International Journal of Modern Electronics and Communication Engineering (IJMECE) ISSN: 2321-2152
Volume No.-7, Issue No.-1, January, 2019

Analysis of Environmental Data Using Pattern Mining techniques

Dr. B. Lavanya
Department of Computer Science
University of Madras, Chennai-600025
lavanmu@gmail.com

V. Janani bai
Department of Computer Science
University of Madras, Chennai-600025

Abstract: The writing to the paper entitled "Analysis of environmental data using pattern mining techniques" is aimed at studying a pattern mining techniques that determine to convert enlarged data from high dimensional space to a lesser dimensional space to improve performance measurement, those attributes determine the quality of environment more suitable for agriculture using the unsupervised learning algorithm. This paper includes the different data mining techniques to diminish the dimensionality of dataset and environment dataset. The four techniques involve in this paper namely Apriori, K-Means Clustering, Frequent Pattern Tree and PCA algorithms used in environmental dataset. The dataset contains record of 543 columns and 12 rows, observations were acquired by estimating the variables of this time and frequent domain from pattern evaluated.

Keywords: K-means Clustering, Apriori Algorithm, Frequent Pattern Tree (FTP), Principle Component analysis (PCA).

I. INTRODUCTION

Data mining used to process the valid data and extract the information from interesting patterns or non-trivial and implicit. The data mining called as knowledge extraction, data dredging, business intelligence, pattern analysis, data archaeology, knowledge discovery database, informative harvesting and data analysis. The database handles the large amount of data and some of the data are hidden in database because data mining access the known data or relevant data in database. This kind of data used in real world application like population study, banking, sales, finance, production, employment etc. Information is retrieved using data mining tool and Finally visualizes the data as discovered knowledge.

Data mining is the most valuable and widely to exp
lain the process and analysis of large quantity of data to acquire novel, potentially useful & intelligent patterns from the database, have way or record known in fort, but cannot tackle the uncertainties of the fueled due to lack of look to use this known inform. DM (knowledge discovery from data) extracts useful patterns or knowledge from huge amount of data**.**

Data analysis & decision support

1. Market analysis & management
2. Risk analysis & management
3. Fraud detection & decision of unusual patterns (outlines)

II. LITERATURE SURVEY

In order of give a literature survey, searched for publications on HAR in view of wearable accelerometers. The information analyzed were Ankit Soni, NeesJan van Eck, Uzay Kaymak [5] One of the classification methods Data Visualization used for visualize textual data is the concept map. Data mining techniques using concept maps are often used in acquaintance domain apparition to elaborate the debate including analysis of the relationship between other commonly used methods with the visualization method.

Dr.Sankar Rajagopal [8] Uses clustering segmentation analysis methods as one of the most momentous methods. It is used in student's promotion methodology driven studies using statistical methodologies of statistics neural network genetic algorithm (GA). Janardan, Shikha Mehta[9] enhancing drift detection methods. These methods uses algorithm for classification of streaming data, the massive amount of data is being generated on diverse applications across the internet in every 60 seconds.
TB. Ai Munandar, Harsiti, Roy Amrullah Ritonga[10] In their proposed system, C4.5 classification method is used to learn and analyze the statistical data. Also Decision tree algorithm is used to analyze the testing phase to develop the functionality of the system.
Victoria Kayser,Knut Blind [13], One of the text mining methods can be seen as disciplined of improving insight methods. This technology diffuses user reception, combines road mapping and text mining to extend internal views.

Said Nawar, Abdul M.Mouazen[14] Proposes Multivariate Adaptive Regression Spine (MARS) methods for the mainstream of categorization of data sets scales with used the attributes of Lowest model performance. Paresh tenna[15] Apriori Algorithm and Association rule mining is becomes one of the significant responsibilities for this concept, the aim of the work to analyze the existing work for frequent pattern mining and evaluate the performance by comparing the Apriori algorithms to find out the candidate generation and tree pruning.

Gessica G.Burke[16] Proposes concept map process is to using traditionally involves six types of steps: preparation, generation, structuring, representation, interpretation and utilization. Chaman lal[17] Proposes that method principle component analysis is using synthetic data and real health datasets from UCI repository, the prediction model that have both a better fit and reduce number of attribute then those produced by using standard logistic regression alone.

III. FREQUENT PATTERN TREE

FP Tree is an efficient and scalable methods for data mining and complete set of frequent patterns by pattern growth, using an extended tree structure for storing compressed and set of information about frequent pattern names and tree structure. This section presents a comprehensive survey, helps to find frequent itemsets from huge amount of datasets.

FP Tree approach combines tree based pattern for compressed representation of the database more tree structure and efficient counting. The tree is used to represent the conditional data types are transaction data type format, the transaction is T and Ti. The FP tree is based on prefix based tree format, the itemset of the nodes that are created and support of the itemset that are the path from the node defined by the root u. the FP Tree is used for effective memory management

IV. PROPOSED WORK

The proposed works analysis the attributes to determine the quality of environment dataset to find their suitability for agriculture. The dataset was collected from Environmental (EDS) database. Those agriculture data and their associated data expression signatures in environmental dataset, are studied and analyzed using the following algorithms,

- Apriori method
- K-means clustering
- Frequent Pattern Mining
- Principle Component Analysis (PCA).

The overview of the proposed work explains the environmental dataset contains 12 attributes, using the mention algorithms, to choose the perfect features for agricultural attributes to determine the quality of environment more suitable for agriculture. In this paper, briefly described the proposed strategy for how, the four algorithm cluster and different pattern mining are Apriori, k - means, frequent pattern mining, principle component analysis, the main contributions of this chapter have been discussed the proposed four stage of the pattern mining of following chapter.

V. DATA DESCRIPTION

It includes physical, chemical and other natural forces. living things live in their environment. They constantly interact with it and adapt themselves to the environment condition. In the environment there are different interactions between field soil, field air, field rh, field soil wc, forest soil, forest air, forest rh, forest soil wc, and other living and non-living things. The dataset contains record of observations were acquired by estimating the variables of this time and frequent domain from pattern evaluated. The Environmental Datasets represent measurements taken in 3 distinct environmental settings at the University of Toronto Mississauga campus.

The data are separated into pond, field and forest data. Data are collected at these sites using HOBO U30 data loggers equipped with sensors monitoring soil moisture, soil temperature, air temperature and relative humidity. Data are collected hourly and downloaded monthly. In the environment there are different interactions between animals, plants, soil, water, and other living and non-living things. The dataset contains record of 743 columns and 12 rows, observations were acquired by estimating the variables of this time and frequent domain from pattern evaluated.

VI. METHODS

i. Apriori Algorithm

Apriori algorithm is a unsupervised learning algorithm that is widely used in data mining. Apriori works by generating associations rules between itemsets. It is widely used in market basket analysis and understanding the customer buying behavior. Apriori algorithm is easy to implement but also it's computationally expensive. The candidate one itemset generation based on the procedure that during the first scanning of the database count of each item and which will be updated in the column support count. To find frequent one item set, Compare support count of each item in candidate itemset with user defined minimum threshold, not satisfying items to be removed.

Even though the Aprioi algorithm based on candidate generation **it** is simple and efficient but it requires multiple passes over the database which leads to spent high running time and results in increased cost and waste of time. It spent more time for searching in the database and performing transactions for frequent itemsets. Apriori is inefficient in terms of memory requirement when large numbers of transactions are in consideration. The frequent itemset one to be used for two itemset candidate generation by join operation. The same procedure will be repeated till find k-frequent itemset generation. Aprioi algorithm in environmental dataset, to detect the attribute more suitable for agriculture and to find the quality of the environment.

The environmental dataset to start with scanning the database for every frequent item sets transaction. Although to generate the candidate itemsets singletons, pair, triple etc. And also count the number of frequent itemsets. After applying the minimum threshold value for candidate itemsets (C1) table. To find the frequent itemsets (L1), like this process going on when the candidate itemsets equal to 1, For example here I taken minimum threshold value = 0.01

- Apriori Algorithm determine itemsets in Environmental dataset
- Apriori algorithm scans the entire dataset in every iteration to calculate the support of each itemset
- In this algorithm frequently obtained itemset at a time by frequent mining subsets and scan the transaction database from the attributes
- In the environment dataset 4 itemset are extracted when used with Apriori algorithm

➢ Forest_soil_wc

- Pond_air_temp_c
- Pond_rh_wc
- Pond_soil_wc

- In this frequently obtained attributes are Forest soil wc and Pond air temp c

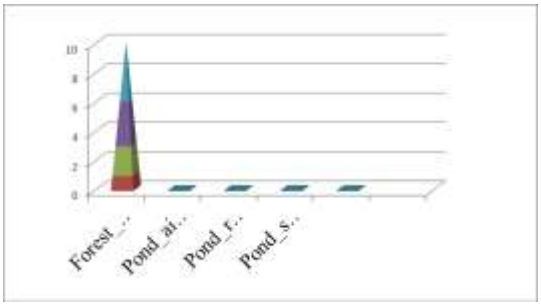


Fig 1: Apriori Algorithm Column Chart

The top most rule shows us that attribute are Forest_soil_wc, Pond_air_temp_c, Pond_rh_wc, Pond_soil_wc these attributes to determine the quality of environment dataset to find the suitable attribute for agriculture land**.**

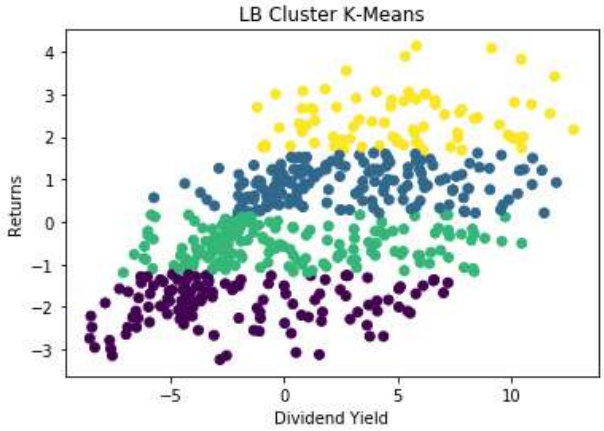
ii. K-Means Clustering

K – Means clustering algorithm is used to find the minimum distance between cluster and dataset. To analyze the complete dataset and used to find the centroid value. And also calculate Euclidean distance between two centroid. Based upon the threshold value and it can be divided into number of cluster. Each Cluster split over the datasets and relationship with centroid value. Centroid is distance between one cluster to another cluster values. The working process of k-means cluster for given below example. Here I taken number of cluster = 4.
Euclidean distance formula,

$$\sqrt{((x1-y1)^2 + (x2-y2)^2}$$

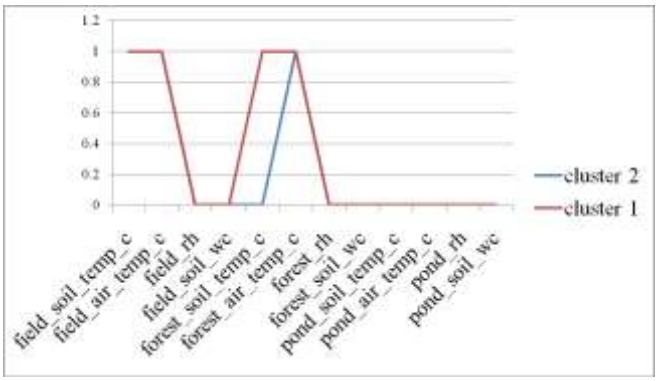
1. Ɒ – Dimensional vector, need to classify the vector into ‘k’ category
2. The centroid to the smallest distance is treated as the identified group.
3. Table: Environmental dataset using k—means clustering
4. In the 8 attribute only 2 attributes are repeated in the clusters that is Filed_soil and Field _air from the give number of clusters = 4

Fig 2: Environmental dataset using k—means clustering



In the 8 attribute only 2 attributes are repeated in the clusters that are Filed_soil and Field _air from the attributes and the number of cluster is 4.

- Consider the setoff normalized the marks ranges from 0 to 1 scored by the 100 student in the class for particular subject.
- Each mark is treated as the vector with 1-D there is the need to classify the collected marks into 4 groups for allocating 4 grades.
- Classify the marks & identify the grades for the individual marks using the initialize 2 centroid.
- Find the mean of the marks collected in the particular grades say ‘1’.
- This is treated as the centroid corresponding to the grades 1 used for the next iteration.
- This process is repeated to compute the new sets of centroid corresponding to the individual grade



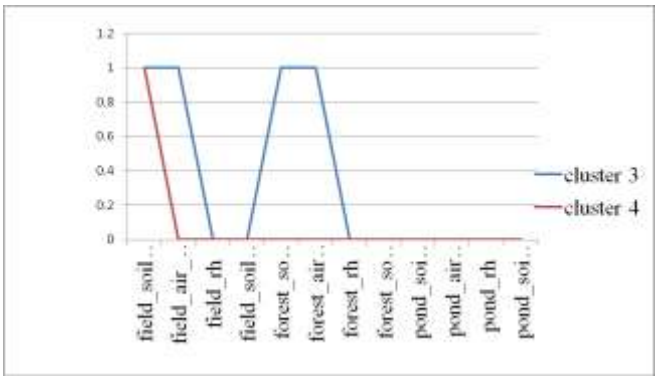


Fig 3 & 4: K-Means Column Chart

- The given number of clusters is 4.
- From the 12 attribute only 2 attributes are repeated in the clusters that is Filed_ soil_wc and Field _air
- Forest_rh, field_rh and field_soil_wc is in Cluster1.
- Forest_soil_temp_c is in Cluster 2.
- The attribute Field_rh , Field _soil_wc and forest_rh are in the cluster 3.
- In cluster 4, Field _air_temp_c attribute is present.

iii. Frequent Pattern Tree

The frequent pattern mining algorithm defined as follows. It is a text file, where each line represents a frequent itemset. On each line, the items of the itemset are first listed. Each item is represented by an integer and it is followed by a single space. After, all the items, the keyword "#SUP:" appears, which is followed by an integer indicating the support of the itemset, expressed as a number of transactions. The first line indicates the frequent itemset consisting of the item 1 and it indicates that this itemset has a support of 3 transactions.

- FP Tree is an efficient and scalable method for complete set of frequent patterns by pattern growth using extended tree structure in environmental dataset
- The environmental data is used to represent the conditional data types are transaction data type format, the transaction is T and Ti.
- As per FP tree we used 12 attributes, finally we obtained 1 attribute frequently, that is

➢ Field_soil_WC,

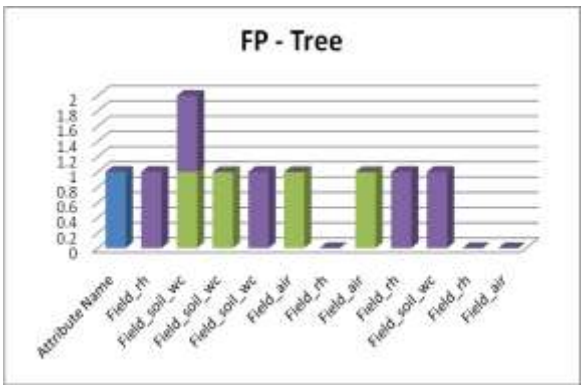


Fig 5: FP – Tree

Environmental dataset is the difference in the usage of the variable name in the transaction process. From transactions, only the number of rows (transactions) and cols (items) are printed. The result of fig 5(data) and item Frequency Plot (data, support = 0.1)

Principle Component Analysis

Principal Component Analysis (PCA) is a simple yet popular and useful linear transformation technique that is used in numerous applications, such as stock market predictions, the analysis of expression data, and many more. The sheer size of data in the modern age is not only a challenge for computer hardware but also a main bottleneck for the performance of many machine learning algorithms. The main goal of a PCA analysis is to identify patterns in data; PCA aims to detect the correlation between variables.

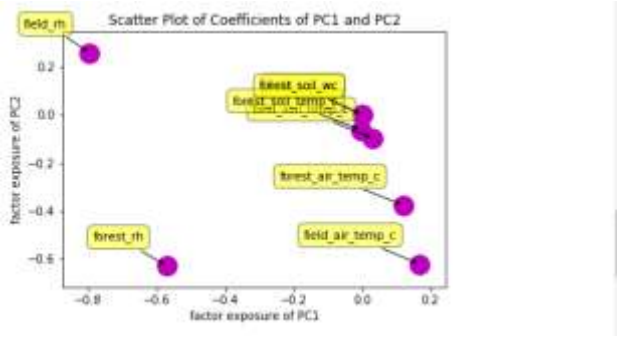


Fig 6: Principle component analysis algorithm

- The number of timestamps is 542.
- The number of stocks is 8.
- 98.59% of the variance is explained by the first 2 PCs

[542 rows x 2 columns] Factor 1 factor 2

Table 1: PCA

Attributes Names	Factor 1	Factor 2
field_air_temp_c	0.03209	0.097145
field_rh	0.168229	0.620055
field_soil_wc	0.794569	0.258494
forest_soil_temp_c	0.000267	0.000436
forest_air_temp_c	0.002895	0.062263
forest_rh	0.120914	0.375092
forest_soil_wc	0.569825	0.628253

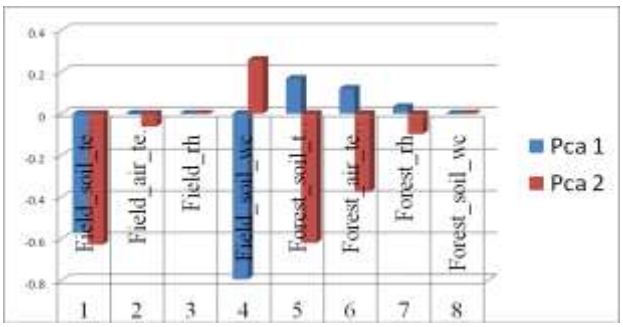


Fig 7: PCA 3 D Area Chart Pca1 and pca2

The Fig 7 shows that attributes can be explained by the first principal component alone. The second, third, fourth and fifth attributes share almost equal amount of information. Comparatively 4th and 5th attribute share

less amount of information as compared to the rest of the Principal components.

Table 2: Result of Apriori, K-Means Clustering, Frequent Pattern Tree And Pca Algorithms

Attribute	Field Soil	Field Air	Field Rh	Field Soil wc	Forest soil temp	Forest air temp	Forest rh	Forest soil wc	Pond soil	Pond air	Pond rh	Pond soil_wc
Apriori	Forest soil wc	Pond air temp c	Pond rh wc	Pond soil wc	Forest soil wc	Pond air temp c						
K-means	Field soil wc	Field air	Field soil Wc	Field air								
FP-Tree	Field rh	Field soil wc	Field soil wc	Field soil wc	Field rh	Field air	Field soil	Field air	Field soil wc	Field rh	Field air	Field rh
PCA	Field soil Wc	Field air	Field _rh	Field soil wc	Forest air temp	Field_ rh	Forest soil temp					

Table 3: Comparison Result Table for Four Methods

S. No	Methods Name	Repeated Attributes	Comparison Result for Four Methods
1	Apriori Algorithm	Forest soil wc Pond air temp c	Field soil wc
2	K – Means Algorithm	Field soil wc Field air	
3	FP – Tree Algorithm	Field soil wc Field air Field rh	
4	PCA - Algorithm	Field soil wc Forest soil temp c Field soil temp c	

VII. CONCLUSION AND FUTURE SCOPE

This paper discusses about the land the attribute determine the quality of environment more suitable for agriculture, this work explains the environmental dataset contains 12 attributes, using the above mention algorithms to choose the perfect features for agriculture to determine the quality of environment more suitable for agriculture.

The proposed work can also be extended to analyze the soil and other factors for the crop and to increase the crop production under the different climatic conditions

REFERENCES

[1] Amirah Mohamed Shahiri, Wahidah Husain,Nur aini Abdul Rashid, “A review on predicting student’s performance using data mining techniques”, Elsevier, Procedia Computer Science 72 (2015) 414 – 422, The third information systems international conference, DOI: 10.1016/j.procs.2015.12.157

[2] Cleiton Ferraro Dos Santos, Flavio Piechnicki,Eduardode FreitasRocha Loures, Eduardo Alves Portela Santos, “Mapping the conceptual relationship among data analysis knowledge generation and decision making in industrial process”, Elsevier, 27th International Conference on Flexible Automation and Intelligent Manufacturing FAIM 2017, 27-30 June 2017, DOI: 10.1016/j.promfg.2017.07.305

[3] Amjad Adu Saa, "Educational data mining & students performance prediction", (IJACSA) International Journal of Advanced computer Science and Applications, Vol. 7, No. 5, 2016.

[4] Ali Daud, Naif Radi Alijohani, Rabeeh Ayaz Abbasi, Miltiadis D.Lytras, Farhat Abbas, Jalal S.Alowibdi "Predicting student performance using advanced learning analytics" International World Wide Web Conference Committee (IW3C2), Perth, Australia – April 03 – 07, 2017.

[5] Ankit Sonia, Nees Jan Van ECK, Uzay Kaymak, "Prediction of stock price movements based on concept map information" IEEE symposium on Computational Intelligence in Multi-criteria Decision Making (MCDM 200

[6] R.Sumitha, E.S.Vinothkumar, "Prediction of students outcomes using data mining techniques" International Journals of Scientific Engineering and Applied science (IJSEAS) volume -2 issue - 6, June 2016.

[7] Anal ACHARYA, Devadatta SINHA, St.Xavier's, "An educational data mining approach to concept map construction for web based learning" Elsevier, Informatica Economica Vol. 21,no. 4/2017,
DOI: 10.12948/issn14531305/21.4.2017.04

8] Dr. Sankar Rajagopal, "Customer Data Clustering Using Data Mining Technique", International Journal of Database Management Systems (IJDMS) Vol.3, No.4, November 2011

[9] Janardan, Shikha Mehta, "Concept drift in streaming data classification algorithms plarforms and issues", Elsevier, Information Technology and Quantitative Management (ITQM2017), DOI: 10.1016/j.procs.2017.11.440.

[10] Tb. Ai Munandar, Harsiti, Roy Amrullah Ritonga, "Mapping concept of equitable regional development using C4.5 classification methods", International Journal of Advanced Research in computer science volume 3, No.2, March -April 2012

[11] Evangelia Gouli, Agoritsa Gogoulou, Kyparisia Papanikolaou & Maria Grigoriadou, "Evaluating learning's knowledge level on concept mapping tasks", IEEE International Conferenc On Advanced Learning Technologies (ICALT'05),
[12] Forough Farazzmanesh (Isvand), Moniresh Hosseini, "Analysis of business cluster's value nework using data techniques", 15/Jun/2017.

[13] Victoria Kayser, Knut Blind, "Extending the knowledge base of the foresight the contribution of text mining", Elsevier, Technological Forecasting & Social change116(2017)208-215, DOI.org/10.1016/j.techfore.2016.10.017

[14] Said Nawar, Abdul M. Mouazen Catena, "Predictive performance of mobile vis - near infrared spectroscopy for
key soil properties at different geographical scales by using spiking and data mining techniques", Elsevier, 151 (2017) 118-129.

[15]Prof. paresh tenna, Dr.yogesh ghodasara,, "Foundation of frequent pattern mining algorithm's implementation "international journal of computer trends and technology (IJCTT) 4, 7-july2013

[16]Jessica g.Burke, Particia O,Campo, Geri L.Pear An introduction to concept mapping as participatory Public Health Research Methods, University of North Dakoda on May 16, 2015

[17]Chaman Lal sabharwal and bushra anjum Dara reduction and regression using Principle Component analysis in Qualitative Spatial Reasoning and health informatics. Missouri University of science and technology, Rolla, MO-63128,USA.

AUTHOR BIOGRAPHY

V. Janani Bai – Working as Guest Lecturer in Kumararani Meena Muthiah College of Arts and Science, 4 Cresent Avenue Road, Gandhi Nagar, Adyar, Chennai, Tamil Nadu – 600020, I Completed M.SC(CS) Loganatha Narayanaswamy College of Arts and Science Ponneri, M.Phil(CS) From University Of Madras, Guindy Campus (opp. Birla Planetarium), NanoScience Building IIIrd Floor, Chennai 600025.