Spatial Data Mining for Earthquake Significance Classification: Exploring Geospatial Insights

Gollapalli Ganga Srinivas, Gadde Madhukar, Gadde Maruti Mahesh, Uppuluri Bogesh, and Rajiv Senapati

Department of CSE, SRM University, Andhra Pradesh, India
<gangasrinivas_g,madhukarsaibabu_g,maheshchowdhary_g,bogesh_uppuluri,rajiv.s>
@srmap.edu.in

Abstract. Spatial data mining is the data collected from physical real-life locations containing map data, Image data, and graph data. Spatial data mining focuses on extracting patterns and information from geographical datasets. In this paper, we have presented earthquake significance classification using Logistic Regression(LR), Decision Tree(DT), Random Forest(RF), Gradient Boosting(GB), Support Vector Machine(SVM), K-Nearest Neighbors(K-NN), Gaussian Naive Bayes, Neural Networks(NN), AdaBoost, and Bagging. Further, the performances of those algorithms were evaluated using F1-score, Recall, Precision, and accuracy. From this analysis it is observed that, RF and Bagging performing better as compared to other techniques.

Keywords: Earthquake · Machine Learning · Spatial data mining · classification

1 Introduction

Spatial data refers to the description of objects in the form of maps, GPS coordinates and satellite imagery etc. It is utilized in fields such as logistics, environmental science, urban planning, transportation and public health. Machine learning can be applied in spatial databases for solving various problems like healthcare, finance, retail, manufacturing, education, environmental science and many more. It influences patient diagnosis in healthcare, guides stock market analysis in finance, optimizes marketing strategies in retail, refines manufacturing processes, improves education by predicting student success, assists environmental scientists in understanding climate change, optimizes transportation systems, aids in scientific research, enhances network security, influences pharmaceutical sales, informs insurance decisions, empowers research across disciplines, and plays a pivotal role in IoT applications. In contrast, spatial data mining leaves an indelible mark on areas such as urban planning, natural resource management, epidemiology, agriculture, emergency response, crime analysis, environmental monitoring, geological exploration, public health planning, climate change modeling, transportation, natural resource exploration, agricultural yield prediction,

tourism and location-based services, wildlife conservation, infrastructure planning, transportation safety, historical preservation, satellite imagery analysis, and market analysis. These applications capitalize on spatial data's unique attributes to make more informed decisions, protect the environment, and ensure public safety. In contrast, spatial data mining is attuned to data rich in spatial or geographic components, delving deep into the spatial context to unveil patterns influenced by proximity, distance, and spatial relationships. In this paper, we have presented earthquake significance classification using machine learning approach. The followings are some of the contributions from this paper.

- This research underscores the importance of algorithm selection in solving classification problems, highlighting the impact of data-driven insights on safety and resource optimization.
- Random Forest exhibited the highest accuracy (0.898) and a well-balanced F1 score (0.75) for classifying earthquake significance, making it an efficient choice among the evaluated algorithms.
- The study underscores the broader applications of spatial data mining in real-world scenarios, ranging from disaster management to resource allocation, showcasing its versatility and impact in diverse fields.

2 Literature Review

The literature review provides an overview of studies that have explored data mining and machine leaning. These summaries focus on aspects of the subject. One particular study, referenced as [1], The authors address the challenges of dealing with large spatial data sets and introduce the Peano Count Tree (P-tree) structure for lossless and compressed data representation. They propose the P-tree based Association Rule Mining (PARM) algorithm, comparing it with FP-growth and Apriori algorithms. The paper discusses potential applications in precision agriculture, resource discovery, and more. It offers a novel approach to extracting valuable patterns and rules from spatial data.In [2],the paper proposes GPU-based algorithms for efficient colocation pattern mining, overcoming the limitations of existing sequential and Map-reduce-based approaches. These algorithms include a novel cell-aggregate-based upper bound filter and two refinement algorithms. The paper also presents theoretical analyses and GPU profiling to optimize the algorithms, achieving significant speedup compared to CPU-based implementations.In [3], The paper discusses a novel approach to mine colocation patterns in urban spaces, taking into account the constraints of a road network. In [4], the review focuses on addressing the challenges associated with spatial data, including spatial autocorrelation, spatial heterogeneity, limited ground truth, and multiple scales and resolutions. It covers various application areas, such as earth science, urban informatics, geosocial media analytics, and public health. The paper categorizes spatial prediction methods based on the challenges they address, discusses their underlying assumptions, and compares their advantages and disadvantages. The goal is to help researchers and practitioners in various domains choose suitable techniques for solving spatial

prediction problems and to identify research opportunities in the field. In [5] this paper addresses the challenges of discovering frequent associations in large databases by introducing innovative algorithms that aim to minimize I/O and computation costs through the use of efficient data structures and decomposition techniques.The paper presents six new algorithms that combine the above features, depending on the database format, decomposition technique, and search procedure employed. In [6] the concept of colocation patterns and rules, their application in various domains, and two primary categories of approaches for discovering these rules: spatial statistics-based methods and data mining-based methods. These methods are essential for understanding and analyzing the relationships between different types of spatial features and their proximity. In [7] The authors focus on characteristics expressed in terms of features that are closely associated with the clusters. In their context, a feature is defined as a closed curve (polygon) representing natural or man-made places of interest, such as lakes, parks, schools, golf courses, and shopping centers. In [8] In this paper we delve into the possibility and practicality of implementing a shared ontology to enhance the discovery of resources, within a European Spatial Data Infrastructure (SDI). The term "SDI" refers to a system that includes metadata, spatial data sets, spatial data services, network services, agreements on data sharing, access and coordination mechanisms—all designed to work. From a standpoint an SDI is seen as a platform infrastructure for geospatial data and services based on open standards.The primary motivation behind this research lies in the fact that a European SDI's inherently multilingual and encompasses information communities. This means that there are perspectives when it comes to information ranging from those who produce metadata to providing applications and end users. The model described in this paper aims to represent these viewpoints reflecting the intricacies involved in discovering multilingual information, within an SDI.In [9], the journal tells the importance of analyzing co-patterns in both point and extended spatial features is highlighted. It emphasizes how this analysis can be applied in fields, like healthcare, urban planning and environmental monitoring. These identified patterns play a role in decision making processes. Help address real world challenges related to disease control, urban management and environmental health. To bridge the existing gap another journal aims to provide a review of spatial prediction methods. These methods are categorized based on the challenges they address. In [10], explores the underlying assumptions of each method, delves into their foundations and evaluates their advantages and disadvantages. The objective is to offer researchers a resource that aids them in choosing suitable techniques for their specific application domains. Various other related ML techniques that were discussed in [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26] is also applicable in the health care sector for different applications.

3 Proposed Methodology

3.1 Dataset

The dataset provides information about earthquake events which consists of various attributes. These attributes include the timestamp (‘time‘) of each earthquake, its geographical coordinates in terms of latitude and longitude (‘latitude‘ and ‘longitude‘), and the depth (‘depth‘) at which the earthquake originated. The dataset also records the earthquake's magnitude (‘mag‘) and the type of magnitude measurement method employed (‘magType‘). Information about the seismic monitoring network, the number of reporting seismometer stations (‘nst‘), the gap between these reporting stations (‘gap‘), and the closest distance to a reporting station (‘dmin‘) is also available. Additionally, the dataset contains details about the signal quality represented as the root mean square (‘rms‘). Each earthquake event is uniquely identified by an ‘id‘. The dataset provides location descriptions (‘place‘) and categorizes earthquake events by type (‘type‘). Geospatial accuracy is accounted for with attributes such as horizontal error (‘horizontalError‘), depth error (‘depthError‘), and magnitude error (‘magError‘). The dataset also captures the number of stations contributing to magnitude information (‘magNst‘). The reporting status of each earthquake event (‘status‘) and the sources of location (‘locationSource‘) and magnitude information (‘magSource‘) are included. This dataset enables in-depth analysis and modeling of earthquake events with a wide range of attributes and information sources.

3.2 Problem Statement

To classify earthquakes into two categories: ”significant” and ”not significant.” The significance is determined based on the magnitude (‘mag‘) of the earthquake. Earthquakes with a magnitude greater than or equal to 5.0 are considered ”significant,” while those below 5.0 are labeled as ”not significant.”

3.3 Earthquake Significance Classification Methods

Logistic Regression Logistic regression is a method that helps us determine the likelihood of an event belonging to a category. In this case we are using regression to estimate the probability of an earthquake being significant, by considering factors. The mathematical equation, for regression is expressed as follows;

$$p = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}} \tag{1}$$

Where, p represents the probability of an earthquake being significant, X represents the vector of input features (latitude, longitude, depth, etc.), β represents the vector of coefficients for each feature, β_0 represents the intercept term. The logistic regression model computes the probability as: e is the base of the natural logarithm (approximately 2.71828).

Decision Tree Decision trees are structures that resemble trees and are used for making decisions and classifying tasks. In the context of this problem a decision tree can be utilized to establish rules based on features in order to classify earthquakes as either significant or not.

Random Forest Random Forest is a technique that combines decision trees to enhance the accuracy and reliability of classification. A Random Forest comprises a collection of decision trees typically generated using bagging.

Gradient Boosting Gradient Boosting is a technique that sequentially builds a group of decision trees focusing on improving the classification of instances that were previously misclassified using the following equation.

$$F(x) = F_{previous}(x) + \alpha()Treek(x) \tag{2}$$

where F previous is the previous model, α is the learning rate, and Tree k is the new tree.

SVM (Support Vector Machine) Support Vector Machines (SVMs) belong to a group of machine learning algorithms that are designed to identify the hyperplane, for separating instances into distinct classes. The hyperplane equation in a binary classification problem can be represented as:

$$p \times x + q = 0 \tag{3}$$

where, "p" is the weight vector that determines the orientation of the hyperplane. "x" is the feature vector of an instance. "q" is the bias term.

K-Nearest Neighbors The K Nearest Neighbors (KNN) algorithm is an easy-to-understand classification method. It assigns a class label to an instance by considering the majority class, among its k neighbors. Mathematical Formula: Let x be the instance to be classified. Let $x_1, x_2, ..., x_k$ be the k nearest neighbors of x. Let $y_1, y_2, ..., y_k$ are the class labels of the k nearest neighbors. The class with the highest sum of indicator functions id given by

$$y = argmax(sum_i(1[y_i = c])) \tag{4}$$

Gaussian Naive Bayes Naive Bayes is an algorithm that utilizes probability and Bayes theorem to determine the likelihood of an earthquake falling into a category based on the distribution of its characteristics.

$$P(S|X) = (P(X|S) * P(S))/P(X) \tag{5}$$

Where, P(S — X) is the posterior probability that the earthquake is "significant" given the features X. P(X — S) is the likelihood of observing the features X given that the earthquake is "significant." P(S) is the prior probability of an earthquake being "significant."

Neural Network Neural networks, which are a type of network commonly employed for classification purposes, are composed of interconnected nodes (also known as neurons). Mathematically, the operation at each node can be described as follows: For a hidden layer node h_j (j-th node in the hidden layer): $h_j = Activation(W_j1 * x_1 + W_j2 * x_2 + ... + W_jn * x_n + b_j)$ For the output layer node o_k (k-th node in the output layer): $o_k = Activation(W_k1 * h_1 + W_k2 * h_2 + ... + W_km * h_m + b_k)$ where W_ji represents the weight associated with the connection between the i-th node in the previous layer and the j-th node in the current layer. x_i represents the input feature $i.b_j$ and b_k represent the biases for the j-th node in the hidden layer and the k-th node in the output layer, respectively.

AdaBoost (Adaptive Boosting) AdaBoost, also known as Adaptive Boosting is a technique used in machine learning that combines classifiers together to form a powerful classifier. AdaBoost is an ensemble learning algorithm that combines weak learners iteratively, assigning increasing weights to misclassified instances. The final prediction is a weighted sum of weak learners, enhancing model performance by focusing on difficult-to-classify data points.

Bagging Bagging, which stands for Bootstrap Aggregating, is a technique used in learning. It involves combining the predictions from models that have been trained on subsets of the data. The main goal of bagging is to reduce variability and enhance the stability and accuracy of predictions by averaging or voting on the outputs of these models. Lets denote the training dataset as D and the base model as H. Bagging creates B bootstrap samples $D1, D2, ...DB$ and trains B base models $H1, H2...HB$. When making predictions, for instance x. For classification,Bagging combines the predictions using majority voting.

Let $p_i(x)$ represent the prediction made by base model H_i for instance x where $i = 1, 2...B$. The classification prediction $P(x)$ is determined by selecting the class c that has the sum of $p_i(x)$ over all base models. For regression; Bagging combines the predictions through averaging. Let $y_i(x)$ be the prediction made by base model H_i, for instance x where $i = 1, 2...B$. The final regression prediction $Y(x)$ is calculated as $(1/B)$ multiplied by the formula consists of adding up all the outputs $(y_i(x))$ from each base model.

4 Result Analysis

In this section, earthquake significance is predicted using Logistic Regression(LR), Decision Tree(DT), Random Forest(RF), Gradient Boosting(GB), Support Vector Machine(SVM), K-Nearest Neighbors(K-NN), Gaussian Naive Bayes, Neural Networks(NN), AdaBoost, and Bagging. The performance of these machine learning algorithms are evaluated through accuracy, precision, recall, and F1-Score as follows.

$$Accuracy = (TP + TN)/(TP + FN + FP + TN) \tag{6}$$

$$Precision = TP/(TP + FP) \tag{7}$$

$$Recall = TP/(TP + FN) \tag{8}$$

$$F1 - Score = 2(P \times R)/(P + R) \tag{9}$$

where, TP, TN, FP, FN are True Positive, True Negative, False Positive, False Negative respectively. The performance of those techniques are presented in Table 1 and Figure 1.

Table 1: Result Analysis

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.864	0.78	0.55	0.65
Decision Tree	0.846	0.66	0.66	0.66
Random Forest	0.898	0.84	0.68	0.75
Gradient Boosting	0.891	0.84	0.65	0.73
SVM	0.872	0.86	0.52	0.65
K-Nearest Neighbors	0.879	0.79	0.63	0.7
Gaussian naive bayes	0.839	0.65	0.63	0.64
Neural Network	0.873	0.87	0.51	0.65
ADA Boost	0.88	0.8	0.63	0.7
Bagging	0.891	0.82	0.67	0.74

Among the models evaluated, Random Forest stands out with the highest accuracy of 0.898, indicating strong overall performance. It achieves a balanced F1 Score of 0.75, showcasing effective precision and recall. Gradient Boosting follows closely with an accuracy of 0.891, displaying high precision but a relatively lower recall of 0.65, resulting in an F1 Score of 0.73. K Nearest Neighbors also performs well, demonstrating accuracy at 0.879 and a balanced F1 Score of 0.70. Notably, the Neural Network exhibits a relatively high accuracy of 0.864, but it's recall of 0.51 suggests a challenge in identifying positive instances. The Decision Tree, SVM, and Gaussian Naive Bayes show moderate performance, with accuracies ranging from 0.839 to 0.846 and F1 Scores from 0.64 to 0.66. These metrics provide insights into the strengths and weaknesses of each model, guiding considerations for their specific applications in the context of the task at hand.

Random Forest appears to perform well with the highest accuracy (0.898) and reasonably balanced F1-scores for both significant and non-significant earthquakes. It can be considered as an efficient choice for this classification problem. Random Forest combines multiple decision trees to make accurate predictions and is less prone to overfitting.

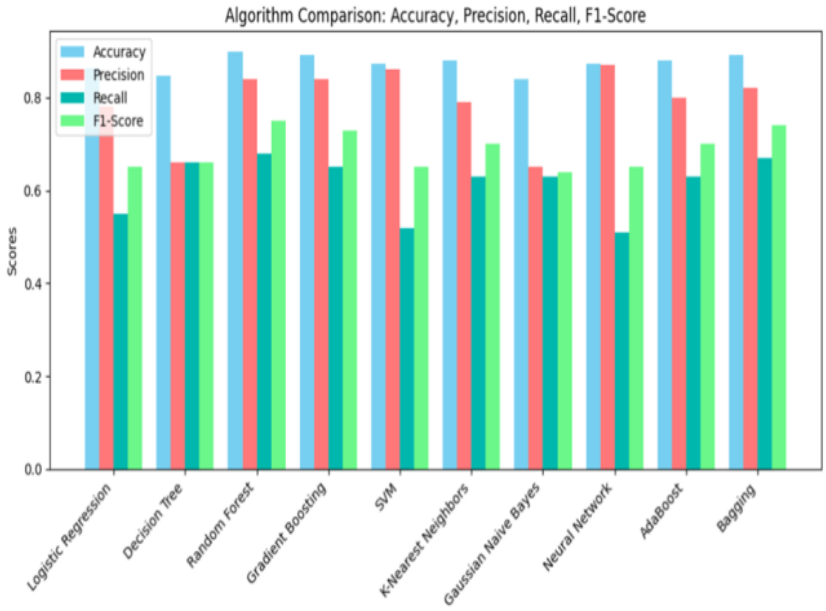

Fig. 1: Result Analysis

5 Conclusion

The study presented here is about spatial data mining on earthquake significance using attributes like magnitude. Spatial data mining plays a pivotal role in revealing patterns and insights within data, facilitating informed decision-making across various industries. In this research ten classification algorithms are used to tell the earthquakes are significant or not significant. The Decision Matrix shows the performance of each algorithm on metrics such as F1-score, accuracy, precision, and recall. Random forest appears to perform well with the highest accuracy and reasonably balanced F1 scores for both significant and nonsignificant earthquakes. Hence it is considered as best algorithm to find out if the earthquake is significant or not. In summary, this study underscores the efficacy of spatial data mining in tackling real-world challenges and emphasizes the critical role of algorithm selection in addressing classification problems. The research findings distinctly illustrate the significant impact on areas such as disaster response and earthquake forecasting.

References

[1] Qin Ding, Qiang Ding, and William Perrizo. Parm—an efficient algorithm to mine association rules from spatial data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 38(6):1513–1524, 2008.

[2] Arpan Man Sainju, Danial Aghajarian, Zhe Jiang, and Sushil Prasad. Parallel grid-based colocation mining algorithms on gpus for big spatial event data. *IEEE Transactions on Big Data*, 6(1):107–118, 2018.

[3] Mengjie Zhou, Tinghua Ai, Guohua Zhou, and Wenqing Hu. A visualization method for mining colocation patterns constrained by a road network. *IEEE Access*, 8:51933–51944, 2020.

[4] Zhe Jiang. A survey on spatial prediction methods. *IEEE Transactions on Knowledge and Data Engineering*, 31(9):1645–1664, 2018.

[5] Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, 12(3):372–390, 2000.

[6] Yan Huang, Shashi Shekhar, and Hui Xiong. Discovering colocation patterns from spatial data sets: a general approach. *IEEE Transactions on Knowledge and data engineering*, 16(12):1472–1485, 2004.

[7] Edwin M Knorr and Raymond T Ng. Finding aggregate proximity relationships and commonalities in spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):884–897, 1996.

[8] Paul C Smits and Anders Friis-Christensen. Resource discovery in a european spatial data infrastructure. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):85–95, 2006.

[9] Yong Ge, Zijun Yao, and Huayu Li. Computing co-location patterns in spatial data with extended objects: a scalable buffer-based approach. *IEEE Transactions on Knowledge and Data Engineering*, 33(2):401–414, 2019.

[10] Zhe Jiang. A survey on spatial prediction methods. *IEEE Transactions on Knowledge and Data Engineering*, 31(9):1645–1664, 2018.

[11] Abhiram Shri Chakravadhanula, Jaswanth Kolisetty, Karthik Samudrala, Bharat Preetham, and Rajiv Senapati. Novel decentralized security architecture for the centralized storage system in hadoop using blockchain technology. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, pages 1–4. IEEE, 2022.

[12] Anasuya Sahoo and Rajiv Senapati. A parallel approach to partition-based frequent pattern mining algorithm. In *Intelligent Systems*, pages 93–102. Springer, 2022.

[13] Anasuya Sahoo and Rajiv Senapati. A novel approach for distributed frequent pattern mining algorithm using load-matrix. In *2021 International Conference on Intelligent Technologies (CONIT)*, pages 1–5. IEEE, 2021.

[14] Pranaya Pournamashi Patro and Rajiv Senapati. Advanced binary matrix-based frequent pattern mining algorithm. In *Intelligent Systems*, pages 305–316. Springer, 2021.

[15] Anasuya Sahoo and Rajiv Senapati. A boolean load-matrix based frequent pattern mining algorithm. In *2020 International Conference on Artificial Intelligence and Signal Processing (AISP)*, pages 1–5. IEEE, 2020.

[16] Sricharan Muttineni, Siddhesh Yerramneni, Bharath Chandra Kongara, Gowtham Venkatachalam, and Rajiv Senapati. An interactive interface for patient diagnosis using machine learning model. In *2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, pages 1–5. IEEE, 2022.

[17] GSK Ganesh Prasad, A Ajay Chowdari, Kaligithi Pritham Jona, and Rajiv Senapati. Detection of ckd from ct scan images using knn algorithm and using edge detection. In *2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, pages 1–4. IEEE, 2022.

[18] Chaitanya Datta M and Rajiv Senapati. An adoptive heart disease prediction model using machine learning approach. In *2022 OITS International Conference on Information Technology (OCIT)*, pages 49–54. IEEE, 2022.

[19] Rajiv Senapati. A novel classification-based parallel frequent pattern discovery model for decision making and strategic planning in retailing. *International Journal of Business Intelligence and Data Mining*, 23(2):184–200, 2023.

[20] Chaitanya Datta Maddukuri and Rajiv Senapati. Hybrid clustering-based fast support vector machine model for heart disease prediction. In *International Conference on Machine Learning, IoT and Big Data*, pages 269–278. Springer, 2023.

[21] KVNS Raviteja, KVBS Kavya, R Senapati, and KR Reddy. Machine-learning modelling of tensile force in anchored geomembrane liners. *Geosynthetics International*, pages 1–17, 2023.

[22] Sarath Chandra Manda, Sricharan Muttineni, Gowtham Venkatachalam, Bharath Chandra Kongara, and Rajiv Senapati. Image stitching using ransac and bayesian refinement. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–5, 2023. https://doi.org/10.1109/CONIT59222.2023.10205634.

[23] Siddhesh Yerramneni, Kotta Sai Vara Nitya, Sirikrishna Nalluri, and Rajiv Senapati. A generalized grayscale image processing framework for retinal fundus images. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–6, 2023. https://doi.org/10.1109/CONIT59222.2023.10205834.

[24] Karthik Samudrala, Jaswanth Kolisetty, Abhiram Shri Chakravadhanula, Bharat Preetham, and Rajiv Senapati. Novel distributed architecture for frequent pattern mining using spark framework. In *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pages 1–5, 2023. https://doi.org/10.1109/CONIT59222.2023.10205903.

[25] M Chaitanya Datta, B Venkaiah Chowdary, and Rajiv Senapati. Multi disease prediction using ensembling of distinct machine learning and deep learning classifiers. In *International Conference on Soft Computing and its Engineering Applications*, pages 245–257. Springer, 2023.

[26] Venkaiah Chowdary B, Chaitanya Datta M, and Rajiv Senapati. An improved cardiovascular disease prediction model using ensembling of diverse machine learning classifiers. In *2023 OITS International Conference on Information Technology (OCIT)*, pages 329–333, 2023. https://doi.org/10.1109/OCIT59427.2023.10430692.