Applications of FP-Growth and Apriori Algorithm for Mining Fuzzified Spatial Dataset

Puneet Matapurkar, Saurabh Shrivastava

***Abstract:** Spatial data, also called geospatial data, is term needed to describe data linked to or containing knowledgeable data about a particular location on Earth's surface. Spatial data mining's primary goal is to uncover hidden complicated information from spatial & non-spatial information in spite of their enormous quantity and find the spatial relations density. Spatial Data Mining techniques, however, continue to be an expansion of individuals utilized in standard data mining. Spatial Data is an extremely challenging area since enormous quantities of spatial data have been obtained from the remote sensed to the GIS (Geographic Information Systems), ecological estimation, computer cartography, planning and many more. In a given paper, we only focus on an essential type of spatial vagueness termed as spatial fuzziness. Spatial fuzziness intakes the property of several spatial objects in certainty which don't contain boundaries of sharp type and interiors or whose boundaries as well as interiors can't be determined in precise form. This paper provides the method for finding fuzzy spatial data of association rule. Association rules provided valuable data in the assessment of important correlations observed in big databases. Compared to the previous research work, the current approach for there search highlights the superiority over the same dataset in terms of time taken and generated rules. The rules generated tell about the occurrence of attributes. The results show that the current research is more efficient than that of the previous work and also less time-consuming.*

***Keywords:** Data Mining, Spatial Data Mining (SDM), Association Rule Mining (ARM), Apriori, FP-Growth, Spatial Data, Fire Data.*

I. INTRODUCTION

Advancement in the database technology and techniques for collecting data consists of barcode reader, remote sense, satellite telemetry, etc. have gathered huge data in large datasets. These rapidly increasing data generate the need for discovering the data, leading to a successful evolving sector named data mining which is the one step of the KDD process. The finding of knowledge in datasets could have described as the finding from big databases of interestingness, hidden, & which was formerly not known knowledge. Data mining is the integration of multiple areas that are database systems, Machine Learning methods, data representation, Statistics &hypothesis of information.Though numerous data mining learning has been conducted into transactional & relational datasets, data mining is also useful in many different kinds of datasets like spatial dataset, sequential databases, object-oriented related dataset, Multimedia relevant dataset and so on. The main focal point is to study about SDM techniques, explicitly discovering interested spatial data understanding**.** Spatial Data is the information of substance occupying spaces. Spatial dataset keeps among such substance spatial objects represent via kinds of spatial data & their relationships. Spatial data holds topologically information which structured through constructions of spatial indexing and accessed through techniques of spatial access. These separate characteristics of spatial database present difficulties and provide possibilities from spatial data for mining information. Spatial data mining, or known spatial database ledge discovery, related the removal for implied data, spatial relationships, or another type of patterns that haven't unambiguously kept into spatial datasets. Machine learning method with statistics and database system has studied in the related work which is done in the existing work to revealing the data from a large database. In addition, it gives a great contribution to the field of SDM from spatial databases. The SDM method has performed outstandingly on large datasets and also useful to find data type complexity of spatial data. The most important job of the SDM technique is a spatial accessing method. It is possible to apply spatial data mining techniques to extract interesting and regular understanding from big spatial databases. They can be used in specific to understand spatial data, to discover interactions between spatial and non-spatial information, to build spatial knowledge bases, to optimize queries, to reorganize information in spatial databases, to capture general features in an easy as well in a short way, and so on. Spatial data knowledge could have different types, such as characteristic and discriminating laws, mining & explanation of well-known buildings or clusters, spatial associations & others.

II. BACKGROUND OF SPATIAL DATA MINING

The most popular method for evaluating spatial data was the statistical spatial analysis. Statistics study defined as well-exploring field & thus a big range of the existing algorithms, as well as multiple methods for optimization. It handles numerical data very well and generally develops a practical model of spatial phenomenon. This strategy has the main drawback of postulation that the spatially distributed information is statistical independent. This creates issues as numerous spatial data is in reality unified, specifically their adjacent substance influence spatial objects.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author
Puneet Matapurkar*, Department of Mathematical sciences and computer applications, Bundelkhand university, Jhansi (Uttar Pradesh) India. E-mail: pmatapurkar.mca@gmail.com
Dr. Saurabh Shrivastava, Department of Mathematical sciences and computer applications, Bundelkhand university, Jhansi (Uttar Pradesh) India. E-mail: hanu.saurabh@gmail.com

Retrieval Number: B3866129219/2019©BEIESP
DOI: 10.35940/ijeat.B3866.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Regression models can be used to some extent to relieve this issue with spatially lag types of reliant variable. Unfortunate, this complicates the entire module procedure & could just achieve via professionals through a reasonable quantity of understanding of the domains& statistics capability. On the other hand, for the evaluation of spatial data, it is not the kind of method that we want to present to end customers. In addition, the statistical approach can not very well model nonlinear laws, &representative value such as name is badly treated. Furthermore, the statistics technique does not operate completely with incomplete information.

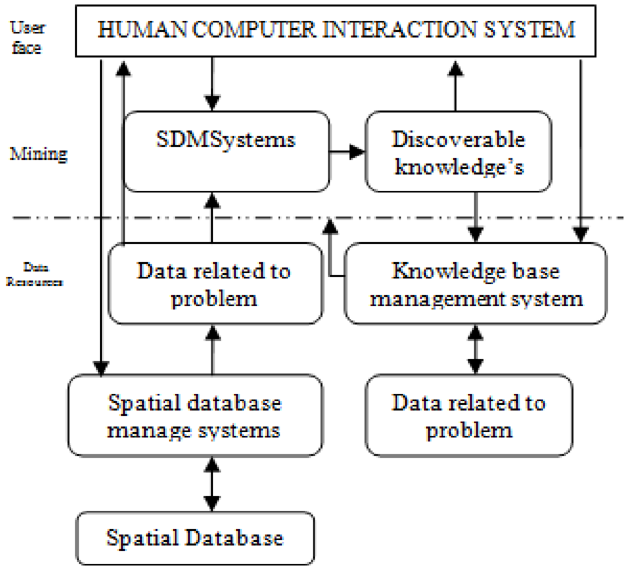


Fig. 1.SDM efficient structure

Another issue associated with statistical spatial analysis is the calculation of the costly result. Researchers have suggested numerous techniques to discover information from big databases with the introduction of information mining. Most of them focus on databases for relationships or transactions. These techniques have sought to combine mature fields such as machine learning, databases, and statistics. Such learning set a basis for the mine of spatial data. Machine learning methods are commonly used in SDM. Another technique to discover understanding in spatial databases has also been expanded. We did some frequently utilized terms in SDM in the next part.

A. Primitives of Spatial Data Mining Rules

In particular, different types of rules can be found from databases. For instance, it is possible to exploit distinctive rules, discriminating rules, association rules, or divergence rules & development. A rule of spatial character is a general spatial data description. A rule telling the overall cost variety of buildings in different geographical areas in a town, for instance, is a rule of spatial character. A spatial discriminating principle is common characteristics portrayal that discriminates else contrast spatial data classes from each additional class such as compare the range of housing cost into distinct geographic areas. At last, a rule for a spatial association is a rule that illustrates the involvement of single/group of features in spatial databases by another set of features.

Thematic maps: This represents only one or some of the attributes from spatial distribution. It varies from the general or referenced map wherever the primary goals are current objects location with respect to another spatial object. This map can subsist to detect distinct rules. Like, while analyzing a geographic region's general weather pattern. There are 2 manners in which thematic maps can be represented: raster and vector. It has pixel connected with different values of the attributes in the raster image form. Such as, a map might include spatial object altitude coded as pixel intensity (or color). To representing vector geometry is used as a spatial object, most frequently the representation of the border along with the thematic characteristics. For example, the boundaries point & the equivalent altitude values can represent a park.

Image databases: These are unique types of spatial databases in which information consists almost completely of images or pictures. Image databases have used in a remote sense, medical image, etc. They are generally stored in grid arrays that represent the strength of the picture into more than one spectral- sequence.

B. Spatial data structures, computations& Queries

Any one author cannot submit more than 05 papers for the same volume/issue. The authors of the accepted manuscripts will be given a copyright form and the form should accompany your final submission. It is noted that:

- Each author profile along with photo (min 100 word) has been included in the final paper.
- Final paper is prepared as per journal the template.
- Contents of the paper are fine and satisfactory. Author (s) can make rectification in the final paper but after the final submission to the journal, rectification is not possible.

III. LITERATURE REVIEW

Recent research on knowledge discovery and spatial data mining is highlighted in the article. We assessed several kinds of literature on spatial data features, prevalent methods in SDM, methods engaged with SDM& Spatial ARM. This study concludes through multiple perspectives upon important effort in SDM as well latest study in Spatial ARM [1].

This paper [2] deals with the research of different information mining clustering algorithms and focuses on clustering basics, requirements, and classification, issue area, and clustering algorithms implementation region. They studied in detail the various types of clustering methods in this document and summarized it. We included definition, requirement, clustering techniques implementation. We also provide details with the advantages and disadvantages of the classification of clustering techniques and their respective algorithms. This article, therefore, offers a fast overview of the various clustering methods in information mining.

This paper [3] attempts to clarify the distinct duties of spatial data mining as well. It also describes how we explain a specific spatial data assignment. The challenge is to take up as a significant area of spatial data mining, etc.

This article summarizes the latest work on SDM, ranging from a systematic SDM framework to SDM methods, guidelines on Spatial Association, etc. This demonstrates the promising field of spatial data mining and many difficult problems.

The purpose of this document is to present GIS & SDM (spatial data mining), GIS & SDM instruments, procedural methods, difficulties as well as the main job of spatial ARM in GIS big dataset [4].

The applicability of FIM methods on the MapReduce platform is investigated in this article. We are introducing two fresh techniques for taking out a big dataset. Dist-Eclat centered on velocity whilst Big FIM has optimized for running upon real big dataset. They demonstrate scalable techniques in our studies [5].

Zhou et al. [6] suggest using frequency ranges to balance PFP groups. PFP's grouping approach is neither memory effective nor velocity effective. Some nodes can read nearly the entire dataset into the memory space, which is considered as many excessive in the Big Data zone. They suggest balancing the allocation with singletons for quicker execution, but as they discussed additional into a document, explore space partitioning with individual items that are not to be the most effective manner.

Malek et al. [7] are proposing an approximate FIM technique using K-Medoids for clustering operations also using the delegate operations of the clusters as candidate items. The writers introduced a variant of MapReduce that parallels the step of support counting.

The techniques of fuzzy sets and association rules were provided by David Olson and Yanhong Li [10], given the differential emphasis of consumers on each attribute through fuzzy areas. Discovered rules are articulated in a more understandable natural language for beings. This research used a synthetically deal though real dataset to demonstrate an efficiency of the suggested strategy. The existing method was not of so good adapted to extracting Fuzzy-weighted association rules by the amount of dataset. This method of both fuzzy-sets also association rules have worked together in this job to allow distinct emphasis on each.

In a single set of an entity, they retained similar objects and looked after no two things in a system of co-location belong to an individual group. Place proximity depends on distance measurement from Euclidean. A technique based on Map-Reduce is implemented in this paper [11] to discover all patterns of co-location from a geo dataset spread between nodes. The visualization of tests on larger data sets is also shown.

In this paper [12], Incremental topological association rules for the mining of spatial datasets are suggested using probabilistic methodology. The geo data set is read and passed by an incremental association rule discovery algorithm based on probability to refute the laws of topological spatial associations. The assumption here that the information store is dynamically updated for each time period is that the laws are derived from the spatial database. This paper uses this dynamic nature to deal with the proposed incremental topological rule mining process..

This paper [13] suggests a modern method for creating a service-oriented architecture for weather information systems that uses these data mining techniques to forecast weather. This can be achieved by using Artificial Neural Network and Decision Tree Algorithms and Real-Time Meteorological Data. In order to generate classification rules for mean weather variables, the algorithm provided the best results. Such data mining techniques have been shown to be appropriate for weather forecasting.

This paper's [14] scope was limited to forecasting for seven days the maximum temperature and minimum temperature gave the weather information for the past two days. Using a linear regression model and a variant on a functional regression model, the latter is capable of capturing climate patterns. Both of our models were outperformed by professional weather forecasting services, although the discrepancy between our models and the professional models fell rapidly for later-day forecasts, and our models might outperform professional models for even longer time scales.

This work [15] uses the techniques of data mining to analyze soil datasets. These data mining algorithms are used for classification purposes to test the soil datasets. In this analysis, the different techniques of data mining are used and compared.

An analysis [16] aims to obtain a view of modern technologies for smart farming large-scale information and to recognize the socio-economic challenges to be tackled. Using a structured approach, a theoretical framework for research was developed which can also be used for future study on this topic. The review shows that Smart Farming's range of large-scale applications reaches beyond primary production and affects the entire food supply chain.

IV. PROPOSED METHODOLOGY

It is proposed to modify the existing approaches and develop new approaches for Association Rule Mining. On spatial data, in order to overcome some of the limitations of the existing approaches.

- Apriori algorithm is slow; in order to improve the speed, it is proposed to make new approaches like Karnaugh map, neural networks, etc. To develop new models and algorithms for Association rule mining for spatial data.
- Due to the presence of uncertainty in relationships among the fields/attributes of the data the existing algorithms lead to the generation of spurious patterns in order to address this challenge of uncertainty. New Fuzzy set-based approaches will be proposed which can prune the spurious patterns.

A. Association Rule Mining

Usually, association rule has mandatory to assure minimum support specific by the user & at the same time minimum confidence specified by the user. Generation of association rules is generally split into two distinct steps:

1. First, to locate all common item sets in a database, minimal assistance is implemented.

2. Secondly, these all frequent item-sets with a minimized level of confidence has utilized to establish rules [3].

Association Rule Mining is a method familiar with how most association rules mining in a multitude of respects in the connection between the information.

Association mining has been implemented that discovers dependencies among the attribute values also it has to emerge as a popular section of studies.

It is the most well-known method for the research to discover interesting relationships among attributes in big datasets in data mining notion.
The rapid development of information extraction from big transactional data sets fuelled the demand for the discovery of knowledge and associative relationships between products. Mining has played a significant part in defining the most frequently transacted items in the latest previous association rule and creates associative laws between different items [9].

There are two essential fundamental techniques for association Rules, Supp(s) & Conf(C). Because the database is enormous also consumers only care regarding the repeatedly bought substances, users frequently determine S & C thresholds to reveal those all rules whose are not of so exciting or necessary.

Confidence(C)described as given no. of transactions % or the fraction containing A with B for entirety records containing A.

A different parameter has put in place for defining the strong association among A & B item sets like confidence, support, and interest. These parameter shave defined on the probabilistic model based on the following:

S (A->B) =P(A,B), or the transactions % into the database which includes both A & B
C (A->B)=P(A, B)/P(A), or the transactions % that have B in the transaction that has A
Interest (A->B) = P(A,B)/P(A) P(B) represent statistical independence analysis[3].

Lift is the rise in B's selling percentage when selling A. Lift's mathematical formula is the following.
Lift (A⇒B) = (Confidence (AB)) / (Support (A))

B. FP-Growth

Two significant steps are needed to find the co-located pattern
1) Spatial data conversion
a) Representing the input
b) Spatial data information
2) Mining of co-location pattern
a) FP-Tree building
b) FP-Tree mine
c)

If one time the FP-tree is built, the frequent patterns mining from the compressed tree array is carried out using the procedure of FP tree mining described in [8]. FP-Growth works in the manner that divides & conquers. The databases firstly scanned extract a list of the frequent-items using these items sorted in decreasing order of frequency. Then the database has packaged in a frequency-patterns tree (or FP-tree), that receives data about item associations, according to frequency-descending list. The FP-tree is mined by constructing its conditional form of a base pattern (a "sub-database," which involves prefix routes set inside FP tree co-occurred amid suffix pattern), after that built their Conditional FP tree, also implementing recursively mining upon this tree. Pattern development has achieved via combining suffix patterns with frequently occurred patterns acquired as of the Conditional FP-Tree.

C. Steps of Proposed Algorithm

step 1: Load Fire Dataset
step 2: Preprocess the dataset
a) During preprocessing convert dataset into fuzzy data
b) Divide each attribute into four parts
step 3: Apply the Apriori algorithm
a) Generation of Frequent Itemset- Generates those all item sets whose support value is equal or greater than to mins up value.
b) Generation of Rule- Generates rules from each frequent itemset whose have high confidence value, where each rule is a binary partitioning of a frequent itemset
Whereas,
Support=0.1, 0.2, 0.3….
Confidence=0.9
step 4: Apply FP-Growth
a) FP tree building
b) FP-Tree Mining
step 5: Number of rules generated
Comparison of the results

V. RESULTS AND DISCUSSIONS

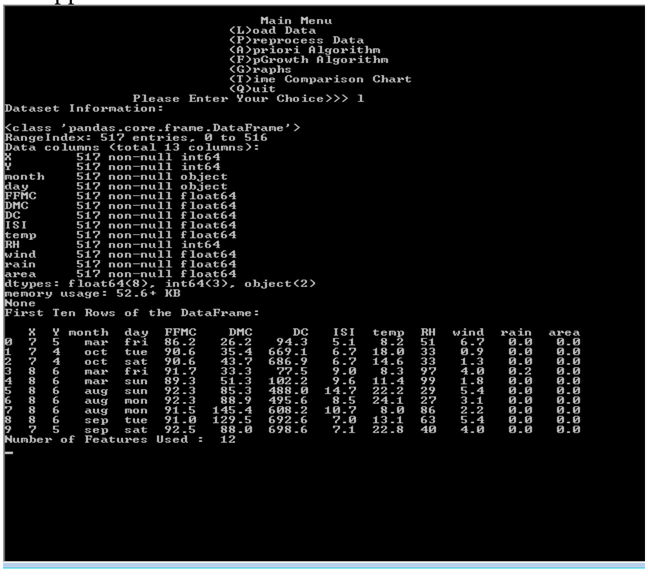
The simulation platform is Python 3.6. In this research work, the previous algorithm used was Association rule mining using the Apriori algorithm. The data set used previously was in the numerical form which is later converted into fuzzified. Fuzzification is to divide the continuous quantity in the fuzzy domain into several levels, according to the requirement; each level can be regarded as a fuzzy variable and corresponds to a fuzzy subset or a membership function.

A. Dataset

The Forest Fire Weather Index (FWI) is Canada's fire hazard classification scheme and contains six parts: buildup index (BUI), drought code (DC), duff moisture code (DMC), fine fuel moisture code (FFMC) FWI and ISI. Started parts work with the fuel code: in this FFMC indicates surface litter of moisturing contents and affects ignition & spreading fire, whereas DMC & DC signify the shallow moisture contents with the deep crude layer that influences the fire intensity. The ISI is a score that correlates by the distribution of flame velocity, while the BUI is the quantity of fuel available. The fuel moisture codes also involve previous weather conditions memory (time lag): 16 hours for FFMC, 12 days for DMC and 52 days for DC.
This data is available at:
https://archive.ics.uci.edu/ml/datasets/forest+fires
Fuzzifying the spatial dataset: Every attribute of a dataset is divided into min-max range example 0-100. The data is then divided into 4 specific ranges example area large, area small, etc as shown in figure 3.

B. Results visualization

1. Support - 0.1 Confidence - 0.9



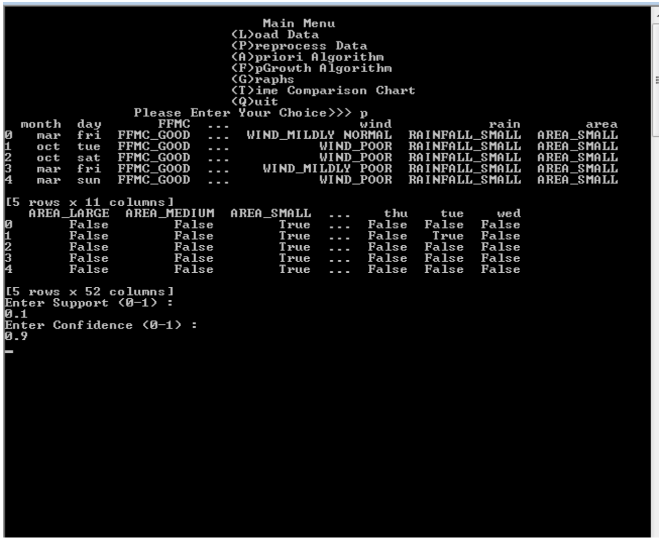
```
                                   Main Menu
                                (L)oad Data
                                (P)reprocess Data
                                (A)priori Algorithm
                                (F)pGrowth Algorithm
                                (G)raphs
                                (T)ime Comparison Chart
                                (Q)uit
                     Please Enter Your Choice>>> l
Dataset Information:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 517 entries, 0 to 516
Data columns (total 13 columns):
X         517 non-null int64
Y         517 non-null int64
month     517 non-null object
day       517 non-null object
FFMC      517 non-null float64
DMC       517 non-null float64
DC        517 non-null float64
ISI       517 non-null float64
temp      517 non-null float64
RH        517 non-null int64
wind      517 non-null float64
rain      517 non-null float64
area      517 non-null float64
dtypes: float64(8), int64(3), object(2)
memory usage: 52.6+ KB
None
First Ten Rows of the DataFrame:

   X  Y month  day  FFMC    DMC     DC   ISI  temp  RH  wind  rain  area
0  7  5   mar  fri  86.2   26.2   94.3   5.1   8.2  51   6.7   0.0   0.0
1  7  4   oct  tue  90.6   35.4  669.1   6.7  18.0  33   0.9   0.0   0.0
2  7  4   oct  sat  90.6   43.7  686.9   6.7  14.6  33   1.3   0.0   0.0
3  8  6   mar  fri  91.7   33.3   77.5   9.0   8.3  97   4.0   0.2   0.0
4  8  6   mar  sun  89.3   51.3  102.2   9.6  11.4  99   1.8   0.0   0.0
5  8  6   aug  sun  92.3   85.3  488.0  14.7  22.2  29   5.4   0.0   0.0
6  8  6   aug  mon  92.3   88.9  495.6   8.5  24.1  27   3.1   0.0   0.0
7  8  6   aug  mon  91.5  145.4  608.2  10.7   8.0  86   2.2   0.0   0.0
8  8  6   sep  tue  91.0  129.5  692.6   7.0  13.1  63   5.4   0.0   0.0
9  7  5   sep  sat  92.5   88.0  698.6   7.1  22.8  40   4.0   0.0   0.0
Number of Features Used :  12
_
```

Fig. 2.Initial data

Figures 2 and 3 show the results of numerical and fuzzified dataset respectively.

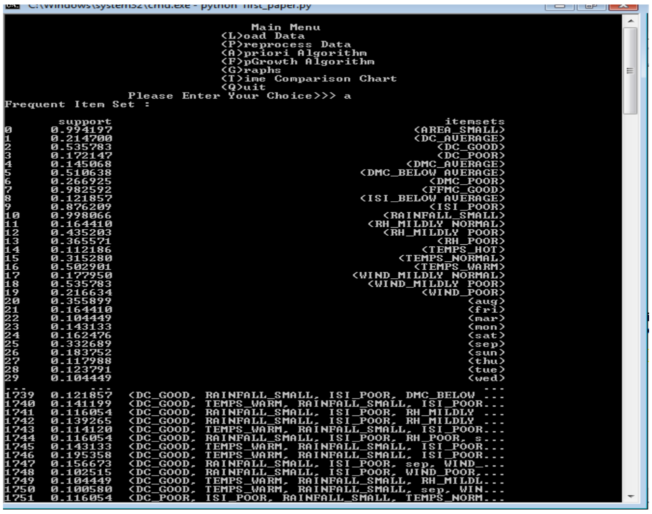


```
                                   Main Menu
                                (L)oad Data
                                (P)reprocess Data
                                (A)priori Algorithm
                                (F)pGrowth Algorithm
                                (G)raphs
                                (T)ime Comparison Chart
                                (Q)uit
                     Please Enter Your Choice>>> p
  month  day       FFMC  ...                wind          rain        area
0   mar  fri  FFMC_GOOD  ...  WIND_MILDLY NORMAL  RAINFALL_SMALL  AREA_SMALL
1   oct  tue  FFMC_GOOD  ...           WIND_POOR  RAINFALL_SMALL  AREA_SMALL
2   oct  sat  FFMC_GOOD  ...           WIND_POOR  RAINFALL_SMALL  AREA_SMALL
3   mar  fri  FFMC_GOOD  ...    WIND_MILDLY POOR  RAINFALL_SMALL  AREA_SMALL
4   mar  sun  FFMC_GOOD  ...           WIND_POOR  RAINFALL_SMALL  AREA_SMALL

[5 rows x 11 columns]
   AREA_LARGE  AREA_MEDIUM  AREA_SMALL  ...    thu    tue    wed
0       False        False        True  ...  False  False  False
1       False        False        True  ...  False   True  False
2       False        False        True  ...  False  False  False
3       False        False        True  ...  False  False  False
4       False        False        True  ...  False  False  False

[5 rows x 52 columns]
Enter Support (0-1) :
0.1
Enter Confidence (0-1) :
0.9
_
```

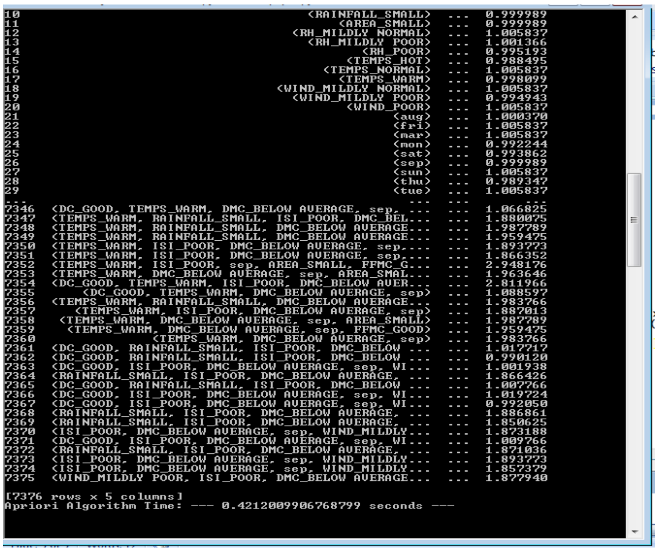
Fig. 3.Fuzzified data into 4 categories



```
C:\Windows\system32\cmd.exe - python first_paper.py
                                   Main Menu
                                (L)oad Data
                                (P)reprocess Data
                                (A)priori Algorithm
                                (F)pGrowth Algorithm
                                (G)raphs
                                (T)ime Comparison Chart
                                (Q)uit
                     Please Enter Your Choice>>> a
Frequent Item Set :

       support                                                 itemsets
0     0.994197                                             (AREA_SMALL)
1     0.214700                                             (DC_AVERAGE)
2     0.535783                                                (DC_GOOD)
3     0.172147                                                (DC_POOR)
4     0.145068                                            (DMC_AVERAGE)
5     0.510638                                      (DMC_BELOW AVERAGE)
6     0.266925                                               (DMC_POOR)
7     0.982592                                              (FFMC_GOOD)
8     0.121857                                      (ISI_BELOW AVERAGE)
9     0.876209                                               (ISI_POOR)
10    0.998066                                         (RAINFALL_SMALL)
11    0.164410                                     (RH_MILDLY NORMAL)
12    0.435203                                       (RH_MILDLY POOR)
13    0.365571                                                (RH_POOR)
14    0.112186                                             (TEMPS_HOT)
15    0.315280                                          (TEMPS_NORMAL)
16    0.502901                                            (TEMPS_WARM)
17    0.177950                                   (WIND_MILDLY NORMAL)
18    0.535783                                     (WIND_MILDLY POOR)
19    0.216634                                             (WIND_POOR)
20    0.355899                                                   (aug)
21    0.164410                                                   (fri)
22    0.104449                                                   (mar)
23    0.143133                                                   (mon)
24    0.162476                                                   (sat)
25    0.332689                                                   (sep)
26    0.183752                                                   (sun)
27    0.117988                                                   (thu)
28    0.123791                                                   (tue)
29    0.104449                                                   (wed)
...        ...                                                      ...
1739  0.121857  (DC_GOOD, RAINFALL_SMALL, ISI_POOR, DMC_BELOW ...
1740  0.141199  (DC_GOOD, TEMPS_WARM, RAINFALL_SMALL, ISI_POOR...
1741  0.116054  (DC_GOOD, RAINFALL_SMALL, ISI_POOR, RH_MILDLY ...
1742  0.139265  (DC_GOOD, RAINFALL_SMALL, ISI_POOR, RH_MILDLY ...
1743  0.114120  (DC_GOOD, TEMPS_WARM, RAINFALL_SMALL, ISI_POOR...
1744  0.116054  (DC_GOOD, RAINFALL_SMALL, ISI_POOR, RH_POOR, s...
1745  0.143133  (DC_GOOD, TEMPS_WARM, RAINFALL_SMALL, ISI_POOR...
1746  0.195358  (DC_GOOD, TEMPS_WARM, RAINFALL_SMALL, ISI_POOR...
1747  0.156673  (DC_GOOD, RAINFALL_SMALL, ISI_POOR, sep, WIND_...
1748  0.102515  (DC_GOOD, RAINFALL_SMALL, ISI_POOR, WIND_POOR,...
1749  0.104449  (DC_GOOD, TEMPS_WARM, RAINFALL_SMALL, RH_MILDL...
1750  0.100580  (DC_GOOD, TEMPS_WARM, RAINFALL_SMALL, sep, WIN...
1751  0.116054  (DC_POOR, ISI_POOR, RAINFALL_SMALL, TEMPS_NORM...
```

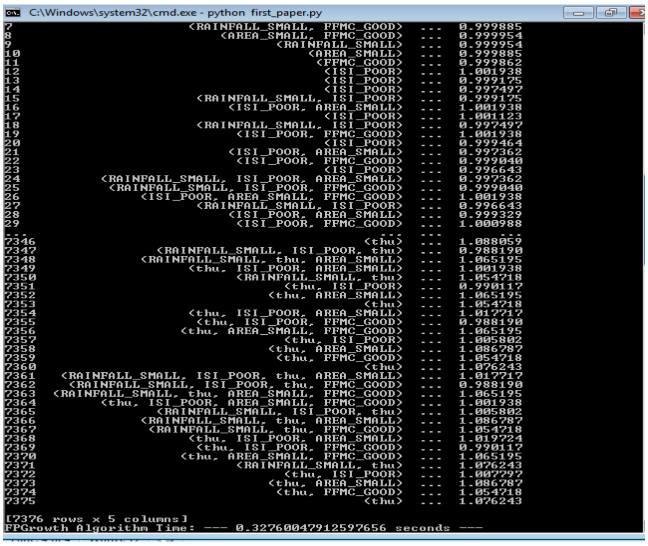
Fig. 4.Generation of rules



```
10                                     (RAINFALL_SMALL)  ...   0.999989
11                                         (AREA_SMALL)  ...   0.999989
12                                   (RH_MILDLY NORMAL)  ...   1.005837
13                                     (RH_MILDLY POOR)  ...   1.001366
14                                            (RH_POOR)  ...   0.995193
15                                          (TEMPS_HOT)  ...   0.988495
16                                       (TEMPS_NORMAL)  ...   1.005837
17                                         (TEMPS_WARM)  ...   0.998099
18                                 (WIND_MILDLY NORMAL)  ...   1.005837
19                                   (WIND_MILDLY POOR)  ...   0.994943
20                                          (WIND_POOR)  ...   1.005837
21                                                (aug)  ...   1.000370
22                                                (fri)  ...   1.005837
23                                                (mar)  ...   1.005837
24                                                (mon)  ...   0.992244
25                                                (sat)  ...   0.993862
26                                                (sep)  ...   0.999989
27                                                (sun)  ...   1.005837
28                                                (thu)  ...   0.989347
29                                                (tue)  ...   1.005837
...                                                 ...  ...        ...
7346  (DC_GOOD, TEMPS_WARM, DMC_BELOW AVERAGE, sep, ...  ...   1.066825
7347  (TEMPS_WARM, RAINFALL_SMALL, ISI_POOR, DMC_BEL...  ...   1.880075
7348  (TEMPS_WARM, RAINFALL_SMALL, DMC_BELOW AVERAGE...  ...   1.987789
7349  (TEMPS_WARM, RAINFALL_SMALL, DMC_BELOW AVERAGE...  ...   1.959475
7350  (TEMPS_WARM, ISI_POOR, DMC_BELOW AVERAGE, sep,...  ...   1.893773
7351  (TEMPS_WARM, ISI_POOR, DMC_BELOW AVERAGE, sep,...  ...   1.866352
7352  (TEMPS_WARM, ISI_POOR, sep, AREA_SMALL, FFMC_G...  ...   2.948176
7353  (TEMPS_WARM, DMC_BELOW AVERAGE, sep, AREA_SMAL...  ...   1.963646
7354  (DC_GOOD, TEMPS_WARM, ISI_POOR, DMC_BELOW AVER...  ...   2.811966
7355      (DC_GOOD, TEMPS_WARM, DMC_BELOW AVERAGE, sep)  ...   1.088597
7356  (TEMPS_WARM, RAINFALL_SMALL, DMC_BELOW AVERAGE...  ...   1.983766
7357     (TEMPS_WARM, ISI_POOR, DMC_BELOW AVERAGE, sep)  ...   1.887013
7358   (TEMPS_WARM, DMC_BELOW AVERAGE, sep, AREA_SMALL)  ...   1.987789
7359    (TEMPS_WARM, DMC_BELOW AVERAGE, sep, FFMC_GOOD)  ...   1.959475
7360               (TEMPS_WARM, DMC_BELOW AVERAGE, sep)  ...   1.983766
7361  (DC_GOOD, RAINFALL_SMALL, ISI_POOR, DMC_BELOW ...  ...   1.017717
7362  (DC_GOOD, RAINFALL_SMALL, ISI_POOR, DMC_BELOW ...  ...   0.990120
7363  (DC_GOOD, ISI_POOR, DMC_BELOW AVERAGE, sep, WI...  ...   1.001938
7364  (RAINFALL_SMALL, ISI_POOR, DMC_BELOW AVERAGE, ...  ...   1.866426
7365  (DC_GOOD, RAINFALL_SMALL, ISI_POOR, DMC_BELOW ...  ...   1.007766
7366  (DC_GOOD, ISI_POOR, DMC_BELOW AVERAGE, sep, WI...  ...   1.019724
7367  (DC_GOOD, ISI_POOR, DMC_BELOW AVERAGE, sep, WI...  ...   0.992050
7368  (RAINFALL_SMALL, ISI_POOR, DMC_BELOW AVERAGE, ...  ...   1.886861
7369  (RAINFALL_SMALL, ISI_POOR, DMC_BELOW AVERAGE, ...  ...   1.850625
7370  (ISI_POOR, DMC_BELOW AVERAGE, sep, WIND_MILDLY...  ...   1.873188
7371  (DC_GOOD, ISI_POOR, DMC_BELOW AVERAGE, sep, WI...  ...   1.009766
7372  (RAINFALL_SMALL, ISI_POOR, DMC_BELOW AVERAGE, ...  ...   1.871036
7373  (ISI_POOR, DMC_BELOW AVERAGE, sep, WIND_MILDLY...  ...   1.893773
7374  (ISI_POOR, DMC_BELOW AVERAGE, sep, WIND_MILDLY...  ...   1.857379
7375  (WIND_MILDLY POOR, ISI_POOR, DMC_BELOW AVERAGE...  ...   1.877940

[7376 rows x 5 columns]
Apriori Algorithm Time: --- 0.4212009906768799 seconds ---
```

Fig. 5.Generated frequent rules through apriori



```
C:\Windows\system32\cmd.exe - python first_paper.py
7                           (RAINFALL_SMALL, FFMC_GOOD)  ...   0.999885
8                               (AREA_SMALL, FFMC_GOOD)  ...   0.999954
9                                      (RAINFALL_SMALL)  ...   0.999954
10                                         (AREA_SMALL)  ...   0.999885
11                                          (FFMC_GOOD)  ...   0.999862
12                                           (ISI_POOR)  ...   1.001938
13                                           (ISI_POOR)  ...   0.999175
14                                           (ISI_POOR)  ...   0.997497
15                           (RAINFALL_SMALL, ISI_POOR)  ...   0.999175
16                               (ISI_POOR, AREA_SMALL)  ...   1.001938
17                                           (ISI_POOR)  ...   1.001123
18                           (RAINFALL_SMALL, ISI_POOR)  ...   0.997497
19                                (ISI_POOR, FFMC_GOOD)  ...   1.001938
20                                           (ISI_POOR)  ...   0.999464
21                               (ISI_POOR, AREA_SMALL)  ...   0.997362
22                                (ISI_POOR, FFMC_GOOD)  ...   0.999040
23                                           (ISI_POOR)  ...   0.996643
24               (RAINFALL_SMALL, ISI_POOR, AREA_SMALL)  ...   0.997362
25                (RAINFALL_SMALL, ISI_POOR, FFMC_GOOD)  ...   0.999040
26                    (ISI_POOR, AREA_SMALL, FFMC_GOOD)  ...   1.001938
27                           (RAINFALL_SMALL, ISI_POOR)  ...   0.996643
28                               (ISI_POOR, AREA_SMALL)  ...   0.999329
29                                (ISI_POOR, FFMC_GOOD)  ...   1.000988
...                                                 ...  ...        ...
7346                                              (thu)  ...   1.088059
7347                     (RAINFALL_SMALL, ISI_POOR, thu)  ...   0.988190
7348                   (RAINFALL_SMALL, thu, AREA_SMALL)  ...   1.065195
7349                         (thu, ISI_POOR, AREA_SMALL)  ...   1.001938
7350                               (RAINFALL_SMALL, thu)  ...   1.054718
7351                                    (thu, ISI_POOR)  ...   0.990117
7352                                  (thu, AREA_SMALL)  ...   1.065195
7353                                              (thu)  ...   1.054718
7354                         (thu, ISI_POOR, AREA_SMALL)  ...   1.017717
7355                          (thu, ISI_POOR, FFMC_GOOD)  ...   0.988190
7356                        (thu, AREA_SMALL, FFMC_GOOD)  ...   1.065195
7357                                    (thu, ISI_POOR)  ...   1.005802
7358                                  (thu, AREA_SMALL)  ...   1.086787
7359                                   (thu, FFMC_GOOD)  ...   1.054718
7360                                              (thu)  ...   1.076243
7361     (RAINFALL_SMALL, ISI_POOR, thu, AREA_SMALL)  ...   1.017717
7362      (RAINFALL_SMALL, ISI_POOR, thu, FFMC_GOOD)  ...   0.988190
7363    (RAINFALL_SMALL, thu, AREA_SMALL, FFMC_GOOD)  ...   1.065195
7364          (thu, ISI_POOR, AREA_SMALL, FFMC_GOOD)  ...   1.001938
7365                     (RAINFALL_SMALL, ISI_POOR, thu)  ...   1.005802
7366                   (RAINFALL_SMALL, thu, AREA_SMALL)  ...   1.086787
7367                    (RAINFALL_SMALL, thu, FFMC_GOOD)  ...   1.054718
7368                         (thu, ISI_POOR, AREA_SMALL)  ...   1.019724
7369                          (thu, ISI_POOR, FFMC_GOOD)  ...   0.990117
7370                        (thu, AREA_SMALL, FFMC_GOOD)  ...   1.065195
7371                               (RAINFALL_SMALL, thu)  ...   1.076243
7372                                    (thu, ISI_POOR)  ...   1.007797
7373                                  (thu, AREA_SMALL)  ...   1.086787
7374                                   (thu, FFMC_GOOD)  ...   1.054718
7375                                              (thu)  ...   1.076243

[7376 rows x 5 columns]
FPGrowth Algorithm Time: --- 0.32760047912597656 seconds ---
```

Fig. 6.Generated frequent rules through FP-Growth

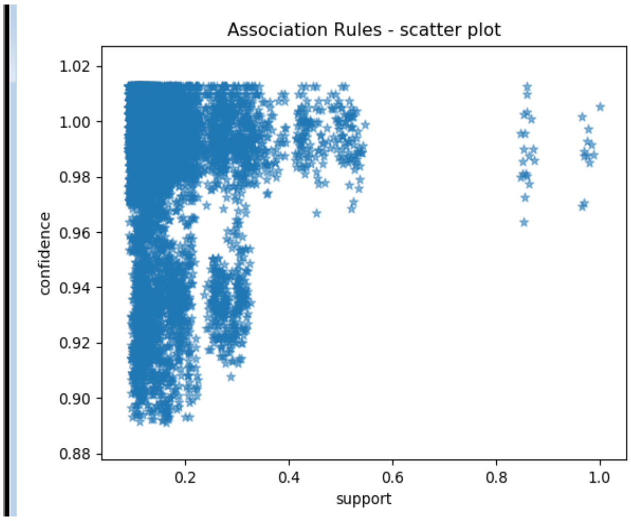


Fig. 7.Scatter plot of association rules

Retrieval Number: B3866129219/2019©BEIESP
DOI: 10.35940/ijeat.B3866.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

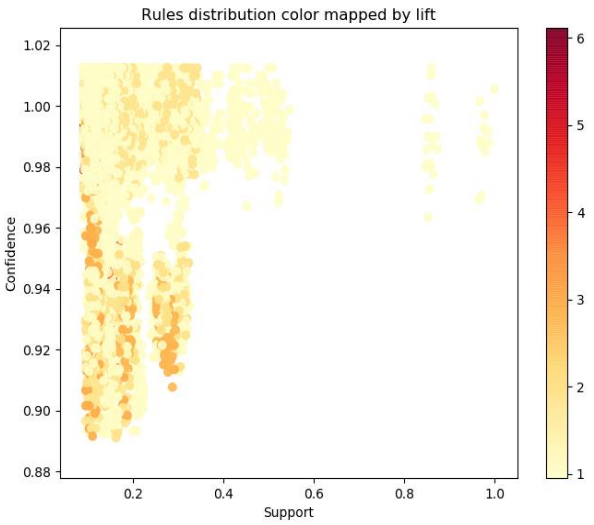


Fig. 8.Scatter plot of association rules with respect of Support, Confidence and Lift

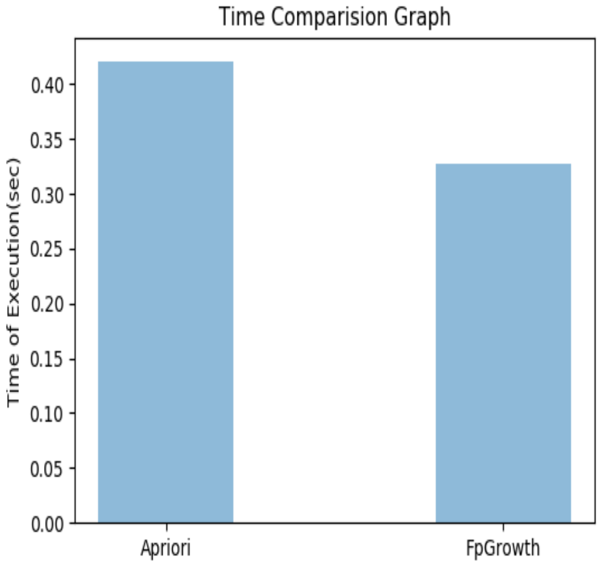


Fig. 9.Time comparison graph between Apriori and FP-Growth

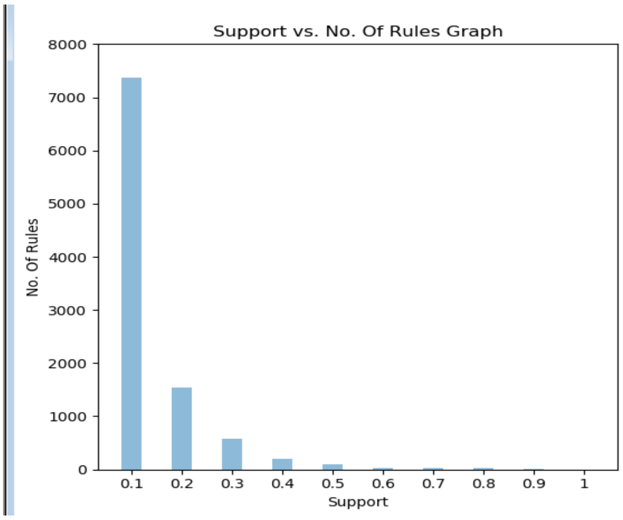


Fig. 10. Graph of the number of rules generated over different support count

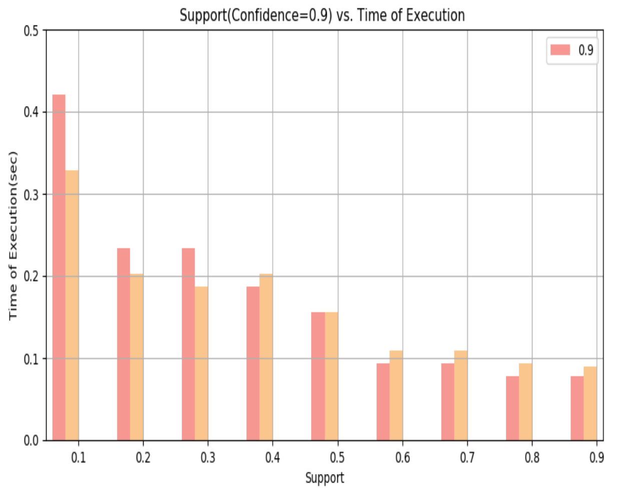


Fig. 11. Time Comparison graph at various support count

Figure 11 shows the comparison graph of varying support and constant confidence value

VI. CONCLUSION

It's likely to make use of the spatial form of data mining methods to obtain exciting and regular understanding from big spatial databases. In this paper, we have worked on spatial data (i.e. Fire dataset) which is fuzzified the attributes. These attributes categorized into four categories according to their characteristics. Apriori algorithm and the FP-Growth algorithm have used on this dataset. Both algorithms generate the same number of rules but the execution time has differed. The FP-Growth algorithm will be superior in performance in comparison to an existing Apriori algorithm for association rule mining on spatial data. The fuzzified dataset will help to categorize the data according to the need of the dataset. The proposed work will contribute to the new knowledge in the field of Association rule mining and finally data mining. In the future, we can try to apply a new technique of pattern generation which is more efficient than that of FP growth. Also, we can apply any other algorithm of fuzzification on a dataset and we can also use any other spatial dataset for new research purposes.

REFERENCES

1. Asmita Bist and Mainaz Faridi, "A Survey: On Spatial Data Mining", International Journal of Engineering Trends and Technology, Vol. 46, No. 6, pp. 327, April 2017, http://www.ijettjournal.org. (IJETT)
2. Deepti Sisodia, Lokesh Singh, Sheetal Sisodia, and khushboo Saxena, "Clustering Techniques: A Brief Survey of Different Clustering Algorithms", International Journal of Latest Trends in Engineering and Technology, Vol.1, Issue 3, Sept 2012. (IJLTET)
3. T. Kalaivani, P. Mangaiyarkarasi, S. Ramya & G. Anuratha, "An Overview of Spatial Data Mining", Imperial Journal of Interdisciplinary Research, Vol. 2, Issue 11, 2016, http://www.onlinejournal.in. (IJIR)
4. Hemlata Goyal, Chilka Sharma, and Nisheeth Joshi, "An Integrated Approach of GIS and Spatial Data Mining in Big Data", International Journal of Computer Applications, Vol. 169, No.11, July 2017. (published)

5. Moens S., Aksehirli E., and Goethals B., “Frequent Itemset Mining for Big Data”, *IEEE Int. Conf. on Big Data*, pp.111-118, 2013. (Conference proceedings)
6. L. Zhou, Z. Zhong, J. Chang, J. Li, J. Huang, and S. Feng, “Balanced parallel FP-Growth with MapReduce”, *In Proc. YC-ICT*, pp. 243–246, 2010. (Conference proceedings).
7. M. Malek and H. Kadima, “Searching frequent itemsets by clustering data: Towards a parallel approach using MapReduce”, *In Proc. WISE 2011 and 2012 Workshops, Springer Berlin Heidelberg*, pp. 251–258, 2013. (Workshop proceedings)
8. Jiawei Han, Jian Pei, Yiwen Yin, and Running Mao, Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. vol.8, no.1, 2004. (Book)
9. R.Z. Inamul Husain and S.K. Srivastava, "A Study of Different Association Rule Mining Techniques", IJCA, Vol. 108, No. 16, December 2014. (published)
10. David Olson and Yanhong Li, Mining Fuzzy Weighted Association Rules. HICSS, pp. 53-53, 2007. (Book)
11. Sandipan Maiti and R.B.V. Subramanyam, “Mining co-location patterns from distributed spatial data”, Journal of King Saud University - Computer and Information Sciences, 19 August 2018. (Available online)
12. Y. Jayababua, G.P.S. Varmab, and A. Govardhanc, “Incremental topological spatial association rule mining and clustering from geographical datasets using probabilistic approach”, Journal of King Saud University - Computer and Information Sciences, Vol. 30, Issue 4, pp. 510-523, October 2018. (Available online)
13. Zhan Jie Wang and A. B. M. Mazharul Mujib, “The Weather Forecast Using Data Mining Research Based on Cloud Computing”, IOP Conf. Series: Journal of Physics: Conf. Series 910, 2017, DOI: 10.1088/1742-6596/910/1/012020.
14. Mark Holmstrom, Dylan Liu, Christopher Vo, “Machine Learning Applied to Weather Forecasting“, International Journal of Forecasting, vol. 12, no. 1, pp. 57-71, 1996. (published)
15. S. S. Baskar, L. Arockiam, and S. Charles, “Applying Data Mining Techniques on Soil Fertility Prediction”, International Journal of Computer Applications Technology and Research, Vol. 2, Issue 6, pp. 660 – 662, 2013, www.ijcat.com. (Published)
16. Sjaak Wolfert, An Ge, CorVerdouwMarc-JeroenBogaardt, “Big Data in Smart Farming – A review”, Agricultural Systems, Vol. 153, pp. 69-80, May 2017. (Book)

AUTHORS PROFILE

Puneet Matapurkar, received a Master degree in computer application (MCA) from Rajiv Gandhi prodyogiki Vishwavidyalaya (RGPV) Bhopal (M.P.) and currently pursuing his Ph.D. Degree in computer science from Bundelkhand university Jhansi(U.P.). His research interests are in the field of data mining and machine learning algorithms.

Dr. Saurabh Shrivastava, received a Ph.D. degree in computer science from Bundelkhand university Jhansi(U.P.).His research interests are in the field of Soft computing techniques in data mining and machine learning.He has more than 20 publications to his credit in international journals and conferences.