Regression Tree Based Correlation Technique in Spatial Data Classification

P.D.Sheena Smart[a], K.K.Thanammal[b], S.S.Sujatha[c]
[a] Reg.No: 18123152282021, Research Scholar, Department of Computer Science, S.T.Hindu College, Nagercoil, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli - 627012, Tamilnadu India.
[b,c] Associate Professor, Department of Computer Science, S.T.Hindu College, Nagercoil, Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli - 627012, Tamilnadu, India.
[a]sheena2021@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Data mining is the process of discovering useful patterns from large geo-spatial datasets with the help of machine learning methods. . The machine learning methods plays an important role for data analytics modeling and visualization. Geo-spatial data is a significant task in many application domains, such as environmental science, geographic information science, and social networks. However, the existing spatial pattern discovery and prediction techniques failed to predict the event accurately with minimum error and time consumption. . In this paper, a novel Pearson Correlated Regression Tree-based Affine Projective spatial data Classification (PCRT-APSDC) technique is proposed to improve the spatial data classification and minimize error based on the Affine Projective classification technique. The proposed algorithm employs a fuzzy rule procedure that constructs the regression tree. The fuzzy rule is applied for linking the inputs (i.e. spatial data) with the outputs (i.e. classification results). Our goal is to classify the data into two subsets such as fired region and non-fired region. Experimental evaluation is carried out using a forest fire dataset with different factors such as classification accuracy, false-positive rate, and classification time. The results confirm that the proposed technique predicts the fired region with increased spatial data classification accuracy and minimized time as well as false-positive rate than the state-of-the-art methods.

Keywords: Spatial data, machine learning, classification, Regression tree, Fuzzy rule

1. Introduction

Spatial data mining is mining knowledge from huge amounts of spatial data. It is extracting knowledge from spatial data like Geographic Information Systems whose information is related to geographic locations. Spatial data are data that comprise the location characteristics that are stored in databases called spatial databases. Spatial data mining is the process of applying various mining techniques such as clustering and classification to a spatial database to extract significant patterns from the spatial data. One of the most important data mining methods is classification. Classification is the task of categorizing the objects from the spatial database into different classes in such a way that the data in one class are similar and it has common features. The traditional techniques have been applied for mining the spatial data but it takes higher complexity. Therefore, the classification task helps in discovering and extracting the interesting patterns from the spatial dataset with lesser complexity.

A Logitboost Ensemble-based Decision Tree (LEDT) method was developed in [1] for mapping the forest fire vulnerability. But the designed method failed to improve the classification accuracy. Graph Convolutional Neural Network (GCNN) architecture was introduced in [2] to evaluate the graph-structured spatial data. Though the GCNN improves the accuracy, the classification time was not minimized.

A support vector machine (SVM) algorithm was introduced in [3] with a data modeling technique to estimate forest fire burning for area approximation. But, the error and false positive rate were not minimized by the SVM algorithm. A Naive Bayes classification method was introduced in [4] for the fire alarm system. The Naive Bayes classification method was introduced with better prediction accuracy results in a data training set based on the smoke source. But, the time complexity was not minimized by the Naive Bayes classification method.

An adaptive ensemble method was introduced in [5] to improve the classification performance of the spatial characteristics of the imbalance data. But the designed method failed to minimize the time complexity in the spatial data classification. A Map-Reduce based approach was developed in [6] to find the entire co-location patterns from a spatial dataset. The designed approach reasonably minimizes the execution time for pattern mining but the accuracy was not improved.

Three different methods were introduced in [7] to reduce the land cover classification problems. The methods improve the classification accuracy but the mapping problem was not solved. A random forests (RFs) classifier was developed in [8] for mapping the land cover through the classification of remote sensing big data. The designed classifier minimizes the classification error but the classification time was not reduced.

A stacked sparse autoencoder was introduced to learn the high-level features and spatial data classification was performed in [9] using a random forest classifier. The designed classifier failed to achieve more robust

performance. A Spatio-temporal data classification method was developed in [10] with multidimensional patterns. But the method failed to improve the performance of data classification with minimum time.

The integration of remote sensing data and GIS concept was introduced in [11] to find the high-risk of the fired area of forest. But the concept failed to use the machine learning technique for effective risk prediction. The least-squares support vector machines (LSSVM) and artificial bee colony (ABC) optimization were introduced in [12] for spatial prediction and mapping of landslides. The designed model failed to minimize the prediction error.

Problem Statement

The issue of spatial data analysis is major concern in the spatial data mining by using huge sizes of the database. Recently, numerous research works are developed for spatial data classification with aid of dissimilar data mining techniques. But, the classification accuracy of existing works was not adequate. The conventional techniques were introduced for spatial data mining although it takes higher complexity. However, the classification time was higher. But, the false positive rate was not reduced. To resolve the issues, a novel Pearson Correlated Regression Tree-based Affine Projective spatial data Classification (PCRT-APSDC) technique is introduced.

The PCRT-APSDC technique employs the fuzzy rule-based classification algorithm and works with multiple spatial data. The multiple spatial data are positioned on the dimensional space and projected the data into different subsets. The Fuzzy rule procedure is used for constructing the regression tree to classify the input data into different classes with minimum error based on the Pearson correlation measure. The Pearson correlation is measured between the training features and the testing features. After the classification, neighboring spatial data paths in the constructed tree are identified by computing the stress function based on the distance measure.

Contribution

A PCRT-APSDC technique is developed for spatial data classification. In comparison with other related works, our proposed technique exhibits improved performance.

The major contributions are described as follows.

- The PCRT-APSDC technique is introduced to improve the spatial data classification accuracy and minimize the time. This contribution is achieved by an affine spatial projection which is the process of mapping the total data into different subsets using a fuzzy rule-based regression tree classification technique. The internal node in the regression tree measures the relationship between the training features and testing features using the Pearson correlation coefficient. Based on the correlation value, the data are classified into different subsets with minimum time.
- Fuzzy rule-based classification is used for discovering the fired region in the forest according to the correlation value.
- The gradient descent function is applied after the spatial data classification to minimize the training error. This helps to reduce the false-positive rate. In the tree, the neighboring spatial data path is identified by calculating the stress function. The stress is the distance function that is measured between the nodes in the tree.

The paper is organized as follows. Related works are presented in Section 2. The problem definition and proposed methodology Pearson Correlated Regression Tree-based Affine Projective Spatial Data Classification (PCRT-APSDC) is presented in Section 3. In Section 4, experimental evaluation and parameter settings are presented and the Performance analyses of different parameters using three different classification techniques are described in Section 5. Finally, the conclusion of the paper is presented in Section 6.

2. Related Works

A Random Subspace (RSS) and Classification and Regression Trees (CART) was developed in [13] for forecasting the landslides with the help of spatial data. The designed hybrid technique failed to minimize the prediction error. A novel approach that uses Extinction filters were designed in [14] that accurately extract spatial and contextual information from remote sensing images. However, this is not applicable to conventional Attribute profiles. Four dissimilar classification algorithms were introduced in [15] for identifying the burned areas on a global scale. The performance of the classification accuracy remained unaddressed. The stacked sparse autoencoder (SSAE) was developed in [16] for classifying the data based on the local spatial information. Though the method improves the classification accuracy, the false positive rate was not minimized.

A GIS-based machine learning technique was introduced in [17] for groundwater nitrate concentration based on spatial data. But, the spatial data classification time was not minimized. A Bayesian spatial generalized linear mixed model (SGLMM) was enveloped in [18] to classify the spatial data. The designed model has a higher complexity in the spatial data classification. A formal concept analysis (FCA) was presented in [19] for the

dynamic classification of spatial data. But the classification error was not minimized. Machine-Learning models were developed in [20] for improving the predictive performance with spatial data. Though the designed model improves the prediction accuracy, the prediction time was not minimized. An Extreme Learning Machine (ELM) was introduced in [21] for classifying the spatial environmental data. The ELM minimizes the mean square error but the performance of time complexity remained unsolved.

A Differential Flower Pollination (DFP) and mini-match backpropagation (MnBp) was introduced in [22] for predicting the forest fire danger using spatial data. But the advanced machine learning or soft computing techniques was not used to increase the forest fire danger prediction. An artificial neural network was developed in [23] for predicting forest fires using a multilayer perceptron. The designed network minimizes the global error at the output layer but time complexity was not minimized.

Piecewise linear regression and predictive modeling was introduced in [24] for data management systems (DMS) predictive analytics. A novel multifeature dictionary learning algorithm (MF-SADL) [25] was developed for hyperspectral image classification. However, the classification accuracy was not improved. Deep neural network (DNN) was introduced in [26] to extract the features for improving the accuracy. But, the classification time was not reduced. A new mining paradigm named spatial-temporal fluctuating patterns (STFs) was introduced in [27] for determining frequent patterns from the spatial-temporal data. A spectral clustering approach was designed in [28] for multivariate geostatistical data. However, it failed to focus the classification accuracy.

3. Methodology

The size of the spatial dataset is also growing significantly in recent days. The problem of spatial data analysis is difficult for human beings since it has large sizes of the database that requires novel techniques to discover the patterns. Moreover, analyzing such a database is more time-consuming and provides errors since the spatial data structure is more complex than the ordinary database. Therefore, spatial data mining is a difficult and complex task to discover interesting patterns from this database.

Therefore an efficient data mining techniques called classification is employed for solving the above issues. Based on the motivation, the proposed Pearson Correlated Regression Tree-based Affine Projective spatial data Classification (PCRT-APSDC) technique is developed to improve the spatial data classification accuracy with minimal error rate.

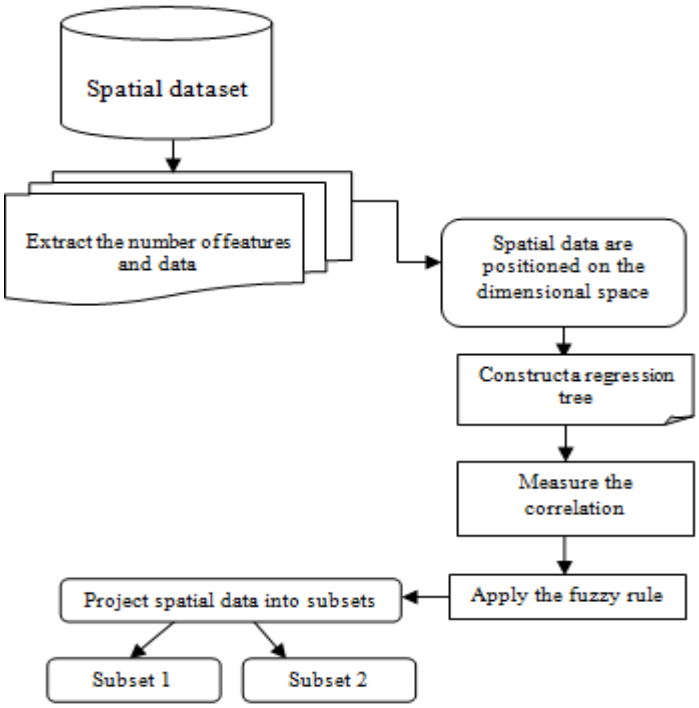

Fig.1 Flow Process of PCRT-APSDC technique

Fig.1 shows the flow process of the proposed PCRT-APSDC technique to classify the spatial data with minimum time. The spatial dataset (i.e. forest fire dataset) includes a number of attributes (i.e. features) $a_1, a_2, a_3 \ldots . a_n$ and each attributes contains the set of data $\{sp_1, sp_2, sp_3, \ldots . sp_n\}$. By applying the forest fire dataset, the burned area is predicted based on the classification. Initially, the number of data are collected from the

dataset. After collecting the data, the classification is performed using Pearson correlated regression tree. The classification process of the proposed PCRT-APSDC technique is described in the following subsection.

3.1 Pearson Correlated Regression Tree-Based Affine Projective Spatial Data Classification

The multiple spatial values of the data are positioned on the dimensional space. In mathematical, the affine spatial projection is the process of mapping the total dataset into different subsets based on the fuzzy rule. Here, the total dataset represents the number of data taken from the spatial dataset and the subsets denote the classification outcomes. The proposed PCRT-APSDC technique performs the classification through the fuzzy rules.

The Pearson Correlated Regression Tree is a machine learning technique and the flow-chart-like structure is used to classify the given dataset into two classes such as fired region or non-fired region. A regression tree includes three types of nodes such as root node, internal node, and leaf node. The topmost node in a decision tree is the root node where the decision is taken by applying the fuzzy rules. Each internal (non-leaf) node performs a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node provides the class labels. The root node in the tree measures the correlation between the features and then the fuzzy rule is applied to classify the data.

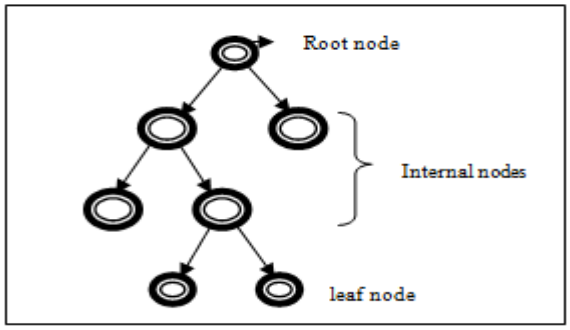

Fig. 2 Structure of the Regression tree

Fig.2 illustrates a structure of the regression tree to classify the spatial data into a fired region or non-fired region. For each node in the tree, the correlation between the training features and the testing features (i.e. forest fire causing features) is measured using the Pearson correlation. The Pearson correlation is measured as follows,

$$\beta = \frac{n*\sum F_i*F_r-(\sum F_i)(F_r)}{\sqrt{[n*\sum {F_i}^2-(\sum F_r)^2][n*\sum {F_r}^2-(\sum F_r)^2]}} \qquad (1)$$

In (1), β denotes a correlation coefficient and 'n' represents several features. $\sum F_i * F_r$ denotes a sum of the product of paired score of two features, $\sum {F_i}^2$ represents a squared score of F_i and $\sum {F_r}^2$ represents a squared score of F_r. The correlation coefficient (β) provides the two results such as '+1' and '-1'. The coefficient provides '+1' indicates a positive correlation and it provides '-1' which represents the negative correlation between two features.

3.2 Fuzzy Rule-Based Classifications

After finding the correlation between the features, the fuzzy rule is applied to the nearest neighbor dimensional space for classifying the spatial data. The fuzzy rule is used for connecting the inputs (i.e. spatial data) with the outputs (i.e. classification results). The rules are formulated using algorithmic formalism are *IF* (condition) and *THEN* (conclusion). The condition part checks the correlation value between the features and the conclusion part provides the desired classification results.

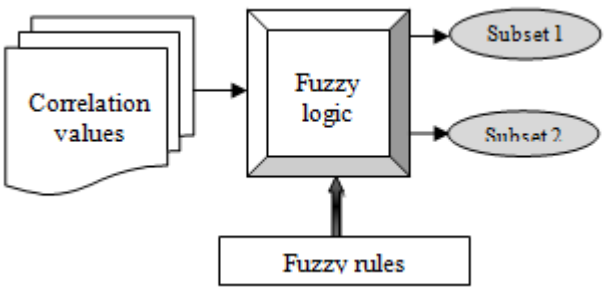

Fig. 3 Fuzzy Rule-Based Classifications

Fig.3 shows the fuzzy rule-based classification to identify the fired region in the forest-based on the correlation value. By the established rules, the data are classified into two subsets such as fired region and non-fired region based on the correlation values. The output of the regression tree is given below,

$$y = \begin{cases} \beta = +1, & fired\ region \\ \beta = -1, & non - fired\ region \end{cases} \quad (2)$$

In (2), 'y' represents the classification output. In this way, the total dataset is projected into two different subsets. After the classification, the error is computed to minimize the incorrect data classification. The training error is calculated using the following mathematical equation,

$$E_t = (y_p - y\,)^2 \quad (3)$$

In (3), E_t denotes a training error, y represents the actual classification and y_p represents the predicted classification results. The gradient descent function is used to minimize the error in the classification process,

$$f(x) = \arg\min E_t \quad (4)$$

In (4), $f(x)$ represents the gradient descent function, $arg\ min$ denotes an argument of the minimum function E_t denotes a training error. In this way, all the data are classified and predicts the fired region in the forest. After the classification, the frequent and persistent soft cycle's path in the tree is identified with spatial data to speeds up the tree construction process by computing the stress function. The stress function is calculated in terms of distance. The distance with spatial data points computes the stress function to identify the soft cycle neighboring spatial data paths in the constructed tree. Let us consider the coordinates of the two nodes represented as (x_1, y_1) and (x_2, y_2) in the two-dimensional space. The distance between the nodes in the tree is computed as follows,

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (5)$$

In (5), d represents the distance between the nodes. The minimum distance is used to find the neighboring spatial data paths in the tree. This helps to accurately find the neighboring fired area in the forest with minimum time. The algorithmic procedure of the proposed PCRT-APSDC technique is described as follows.

Algorithm 1 Pearson Correlated Regression Tree-based Affine Projective spatial data Classification

Input: Spatial dataset, number of attributes (i.e. features) $a_1, a_2, a_3 \ldots\ldots a_n$ and spatial data $\{sp_1, sp_2, sp_3, \ldots. sp_n\}$.

Output: Improved spatial data classification accuracy

Begin

1. Position spatial data $\{sp_1, sp_2, sp_3, \ldots. sp_n\}$ in dimensional space
2. Construct regression tree R_r with nodes
3. **for** each data sp_i
4. Measure correlation β
5. **if** $(\beta = +1)$ **then**
6. Positive correlation between training features and fired features

Algorithm 1 describes the process of Pearson Correlated Regression Tree-based Affine Projective spatial data classification with minimum error. The spatial data are positioned in the given dimensional space. Then mapping from the input dataset into different subsets is performed by constructing the regression tree. The regression tree-based classification is performed through the correlation between the training features data and testing features data. If the two features are highly correlated, then the data are classified into one subset. Otherwise, the data are classified into another subset. Followed by, the classification error is calculated and minimized using gradient descent function. This helps to improve the spatial data classification accuracy and minimizes the error rate. Finally, the neighboring spatial data paths are identified through the distance function to find the neighboring fired paths in the forest with minimum time.

The above algorithm is implemented in the experimental evaluation to show the performance of the proposed PCRT-APSDC technique.

4. Experimental Evaluation And Parameter Settings

Experimental evaluations of proposed PCRT-APSDC technique and existing methods namely LogitBoost Ensemble-based Decision Tree (LEDT) [1] and Graph Convolutional Neural Network (GCNN) [2], Support Vector Machine (SVM) Algorithm [3] and Naive Bayes classification Method [4] are implemented using Java language.

The experiments are carried out with different parameters given below:

- classification accuracy
- false-positive rate
- classification time

4.1 Datasets

In this section, Forest Fires Dataset [29] is taken from the UCI machine learning repository. The main aim of the dataset is to predict the burned region of forest fires, in the northeast area of Portugal with the help of meteorological data. The dataset comprises 517 instances and 13 attributes. The associated task of the dataset is the regression. The attributes characteristics are real and the dataset characteristics are multivariate. The proposed PCRT-APSDC technique uses holdout method for performing the cross-validation process. The input dataset is separated into two sets such as training set and testing set. Most of the data is used for training (i.e., 60 percentage of data) and a smaller portion of the data is taken for testing i.e., 40 percentage of data. To conduct the experiment, the number of spatial data (i.e. instances) considered in the range from 50-500 from the forest fires dataset.

The experiments are carried out with different parameters given below:

- classification accuracy
- false-positive rate
- classification time

5. Results And Discussion

The experimental results of the proposed PCRT-APSDC technique and existing methods namely LEDT [1], GCNN [2], SVM Algorithm [3], and Naive Bayes classification Method [4] are discussed in this section with different parameters such as classification accuracy, false positive rate and classification time. Performance results are evaluated with the help of graphical representations. For each subsection, the sample mathematical computation is presented.

5.1 Performance Results of Classification Accuracy

The classification accuracy is defined as the ratio of a number of data correctly classified for predicting the burned area to the total number of spatial data. The formula for calculating the classification accuracy is given below,

$$Classification\ Accuracy = \left(\frac{Number\ of\ data\ correctly\ classified}{n}\right) * 100 \quad (6)$$

In (6), 'n' refers to the total number of spatial data. The classification accuracy is measured in the unit of percentage (%). The classification accuracy result using the PCRT-APSDC Technique is compared with the four conventional methods LEDT [1], GCNN [2], SVM Algorithm [3] and Naive Bayes classification Method [4]. The performance result analysis of classification accuracy is shown in Table 1.

Table 1. Comparision of Classification Accuracy (%) with respect to PCRT-APSDC, LEDT, GCNN, SVM and Naïve Bayes

Number of data	**Classification accuracy (%)**				
	PCRT-APSDC	LEDT	GCNN	SVM	Naive Bayes
50	86	66	74	60	58
100	90	68	76	63	57
150	94	75	83	66	60
200	92	80	85	69	63
250	93	82	88	67	61
300	92	82	87	65	59
350	91	81	86	68	57
400	94	79	85	70	55
450	92	77	86	72	58
500	90	77	84	74	60

Fig.4 illustrates the experimental results of classification accuracy versus a number of spatial data.

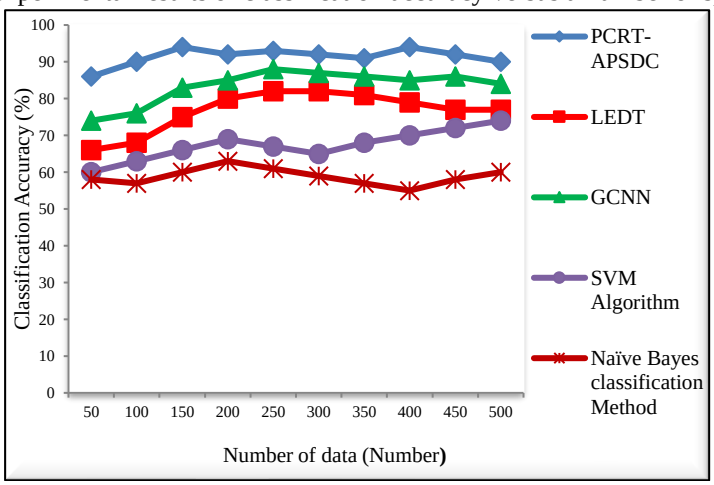

Fig.4 Performance results of classification accuracy

For the experimental evaluation, the spatial data are taken in the range from 50 to 500. Totally ten results of classification accuracy are obtained with various input data as shown in fig.4. The graphical results clearly show that the classification accuracy is found to be higher using the proposed PCRT-APSDC technique as compared to the conventional technique. This significant improvement is achieved by projecting the total spatial data into

different subsets. The mapping of the spatial data is carried out using the regression tree. The spatial data are collected from the forest fire dataset. Then the correlations of training features with the fire testing features are measured to classify the given instance (i.e. data) into the burned area. This helps for the proposed PCRT-APSDC technique to improve the number of spatial data correctly classified and effectively predicts the burned area in the given location. Besides, the neighboring burned area also identified using the PCRT-APSDC technique by measuring the distance between the nodes in the tree. The statistical results confirm that the classification accuracy of the proposed PCRT-APSDC technique is higher than the existing methods. Let consider the 50 spatial data, 43 data are correctly classified using the PCRT-APSDC technique and their percentage is 86%. Similarly, the 33, 37, 30, and 28 data are correctly classified by the existing LEDT, GCNN, SVM Algorithm, and Naive Bayes classification Method, and their classification accuracy percentages are 66%, 74%, 60%, and 56% respectively.

The proposed classification results are compared to the classification accuracy of the existing technique. The comparison results show that the proposed PCRT-APSDC technique improved the classification accuracy by 20%, 10%, 36%, and 56% than the LEDT, GCNN, SVM Algorithm, and Naive Bayes classification Method respectively.

5.2 Performance Results of False-Positive Rate

The false-positive rate is defined as the ratio of a number of data incorrectly classified to the total number of spatial data. The false-positive rate is mathematically calculated as follows,

$$false - positive\ rate = \left(\frac{Number\ of\ data\ incorrectly\ classified}{n}\right) * 100 \quad (7)$$

In (7), 'n' refers to the total number of spatial data. The false-positive rate is measured in the unit of percentage (%). The experimental result of the false-positive rate using the PCRT-APSDC Technique is compared with four state-of-the-art methods LEDT [1], GCNN [2], SVM Algorithm [3], and Naive Bayes classification Method [4]. The tabulation result analysis of the false-positive rate is demonstrated in Table 2.

Table 2. Comparision of false-positive rate (%) with respect to PCRT-APSDC, LEDT, GCNN, SVM and Naïve Bayes

Number of data	**False-positive rate (%)**				
	PCRT-APSDC	LEDT	GCNN	SVM	Naive Bayes
50	14	34	26	40	44
100	10	32	24	37	43
150	6	25	17	34	40
200	8	21	15	31	37
250	7	18	12	33	39
300	8	18	13	35	41
350	9	19	14	32	43
400	7	22	15	30	45
450	8	23	14	28	42
500	10	23	16	26	40

As shown in fig.5, the performance result of the false-positive rate is illustrated with the number of spatial data.

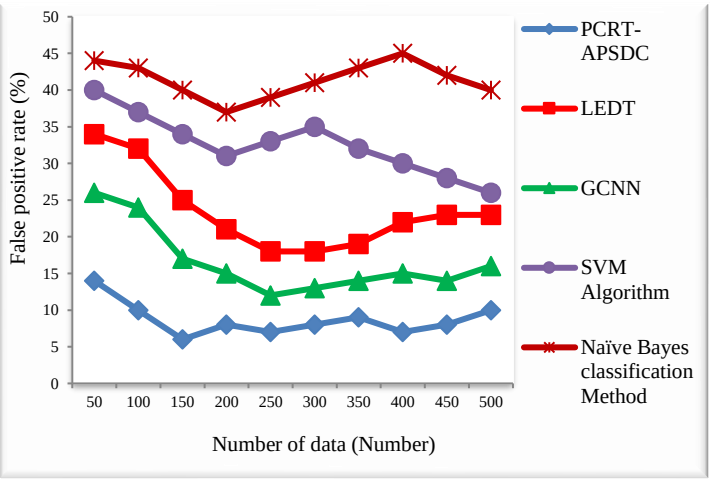

Fig.5 Performance results of the false-positive rate

The false-positive rate is the number of spatial data that are incorrectly classified. The false-positive rates of five different methods namely PCRT-APSDC, LEDT [1], GCNN [2], SVM Algorithm [3], and Naive Bayes classification Method [4] are represented by the five different colors of lines as shown in fig. 5. The false-positive rate of the proposed PCRT-APSDC technique is minimized as compared to existing results. The reason behind the classification error is minimized by using a gradient descent function. By applying the gradient descent function after the classification, the training error is minimized.

Also, the fuzzy rule is applied to the tree structure to classify the given spatial data with the help of the correlation between the features, from which, the fire in the specific location, as well as the neighboring area, is predicted with minimum error through the efficient classification results. There are 10 different results of the false-positive rate which are obtained for each technique. The results of the proposed PCRT-APSDC technique are compared to the results of the existing classification methods. Hence, the average false-positive rate is found to be lesser using the proposed PCRT-APSDC technique by 62% when compared to LEDT, 47% as compared to GCNN , 73% as compared to SVM Algorithm, and 79% as compared to Naive Bayes classification Method.

5.3 Performance Results of Classification Time

The classification time is defined as the amount of time required to classify the spatial data. The classification time is mathematically calculated as follows,

$$CT = n * T\ [classifying\ single\ data] \qquad (8)$$

From equation (8), CT represents the classification time, n denotes the number of spatial data, and 'T' represents the time taken for classifying a single spatial data. The classification time is measured in the unit of milliseconds (ms). The experimental result of time complexity using the PCRT-APSDC Technique is

compared with four state-of-the-art methods LEDT [1], GCNN [2], SVM Algorithm [3], and Naive Bayes classification Method [4]. The tabulation result analysis of classification time is demonstrated in Table 3.

Table 3. Comparision of Classification Time (ms) with respect to PCRT-APSDC, LEDT, GCNN, SVM and Naïve Bayes

Number of data	**Classification time (ms)**				
	PCRT-APSDC	LEDT	GCNN	SVM	Naive Bayes
50	11	19	16	22	25
100	16	30	23	35	41
150	23	44	35	49	53

200	26	54	40	60	65
250	30	65	48	75	79
300	33	69	54	80	85
350	42	70	60	89	96
400	56	76	68	95	104
450	63	81	72	105	119
500	70	100	90	112	124

Fig.6 depicts the performance results of the classification time graph with respect to a number of spatial data.

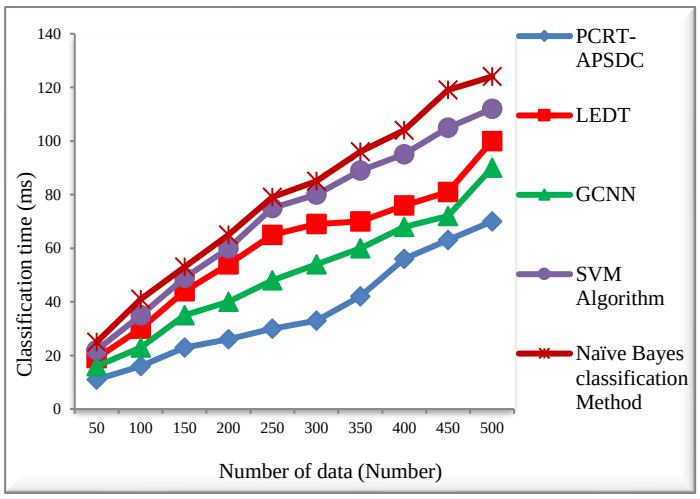

Fig.6 Performance results of classification time

The number of spatial data is taken as input in the 'x' axis and corresponding results of classification time are obtained at the 'y' axis. The average classification time here refers to the time taken to classify the data into the fired region or non-fired region. From the figure, it is inferred that a number of spatial data are directly proportional to the average processing time. In other words, while increasing the number of spatial data, classification time consumption gets increased using all five classification techniques. But the classification time is found to be lesser using the proposed PCRT-APSDC technique. The tree-based classification technique effectively classifies the data through the fuzzy rule with minimum time. The distance between the nodes in the tree is calculated to find the neighboring fired region with lesser time. Let us take '$50'$ data for experimentation, the time is taken for classifying the data being $11ms$' using the proposed PCRT-APSDC technique, whereas '$19ms$', '$16ms$', '$22ms$' and '$25ms$' time taken by existing techniques LEDT [1], GCNN [2], SVM Algorithm [3] and Naive Bayes classification Method [4]. As a result, the overall classification time of the proposed PCRT-APSDC technique is lesser as compared to the existing classification techniques. Hence the comparison of ten different results is found that the proposed PCRT-APSDC technique minimizes the classification time by 41%, 29%, 50%, and 55% when compared to LEDT, GCNN, SVM Algorithm, and Naive Bayes classification Method respectively. The above discussion of various parameter results observed that the proposed PCRT-APSDC technique effectively performs the spatial data classification with higher accuracy and lesser time consumption as well as false-positive rate.

6. Conclusion

The main goal of our research is to develop an efficient machine learning technique called PCRT-APSDC for spatial data mining by projecting the total dataset into different subsets with minimum time consumption. The spatially distributed data in the given dimensional space is correctly classified to extract useful patterns. In the mapping phase, the input spatial data is mapped into the different subsets through the fuzzy rule-based tree construction. Pearson correlations between the features are measured and the fuzzy rules are applied for constructing the tree with different nodes. The leaf nodes in the tree provide the final classification results. Therefore, the regression tree minimizes the classification error using the gradient descent function. Experimental evaluation is carried out using a forest fire dataset with three different parameters such as classification accuracy, false-positive rate, and classification time. The results illustrate that the spatial data classification accuracy is improved with minimum time complexity and false-positive rate when compared to the state-of-the-art methods.

References

1. Tehrany M S, Jones S, Shabani F, Martínez-Álvarez F, Bui D T.: A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial data. Theor. Appl. Climatol. Springer, 137, 637–653 (2019)
2. Yan X, Ai T, Yang M, Yin H.: A graph convolutional neural network for classification of building patterns using spatial vector data. ISPRS J. Photogramm. Remote Sens. Elsevier, 150, 259-273 (2019)
3. Kerdprasop N, Poomka P, et al.: Forest Fire Area Estimation using Support Vector Machine as an Approximator. International Joint Conference on Computational Intelligence, 269-273 (2018)
4. Putrada A M, Abdurohman M and Putrada A G : Increasing Smoke Classifier Accuracy using Naïve Bayes Method on Internet of Things. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 4, 19-26 (2018)
5. Wang L, Zhao L, Gui G, Zheng B, and Huang R. : Adaptive Ensemble Method Based on Spatial Characteristics for Classifying Imbalanced Data. Sci. Program., Hindawi, 1-8 (2017)
6. Maiti S and Subramanyam R B V.. Mining co-location patterns from distributed spatial data. J. King Saud Univ., Comp. & Info. Sci., Elsevier, 1-10 (2018)
7. Lavreniuk M S, Skakun S V, Shelestov A J, YalimovB Y , Yanchevskii S L, YaschukD J, and Kosteckiy A I. : Large-Scale Classification Of Land Cover Using Retrospective Satellite Data. Cybern Syst Anal., Springer, 52, 1, 127-138 (2016)
8. Zhang X M, He G J,. Zhang Z M, Peng Y and Long T F.: Spectral-spatial multi-feature classification of remote sensing big data based on a random forest classifier for land cover mapping. Cluster Comput., Springer, 20, 2311–2321 (2017)
9. Wan X, Zhao C, Wang Y and Liu W. : Stacked sparse autoencoder in hyperspectral data classification using spectral-spatial, higher-order statistics and multifractal spectrum features. Infrared Phys Technol , Elsevier, 86, 77-89 (2017)
10. Pitarch Y , Ienco D , Vintrou E, Bégué A, Laurent A, Poncelet P, Sala M, Teisseire M.: Spatio-temporal data classification through multi dimensional sequential patterns: Application to crop mapping in complex landscape. Eng. Appl. Artif. Intell, 37, 91–102 (2015)
11. Adab H, Kanniah K D , Solaimani K.: Modeling forest fire risk in the northeast of Iran using remote sensing and GIS techniques. Nat Hazards , Springer, 65, 1723–1743 (2013)
12. Bui D T, Tuan T A, Hoang N D, Thanh N Q, Nguyen D B, Liem N V, Pradhan B.: Spatial prediction of rainfall-induced landslides for the Lao Cai area (Vietnam) using a hybrid intelligent approach of least squares support vector machines inference model and artificial bee colony optimization. Landslides, Springer, 14, 447–458 (2017)
13. Thai B, Prakash P I, Bui D T.: Spatial prediction of landslides using a hybrid machine learning approach based on Random Subspace and Classification and Regression Trees. Geomorphology, Elsevier, 303, 256-270 (2018)
14. Ghamisi P, Souza R, Benediktsson J A, Zhu X X, Rittner L, and Lotufo R A : Extinction Profiles for the Classification of Remote Sensing Data. IEEE Trans Geosci Remote Sens. 54, 5631-5645 (2016)
15. Ramo R, García M, Rodríguez D, Chuvieco E.: A data mining approach for global burned area mapping", Int J Appl Earth Obs Geoinf . Elsevier, 73, 39-51(2018)
16. Zhang L, MaW, Zhang D.: Stacked Sparse Autoencoder in PolSAR Data Classification Using Local Spatial Information. IEEE Geosci. Remote. Sens, 13, 1359 – 1363 (2016)
17. Knoll L, Breuer L, Bach M.: Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Sci. Total Environ. Elsevier, 668, 1317-1327 (2019)
18. Berrett C and .Calder C A. : Bayesian spatial binary classification. Spat. Stat., Elsevier, 16, 72-102 (2016)
19. Chen Y, Zhou J, Wilson J P, Wu J, Wu Q, Yang J. : A Dynamic Classification Pattern of Spatial Statistical Services Using Formal Concept Analysis. Geogr. Anal, 50, 454-476 (2018)
20. Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A. : Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol Modell, Elsevier, 406, 109-120 (2019)

21. Leuenberger M and Kanevski M.: Extreme Learning Machines for spatial environmental data. Comput Geosci., Elsevier, 85, 64-73 (2015)
22. Bui D T, Le H V, Hoang N D.: GIS-based spatial prediction of tropical forest fire danger using a new hybrid machine learning method. Ecol. Inform., Elsevier, 48, 104-116 (2018)
23. Safi Y and Bouroumi A.: Prediction of Forest Fires Using Artificial Neural Networks. Appl. Math. Sci., 7, 271 – 286 (2013)
24. Anagnostopoulos, C., Triantafillou, P.: Large-scale predictive modeling and analytics through regression queries in datamanagement systems. Int J Data Sci Anal, Springer, 9, 17–55 (2020)
25. Zhang, H., Yang, M., Yang, W., Lv, • J.: Spatial-aware hyperspectral image classification via multifeature kernel dictionary learning. Int J Data Sci Anal, Springer, 7, 115–129 (2019)
26. Endo, Y., Toda, H., Nishida, K., Ikedo, J.: Classifying spatial trajectories using representation learning. Int J Data Sci Anal, Springer, 2, 1–11 (2016)
27. Teng, S., Ou, C.,Chuang, K.: On the discovery of spatial-temporal fluctuating patterns. Int J Data Sci Anal, Springer, 8, 57–75 (2018)
28. Fouedjio, F.: A spectral clustering approach for multivariate geostatistical data. Int J Data Sci Anal, Springer, 4, 301-312 (2017)
29. [dataset] Forest Fires Dataset: UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/forest+fires, Cortez P and Morais A. , last accessed on 12 September 2018.