded.

II. Types of Gaussian Process Regression Algorithms

Gaussian process regression (GPR) models are nonparametric kernel-based probabilistic models with a finite collection of random variables with a multivariate distribution. Every linear combination is evenly distributed. The concept of Gaussian processes is named after Carl Friedrich Gauss because it is based on the notion of the Gaussian distribution to be an infinite-dimensional generalization of multivariate normal distributions. Gaussian processes are utilized in statistical modeling, regression to multiple target values, and analyzing mapping in higher dimensions. For each GPR model we will be (1) Training a data set with GPR models such as Rational Quadratic GPR, Squared Exponential GPR, Matern 5/2 GPR, and Exponential GPR (2) Plotting the behavior of each algorithm figuring out the RSME, R-Squared Value, MSE, MAE, Prediction Speed, Training Time, and (3) Analyzing the results of each Gaussian process regression to see the similarities and differences of the data. The purpose of these trials is to see if we can find some interesting behaviors, so we can find different methods to optimize GPR models. Shown below are the different behaviors of each GPR.

A. Rational Quadratic GPR

The Rational Quadratic GPR kernel allows us to model data varying at multiple scales. The Rational Quadratic GPR algorithm is used in spatial statistics, geostatistics, machine learning, image analysis, and other fields where multivariate statistical analysis is conducted on metric spaces. The algorithm of the rational quadratic GPR is illustrated as follows.

Algorithm of the Rational Quadratic GPR

Input:

1. A training data set of the form:

$\{(x_i,y_i);\ i=1,2,\ldots,n\}$

where $x_i \in R^d$ and $y_i \in R$

2. A linear regression model of the form:

$y=x^T\beta+\varepsilon$

Procedure:

1. Let the given training data set of n points be in the form of:

$\{(x_i,y_i);\ i=1,2,\ldots,n\}$

where $x_i \in R^d$ and $y_i \in R$

2. A linear regression model of the form:

$y=x^T\beta+\varepsilon$

3. The linear regression model, where K(X, X) is parametrized looks as follows:

$$\mathbf{K(X,X)}=\begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & \ldots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \ldots & k(x_2,x_n) \\ \vdots & \vdots & \vdots & \vdots \\ k(x_n,x_1) & k(x_n,x_2) & \ldots & k(x_n,x_n) \end{pmatrix}$$

4. The Rational GPR Model becomes:

$$k(x_i,x_j|\theta) = {\sigma_f}^2\left(1+\frac{r^2}{2\propto {\sigma_l}^2}\right)$$

where:

$$r=\sqrt{(x_i-x_j)^T(x_i-x_j)}.$$

θ is the maximum a posteriori estimates. σ_f is the signal standard deviation. $\propto$ is the non-negative parameter of the covariance.

The inferential results are dependent on the values of the hyperparameters θ defining the model's behavior. It is commonly used to define the statistical covariance between measurements made at two points, which are d units distant from each other. The covariance only depends on distances between points, which are stationary. If the distance is Euclidean distance, the rational quadratic covariance function is called isotropic. The advantage of the Rational Quadratic GPR algorithm is on the large data sets if the interpolating functions are smooth the results are less likely to produce error. If the functions have any discontinuities length scale will end up being extremely short and posterior mean will have 'ringing' effects. If the data set is more than two-dimensions, it may be hard to detect errors. The obvious sign there are errors in higher dimensions is the length scale never becomes smaller. This is a classic sign of model misspecification.

B. Squared Exponential GPR

Square Exponential GPR is a function space expression of a radial basis function regression model with infinitely many basis functions. The Squared Exponential GPR is identical to the Exponential GPR except that the Euclidean distance is squared. A fascinating feature utilizing the Square Exponential GPR is it replaces inner products of basis functions with kernels. The advantage to this feature is handling large data sets in higher dimensions will unlikely produce huge errors. Also, it handles discontinuities well. The algorithm of the squared exponential GPR is illustrated as follows.

Algorithm of the Square Exponential GPR

Input 1 and 2 and Procedure 1-3 are the same as the Rational Quadratic GPR.

Procedure 4. The Square Exponential GPR Model becomes:

$$k(x_i,x_j|\theta) = {\sigma_f}^2 \exp\left[-\frac{1}{2}\frac{(x_i-x_j)^T(x_i-x_j)}{{\sigma_l}^2}\right]$$

where:

$$r=\sqrt{(x_i-x_j)^T(x_i-x_j)}$$

C. Matern 5/2 GPR

The Matern 5/2 kernel takes spectral densities of the stationary kernel and create Fourier transforms of RBF kernel. The Matern 5/2 kernel does not have concentration of measure problems for high dimensional spaces. Sample functions from Matérn 5/2 forms are $|\nu - 1|$ times differentiable. Thus, the hyperparameter ν can control the degree of smoothness. The algorithm of the matern 5/2 GPR is illustrated as follows.

Algorithm of the Matern 5/2GPR

Input 1 and 2 and Procedure 1-3 are the same as the Rational Quadratic GPR.

Procedure 4. The Matern 5/2 GPR Model becomes:

$$k(x_i,x_j|\theta) = {\sigma_f}^2\left(1+\frac{\sqrt{3r}}{\sigma_l}\right)\exp\left(-\frac{\sqrt{3r}}{\sigma_l}\right)$$

where:

$$r=\sqrt{(x_i-x_j)^T(x_i-x_j)}$$

D. *Exponential GPR*

Exponential GPR is identical to the Squared Exponential GPR except that the Euclidean distance is not squared. Exponential GPR replaces inner products of basis functions with kernels slower than the Squared Exponential GPR. The Exponential GPR handles smooth functions well with minimal errors, but with discontinuities it does not handle well. The algorithm of the median exponential GPR is illustrated as follows.

Algorithm of the Exponential GPR
Input 1 and 2 and Procedure 1-3 are the same as the Rational Quadratic GPR.
Procedure 4. The Exponential GPR Model becomes: $k(x_i, x_j\|\theta) = {\sigma_f}^2 \exp(-\frac{r}{\sigma_l})$ where: $r = \sqrt{(x_i - x_j)^T (x_i - x_j)}$

III. DATA SET

In Section IV A-D, the banknote authentication data set from the UCI Machine Learning Repository [6] will be used to demonstration the simulation results of the regression models, as shown. There are 1,372 observations with 4 input variables and 1 output variable. The banknote authentication classification involves identifying and classifying counterfeit Banknotes from authentic ones using features or attributes collected from a photograph. It is a binary classification problem, i.e. class (0 for authentic, 1 for inauthentic). In addition, in Section IV E, we will use 19 other data sets from the UCI Machine Learning Repository to compare the error rate of Gaussian process regression (GPR) models with other methods.

IV. EXPERIMENTAL RESULTS

A. *Explore Data and Results in Response Plot*

After a regression model is trained, the regression model results can be displayed by the response plot, i.e. the predicted response versus record number. Holdout or cross-validation is used, thus each prediction is obtained using a model that was trained without using the corresponding observation. Therefore, these predictions are the predictions on the held-out observations. 80% of the data is used to train the network and the remaining 20% data points are used as the testing data.

The response plot of rational quadratic GPR, squared exponential GPR, matern 5/2 GPR, and exponential GPR are shown in Fig. 1, Fig. 2, Fig. 3, and Fig. 4, respectively.

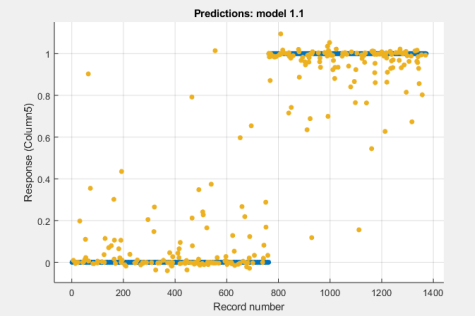

Fig. 1 The response plot of rational quadratic GPR.

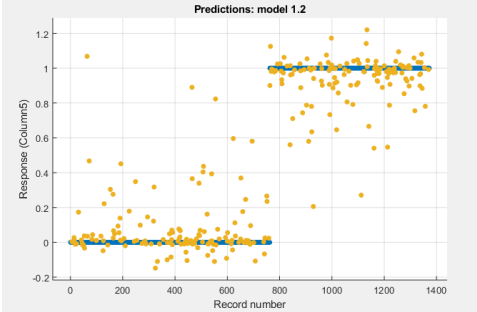

Fig. 2 The response plot of squared exponential GPR.

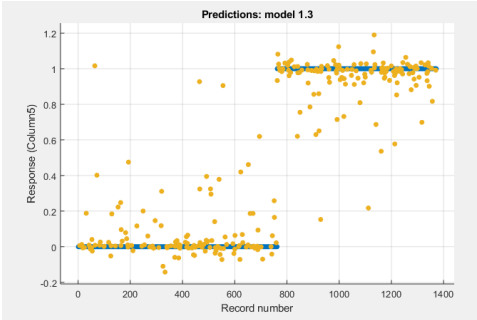

Fig. 3 The response plot of Matern 5/2 GPR.

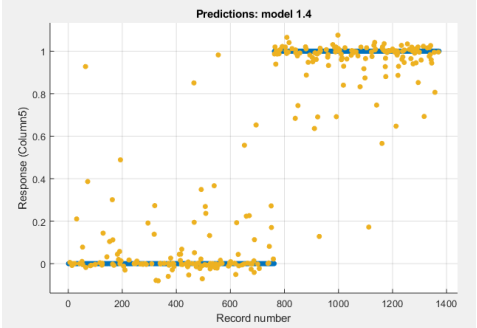

Fig. 4 The response plot of Exponential GPR.

B. *Predicted vs. Actual Response*

The Predicted vs. Actual plot is used to check model performance after training a model. Use this plot to understand how well the regression model makes predictions for different response values.

When the plot is open, the predicted response of our model is plotted against the actual, true response. A perfect regression model has a predicted response equal to the true response, so all the points lie on a diagonal line. The vertical distance from the line to any point is the error of the prediction for that point. A good model has small errors, and so the predictions are scattered near the line. Usually a good model has points scattered roughly symmetrically around the diagonal line. If we can see any clear patterns in the plot, it is likely that we can improve the model.

The predicted vs. actual plot of rational quadratic GPR, squared exponential GPR, matern 5/2 GPR, and exponential GPR are shown in Fig. 5, Fig. 6, Fig. 7, and Fig. 8, respectively.

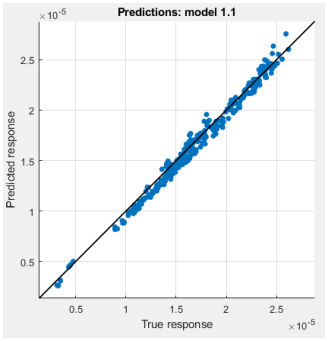

Fig. 5 The Predicted vs. Actual plot of linear SVM.

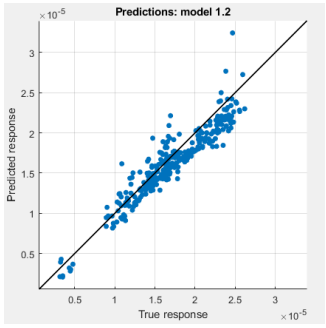

Fig. 6 The Predicted vs. Actual plot of quadratic SVM.

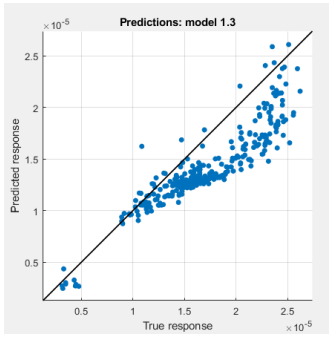

Fig. 7 The Predicted vs. Actual plot of cubic SVM.

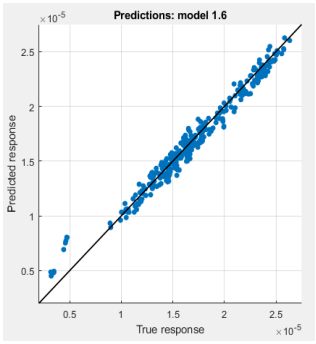

Fig. 8 The Predicted vs. Actual plot of coarse Gaussian SVM.

C. *Evaluate Model Using Residuals Plot*

We further evaluate the model performance by using the residuals plot after training a model. The residuals plot displays the difference between the predicted and true responses. Usually a good model has residuals scattered roughly symmetrically around 0. If we can see any clear patterns in the residuals, it is likely that we can improve the model. We especially look for the following patterns:

- Residuals are not symmetrically distributed around 0.
- Residuals change significantly in size from left to right in the plot.
- Outliers occur, that is, residuals that are much larger than the rest of the residuals.
- Clear, nonlinear pattern appears in the residuals.

The residual plots of rational quadratic GPR, squared exponential GPR, matern 5/2 GPR, and exponential GPR are shown in Fig. 9, Fig. 10, Fig. 11, and Fig. 12, respectively.

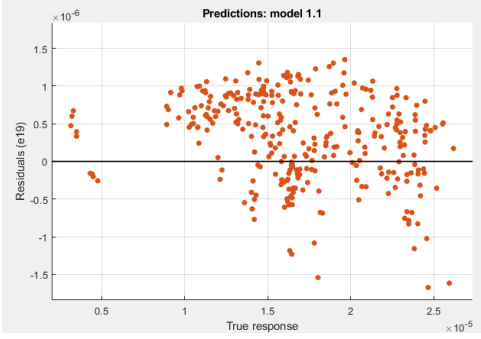

Fig. 9 The residuals plot of rational quadratic GPR.

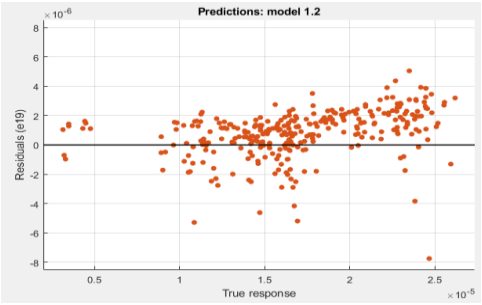

Fig. 10 The residuals plot of squared exponential GPR.

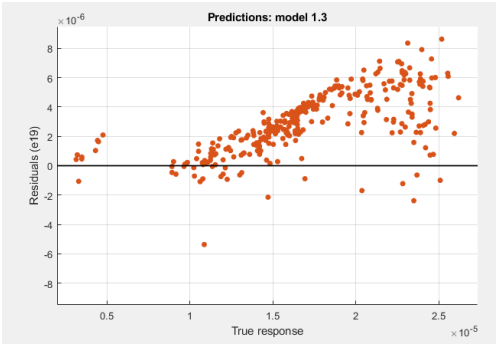

Fig. 11 The residuals plot of matern 5/2 GPR.

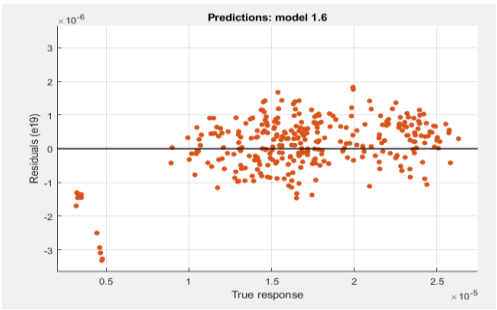

Fig. 12 The residuals plot of exponential GPR.

D. *Model Statistics*

The model parameters are very useful and important to evaluate the performance of different models. For each Gaussian process regression (GPR) algorithm, after the network has been well trained, we evaluate the performance of each featured subset. The comprehensive comparison is shown in Table 1.

Table 1 Comparison of Gaussian Process Regression (GPR) Models on Bank Note Dataset

	RSME	**R-Sq**	**MSE**	**MAE**	**Train Time (sec)**
Rational Quadratic	0.166	0.89	0.0274	0.0704	58
Square Exponential	0.181	0.87	0.0326	0.934	8.3
Matern 5/2	0.172	0.88	0.0295	0.0794	10
Exponential	0.165	0.89	0.0271	0.0698	21

The performance of difference GPR based models are compared using the following model statistics.

- RMSE (Root mean square error). The RMSE is always positive and its units match the units of the response. Look for smaller values of the RMSE.
- R-Squared. Coefficient of determination. R-squared is always smaller than 1 and usually larger than 0. It compares the trained model with the model where the response is constant and equals the mean of the training response. If the model is worse than this constant model, then R-Squared is negative. Look for an R-Squared close to 1.
- MSE (Mean squared error). The MSE is the square of the RMSE. Look for smaller values of the MSE.
- MAE (Mean absolute error). The MAE is always positive and similar to the RMSE, but less sensitive to outliers. Look for smaller values of the MAE.

E. Error Rate Comparison of Gaussian Process Regression (GPR) models with Other Methods

We further apply our GPR classifiers to 19 real world datasets from UCI Machine Learning Repository [7]. Table 2 presents the classification error rates in percentage for these 19 UCI datasets in comparison to the ENN, classic KNN, naive Bayes, linear discriminant analysis (LDA), and the classic multilayer perceptron (MLP) neural network. It shows that ENN always performs better than KNN, and in 17 out of these 19 datasets.

V. CONCLUSION

We propose a Gaussian process regression (GPR) algorithm with variable models to adapt to numerous pattern recognition data for classification. For each GPR algorithm it reveals classification accuracy and minimum feature number objectives. After the network has been well trained, we evaluate the performance of each featured subset. The response plot, predicted vs. actual plot, and residuals plot of rational quadratic GPR, squared exponential GPR, matern 5/2 GPR, and exponential GPR are demonstrated. In addition, a comprehensive comparison of these models is performed in terms of root mean square error, R-squared, mean squared error, and mean absolute error. Furthermore, the classification error rates of these four GPR based models are in comparison to the extended nearest neighbor (ENN), classic k-nearest Neighbor (KNN), naive Bayes, linear discriminant analysis (LDA), and the classic multilayer perceptron (MLP) neural network. The excellent experimental results demonstrated that the Gaussian process regression models provide a very promising feature selection solution to numerous pattern recognition problems. The algorithm is able to learn from the global distribution, therefore improving pattern recognition performance.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation (NSF) grants: HRD #1505509, HRD #1533479，and DUE #1654474.

REFERENCES

[1] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang and D. Tao, "Simultaneous Spectral-Spatial Feature Selection and Extraction for Hyperspectral Images," in IEEE Transactions on Cybernetics, vol. 48, no. 1, pp. 16-28, Jan. 2018.

[2] M. Liu, C. Xu, Y. Luo, C. Xu, Y. Wen and D. Tao, "Cost-Sensitive Feature Selection by Optimizing F-Measures," in IEEE Transactions on Image Processing, vol. 27, no. 3, pp. 1323-1335, March 2018.

[3] X. Wen, L. Shao, W. Fang and Y. Xue, "Efficient Feature Selection and Classification for Vehicle Detection," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 25, no. 3, pp. 508-517, March 2015.

[4] C. Cao, R. L. Tutwiler and S. Slobounov, "Automatic Classification of Athletes With Residual Functional Deficits Following Concussion by Means of EEG Signal Using Support Vector Machine," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 16, no. 4, pp. 327-335, Aug. 2008.

[5] Vapnik, V. The Nature of Statistical Learning Theory. Springer, New York, 1995.

[6] M. Lichman. (2013). UCI Machine Learning Repository. School of Information and Computer Science. Irvine, CA: Univ. California. Available: http://archive.ics.uci.edu/ml/.

[7] B. Tang and H. He, "ENN: Extended Nearest Neighbor Method for Pattern Recognition," IEEE Computational Intelligence Magazine, vol.10, no.3, pp.52 - 60, Aug. 2015.

Table 2 Error Rate Comparison of Gaussian Process Regression (GPR) models with Other Methods

Dataset	Rational Quadratic	Square Exponential	Matern 5/2	Exponential	ENN	KNN	Naïve Bayes	LDA	Neural Network
Ionosphere	**0.363** ± 2.18	0.345 ± 1.98	0.357 ± 2.07	0.393 ± 2.19	17.35 ± 2.69	18.55 ± 2.94	19.83 ± 2.86	20.68 ± 3.00	18.48 ± 2.90
Vowel	**0.476** ± 2.77	0.539 ± 3.22	0.487 ± 2.57	0.552 ± 3.58	8.50 ± 1.92	11. 73 ± 2.80	43.90 ± 2.98	40.94 ± 1.97	45.17 ± 3.15
Sonar	**0.332** ± 2.56	0.338 ± 2.62	0.333 ± 2.52	0.312 ± 2.38	22.67 ± 3.97	22.49 ± 4.06	29.22 ± 4.16	33.75 ± 5.11	27.24 ± 4.37
Wine	0.170 ± 1.38	**0.170** ± 1.30	0.166 ± 1.36	0.162 ± 1.32	4.49 ± 2.16	7.08 ± 2.20	5.07 ± 1.71	2.58 ± 0.59	7.21 ± 2.88
Breast Cancer	0.377 ± 1.61	0.378 ± 1.63	**0.372** ± 1.56	0.330 ± 1.35	4.04 ± 0.87	4.44 ± 1.07	5.76 ± 1.04	5.68 ± 1.18	4.57 ± 1.17
Haberman	**0.421** ± 3.38	0.424 ± 3.51	0.425 ± 3.54	0.423 ± 3.61	31.32 ± 6.53	32.13 ± 5.79	36.65 ± 10.85	34.63 ± 9.92	37.40 ± 10.58
Breast Tissue	**0.160** ± 1.11	0.170 ± 1.38	0.178 ± 1.78	0.170 ± 1.78	36.71 ± 6.37	42.40 ± 6.19	44.02 ± 6.18	41.24 ± 6.60	67.62 ± 5.22
Movement Libras	**1.52** ± 0.92	1.82 ± 1.52	1.60 ± 1.02	1.62 ± 1.02	26.3 ± 2.88	32.16 ± 2.97	45.41 ± 3.39	39.90 ± 3.31	40.87 ± 4.34
Mammographic Masses	0.318 ± 1.44	0.318 ± 1.78	**0.312** ± 1.44	0.322 ± 1.55	21.16 ± 1.43	22.27 ± 1.55	18.96 ± 1.57	19.17 ± 1.72	49.40 ± 0.29
Segmentation	0.237 ± 1.78	**0.235** ± 1.71	0.245 ± 1.79	0.267 ± 1.23	24.71 ± 3.07	27.85 ± 3.04	12.64 ± 2.93	12.79 ± 2.88	23.06 ± 5.95
ILPD	**0.423** ± 0.362	0.458 ± 3.62	0.439 ± 3.58	0.467 ± 7.58	40.0 ± 3.58	40.91 ± 3.68	26.87 ± 2.69	29.64 ± 3.39	32.09 ± 3.53
Pimma Indians Diabetes	0.423 ± 3.44	0.423 ± 2.44	0.439 ± 2.44	**0.413** ± 3.44	31.22 ± 2.15	33.08 ± 2.69	29.44 ± 2.19	28.29 ± 2.01	25.38 ± 2.77
Knowledge	0.523 ± 3.48	0.477 ± 2.46	0.439 ± 2.43	**0.413** ± 3.22	23.93 ± 4.69	27.11 ± 4.45	12.66 ± 2.45	6.97 ± 2.53	14.42 ± 3.86
Vertebral	26.43 ± 1.44	**22.01** ± 2.48	26.22 ± 2.44	23.56 ± 3.11	35.1 ± 4.83	37.64 ± 5.06	47.93 ± 3.41	36.88 ± 4.83	45.11 ± 3.12
Magic	0.411 ± 1.44	**0.323** ± 1.64	0.339 ± 2.44	0.567 ± 2.44	20.10 ± 0.33	20.42 ± 0.36	25.69 ± 0.61	23.30 ± 0.34	29.62 ± 0.38
Pen Digits	0.423 ± 1.44	0.423 ± 2.45	0.439 ± 2.55	**0.413** ± 3.55	0.74 ± 0.15	0.94 ± 0.17	15.38 ± 0.41	11.22 ± 0.52	11.65 ± 0.70
Faults	0.413 ± 2.44	0.113 ± 2.44	0.344 ± 2.44	0.413 ± 3.44	0.91 ± 0.52	1.65 ± 0.86	**0.00** ± 0.00	**0.00** ± 0.00	**0.00** ± 0.00
Letter	12.17 ± 0.38	11.70 ± 0.15	20.170 ± 0.38	20.17 ± 0.22	**5.60** ± 0.25	7.44 ± 0.25	40.09 ± 0.47	29.80 ± 0.37	28.33 ± 0.52
Spam	0.219 ± 1.27	0.239 ± 1.57	0.231 ± 1.44	**0.217** ± 1.23	10.08 ± 0.59	11.52 ± 0.63	11.52 ± 0.78	9.64 ± 0.61	15.32 ± 1.02