Sammon Projected Feature Selection Based Linear Support Vector Regression for Big Data Predictive Analytics

Anita M[a], and Dr. Shakila S[b]

[a,b]
The Governement Arts College, Department of Computer Science, Trichy. 22,
anitarajkumar040908@gmail.com , shakilamuthusamy@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: In big data analytics, classification is a key problem to be resolved for providing efficient prediction results. Recently, many research works have been designed for big data classification. However, accuracy was not increased with minimal time complexity when considering large dataset as input. In order to handle such issues, Sammon Projected Feature Selection based Linear Support Vector Regression (SPFS-LSVR) Method is introduced. The SPFS-LSVR method is developed to minimize the time and space complexity involved in the classification with better accuracy. At first, SPFS-LSVR method takes big dataset as input where it includes number of features and data for executing predictive analytics. Then theSammon Projection is employed in SPFS-LSVR methodfor performing the feature selection in high dimensional data. Sammon Projection helps in finding the relevant features by mapping the high-dimensional space to the lower-dimensionality space. This in turns, the time and space complexity occurred during the classification is reduced. After that, SPFS-LSVR method uses a Linear Support Vector Regression (LSVR) model for examining the selected features of input data with higher accuracy. The designed model uses hyperplane for producing the exact prediction result with higher accuracy. The LSVR model determines the relationship between independent data (i.e., features of input data) and the dependent data (i.e., prediction outcomes) with the help of Laplace kernel function. From that, the maximum relationship between the input features of data is classified into different classes. This helps in SPFS-LSVR method to enhance the performance of classification with maximum accuracy and minimal error rate. Through the efficient classification performance of big data, proposed SPFS-LSVR method improves the big data predictive analytics process as compared to state-of-the-art works. Experimental evaluation is carried out using big dataset on factors such as prediction time, prediction accuracy, false positive rate and space complexity with respect to number of data.

Keywords: Big data, Sammon Projection, feature selection, Linear Support Vector Regression, high-dimensional space, lower-dimensionality space, prediction outcomes, classification

1. Introduction

Big data analytics is the concept for evaluating vast volume of data. It also examines uncover information such as the hidden patterns and undefined correlations for taking the future business decisions. Classification is a significant tool in big data analytics. Classification is the process of partitioning given data into the diverse relevant classes. Some of the existing methods based on classification provide the better prediction results in big data analytics. But, the existing classification technique unable to improves the prediction accuracy and reduces the prediction time. In order to address these existing problems, several machine learning techniques can be used for big data analytics.

Ensemble learning was introduced in [1] for predicting big time series data. The designed model uses the three regression models for increasing the accuracy of prediction. However, the processing of large data increases the time complexity. Cuckoo-Grey wolf based Correlative Naive Bayes classifier and MapReduce Model (CGCNB-MRM) was developed in [2] to categorize the big data. However, the accuracy of classification was not improved efficiently.

A novel algorithm was implemented in [3] for detecting big time series data with higher accuracy. But, the time complexity was not reduced. In [4], Support Vector Regression (SVR) model was designed for agricultural drought detection using climate data. However, the error rate was not reduced effectively.

A novel prediction model using decision trees, bagging, random forests, and boosting were introduced in [5] to identify the severe rain damage. However, the accuracy was not improved. In [6], an efficient classifier namely online bagging ensemble model was designed to learn the big data stream. The model provided better accuracy but, time consumption was not addressed.

Supervised classification algorithm was designed in [7] for identifying the changes in the impacts of climate conditions on buildings. But, space complexity was not handled. A novel multi-model ensemble approach was introduced in [8] to predict the climate changes. But, the error rate was not effectively minimized.

In [9], self-adaptive stream data classification was presented to categorize the stream data. However, accuracy was not increased with less time. Distributed fuzzy associative classification was designed in [10] for big data classification. But, misclassification rate was not decreased.

In order to resolve the above mentioned existing issues such as lower prediction accuracy, higher prediction time and space complexity and false positive rate in big data classification, SPFS-LSVR method is introduced. The main contribution of SPFS-LSVR method is described in below.

- To achieve improved classification performance for big data, SPFS-LSVR method is proposed in this research work.
- To reduce the prediction time and space complexity during the data classification, Sammon projection based feature selection is performed in SPFS-LSVR method.
- To improve the prediction accuracy with less false positive rate, linear support vector regression model is used where it classifies the input data into either side of the hyperplane using Laplace kernel function.

2. Sammon Projected Feature Selection Based Linear Support Vector Regression Method

Big data analytics is the procedure of investigating large sets of data to determine the patterns and other useful information. In big data analytics, classification techniques play a significant role to provide efficient prediction results. With the process of large volume of data, time complexity during prediction is increased. Recently, many classification methods were designed to predict the future results. But, the existing techniques failed to improve the prediction accuracy and reduce the prediction time. In order to handle these problems in big data, Sammon Projected Feature Selection based Linear Support Vector Regression (SPFS-LSVR) Method is proposed. The proposed SPFS-LSVR Method improves the future prediction performance with better accuracy and time.

In SPFS-LSVR method, Sammon projection is employed to perform feature selection where it maps the features from high dimensional space to the low-dimensional space. This helps for SPFS-LSVR method to choose the more significant features for accurate prediction with minimal time. In addition, Linear Support Vector Regression (LSVR) model is employed in SPFS-LSVR method for categorizing the selected input features of data for future result prediction. This assists to classify the data into different classes with maximum accuracy.

Figure 1 demonstrates an architecture diagram of SPFS-LSVR method for achieving the future prediction with maximum accuracy and minimal time. The SPFS-LSVR method comprises the two processes such as feature selection and classification for predicting future results. At first, the number of features and data are gathered as input from the big dataset. After obtaining the input, the relevant feature selection is performed with the help of Sammon projection. Sammon Projection employed to determine the relevant features by measuring the distance between the input features. This aids to decreases the time complexity in big data prediction.

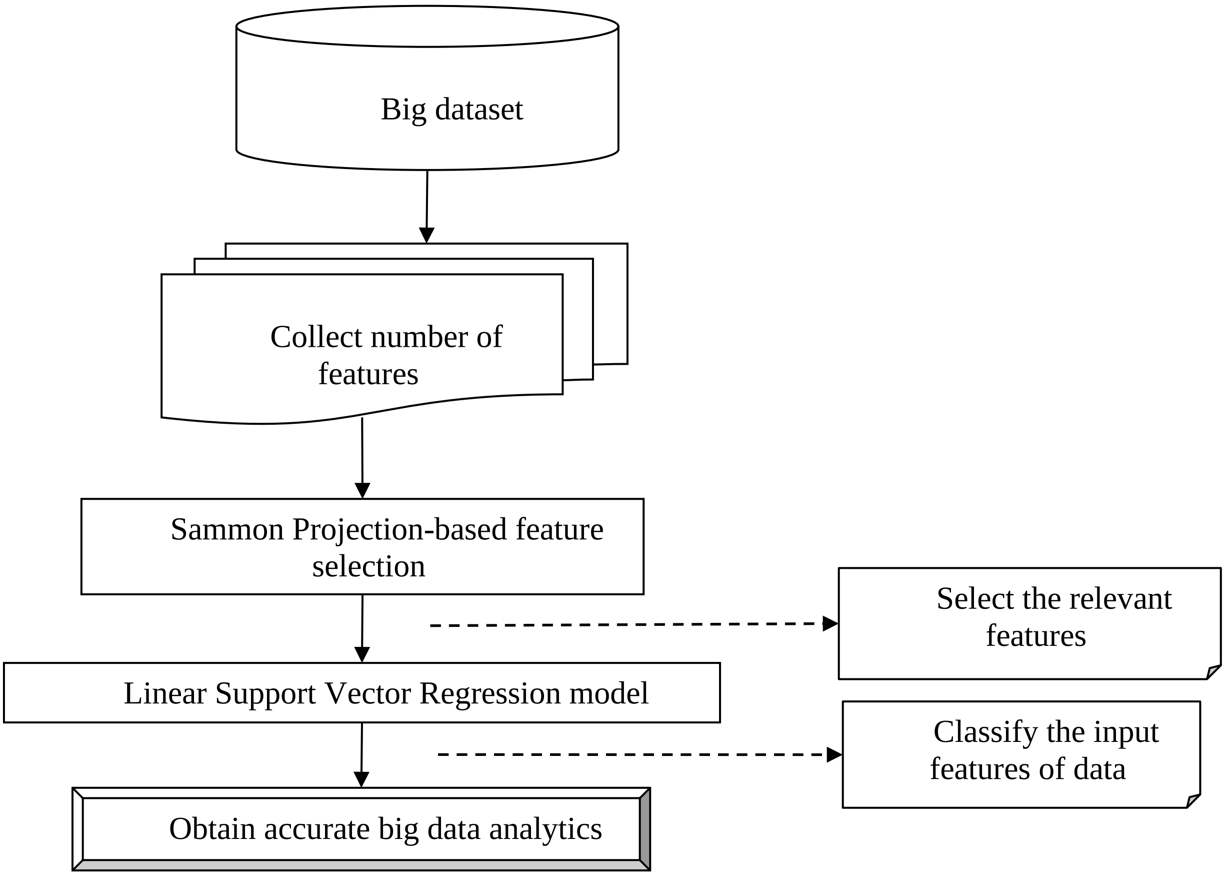
Big dataset
Collect number of features
Sammon Projection-based feature selection
Select the relevant features
Linear Support Vector Regression model
Classify the input features of data
Obtain accurate big data analytics

Figure 1 Architecture Diagram of Sammon Projected Feature Selection based Linear Support Vector Regression Method for Predictive Analytics

Then, the Linear Support Vector Regression model (LSVR) is carried out to analyze the selected features of the input data for predicting results. In LSVR, hyperplane is constructed to provide exact future prediction result.

This helps to increase the performance of prediction with less error and high accuracy. The above explained processes of SPFS-LSVR method are elaborately described in following sections.

1.1 Sammon Projection-based feature selection

Feature selection is the first process in SPFS-LSVR method to predict the big data with less time complexity. The main aim of feature selection is to reduce the space and time complexity involved during the prediction process. The above said goal of feature selection is achieved with the aid of choosing only relevant features from the big dataset. In SPFS-LSVR method, Sammon Projection-based feature selection is carried out to pick the significant features from the other features in the dataset.

Sammon projection is introduced byJohn W. Sammon. Hence the name is called Sammon projection. Sammon projection is a nonlinear projection method for analyzingdifferent kinds of data. The main use of the projection is visualization.It is helpful in preliminary investigation in all geometric pattern identification. Sammon projection is derived based on the multidimensional scaling methods. Sammon projection refers the input features in a lower dimensional space where the distance of data features are protected. Sammon projection in SPFS-LSVR method maps the features into low dimensional space for selecting the relevant features.

By using Sammon projection, the relevant features are easily extracted to lessen the complexity during the prediction process. Let us consider the big dataset BD with number of features.

$$k_i = \{k_1, k_2, k_3, \ldots . k_n\} \in BD \qquad (1)$$

In (1),$k_i$represents a number of features$\{k_1, k_2, k_3, \ldots . k_n\}$, BDdenotes aninput big dataset. Then the Sammon projection is applied on the input features for choosing the relevant features by means of mapping the features in higher-dimensional space to a lower-dimensional space. In SPFS-LSVR, Sammon projection tries to protect the inter-point distancestructure when the features are transformed to a lower-dimensional space. The preservation of this inherent structure is obtained by maintaining the distances between features under projection. Thus, distance between features in two dimensional space is mathematically computed as follows,

$$ED(k_i, k_j) = \sum_{i=1}^{n} |X_i - Y_i| \qquad (2)$$

In the above equation (2), '$ED(k_i, k_j)$' denotes a Euclidean Distance between the two features and 'X, Y' denotes a coordinates in two dimensional space.In this way, the distance between all the features in the big dataset is computed to choose the relevant features using Sammon projection. The Euclidean distance between two features is smaller, the features are projected to a lower-dimensional space and the features are selected as relevant for predictive analysis. If the distance between two features is high, the features are not projected to a lower-dimensional space and the feature is said to be irrelevant.Therefore, the Sammon projection chooses the minimal distance features for mapping it into a lower-dimensional space. It is mathematically expressed as given below.

$$SP \rightarrow arg\ min\ ED(k_i, k_j) \qquad (3)$$

In equation (3), 'SP' denotes the Sammon Projection and '$arg\ min\ ED(k_i, k_j)$' refers the features with minimal distance. Based on the above equation, feature with minimal distance is chosen to map the high-dimensional space to low-dimensional space. With this, the relevant features are obtained in SPFS-LSVR method and thus reduce the time complexity in prediction process. The algorithm for Sammon projection based feature selection is described as follows.

Algorithm 1: Sammon projection based feature selection

```
Input: Big dataset $\boldsymbol{BD}$, Number of features $\boldsymbol{k_1, k_2, k_3, \ldots . k_n}$
Output: Select relevant features
Begin
    1. For each feature $\boldsymbol{k_i}$ in $\boldsymbol{BD}$
    2. Compute the Euclidean Distance between the two features using (2)
    3. Choose the features with less distance using (3)
    4. Project the minimal distance features to the low-dimensional space
    5. End for
End
```

Algorithm 1 explains the process of relevant feature selection using Sammon projection for performing big data predictive analysis. To start with the feature selection process, the number of features from the big dataset is obtained. After that, the Euclidean distance between two features is computed. From that, the features with less distanceare mapped from the high-dimensional space to the lower dimensional space. With this, the features in the lower dimensionality space are taken as relevant for big data predictive analytics. This helps to reduce the space and time complexity in proposed SPFS-LSVR method. Once the relevant features are chosen, the data classification is carried out for providing the efficient predictive results.

2.2 Linear Support Vector Regression Model

After completing the relevant feature selection, data classification is carried out in SPFS-LSVR method to perform predictive analysis. In SPFS-LSVR method, Linear Support Vector Regression (LSVR) model is utilized to analyze the features of input data during the prediction. LSVR model uses the hyperplane to divides the output class labels for providing the accurate prediction results. LSVR model computes the association between independent data (i.e., features of input data)and dependent data (i.e., prediction outcomes)with less time and accuracy. The process of LSVR model for data classification is illustrated in Figure 2.

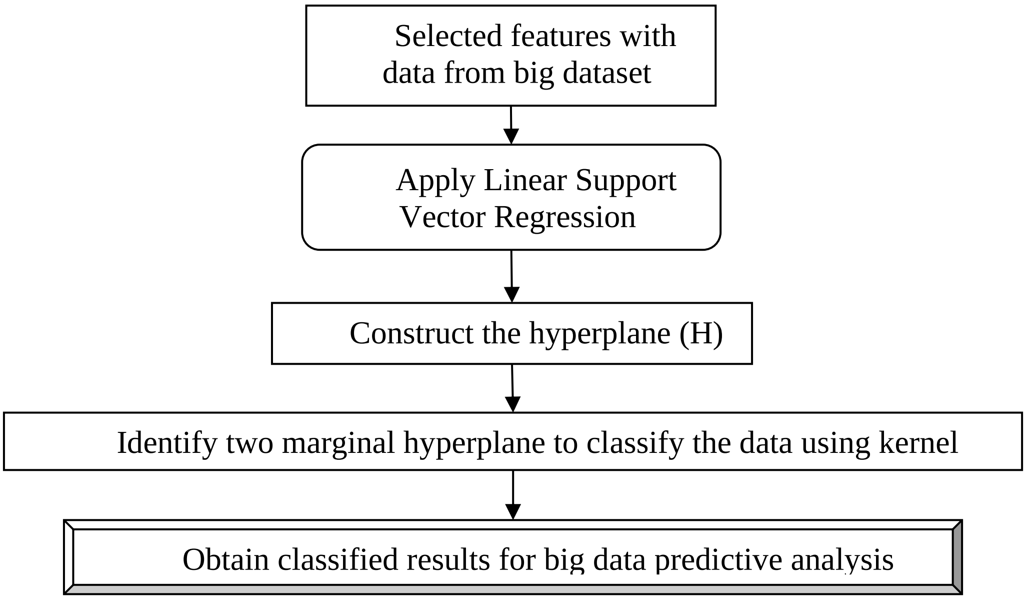
Selected features with data from big dataset
Apply Linear Support Vector Regression
Construct the hyperplane (H)
Identify two marginal hyperplane to classify the data using kernel
Obtain classified results for big data predictive analysis

Figure 2 Linear Support Vector Regression Model for Data Classification

The above figure 2 illustrates the process of LSVR model for classifying the input features of data with higher accuracy and less time. At first, the number of selected features and data is taken as input from big dataset. Then thelinear support vector regression modelis employed in proposed LSVR model to classify the data with selected features. In the classification process, hyperplane is constructed for separating the original vector space into two sets. After that, the Laplace kernel function is applied to compute the association between features of input data to attain the prediction outcomes. According to the relationship between data, the two marginal hyperplane is constructed to categorize the input data. The relationship between two input features of data is high, the data is classified into different classes. From that, the prediction accuracy isimproved with minimal time in SPFS-LSVR method.

The LSVR model considers the set of training samples '$\{(d_1, x_1), (d_2, x_2), \dots (d_n, x_n)\}$' where '$d_n$'represents the number of input data and 'x_n' 'represents the output of classification results. The result of classification provides two classes as$x_i \in \{+1, -1\}$ where $x_i = +1$ indicates the input data is correctly classified into different classes and $x_i = -1$ indicates input data is not classified into different classes. LSVR model uses the hyperplane to classify the input data with selected features. The hyperplane is considered as a decision boundary between the two classes where the data is classified on either side of the decision boundary. Thus, the hyperplane in LSVR model is provided as follows,

$$x_i = \vec{v}.d_i + \vec{a} \qquad (4)$$

$$H \rightarrow \vec{v}.d_i + \vec{a} = 0 \qquad (5)$$

In the above equation (4) and (5), 'H' indicates a hyperplane (i.e. boundary), '$\vec{v}$' indicates the normal weight vector to the hyperplane, '$\vec{a}$' refers a bias and 'd_i' denotes input data. In LSVR, two marginal hyperplanes are taken as lower and upper side of the decision boundary when the training sets are linearly separable. It is given by,

$$m_1 \rightarrow v.d_i + a > 0 \text{ for } d_i = \text{ '}+1\text{'} \ (6)$$

$$m_2 \rightarrow v.d_i + a < 0\text{for } d_i = \text{ '}-1\text{'} \ (7)$$

In the above equation (6) and (7),m_1 and$m_2$represents the lower and upper marginal hyperplanes for classifying the input big data into above and below the boundary. Then, the distance between two hyperplane is provided as,

$$D = \frac{2}{\|\vec{v}\|} \qquad (8)$$

In the above equation (8), 'D' represents a distance between two marginal hyperplane 'm_1' and 'm_2'. The predicted output of (x) LSVR model is obtained with the aid of kernel function which is expressed as follows,

$$x = sign \sum v.K(d_i, d') \qquad (9)$$

In the above expression (9),'x'indicates a predicted classification results, 'v' indicates a training sample weights, 'K' denotes a kernel function which measures the input data relationshipand '$sign$' indicates the output of predicted results either positive or negative. In LSVR model, Laplace kernel function is used to determine the predicted output of classification results. It is obtained by,

$$K = exp\left(-\frac{\|d_i - d'\|}{\sigma}\right) \qquad (10)$$

In the above equation (10), 'K' indicates Laplace kernel function for predicting classification results, 'σ' indicates the deviation and 'd_i' and 'd'' denotes relationship between features of input data. The output ofLaplace kernelpredictionresults is given as,

$$K = \begin{cases} +1 & data\ is\ classified\ into\ different\ classes \\ -1 & data\ is\ not\ classified\ into\ different\ classes \end{cases} \qquad (11)$$

In the above equation (11), the positive results provide the maximum relationship between input data and it is classified into diverse classes using Laplace kernel function. Whereas the negative results present minimal relationshipbetween the input data i.e. data is not classified for big data predictive analysis. The process diagram ofLinear Support Vector Regression model is demonstrated in figure 3.

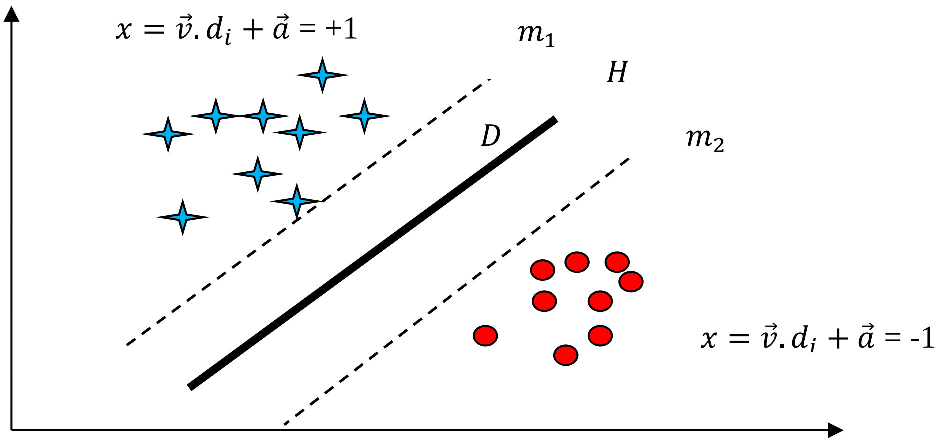
$x = \vec{v}.d_i + \vec{a}$ = +1
m_1
H
D
m_2
$x = \vec{v}.d_i + \vec{a}$ = -1

Figure 3 Process of Linear Support Vector Regression Model

As depicted in above figure 3, the data classification is performed using linear support vector regression model. As represented in above figure 3, the input features of data is classified into positive (upper side) and negative classes (lower side of the hyperplane) with higher accuracy and less time. The algorithm of LSVR model is described as follows.

Algorithm: 2 Linear Support Vector Regression Model

```
Input: Number of input data $d_1, d_2, d_3, .... d_n$ with selected features
Output: Improve the prediction accuracy with less time
Begin
    For each input features of data $d_i$
      Construct hyperplane $H$
      Find two marginal hyperplane $m_1, m_2$
      Calculate distance between $m_1$and  $m_2$
      Compute the output of classifier
      Determine the Laplace kernel function
      If the relationship between $(d_i, d')$ is high
            $x = +1$
          Data is classified into  different classes
    Else
    $x = -1$
          Data is not classified into different classes
    End if
    End for
End
```

Algorithm 2 describes the step by step process of data classification using LSVR model for achieving the better prediction outcomes. The selected relevant features and the data are considered as input for classification process. Then theseparating hyperplane is used in LSVR model in order to label the input features of data. With the help of hyperplane, the original vector is divided into two sets. At every instant, the features of input data are classified on either side of the hyperplane. This is carried out by measuring relationship between the data through the Laplace kernel function. If the relationship between the data is maximum, then the input features of data is categorized into different classes. Otherwise, the data is not classified to the different classes. This in turns, the classification process of LSVR model in SPFS-LSVR method increases the big data predictive analytics with higher accuracy and minimal time.

3. Result and Discussions

In this section, the performance result analysis of SPFS-LSVR method is discussed. The performance of SPFS-LSVR method is compared with existing Ensemble learning [1] and CGCNB-MRM [2] respectively. The effectiveness of SPFS-LSVR method is evaluated along with the following metrics with the help of tables and graphs. At first, performance of big data predictive analytics is carried by measuring the prediction accuracy metric.

Prediction accuracy is defined as the ratio of number of accurate classification of data to the total number of input data. The prediction accuracy is mathematically determined using below mentioned formula,

$$Pre_a = \frac{ACd_i}{d_n} \times 100 \qquad (12)$$

In the above equation (12), 'Pre_a' denotes prediction accuracy, 'ACd_i' denotes the number of accurate classification of data and 'd_n' indicates the total number of data. The value of 'Pre_a' is computed in percentage (%).

Table 1 Comparative analysis of prediction accuracy

Instance ofMeteorological data	Prediction accuracy (%)		
	Existing CGCNB-MRM	Existing Ensemble learning	Proposed SPFS-LSVR method
500	74	80	87
1000	73	78	85
1500	72	77	83
2000	74	79	86
2500	71	77	85
3000	70	75	84
3500	75	80	89
4000	76	82	90
4500	75	81	87
5000	78	84	91

The above table 1 demonstrates the performance analysis of prediction accuracy using three methods with respect to the different number of data. To conduct the experiments, the number of data is taken as 500 to 5000 from El-Nino dataset. The experiment is conducted by comparing the proposed SPFS-LSVR method with existing methods such as Ensemble learning [1] and CGCNB-MRM [2]. From table1, it is observed that the proposed SPFS-LSVR methodsignificantly enhances the number of data that are preciously classified than the other two existing methods. Based on the values in table 1, the graph is plotted for prediction accuracy versus number of data is shown in figure 4.

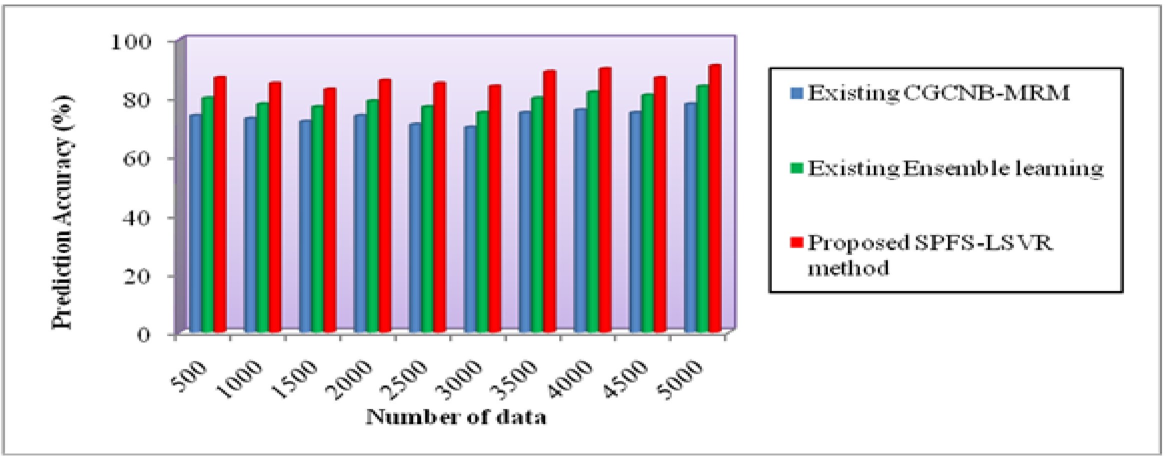

Figure 4 Performance of prediction accuracy

Figure 4 illustrates the experimental results of prediction accuracy using three different classification techniques namely SPFS-LSVR method, Ensemble learning [1] and CGCNB-MRM [2].The numbers of data are taken as input for calculating the prediction accuracy. Number of data is taken in 'x' axis and the prediction accuracy are attained at 'y' axis. The figure confirms that the prediction accuracy is improved using proposed SPFS-LSVR method when compared to existing Ensemble learning [1] and CGCNB-MRM [2]. Let us consider, number of data as 5000, the prediction accuracy of existing Ensemble learning [1] and CGCNB-MRM [2] obtains the prediction accuracy as 84% and 78% respectively. Besides, proposed SPFS-LSVR method attains the prediction accuracy as 91% for big data prediction. This is observed as the highest value of accuracy than the other methods.

The better enhancement of prediction accuracy is achieved with the help of applying both feature selection and classification process. At first, Sammon projection is employed to determine the more significant features for future result prediction. Then the LSVR model is applied is classify the input data by means of estimating the relation between input features of data. This aids to increases the prediction accuracy of SPFS-LSVR method. As a result, the performance of prediction accuracy using proposed SPFS-LSVR method is improved by 9% when compared to existing Ensemble learning [1]. Similarly, the prediction accuracy of proposed SPFS-LSVR method is increased by 17% when compared to existing CGCNB-MRM [2].

In proposed SPFS-LSVR method, prediction time is another metric to analyze the performance of big data prediction. Prediction time is measured as the amount of time utilized for future result prediction through the classification process. The formula for prediction is given by,

$$T_p = d_n \times Td_i \qquad (13)$$

In the above equation (13), 'T_p' indicates the prediction time, d_n indicates the number of data and 'Td_i' denotes the time taken for single data. Prediction time is measured in terms of milliseconds (ms).

Table 2 Comparative analysis of prediction time

Instance of Meteorological data	Prediction time (ms)		
	Existing CGCNB-MRM	Existing Ensemble learning	Proposed SPFS-LSVR method
500	31	27	22
1000	34	30	25
1500	37	33	28
2000	39	36	32
2500	42	38	35
3000	46	43	38
3500	49	46	42
4000	53	49	45
4500	55	51	48

5000	58	54	50

Above table 2 demonstrates the comparison analysis of prediction time for three methods such as proposed SPFS-LSVR method, Ensemble learning [1] and CGCNB-MRM [2]. The different number of data is considered as input which is varied from 500 to 5000. During the experimental conduction, three methods perform big data classification with less amount of time complexity. Comparatively, the prediction time is effectively reduced with the help of proposed SPFS-LSVR method as compared to other existing methods. According to the values in table 2, the graph is plotted between a number of data and prediction time as shown in figure 5.

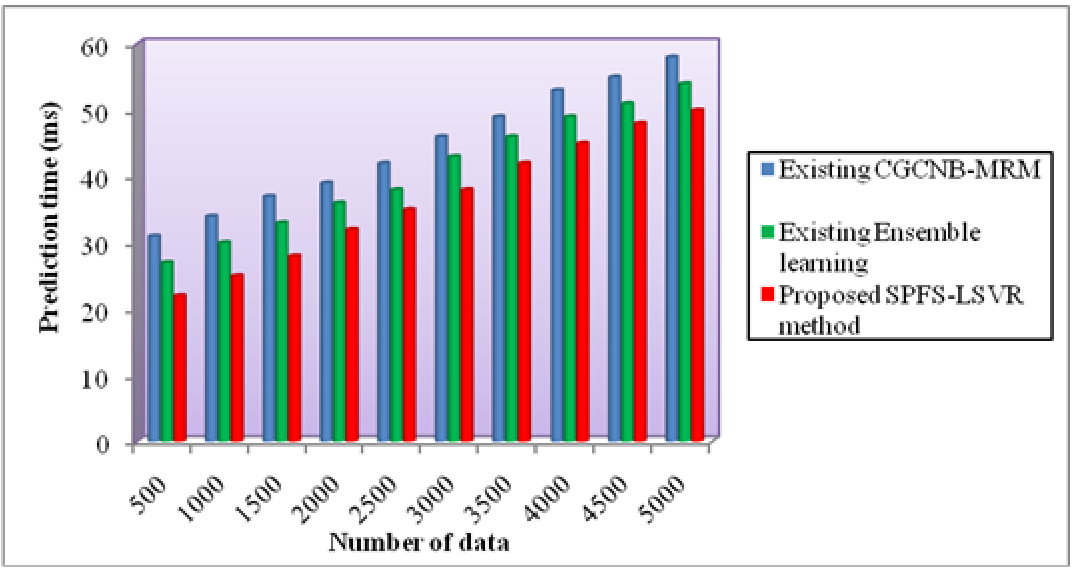

Figure 5 Performance of prediction time

Figure 5 shows the performance results of prediction time based on the number of data. The above figure depicts that the performance results of prediction time with three classification techniques. In the two dimensional graphical outcomes, the number of data is taken for computing the prediction time. In figure 5, the red, green and blue, indicates the performance results of prediction time using three techniques SPFS-LSVR method, Ensemble learning [1] and CGCNB-MRM [2] respectively. The figure evidently shows that the SPFS-LSVR method reduces the prediction time than other two existing classification techniques. For example consider number of data as 500 in first iteration, the prediction time of SPFS-LSVR method, Ensemble learning [1] and CGCNB-MRM [2] is observed as 22 ms, 27 ms and 31 ms respectively. In the above discussion SPFS-LSVR method consumes less time for classifying the input data.

On the contrary to existing works, the proposed SPFS-LSVR method uses the Sammon projection to pick the more relevant features. With this, only minimal number of features is processed to get the future results. In contrast to classifying all the data in big dataset, the LSVR model is categorize the selected features of input data. This in turns, the time required to predict the future result is considerably decreased in proposed SPFS-LSVR method than the existing methods. Therefore, the prediction time of SPFS-LSVR method is reduced by 11% and 19% when compared to existing Ensemble learning [1] and CGCNB-MRM [2] respectively.

The third evaluation metric of SPFS-LSVR method is false positive rate. It is defined as the proportion of number of data incorrectly classified to the total number of data considered for experiments. False positive rate is mathematically computed as given below.

$$FPR = \frac{NICd_i}{d_n} \times 100 \qquad (14)$$

In the above equation (14), 'FPR' refers a False Positive Rate, '$NICd_i$' refers a number of data incorrectly classified and 'd_n' is the total number of data. FPR is determined in terms of percentage (%).

Table 3 Comparative analysis of false positive rate

Instance of Meteorological data	**False positive rate (%)**		
	Existing CGCNB-MRM	**Existing Ensemble learning**	**Proposed SPFS-LSVR method**
500	26	20	13
1000	27	22	15

1500	28	23	17
2000	26	21	14
2500	29	23	15
3000	30	25	16
3500	25	20	11
4000	24	18	10
4500	25	19	13
5000	22	16	9

The above table 3 describes the comparison results of false positive rate for three different methods while taking the different number of data as input with the range of 500 to 5000. As given in table 3, the experiment is conducted for verifying the effectiveness of the proposed SPFS-LSVR method and it is compared with existing Ensemble learning [1] and CGCNB-MRM [2].In all the iterations, the rate of false positives using proposed SPFS-LSVR method is significantly reduced when compared to two existing methods. According to the values in above table, the graph is plotted between a number of data and false positive rate as shown in figure 6.

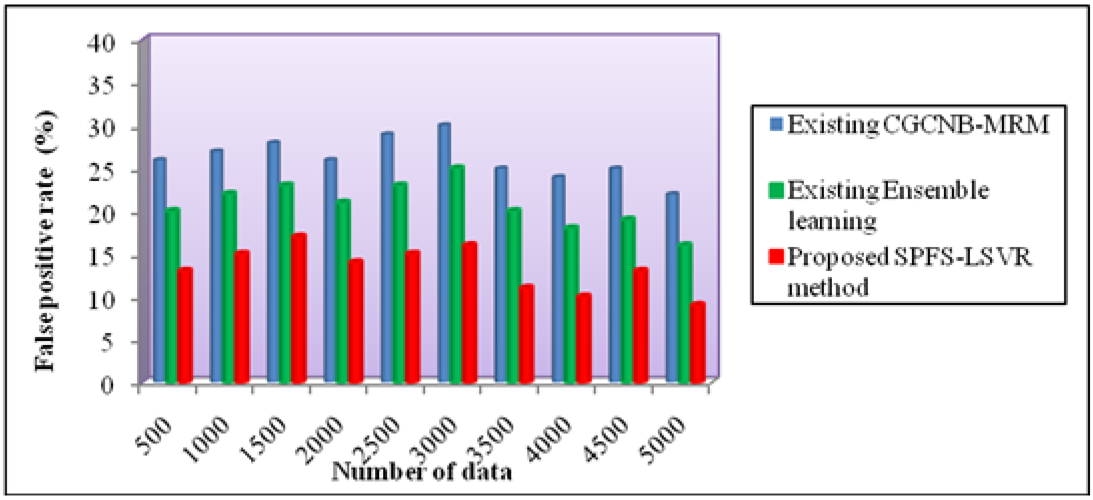

Figure 6 Performance of false positive rate

Figure 6reveals the experimental results of false positive rate based on the number of data. The above figure shows that the performances of the false positive rate with three different techniques. In the above visual comparison chart, the number of data is taken as input and the corresponding false positive results are attained as an output. The graphical results clearlydemonstrate that the false positive rate of proposed SPFS-LSVR method highly minimized than the existing Ensemble learning [1] and CGCNB-MRM [2]. Minimal false positive rate of proposed SPFS-LSVR method is obtained with the application of classification process called LSVR model. During the classification, the hyperplane is used to divide the input dataset into two sets depends on the relationship between two data. From that, data with more association is classified as positive class. Otherwise, the data is classified into negative classes. This aids to decreases the error rate in classification process. As a result, the false positive rate of SPFS-LSVR method is greatly decreased than the existing methods.

When considering the 500 data, the false positive rate of Ensemble learning [1] and CGCNB-MRM [2] is obtained as 20% and 26% whereas, the proposed SPFS-LSVR method obtains the false positive rate as 13%. Likewise, all the iteration results are computed to compare the performance of false positive rate. From that, the proposed SPFS-LSVR method minimizes the false positive rate by 36% when compared to existing Ensemble learning [1] and 50% when compared to existing CGCNB-MRM [2] respectively.

The fourth performance metric of SPFS-LSVR method is space complexity. It is amount of storage space needed to store the data. Space complexity is calculated by,

$$S_{com} = d_n \times MSd_i \qquad (15)$$

In above equation (15), 'S_{com}' represents the space complexity, 'd_n' represents number of data and 'MSd_i' is the memory required for storing single data. Space complexity is calculated in terms of Megabytes (MB).

Table 4 Comparative analysis of space complexity

Instance of Meteorological data	Space complexity (MB)		
	Existing CGCNB-MRM	Existing Ensemble learning	Proposed SPFS-LSVR method
500	22	18	15
1000	25	21	18
1500	27	23	20
2000	29	25	23
2500	31	28	25
3000	33	31	28
3500	36	34	30
4000	39	36	33
4500	42	38	36
5000	45	42	39

Above table 4 shows the comparison analysis of space complexity for three methods such as existingEnsemble learning [1], CGCNB-MRM [2] and SPFS-LSVR method. During the experimental conduction, three methods consume lower memory space for storing the data. Comparatively, space complexity is increased with the help of proposed SPFS-LSVR method as compared to other existing methods. According to the values in table 4, the graph is plotted between a number of data versus space complexity is illustrated in figure 7.

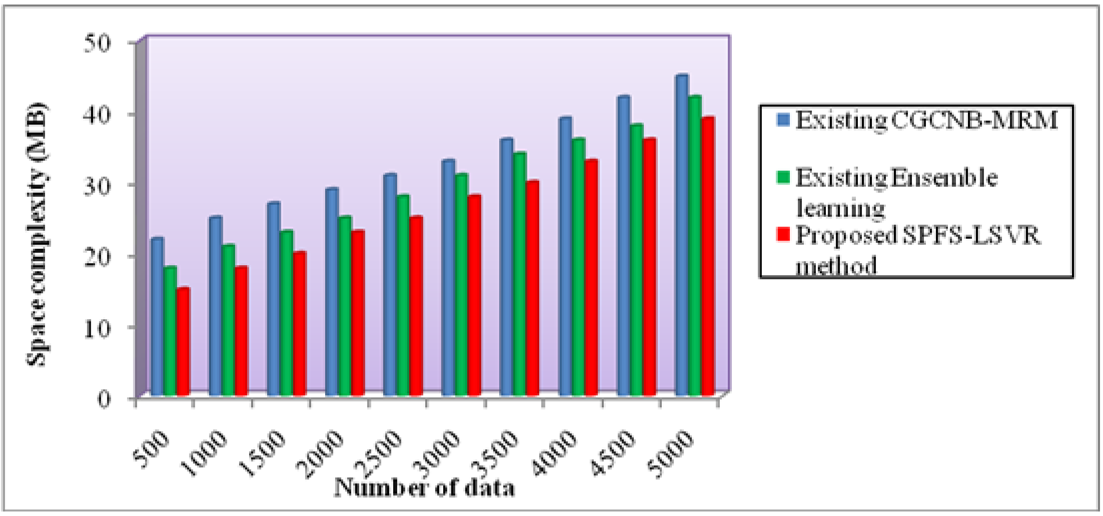

Figure 7 Performance of space complexity

Figure 7 demonstrates the performance result analysis of space complexity versus various numbers of data using three methods namely proposed SPFS-LSVR method, existing Ensemble learning [1] and CGCNB-MRM [2]. As revealed in above figure, SPFS-LSVR method gives lower space complexity when taking big dataset as input than the existing Ensemble learning [1] and CGCNB-MRM [2]. The space complexity of proposed SPFS-LSVR method is attained as 15 MB with the input of 500 data. Whereas, the existing Ensemble learning [1] and CGCNB-MRM [2] provides the space complexity as 18 MB and 22 MB respectively.

The reduction of space complexity is obtained through the process of relevant feature selection in SPFS-LSVR method. Here, the distance between two input features is computed to choose the significant feature for obtaining the prediction outcomes. With the process of lesser number of selected features, the amount of space required to store the data is decreased in SPFS-LSVR method than the existing methods. Thus, the space complexity of proposed SPFS-LSVR method is reduced by 10% and 20% when compared to existing Ensemble learning [1] and CGCNB-MRM [2] respectively.

4. Conclusion

An efficient method called SPFS-LSVR is introduced for big data analytics with higher accuracy and lower time. Developed SPFS-LSVR method uses the Sammon projection and LSVR model for categorizing the big data. From the large volume of data, the more significant features are selected using Sammon projection based feature selection algorithm. This is carried out by mapping the features to the lower-dimensional space from the high-dimensional space through computing theEuclidean distance between two features. With this, the features with less distance are projected to the lower-dimensional space. This aids to lessen the time and space complexity involved in the classification process. In addition, data classification is employed in SPFS-LSVR method by applying LSVR model. In LSVR model, hyperplane is formed to classifying the data with the aid of Laplace kernel function where it finds out the association between input features of data. From that, maximum relationship between the data is labeled into upper side of the hyperplane for classifying it into diverse classes. This in turns, the prediction accuracy is improved with minimal error. The performance of SPFS-LSVR method is evaluated in terms of prediction accuracy, prediction time, space complexity and false positive rate compared with two existing works. The experimental result demonstrates that SPFS-LSVR method provides better performance with an improvement of accuracy and minimization of time and false positive rate as compared to state-of-the-art works.

References

a. A.Galicia, R.Talavera-Llames, A. Troncoso,l. Koprinska, F.Martínez-Álvarez, ‘Multi-step forecasting for big data time series based on ensemble learning’, Knowledge-Based SystemsVolume 163, 1 January 2019, Pages 830-841

b. Chitrakant Banchhor and N. Srinivasu, ‘Integrating Cuckoo search-Grey wolf optimization and Correlative Naive Bayes classifier with Map Reduce model for big data classification’, Data & Knowledge Engineering,Volume 127, May 2020, Pages 1-38

c. R. Talavera-Llames, R. Pérez-Chacón, A. Troncoso, F. Martínez-Álvarez, ‘MV-kWNN: A novel multivariate and multi-output weighted nearest neighbours algorithm for big data time series forecasting’, Neurocomputing, Volume 353, 2019, Pages 56-73

d. Ye Tian, Yue-Ping Xu, Guoqing Wang, ‘Agricultural drought prediction using climate indices based on Support Vector Regression in Xiangjiang River basin’, Science of The Total Environment,Volumes 622–623, 1 May 2018, Pages 710-720

e. Changhyun Choi,Jeonghwan Kim, Jongsung Kim, Donghyun Kim, Younghye Bae, and Hung Soo Kim, ‘Development of Heavy Rain Damage Prediction Model Using Machine Learning Based on Big Data’, Hindawi, Advances in Meteorology, Volume 2018, Pages 1-11.

f. Yanxia Lv, Sancheng Peng, Ying Yuan, Cong Wang, Pengfei Yin, Jiemin Liu, and Cuirong Wang, ‘A Classifier Using Online Bagging Ensemble Method for Big Data Stream Learning’, Tsinghua Science and Technology, Volume 24, Issue 4, 2019, Pages 379 – 388

g. Liu Yang, Kailin Lv, Honglian Li, Yan Liu, Building climate zoning in China using supervised classification-based machine learning, Building and Environment, Volume 171, March 2020,Pages 1-45

h. Getachew Tegegne, Assefa M. Melesse, Abeyou W. Worqlul, ‘Development of multi-model ensemble approach for enhanced assessment of impacts of climate change on climate extremes’,Science of The Total Environment, Volume 704, 2020,Pages 1-43

i. Shizhuo Deng , Botao Wang, Shan Huang, Chuncheng Yue, Jianpeng Zhou, and Guoren Wang, ‘Self-Adaptive Framework for Efficient Stream Data Classification on Storm’, IEEE Transactions on Systems, Man, and Cybernetics: Systems, Volume 50, Issue 1, 2020, pages 123-136

j. Armando Segatori, Alessio Bechini, Pietro Ducange, and Francesco Marcelloni, ‘A Distributed Fuzzy Associative Classifier for Big Data’, IEEE Transactions on Cybernetics, Volume 48, Issue 9,2018, Pages: 2656 – 2669