Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 70(2) 2016 167–172

Extracting Association Rules in Spatial Databases of Agriculture Domain for Land Use Planning

Pranita Singh and S.D. Samantaray
G.B. Pant University of Agriculture & Technology, Pantnagar

Received 27 May 2015; Revised 16 January 2016; Accepted 09 June 2016

SUMMARY

With wide application of remote sensing technologies and automatic data collection tools, tremendous amount of spatial and non spatial data of agriculture domain have been collected and stored in large spatial databases. The extraction and comprehension of the knowledge implied by such huge amount of spatial data pose great challenges to the researches involved in the area of Agri-Informatics and computer intelligence. Use of computational intelligence in agriculture is important challenging area for developing tools for mass impact of socio-economic development. The present work deals with the development of an integrated computer intelligent system to extract association rules from spatial databases of land use and slope which provide decision making reference for land use planning. Apriori algorithm is applied on the transaction database generated after preprocessing and analysis of spatial databases using ENVI and ArcGIS tools. The investigation was conducted on the land use and slope data of Udham Singh Nagar district obtained from Agrometeorology department, G.B.P.U.A. & T., Pantnagar. The implementation was successfully done with the development of GUI using MATLAB. The association rules obtained have been verified by agriculture domain experts.

Keywords: Data mining, Spatial data mining, Association rule, Land use.

1. INTRODUCTION

In recent years, it is observed that data mining has involved a lot of attention from the information industry and society because a huge amount of data is available in various areas and there exists an intense need of transforming this data to some fruitful information and knowledge (Ladner *et al*. 2003). The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases (Sumathi *et al*. 2008).

Spatial database is a database that is optimized to store and query data that is related to objects in space, including points, lines and polygons. There are additional functionalities added to process a spatial database than a normal database. These are typically called geometry or feature. Spatial database can perform a wide variety of spatial operations. The following query types and many more are supported by Open Geospatial Consortium:

Spatial Measurements: Finds the distance between points, polygon area, etc.

Spatial Functions: Modify existing features to create new ones, for example by providing buffer around them, intersecting features, etc.

Spatial Predicates: Allows true/false queries such as is there a residence located within a mile of the area we are planning to build the landfill.

Corresponding author: Pranita Singh
E-mail address: pranitasingh1@gmail.com

Constructor Functions: Creates new features with an SQL query specifying the vertices which can make up line.

Observer Functions: Queries which return specific information about a feature such as the location of the centre of a circle.

Spatial databases required are shape files which are a popular geospatial vector data form for geographic information system software. While a term ‘shape file’ is quite common, a ‘shape file’ is actually a set of several files. Three individual files are mandatory to store the core data that comprises a shape file: .shp, .shx, .dbf.

Data mining plays an important role in many applications and various special tools of data mining such as Clementine and DBMiner are used to perform the data mining task (Han and Kamber 2006). In this paper MATLAB is used for this purpose. It has been employed in many fields because of its powerful data-process function.

Land use is a product of interactions between a society’s cultural background, state and its physical needs on the one hand and the natural potential of land on the other (Balak and Kolarkar 1993). Various studies have been conducted in different areas (Kail Watershed of Central Himalaya (Rawal *et al*. 2014), Mau district of Uttar Pradesh (Singh and Singh 2011), Doon valley in Dehradun Tehsil of Uttarakhand (Tiwari and Khanduri 2011), Kiliyar sub-watershed of Tamilnadu (Poongothai *et al*. 2014) to see the dynamics of land use using GIS and remote sensing techniques. In addition to facilitate sustainable management of the land, land use information may be used for planning, monitoring, evaluation of development, industrial activity or reclamation. Issues driving land use studies include the removal or disturbance of productive land, urban encroachment and depletion of forests (Bhatta 2011). Increasing urban land value was a major driver of farmland development, while rising rural income was a primary driver of conversion of farmland to forests and grasslan (Li *et al.* 2013). Slope gradient was chosen as the index of topography to study their relationship with land use. In this study, we used association rule mining program coded by M-language in MATLAB, performed the association rule mining between land use and slope data and aimed to find out the association between the land use type and slope in the study area. The land use was classified into eight categories: barren land, water bodies, forest, built up land, bushes, agricultural land, river bed and fallow land.

Association rules mining is finding out the patterns whose appearance probabilities are much higher than any other transaction to instruct the decision-making (Liang 2006). In nature, it is a simplification of conditional probability and joint probability. An association rule is an implication formula such as

$$\mathrm{X} \rightarrow \mathrm{Y}$$

In this formula, X and Y are called itemset which is composed of items, and $X \cap Y = \Phi$.

In transaction database, if there are s% transactions that contain $X \cup Y$, then the support of association rule $X \rightarrow \mathrm{Y}$ is s%, and it is recorded as support($X \rightarrow Y$) = s%. Support of $X \rightarrow Y$ is the percentage of the transaction containing $X \cup Y$ in transaction database. In this way, *support* $(X \rightarrow Y) = P\,(X \cup Y)$. If the ratio of transactions containing $X \cup Y$ to transactions containing X is c%, then the confidence of association rule $X \rightarrow Y$ is c% . So, *confidence* $(X \rightarrow Y) = P(Y|X)$.

According to the above concept description and analysis, by taking the land use database, MATLAB program is used to calculate support and confidence of association rules and discover the association between land use type and slope.

It is expected that this study should provide decision-making reference for land use planning and land use structure optimizing in local area and other similar region.

2. METHODOLOGY OF THE PROPOSED WORK

In this research work, extraction of association rules in spatial databases of land use and slope is performed using association rule mining based on the concept of Apriori algorithm. The whole process can be divided into three main phases: pre-processing of data, transformation of data and finding association rules.

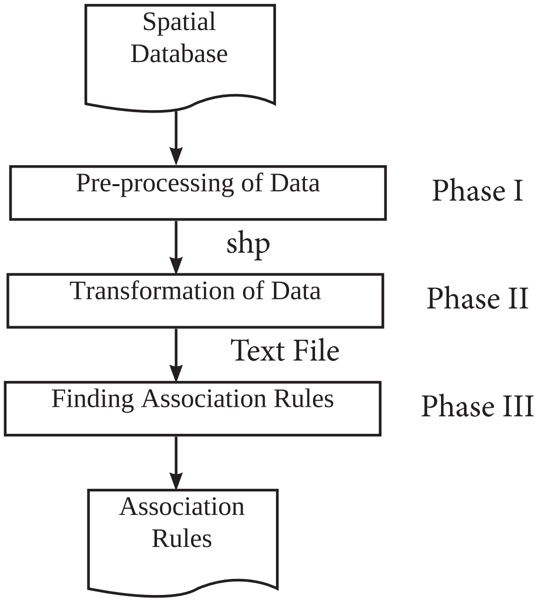


Fig. 1. Schematic showing major phases

2.1 Pre-Processing of Data

The land use data of Udham Singh Nagar obtained from Agrometeorology department is preprocessed using ENVI. Five land use images of December month of 2011 are used. The data is opened in ENVI and classification is done using supervised classification. In classification, entire image has been classified into separate classes (barren land, water bodies, forest, built up land, bushes, agricultural land, river bed and fallow land). In order to provide training ROI were created over the different objects such as forest, water body, built-up land, etc. on the basis of spectral signatures of various objects. After classification, the file is converted from raster to vector and saved as .evf. The .evf so obtained is exported as shape file using ENVI. The slope data of Udham Singh Nagar is in the form of DEM. The DEM file is opened in QGIS and slope is derived from the file.

2.2 Transformation of Data

After preprocessing of databases of land use and slope in ENVI and QGIS, the shape files of land use and slope databases were obtained. These two shape files were then opened in ArcGIS for analysis.

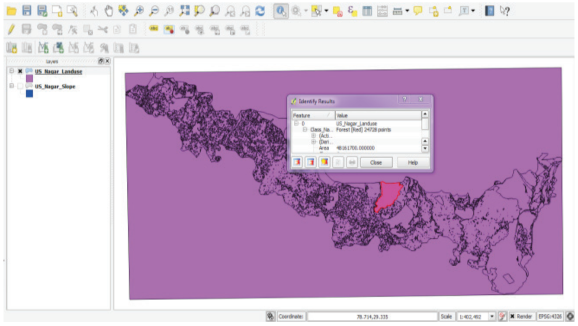


Fig. 2. Land Use Database

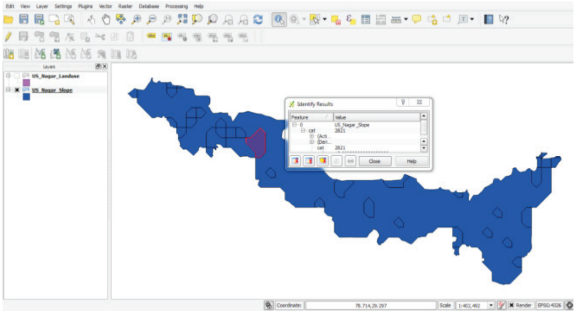
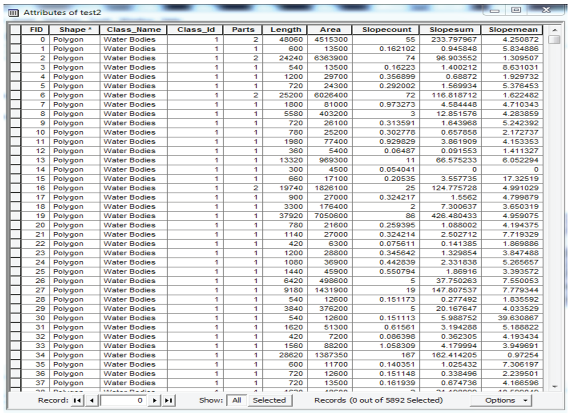


Fig. 3. Slope Database

After opening the shape files of land use and slope databases in ArcGIS, these two databases were merged using spatial join operation of ArcGIS. After merging the two databases, we got the attribute table containing attributes of both databases. The attribute table which was obtained after spatial join operation is shown below.



Attributes of test2

FID	Shape *	Class_Name	Class_Id	Parts	Length	Area	Slopecount	Slopesum	Slopemean
0	Polygon	Water Bodies	1	2	48060	4515300	55	233.797967	4.250872
1	Polygon	Water Bodies	1	1	600	13500	0.162102	0.945848	5.834886
2	Polygon	Water Bodies	1	2	24240	6363900	74	96.903552	1.309507
3	Polygon	Water Bodies	1	1	540	13500	0.16223	1.400212	8.631031
4	Polygon	Water Bodies	1	1	1200	29700	0.356899	0.68872	1.929732
5	Polygon	Water Bodies	1	1	720	24300	0.292002	1.569934	5.376453
6	Polygon	Water Bodies	1	2	25200	6026400	72	116.818712	1.622482
7	Polygon	Water Bodies	1	1	1800	81000	0.973273	4.584448	4.710343
8	Polygon	Water Bodies	1	1	5580	403200	3	12.851576	4.283859
9	Polygon	Water Bodies	1	1	720	26100	0.313591	1.643968	5.242392
10	Polygon	Water Bodies	1	1	780	25200	0.302778	0.657858	2.172737
11	Polygon	Water Bodies	1	1	1980	77400	0.929829	3.861909	4.153353
12	Polygon	Water Bodies	1	1	360	5400	0.06487	0.091553	1.411327
13	Polygon	Water Bodies	1	1	13320	969300	11	66.575233	6.052294
14	Polygon	Water Bodies	1	1	300	4500	0.054041	0	0
15	Polygon	Water Bodies	1	1	660	17100	0.20535	3.557735	17.32519
16	Polygon	Water Bodies	1	2	19740	1826100	25	124.775728	4.991029
17	Polygon	Water Bodies	1	1	900	27000	0.324217	1.5562	4.799879
18	Polygon	Water Bodies	1	1	3300	176400	2	7.300637	3.650319
19	Polygon	Water Bodies	1	1	37920	7050600	86	426.480433	4.959075
20	Polygon	Water Bodies	1	1	780	21600	0.259395	1.088002	4.194375
21	Polygon	Water Bodies	1	1	1140	27000	0.324214	2.502712	7.719329
22	Polygon	Water Bodies	1	1	420	6300	0.075611	0.141385	1.869886
23	Polygon	Water Bodies	1	1	1200	28800	0.345642	1.329854	3.847488
24	Polygon	Water Bodies	1	1	1080	36900	0.442839	2.331838	5.265657
25	Polygon	Water Bodies	1	1	1440	45900	0.550794	1.86916	3.393572
26	Polygon	Water Bodies	1	1	6420	498600	5	37.750263	7.550053
27	Polygon	Water Bodies	1	1	9180	1431900	19	147.807537	7.779344
28	Polygon	Water Bodies	1	1	540	12600	0.151173	0.277492	1.835592
29	Polygon	Water Bodies	1	1	3840	376200	5	20.167647	4.033529
30	Polygon	Water Bodies	1	1	540	12600	0.151113	5.988752	39.630867
31	Polygon	Water Bodies	1	1	1620	51300	0.61561	3.194288	5.188822
32	Polygon	Water Bodies	1	1	420	7200	0.086398	0.362305	4.193434
33	Polygon	Water Bodies	1	1	1560	88200	1.058309	4.179994	3.949691
34	Polygon	Water Bodies	1	1	28620	1387350	167	162.414205	0.97254
35	Polygon	Water Bodies	1	1	600	11700	0.140351	1.025432	7.306197
36	Polygon	Water Bodies	1	1	720	12600	0.151148	0.338496	2.239501
37	Polygon	Water Bodies	1	1	720	13500	0.161939	0.674736	4.166596

Records (0 out of 5892 Selected)

Fig. 4. Merged Attribute Table

After preprocessing, the shape files of land use and slope are opened in ArcGIS for analysis and spatial join is performed to have a merged .dbf file of both the databases. The .dbf file obtained after spatial join contains attributes of both the databases. For applying association rule mining, it is needed to convert the database to the transaction database having binary values. The .dbf file obtained after merging of two databases of land use and slope is first exported to MS Excel, which showed up in a tabular manner. A database application like MS Access could have been used to open the .dbf file, but the size of the database was much larger than access could process. The excel format is relatively small and now the relatively small excel data was exported to MS Access file.

The Access database is now converted to MySQL database. For this an ODBC link was created from MySQL by using MySQL driver for ODBC to link the MS Access database. Finally the database was exported to MySQL. This move was so taken owing to the better querying capabilities as well as good compatibility across multiple platforms.

For further processing our data we first created a primary attribute ID of integer type to uniquely identify each tuple. Next we created separate attributes for each type of land use e.g. forest, fallow land etc. We applied SQL queries as :

ALTER TABLE new ADD COLUMN BARREN_LAND BIT NULL DEFAULT 0

ALTER TABLE new ADD COLUMN FOREST BIT NULL DEFAULT 0

Separate attributes for slope were also created as :

ALTER TABLE new ADD COLUMN SLOPE_0 BIT NULL DEFAULT 0

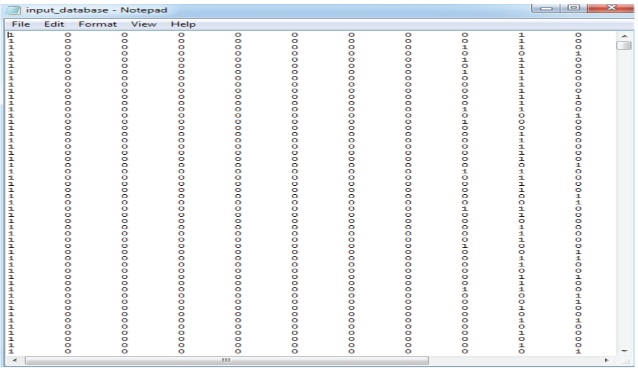
ALTER TABLE new ADD COLUMN SLOPE_1 BIT NULL DEFAULT 0

We set 0 as the default value of the attributes. A bit value of 1 was set to each particular tuple according to the type of land use and slope range. The application of queries is as under :

UPDATE new set BARREN_LAND=1 WHERE Class_Name =Barren Land

UPDATE new set SLOPE_0=1 WHERE Slopemean<=2

We now exported the MySQL database to comma separated values (.csv) format. Next we opened the .csv file in MS Access and exported the database to text file using space as delimiter.



```
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     1     1     0
1     0     0     0     0     0     0     0     0     0     1
1     0     0     0     0     0     0     0     1     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     1     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     1
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     1     1     0
1     0     0     0     0     0     0     0     0     0     1
1     0     0     0     0     0     0     0     1     0     0
1     0     0     0     0     0     0     0     0     0     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     0     1
1     0     0     0     0     0     0     0     1     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     0     1
1     0     0     0     0     0     0     0     0     0     1
1     0     0     0     0     0     0     0     1     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     0     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     1     0     0
1     0     0     0     0     0     0     0     0     0     1
1     0     0     0     0     0     0     0     0     1     1
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     0     0
1     0     0     0     0     0     0     0     0     1     1
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     1     1     0
1     0     0     0     0     0     0     0     0     0     1
1     0     0     0     0     0     0     0     0     0     1
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     0     0
1     0     0     0     0     0     0     0     0     1     1
1     0     0     0     0     0     0     0     0     0     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     0     0
1     0     0     0     0     0     0     0     0     1     0
1     0     0     0     0     0     0     0     0     0     1
```

Fig. 5. Text File

This text file is imported in MATLAB and then converted to MAT file.

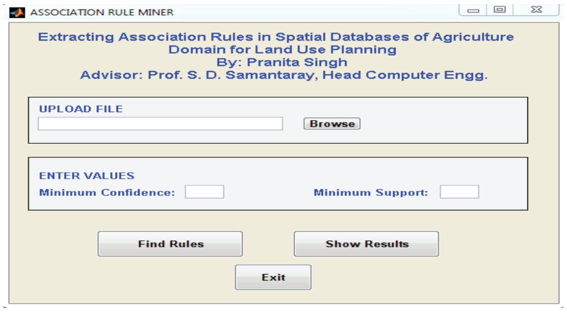
In this study, slope levels 0, 1, 2, 3, 4, 5, and 6 denote slope range that is <2°, 2°- 4°, 4°- 6°, 6°-8°, 8°- 10°, 10°- 12°and >=12°, respectively.

2.3 Finding Association Rules

After exporting the database to text file, the text file is imported in MATLAB using import option. After importing the text file in MATLAB, save the file as MAT file. The text file is saved as MAT file by running 'save filename.mat' command in the command window. The association rule mining algorithm that is used to find association rules between land use and slope is based on the concept of Apriori algorithm.

Apriori algorithm is the originality algorithm of Boolean association rules of mining frequent item sets (Agrawal and Srikant 1994). The purpose of the Apriori Algorithm is to find associations between different sets of data. Each set of data has a number of items and is called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data. Association rule is based mainly on discovering frequent item sets. The core principles of this theory are the subsets of frequent item sets are frequent item sets and the supersets of infrequent item sets are infrequent item sets. Apriori is designed to operate on databases containing transactions. The algorithm is coded in MATLAB (Zheng and

Zhao 2008) and it takes MAT file of transaction database as input. After applying association rule mining on the transaction database, association rules are obtained between land use and slope. The rules are finally saved in a text file.



Extracting Association Rules in Spatial Databases of Agriculture Domain for Land Use Planning
By: Pranita Singh
Advisor: Prof. S. D. Samantaray, Head Computer Engg.

Fig. 6. Developed GUI

The GUI shown in Fig. 6 was created using uicontrol function of MATLAB. The interface has three text fields for entering minimum support value, entering minimum confidence value, and browsing the file. It has four buttons: Find Rules: for finding association rules, Show Results: for displaying the rules along with their support and confidence, Exit: for closing the main window and Browse: for finding the file to be uploaded.

The algorithm is implemented in MATLAB to find the association rules. The results obtained proved that this study method is feasible and will provide a decision support system for land use planning.

3. RESULTS AND DISCUSSION

Table 1 shows the support and confidence of the association rules. A denotes the land use type and B denotes the slope level. A→B_0 denotes the association rule between land use and slope level 0.

Most of the land use types are concentrated at the slope level 0. There are eight kinds of land use types in study area. The supports of the various land use types at this slope level are approximately between 0% and 52%. The support of fallow land is found to be 51.92% which is the highest support value which means a large amount of area is left out for a period to retain its fertility. The land use dataset which we have used is of a particular season. Therefore, the fallow is showing a high support because after one crop land is left idle for a period of time. This land can be utilized in an efficient way to have a high production. The results for agricultural land are somewhat similar to the results of the case study of Mau District, Uttar Pradesh. In Mau district, agriculture is the major land use categories due to the one of fertile soil of the world.

Therefore, production in this area is mainly concentrated on the land where the slope is less than 2°. At this slope level, built up land has a very high confidence which means the population of the district is mainly concentrated on the slope less than 2°.

Table 1. Support and confidence of the association rules

A	A→B_0 s(%) c(%)	A→B_1 s(%) c(%)	A→B_2 s(%) c(%)	A→B_3 s(%) c(%)	A→B_4 s(%) c(%)	A→B_5 s(%) c(%)	A→B_6 s(%) c(%)
Water Bodies	2.76 77.25	1.98 55.45	1.25 35.07	0.69 19.43	0.39 10.90	0.18 5.21	0.44 12.32
Barren Land	4.12 78.38	3.00 57.09	1.68 31.93	1.00 19.03	0.52 10.00	0.32 6.12	0.52 10.00
Forest	3.02 70.07	2.85 66.14	1.62 37.79	1.10 25.59	0.81 18.89	0.44 10.23	0.76 17.71
Built up Land	8.23 83.47	5.97 60.58	3.27 33.21	1.96 19.96	1.08 11.01	0.57 5.85	0.84 8.60
Bushes	1.08 79.01	0.84 61.72	0.51 37.03	0.22 16.04	0.11 8.64	0.11 8.64	0.11 8.64
River Bed	0.27 76.19	0.22 61.90	0.08 23.80	0.05 14.28	0.02 4.76	0.02 4.76	0.03 9.52
Agricultural Land	8.85 81.69	6.89 63.53	3.44 31.76	2.32 21.43	1.47 13.61	0.76 7.04	1.03 9.54
Fallow Land	51.92 80.60	38.34 59.52	21.65 33.62	13.10 20.34	7.79 12.09	4.61 7.16	6.29 9.77

The river bed has the lowest value of support at slope level 0 which means river bed covers least area among all land use types. The confidence of most of the land use types is showing a sharp decrease at slope level 1 but confidence of forest is showing a gradual decrease. The confidence of forest land at slope level 6 is highest among all other land use types which mean forest can be grown more successfully at higher slopes. It is observed that confidence of barren land is also high at slope level 6.

At higher slopes, there are more chances of erosion which is one of the major causes of the barren land.

4. CONCLUSION

The goal of this study was to extract association rules in spatial databases of agriculture domain. The databases of land use and slope obtained from Agrometeorology department of Agriculture College, G.B.P.U.A. & T., Pantnagar were preprocessed and transformed using ENVI, ArcGIS and MySQL software. GUI incorporating Apriori algorithm developed in MATLAB was able to successfully generate association rules out of land use and slope databases. It was found that all land use types mainly concentrate on the land having slope less than 2°. A large amount of land is fallow which can be utilized in an efficient way to have a high production. Therefore, production in this area is mainly concentrated on the land where the slope is less than 2°. The forest can be grown effectively on the higher slopes. The results obtained are expected to provide assistance in land use planning. In future, similar work can be performed for providing decision making reference regarding land use planning in different catchment areas. Other association rule mining algorithms can be applied to the spatial databases of agriculture domain and changes in results can be observed.

REFERENCES

Ladner, R., Petry, F.E. and Cobb, M.A. (2003). Fuzzy set approaches to spatial data mining of association rules. *Trans. GIS.* **7(1)**, 123-138.

Sumathi, N., Geetha, R. and Bama, S.S. (2008). Spatial data mining-techniques trends and its applications. *J. Comput. Appl.* **1(4)**, 28-30.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. 2nd ed. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers.

Balak, R. and Kolarkar, A.S. (1993). Remote sensing application in monitoring land use changes in arid Rajasthan. *Inter. J. Remote Sens.* **14(17)**, 3191-3200.

Rawal, H.S., Rawat, J.S., Kumar, M., Pant, N.C. and Rani, N. (2014). Land use/cover dynamics of Kail Watershed, Central Himalaya, India using Remote Sensing and GIS techniques. *Inter. Journal Adv. Remote Sens. GIS Geog.* **2(2)**, 55-59.

Singh, P. and Singh, S. (2011). Landuse pattern analysis using remote sensing: a case study of Mau District, India. Scholars research library. Archives of Applied Science Research, **(5)**, 10-16.

Tiwari, K. and Khanduri, K. (2011). Land use / land cover change detection in Doon valley (Dehradun Tehsil), Uttarakhand: using GIS & remote sensing technique. *Inter. J. Geomatics Geosci.,* **2(1)**, 34-41.

Poongothai, S., Sridhar, N. and Shourie, R.A. (2014). Change detection of land use/ land cover of a watershed using remote sensing and GIS. *Inter. J. Engg. Adv., Tech*., **3(6)**, 226-230.

Bhatta, B. (2011). *Remote Sensing and GIS*. Second Edition. Oxford University Press. Chapter 12, 396-400.

Li, M., Wu, J. and Deng, X. (2013). Identifying drivers of land use change in China: A spatial multinomial logit model analysis. *Land Eco.,* **89(4)**, 632-854.

Liang, X. (2006). A*lgorithm and Application of Data Mining*. Peking University Press.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In: Proceedings of 1994 International Conference VLDB, Santiago, Chile. pp. 487-499.

Zheng, X. and Zhao, L. (2008). Association rule analysis of spatial data mining based on Matlab: A case of Ancheng Township in China. In: *First International Workshop on Knowledge Discovery and Data Mining*, IEEE. pp. 76-80.

Badal, N. and Bagga, S. (2014). Implementaion of apriori algorithm in MATLAB using attribute affinity matrix. *Inter. J. Adv. Res. Comp. Sci. Software Engg*., **4(1)**, 10-16.

Fayyad, U.M., Piatetsky-Shapiro G. and Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, AAAI Press, Menlo Park, California. pp. 82-88.