JADAVPUR UNIVERSITY

Data Mining Application on Biodiversity

by

ANINDITA ROY

CLASS ROLL NUMBER: 001711002005

REGISTRATION NUMBER: 140966 OF 2017-18

EXAMINATION ROLL NUMBER: M4SWE19009

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF "Master Of Engineering" IN SOFTWARE ENGINEERING Under Supervision of

DR. KARTICK CHANDRA MONDAL

Assistant Professor

in the

Faculty of Engineering & Technology

Department of Information Technology

May, 2019

DEPARTMENT OF INFORMATION TECHNOLOGY

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

Certificate of Submission

I hereby recommend the thesis entitled, "Data Mining Application On Biodiversity ", prepared by Miss. Anindita Roy (CLASS ROLL NUMBER: 001711002005; REGISTRATION NUMBER: 140966 OF 2017-18; EXAMINATION ROLL NUMBER: M4SWE19009), under my supervision, be accepted in partial fulfillment of the requirements for the degree of Master of Engineering in Software Engineering from the Department of Information Technology under Jadavpur University.

_____

Dr. KARTICK CHANDRA MONDAL

Supervisor

Assistant Professor

Department of Information Technology

Jadavpur University

Countersigned by:

_____

Head of the Department

Information Technology

Jadavpur University

_____

Dean

Faculty of Engineering Technology

Jadavpur University

DEPARTMENT OF INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

(Only in case of thesis is approved)

The thesis at instance is hereby approved as creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve this thesis for the purpose for which it is submitted.

_____

Signature of the External Examiner

Signature of the Supervisor

Dr. KARTICK CHANDRA MONDAL

Assistant Professor

Department of Information Technology

Jadavpur University

DEPARTMENT OF INFORMATION TECHNOLOGY

FACULTY OF ENGINEERING AND TECHNOLOGY

JADAVPUR UNIVERSITY

Declaration of Originality

and Compliance of Academic Ethics

I, hereby declare that this thesis work completed by me as a part of the Master of Engineering in Software Engineering course, during the year 2018-2019, under the supervision of Dr. Kartick Chandra Mondal, Assistant Professor, Department of Information Technology, Jadavpur University.

All information, materials, and methods that are not original to this work have been properly referenced and cited. I am the only responsible person if found guilty of plagiarism. I am aware/warned of the plagiarism issues by our supervisor several time during the progress of our thesis.

The idea of the work has been given by my supervisor and guided by him only. The work has been done solely by me and I did not include anybody's work in it. The supervisor has full authority to use, modify, correct and distribute the work presented here.

I also declare that no part of this project has been submitted for the award for any other degree prior to this project by me. Also, I declare that I will not distribute or use the full or part of this project work for the award of any other degree by us in the future.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct

_____

(Signature with Date)

Acknowledgements

I take this opportunity to express my deepest sense of gratitude towards my respected guide Dr. Kartick Chandra Mondal, Assistant Professor, Department of Information Technology, Jadavpur University for giving me a great opportunity to work under his excellent guidance and motivating me throughout. I would like to thank him for giving me ample liberty to conduct research on my field of interest. He has been an inspiration to me for his perfection towards work.

I would also like to thank PhD Scholar in Information Technology Department at Jadavpur University, Mrs. Moumita Ghosh for supporting me technically throughout the thesis work as well as the entire Information Technology Department for giving me liberty to work and complete the thesis on time.

I would like to express my gratitude and love towards my respected parents, my elder brother and my peers for being the main source of inspiration and motivation throughout my career and the making of this thesis.

Anindita Roy

JADAVPUR UNIVERSIY

Abstract

Faculty of Engineering & Technology

Department of Information Technology

Master of Engineering

by Anindita Roy

Population trend of Indian Himalayan medicinal plant species are continuously decreasing due to over harvesting, over exploitation, loss of habitat, unsustainable collection and other threats.Many of medicinal plants species are already listed in IUCN threatened category, conservation of those plants species are required immediately as per their conservation status, habitat type, habitats, distribution area. Continuously increasing of listed threatened plant species increase the biodiversity data which is quite difficult to analysis for ecologist. Data mining application provides a suitable way to mine data and generate knowledge. Association rule mining, a data mining technique which mines the items set which are frequently occurred has been applied to find the correlation in the species conservation data. Application of data mining techniques on the plant's biodiversity data identified new association rules which can help ecologist in case of conservation of the plant biodiversity.

Table of Contents

List of Tables

List of Figures and Illustrations

Chapter 1

Introduction

1.1 Biodiversity

Biodiversity is existing of variation of life on the earth, including different types of species of plants, fungi, animal and other genetics of different ecosystems.It can be measured as how many species are presented in a area.If a less number of species found in a particular biosphere, then can say there are low biodiversity.

1.1.1 Problematic criteria

People are today very much dependent on medicines where as a source of medicines, contribution of medicinal plants plays a major role. As a source of economical property different parts of those medicinal plants are collected unsustainably and excessively, besides of this deforestation for creating buildings, Industries, making furniture, continuous grazing, loss of habitat become the cause of decreasing medicinal plants species and making them extinct, which affects ecosystem, ecological level and result of low biodiversity. Many of medicinal plants species are already listed in IUCN threatened category, conservation of those plants species are required immediately as per their conservation status, habitat type, habitats, area where they are found in. For ecologist it is even more difficult to analyze a large set of data and make relations among the large data items without using any tools. Data mining provides a better way to extract use full information from a large data set by providing different types of techniques like prediction, classification, association, clustering.

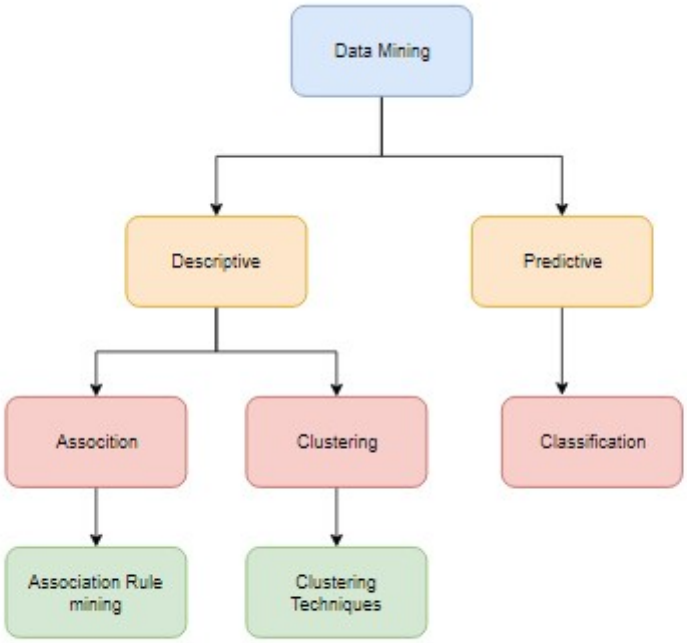


Figure 1.1: Category of Data mining techniques

1.2 What is Data mining?

Data mining is a way of discovering use full information or knowledge from a large homogeneous or heterogeneous data set through generating rules or patterns. In figure 1.1 an overview of different category of data mining techniques is shown. There are different kind of applications of data mining applied in sales and marketing, different business Industries, Bio information, Biodiversity, Health care service, Educational and other organizations.

1.2.1 Main objective

Main goal of this thesis is applying a data mining technique, association rule mining on the species data base to find out relationship between different conservation factors of species by generating rules and analyzing them. Main per-

spective is to show how data mining techniques help to discover knowledge on biodiversity data in case of conserving and cultivating threatened Himalayan medicinal plant in India.

1.2.2 Step-wise procedure

In the first stage, Himalayan medicinal plant species data had been collected from different sources and a new species data base of 40 threatened species objects and 15 attributes had been created. In the 2nd stage, data preprocessing technique had been applied to make the species data base feasible for applying the apriori algorithm on it. In the 3rd stage, apriori method had been applied on it to get the frequent items set and generates the rules. In the last stage, analysis of the rules and generation of some useful information about the cultivation of Indian Himalayan medicinal threatened plant species had been accomplished.

1.3 Organization Of Thesis

This thesis provides an application of data mining on biodiversity data. The chapter 2 discusses some related work on biodiversity and some related work on the application of data mining technique. Some researchers had worked on plant species data using association rule mining which has also been discussed here. The chapter 3 discusses about the tools and techniques which have been used, where apriori algorithm, association rule mining, different applied tools in r has been discussed. Chapter 4 discusses the data base preparation part. From data collection to applying data preprocessing technique and prepare the final species data has been discussed there. In chapter 5, the experimental part has been discussed. It shows the experiments on the species data base by applying association rule mining using apriori algorithm and the summary of results

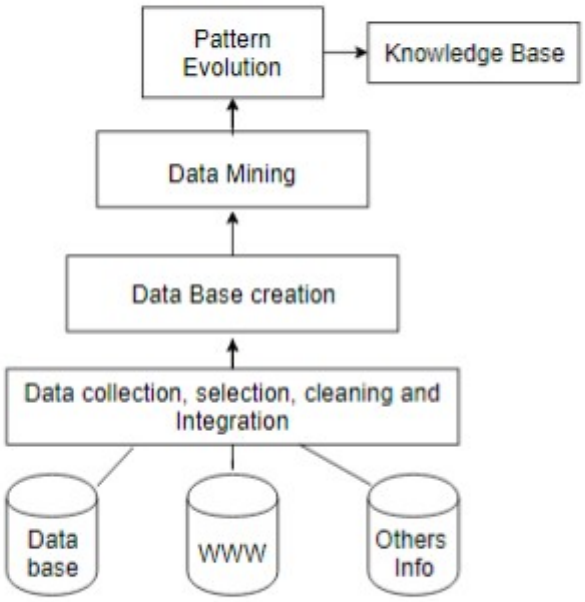


Figure 1.2: Data mining architechture

and analysis of species data base. Visualization of the generated and selected rules using different plots and graph has also been shown there. At the end of chapter 5, final analysis of the data and generation of new correlations in the species data elements for the cultivation and conservation of biodiversity has been discussed. Chapter 6, describes some future prospects and conclude the thesis with the hope that application of data mining technique on bio diversity data will solve the issues of plant diversity.

Chapter 2

Related works

2.1 Approach

A study of medicinal aromatic plants in Indian Himalayan region had found 152 medicinal aromatic plant species in western Himalaya which were categorized in to high medicinal value, demanding plant. From which 43 medicinal plants were identified highly threatened, and an urgent need of taking fast cultivation and fast conservation of those species are required in Western Himalayan region [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi].

Easy availability, low cost of herbal medicines cause of illegal harvesting and excessive extraction of Himalayan medicinal plant make them native in Himalayan region which making biodiversity plants in heavy pressure. To conserve medicinal plant various programs, policy were arranged but did not get any valuable result. A suggestion of making a network with local villagers to exchange information had been thought a way of conservation.For conserving the threatened valuable plant in paper [Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill] suggested to make an organized network in every village where the valuable medicinal plants are illegally harvested and to involve with the local villagers directly for giving them knowledge about various technique of plant conservation had been thought as a effective way of conservation.

A study of threat assessment of high value medicinal plant species on cold habitat region, Johar valley [Pandey et al.(2019)Pandey, Chandra Sekar, Joshi, and Rawal] , Western Himalaya, India identified the threat and the ecological status of high value medicinal plant species. These kinds of species were mostly threatened

due to extraction of medicinal part and the climate change. Ecological distribution of total threatened medicinal plant of this region was listed in this study. By using rapid threat assessment process total 22 high value medicinal plant species were identified in threatened category and from them Betula utilis and Nardostachys jatamanasi were highly threatened and fast conservation of those species was immediately required.

Another work on biodiversity status and distribution about the conservation of the ethno medicinal plants of Indian Himalayan region had been shown in [KUMARI et al.(2012)KUMARI, Chandra JOSHI, and Mohan TEWARI] Gradual decreasing of ethno medicinal plant species creates an urgent concerned of conservation of medicinal plant species in Indian Himalayan region. Lots of medicinal plant species were listed in IUCN threatened category. Species like Habenaria intermedeia, Taxus baccata, Bergenia ciliata, Polygonatum verticillatum, Polygonatum cirrhifolium were going to extinct. Immediate actions were needed in case of conservation of plant species in Indian Himalayan region. Providing information on the local people, society and villagers that how over exploitation of medicinal plant affects ecology, causes destruction of ecology, disturbance of the plant diversity and providing information about the cultivation of plants among the villagers had been tried to conserve the threatened valuable medicinal plant of Indian Himalayan Region.

An Urgent need of conservation of some Medicinal aromatic plant species in Indian Himalayan region had been shown in [Sharma and Kala(2018)]. Population of plant species were continuously declined due to the over harvesting for their high value trend in market place. By analyzing the trend of harvesting of Medicinal aromatic plant, sustainability of those plants had been taken. Depend on different season species had been collected from different altitude range of Himalayan region. The medicinal aromatic plants which are in high trend value are

collected in a unsustainable way which makes them in extinct condition, though by restricting the collectors at that time of harvesting of those extinct species on the specific region of Himalaya had been thought as a way of conservation but that was not suitable because of change of the collectors harvesting cycles and periods. Suitable actions were required to take in case of long term conservation.

2.2 Application

An application of Association rule mining shows how it was effectively work in total productive management for making use full knowledge fast in manufacturing industries [Djatna and Alitu(2015)]. The relationship between an indicator of useful measurement(Overall equipment effectiveness) and the corresponding action required for machine utilization had been identified by generating rules. An application of data mining techniques using apriori method to generate association rule on a database of ichthyoplankton samples from a freshwater reservoir in Legal Amazon in [Silva et at.(2013)] had been described. In this paper, apriori method had been applied on the data base to generate association rules and they had found a relationship between biotic and abiotic factors of their Data base.

Another work on plant species data in paper [Silva et at.(2016)] had been shown how data mining technique work on ecology. Data mining technique had been applied to analysis the spatial distribution of species occurrences where association rule mining technique had been applied to identify the pattern of species co-occurrences. To help ecologist and researcher in case of generating species co-occurrences a tool had been implemented and used to find the pattern of species co-occurrences based on the tropical forest data. Negative and positive

correlation for a large number of species had been identified using this tool.
To analyze the students trend data mining techniques was applied on student data [Rahman and Das(2017)]. Association rule mining and Apriori algorithm was applied to find the hidden pattern from the students data and discover knowledge which was related to students enrollment, subject choosing and their educational performance. Analysis was done based on the educational history of Students with respect to their current educational degree performance that student trend was decreasing or increasing and discovered a new knowledge successfully.

Chapter 3

Algorithm and tools used

3.1 Apriori algorithm

Apriori algorithm is an important aspect in data mining which mines the item set which are frequently appeared in the data base. Apriori algorithm sets a minimum support value and then starts searching from for all the individual element set, that is 1-item set which have at least the given minimum support value and repeats this searching procedure by extending the item set number by 1 in each case, like 2-item set, 3-item set and so on until reaches the maximum size of item set which are frequently occurred on the condition of the minimum support value.

3.1.1 Mining Association rules using apriori

Association rule mining using apriori algorithm works as follows:

- Apriori algorithm is used to generate the frequent item set of maximum size, says I.
- Divides the frequent item set I as a rule(R) such that, $L.H.S \rightarrow R.H.S$.
- For each rule calculate the confidence value as: $confidence(R) = \frac{Support(I)}{Support(L.H.S)}$,
- Rules which do not have the confidence value greater than or equal to the minimum confidence value are discarded.

3.2 Tools and Techniques

3.2.1 Tools in R

Tools which have been used are described below

spocc

Species occurrences data which are recorded in to different biodiversity portals, spocc can be used to collect those data from different biodiversity portals. Used functions: Two main functions of spocc package have been used to collect the latitude and longitude data of each plant species from GBIF portal. Since we want to collect the longitude and latitude values of each specimen, those 2 methods are used.

occ: We got the specimen occurrences for country India from gbif portal using this function. Parameters we passed, each specimen name, country and biodiversity portal's "gbif" as a value of 'from' parameter.

occ2df: We got a combining results for each specimen occurrences data from India in a single data frame from where the longitude and latitude value for each species are collected.

Red

It is an important package which is used to collect data about the IUCN red listed threatened species. It is a statistical software package which is mostly used in ecological science which provides fully data driven open source access .Here to calculate the extent of occurrences of red listed species and the area of occupancy red package has been used. Methods of red package used are described below.

eoo: This method finds the surrounding area of specimen occurrences in km2.

Parameter: Matrix of 2 column of longitude and latitude value is used as a parameter.

aoo: Area of occupancy, which gives grid cell value in $2 \times 2km^2$.

paramete: Matrix of latitude and longitude of each specimen is used as a parameter to find the area of occupancy of each of that specimen .

The matrix of 2 columns of maximum 5 different longitude and latitude values collected by using spocc package for each species can be create for each species using matrix() method as :
$matrix(c('l1','l2','l3','l4','l5','lat','lat2','lat3','lat4','lat5'),$
$nrow = 5, ncol = 2, dimnames = list(c('row1','row2','row3','row4','row5'),$
$c('longitude','latitude')))$, where parameters l1,l2,..,l5 defines longitude values, lat1,lat2,..,lat5 defines latitude values, nrow defines the rows number and ncol defines the columns number 17 and dimnames defines the dimensions name.

Arules

A framework provided by r to represent, manipulate and analyze the transnational data and mines the items set which are frequently appeared. In r, to perform association rules mining arules and arulesViz packages are used.

3.2.2 Association Rule mining

Association rule mining is used to find associations or correlations or relationship between different data items set in the data base. We can mine items sets which are frequently occurred in our data base by using association rule mining. In transnational data base, to get the frequent occurring item sets association rule mining is used often. By applying association rule mining we generate rules from which we can give some predictions.

For a rule say R, is written as $A \rightarrow B$, support, confidence, count and lift can be defined as below.

Support

Support for the above rule R is the percentage of all the transaction where A and B occur together. Support can be defined as the percentage of occurrences of the rule out of total number of occurrences. It is the ratio of the number of transactions which have the items set of the rule R to the total number of transactions.

$$Support(R) = \frac{(Numberoftransactionshaving AandB)}{(Totalnumberoftransactions)}$$

Confidence

Confidence can be defined as the probability or chances of occurrences of B when A occurs. It is calculated as the ratio of occurrences of A and B together to the number of occurrences of A, or, the ratio of the number of transactions which have both A and B together to the total number of transactions which contain A only.

$$Confidence(R) = \frac{(Numberoftransactionscontaining AandBtogether)}{(Numberoftransactionscontaining A)}$$

Lift

Lift measures how likely the right hand side of the rule will happen if the left hand side of the rule happens. Lift value represents the relationship of dependency between A and B. If lift value shows 1, then there are no dependency relationship between A and B, both are independent. If the lift value is less than 1 shows a negative dependency that is mean B is unlikely to be happened if A is happened. Lift value greater than 1 means there are positive dependency between A and B, both are dependent to each other. For rule R if lift value is greater than 1, it means B likely to be happened if A is happened. Lift value can be measured as follow:

$Lift(R) = \frac{Support(A,B)}{Support(A) \times Support(B)}$

Count

Count for a particular rule can be define as the no. of transactions out of total transactions which supports the rule.

Discretize

Arules provides a function discretize to convert the numerical values in to factorize or categorical format. For applying association rule mining all the numerical columns of the species data set were required to convert into categorical data. Discretize function implements some unsupervised methods to factorize the numerical data. Equal width interval method had been used on the numerical data for column 5, 6, 8, 9, 10 and fixed method where range is user specified, had been used for column 11 to convert them into categorical data.

Equal width interval is an unsupervised discretization method. In this method, data is divided into the specified interval with equal width size. For a particular column, this method selects the maximum and minimum value from the list of numerical values and takes the no. of intervals on which user wants to make the category and calculates the width size of interval by dividing the difference of maximum and minimum value into the no. of intervals. If data is divided into k interval then, Size of Width of intervals W(says), can be calculated as

$W = \frac{(Max-min)}{k}$

And the boundaries of interval will be $min + w$, $min + 2w$, ..,$(min + k - 1)w$.

arulesViz

ArulesViz is used to get different visualization of the resultant rules.

Chapter 4

Species Data Base

4.1 Data Collection

I have collected threatened Indian Himalayan medicinal plant species data from different database and by merging and integrated them prepared a new species data base of Indian Himalayan medicinal plant species. The new created species Data table has 40 objects of species with 15 attributes of different records value.

Attributes description and collection

The 2nd attribute, Geographical Distribution or distribution describes the part of Himalaya where the species distributed. NE=North Eastern Himalaya, E=Eastern Himalaya, W=Western Himalaya, NW=North Western Himalaya, TH= Through Out Himalaya. The 3rd attribute, IUCN/threat status, which carries each species threat status according to listed in to IUCN, values are EN, CR, VU, LC, NT where EN= Endangered, CR= Critically Endangered, VU=Vulnerable, NT= Nearly Threatened, LC= Least Concerned.(IUCN 2019) Parts used/uses is the 4th attribute in my data table, the part of the plant species which are used for medicinal purpose are listed here. Those are Root, Rhizomes, Bark, Stem, Leave, wood, Seed, Fruit, Bulb.

Lower elevation limit is the lowest altitude value above the sea level of each specimen(IUCN 2019). Upper elevation limit is the highest altitude value above the sea level,(IUCN 2019). Habitat type specifies the natural environment or type of place of each plant species where it can naturally live and grow, which are F=Forest, Gl=Grassland, Sl=Shrub land/ Shrubberies, R=Rocks or Rocky areas.

Generation length, which specifies the average age of each species when they can able to produce their next Generation. Values of this attribute are collected from IUCN 2019. Extent Of occurrence is defined as the area surrounded all the known sites of occurrences of a specimen. It is measured as the area within a continuous imaginary boundary which surrounded all the occurrences place of species. It may not include the area of its unsuitable habitat. Area of occupancy of each species is defined as the area occupied by the species within its extent of occurrences. Actually Species does not contain the entire area of its extent of occurrence, they only occupy some specific place in its extent of occurrence which is specified as area of occupancy. Conceptually it can be calculated as a grid cells of 2 km where cell size is $4km^2$, area of unsuitable habitat and discontinues are excluded. Habitats or zone or climate can be defined as the area which carries similar type of environmental characteristics or conditions. Propagation: It is the process to produce new plants by using the parts, like root, rhizome, tuber cutting or seed germination. Major threats or main threats or threats of species is defined as the cause of which species are affected and are gradually decreasing.(vikaspedia) The threats may be Loss of habitat or habitat loss or habitat degradation =lh, Over or Unregulated harvesting=oh, Unsustainable or Unregulated collection=usc, Grazing=g, Over exploitation= oe, Habit describes the plants characteristics form in which their life grows. It may include herb, tree, orchid, shrub.

4.1.1 Species Data table

Table 4.1: Species Data Table from column 1 to column 3

Species	**Geographical distribution**	**Conservation status**
Aconitum heterophyllum	NW	EN

Continued on next page

Table 4.1 – continued from previous page

Species	Geographical distribution	Conservation status
Aconitum chasmanthum	W	CR
Aconitum violaceum	W	VU
Angelica glauca	W	EN
Acer caesium	NW	LC
Aquilaria malaccensis	E	CR
Arnebia benthami	NW	CR
Arnebia euchroma	NW	CR
Bergenia ciliata	NE	VU
Bergenia ligulata	W	VU
Bergenia stracheyi	W	VU
Betula utilis	NW	LC
Cayratia pedata	NE	VU
Coptis teeta	E	EN
Dactylorhiza hatagirea	NW	CR
Dioscorea deltoidea	NW	EN
Ephedra gerardiana	NW	EN
Fritillaria cirrhosa	NE	EN
Gentiana kurroo	NW	CR
Gymnocladus assamicus	NE	CR
Habenaria intermedia	NW	EN
Hedychium spicatum	W	VU
Heracleum candicans	NW	EN
Illicium griffithii	NE	EN
Lilium polyphyllum	NW	CR
Malaxis Muscifera	W	VU
Nardostachys jatamansi	TH	CR
Paris polyphylla	NW	EN
Picrorhiza kurrooa	TH	EN

Continued on next page

Table 4.1 – continued from previous page

Species	Geographical distribution	Conservation status
Piper pedicellatum	E	VU
Podophyllum hexandrum	W	EN
Polygonatum cirrhifolium	W	EN
Polygonatum verticillatum	NW	VU
Rheum australe	TH	EN
Saussurea costus	NW	CR
Selinum tenuifolium	NW	NT
Taxus wallichiana	TH	EN
Thalictrum foliolosum	W	VU
Tinospora cordifolia	NW	EN
Valeriana jatamansi	NW	VU

Table 4.2: Species Data Table from column 4 to column 6

PartUsed	LowerElevation Limit(m)	UpperElevation Limit(m)
Root	2400	4500
Root	2100	3600
Root	3600	4800
Root	1800	3700
Bark	2400	3800
Wood	0	1000
Root	2000	3000
Root	3000	4200
Rhizome	1200	3000
Root	1500	3200
Root	3200	4200
Bark, Root	2500	4500

Continued on next page

Table 4.2 – continued from previous page

Part Used	Lower Elevation Limit(m)	Upper Elevation Limit(m)
Leave	1200	1900
Root	1700	2800
bulb	2500	3500
Rhizome	1000	3200
Stem	3700	5300
Bulb	3200	4600
Rhizome	1500	3000
Seedpods	1200	2052
Bulb	1500	2800
Rhizome	1000	2800
Root	2000	3800
Seed, Fruit	1200	1800
Bulb	2100	3000
Bulb	2500	4000
Root	2200	4800
Rhizome	1400	4300
Root	2700	4500
Leaves, Fruit	700	1800
Root	2000	4500
Rhizome	1700	4600
Rhizome	1500	3300
Root	2800	4000
Root	3200	3800
Root	2000	4000
Leaves, Bark	900	3700
Root	1300	3400
Stem, Root	300	600
Root	1800	3000

Table 4.3: Species Data Table from column 7 to column 9

Habitat type	**Generation length(yr)**	**Species Occurrences**
F, Gl	1	54
F, Gl	1	7
Gl, R	1	71
F, Gl, R	1	14
F	-	21
F	50	0
Sl	-	0
R	-	172
F, R	29	19
F, R	-	7
R	-	52
F	-	95
F, Sl	14	30
F, Sl	1	1
Sl	-	83
F	-	132
R	-	171
F	-	46
R, Gl	1	14
F	28	2
F, Gl	-	29
F, Sl	-	51
R	-	25
F	28	0
F	8	20
F, Gl, Sl	1	0
Gl, R	1	71
F	-	26

Continued on next page

Table 4.3 – continued from previous page

Habitat type	Generation length(yr)	Species Occurrences
F	-	41
F	17	20
F	-	82
F, Sl	-	18
F	-	34
R	-	13
Gl	1	29
Sl	-	0
F	30	41
F, Sl	-	51
F	1	20
F, Gl	-	53

Table 4.4: Species Data Table from column 10 to column 12

Extent Of Occurrance(eoo)	Area Of Occupancy(aoo)	zones or habitats
8311	16	Alpine
1190	16	Alpine, subalpine
21818	20	Alpine, subalpine
20087	20	Alpine
9050	16	Temperate
-	-	Subtropical, Tropical
-	-	Alpine
5429	20	Alpine
24574	20	Temperate
0	0	-
12543	20	Alpine

Continued on next page

Table 4.4 – continued from previous page

Extent Of Occurrance(eoo)	Area Of Occupancy(aoo)	zones or habitats
37872	20	Temperate
0	0	Subtropical, Tropical
0	0	Temperate
34291	20	Alpine
0	8	Temperate
10718	20	Alpine
0	4	Alpine
1913	16	Alpine
0	0	Subtropical
3133	12	Temperate
0	4	Temperate
24413	16	Alpine
-	-	Temperate
20494	20	Subtropical
-	-	Temperate
0	8	Alpine
0	8	Subtropical
0	8	Alpine
360842	20	Subtropical, Tropical
664	12	Temperate
0	4	Temperate
133306	20	Subalpine
4589	12	Alpine
2185	12	Subalpine
-	-	Alpine
0	8	Temperate
2091	12	Temperate
0	8	Subtropical, Tropical

Continued on next page

Table 4.4 – continued from previous page

Extent Of Occurrance(eoo)	Area Of Occupancy(aoo)	zones or habitats
101376	16	Temperate

Table 4.5: Species Data Table from column 13 to column 15

Propagation	Major Threats	Habit
Seed, tuber	lh, uc	herb
Seed	lh, oh	herb
Seed	lh, uc	herb
Rhizome, Seed	lh, uc	herb
Seed	oe	tree
Seed	lh	tree
Seed	lh, oe, g	herb
Seed	oe, lh	herb
Rhizome, Seed	oe, lh	herb
Rhizome, Seed	oe, lh	herb
Rhizome, Seed	-	herb
Seed	oe	tree
Seed	lh, oh	climber
Seed	lh, uc	herb
Seed, tuber	oe, lh	orchid
Seed, tuber	oe	herb
Seed	oe	herb
Seed	uc	herb
Seed	lh, oh	herb
Seed	oe	tree
Seed, tuber	lh	herb
Rhizome, Seed	oe	herb

Continued on next page

Table 4.5 – continued from previous page

Propagation	Major Threats	Habit
Rhizome, Seed	oh	herb
Seed	lh, uc	tree
Seed	lh, uc	herb
Bulb	lh, uc	herb
Rhizome, Seed	lh, uc	herb
Seed	oe	herb
Seed	oe	herb
Seed	lh, uc	Shrub
Rhizome, Seed	oe	herb
Rhizome, Seed	oe, lh	herb
Rhizome, Seed	oe, lh	herb
Rhizome, Seed	oe	herb
Rhizome, Seed	lh, uc, oe	herb
Rhizome, Seed	uc	herb
Seed	oe	tree
Seed	oe	herb
Stem, Seed	oe	herb
Rhizome, Seed	oe	herb

4.2 Data Preprocessing

Data cleaning

All the missing values in the data table are assigned as global constant 'Unknown'.

Data Reduction

- The unrelated and irrelevant data was removed at the time of creating data base.
- NA value had been suppressed.

Data transformation

Data transformation is required to convert the format of the value of the data base into a suitable format so that we can apply our desired algorithm.

- The raw data table was transformed into .csv file
- Space or any special character on the attributes section was replaced by '.'
- At the time of converting the data base into data frame, sapply had been applied on data base to convert it in a matrix form. Used as.factor as parameter of sapply to make the all string attributes data base into factor.
- Attributes with numerical values were transformed into continuous variables. Unsupervised discretize() method 'equal width interval' for columns 5,6,8,9,10 had been applied to convert them in to continuous variables and factorize them with a specified labels. For column no. 11, to convert the numerical values to continuous variable, user specified discretize method 'fixed' with specified breaks and labels had been applied.

  Discretization: Before applying apriori, all attributes with numerical values of SpeciesData data table need to be factorized. In my the data table, 5th, 6th, 8th, 9th, 10th and 11th columns are numerical. I had applied unsupervised discretize() method 'equal

width interval' to convert the continuous variable into categorical and factorize them by specified level accordingly.

For each attributes with numerical value, define it into a variable, say,x like x←SpeciesData[, colno] Then use discretize() function into x, to convert the continuous variable into a categorical variable. For attributes no. 5,6,8,9 and 10 unsupervised discretization method 'equal interval width' had been used to discretize the numerical column. For attributes no. 11 'fixed' discretization method with specified range was applied. All the discretized attributes, range and their equivalent or factorized name which had been given at the time of labeling are shown in table.

- data was taken in a data frame and as.factor had been used the data table to factorize remaining item values in the data set.
- Data frame contained the species transaction data had been converted in to transaction object.

Table 4.6: Discretized attributes.

Attributes	**Range**	**Labels**	**Method**
LowerElevationLimit(m)	0.000≤ and<1233.333	Verylow	interval
	1233.333≤ and <2466.667	low	
	2466.6674≤ and≤3700.000	medium	
UpperElevationLimit(m)	600.000 ≤ and < 2166.667	Low	interval
	2166.667≤ and < 3733.333	Medium	
	3733.333≤ and ≤5300.000	High	
GenerationLength(yr)	1.0≤ and<10.8	shortest	interval
	10.8≤ and<20.6	short	
	20.6≤ and<30.4	mid	

Continued on next page

Table 4.6 – continued from previous page

Attributes	Range	Labels	Method
	30.4≤ and<40.2	long	
	40.2≤ and≤50.0	longest	
SpeciesOccurrances	0≤ and<43	toolow	interval
	43≤ and<86	lowOcc	
	86≤ and<129	avg	
	129≤ and≤172	more	
ExtenOfOccurrence	-inf≤ and<500	none	Fixed
	500≤ and<10000	VerySmall	
	10000≤ and<25000	Small	
	25000≤ and<40000	Average	
	40000≤ and<100000	Large	
	100000≤ and≤ inf	VeryLarge	
AreaofOccupancy	0≤ and<5	verylowArea	interval
	5≤ and<10	lowArea	
	10≤ and<15	midArea	
	15≤ and ≤ 20	broad	

4.2.1 Final Species Data table

Table 4.7: Species Data Table from column 1 to column 3

Species	Geographical.distribution	IUCN.status
Aconitum heterophyllum	NW	EN
Aconitum chasmanthum	W	CR
Aconitum violaceum	W	VU
Angelica glauca	W	EN
Acer caesium	NW	LC
Aquilaria malaccensis	E	CR

Continued on next page

Table 4.7 – continued from previous page

Species	Geographical.distribution	IUCN.status
Arnebia benthami	NW	CR
Arnebia euchroma	NW	CR
Bergenia ciliata	NE	VU
Bergenia ligulata	W	VU
Bergenia stracheyi	W	VU
Betula utilis	NW	LC
Cayratia pedata	NE	VU
Coptis teeta	E	EN
Dactylorhiza hatagirea	NW	CR
Dioscorea deltoidea	NW	EN
Ephedra gerardiana	NW	EN
Fritillaria cirrhosa	NE	EN
Gentiana kurroo	NW	CR
Gymnocladus assamicus	NE	CR
Habenaria intermedia	NW	EN
Hedychium spicatum	W	VU
Heracleum candicans	NW	EN
Illicium griffithii	NE	EN
Lilium polyphyllum	NW	CR
Malaxis Muscifera	W	VU
Nardostachys jatamansi	TH	CR
Paris polyphylla	NW	EN
Picrorhiza kurrooa	TH	EN
Piper pedicellatum	E	VU
Podophyllum hexandrum	W	EN
Polygonatum cirrhifolium	W	EN
Polygonatum verticillatum	NW	VU
Rheum australe	TH	EN

Continued on next page

Table 4.7 – continued from previous page

Species	Geographical.distribution	IUCN.status
Saussurea costus	NW	CR
Selinum tenuifolium	NW	NT
Taxus wallichiana	TH	EN
Thalictrum foliolosum	W	VU
Tinospora cordifolia	NW	EN
Valeriana jatamansi	NW	VU

Table 4.8: Species Data Table from column 4 to column 6

PartUsed	LowerElevation.Limit.m.	UpperElevation.Limit.m.
Root	low	High
Root	low	Medium
Root	medium	High
Root	low	Medium
Bark	low	High
Wood	Verylow	Low
Root	low	Medium
Root	medium	High
Rhizome	Verylow	Medium
Root	low	Medium
Root	medium	High
Bark,Root	medium	High
Leave	Verylow	Low
Root	low	Medium
bulb	medium	Medium
Rhizome	Verylow	Medium
Stem	medium	High

Continued on next page

Table 4.8 – continued from previous page

PartUsed	LowerElevation.Limit.m.	UpperElevation.Limit.m.
Bulb	medium	High
Rhizome	low	Medium
Seedpods	Verylow	Low
Bulb	low	Medium
Rhizome	Verylow	Medium
Root	low	High
Seed,Fruit	Verylow	Low
Bulb	low	Medium
Bulb	medium	High
Root	low	High
Rhizome	low	High
Root	medium	High
Leave,Fruit	Verylow	Low
Root	low	High
Rhizome	low	High
Rhizome	low	Medium
Root	medium	High
Root	medium	High
Root	low	High
Leave,Bark	Verylow	Medium
Root	low	Medium
Stem,Root	Verylow	Low
Root	low	Medium

Table 4.9: Species Data Table from column 7 to column 9

Habitat.type	Generation.length.yr.	Species.occurrences
F,Gl	shortest	lowOcc
F,Gl	shortest	tooLow
Gl,R	shortest	lowOcc
F,Gl,R	shortest	tooLow
F	NA	tooLow
F	longest	tooLow
Sl	NA	tooLow
R	NA	more
F,R	mid	tooLow
F,R	NA	tooLow
R	NA	lowOcc
F	NA	avg
F,Sl	short	tooLow
F,Sl	shortest	tooLow
Sl	NA	lowOcc
F	NA	more
R	NA	more
F	NA	lowOcc
R,Gl	shortest	tooLow
F	mid	tooLow
F,Gl	NA	tooLow
F,Sl	NA	lowOcc
R	NA	tooLow
F	mid	tooLow
F	shortest	tooLow
F,Gl,Sl	shortest	tooLow
Gl,R	shortest	lowOcc
F	NA	tooLow

Continued on next page

Table 4.9 – continued from previous page

Habitat.type	Generation.length.yr.	Species.occurrences
F	NA	tooLow
F	short	tooLow
F	NA	lowOcc
F,Sl	NA	tooLow
F	NA	tooLow
R	NA	tooLow
Gl	shortest	tooLow
Sl	NA	tooLow
F	mid	tooLow
F,Sl	NA	lowOcc
F	shortest	tooLow
F,Gl	NA	lowOcc

Table 4.10: Species Data Table from column 10 to column 12

ExtentOfOccurrance.eoo.	AreaOfOccupancy.aoo.	zones.or.habitats
verySmall	broad	Alpine
verySmall	broad	Alpine,subalpine
Small	broad	Alpine,subalpine
Small	broad	Alpine
verySmall	broad	Temperate
NA	NA	Subtropical,Tropical
NA	NA	Alpine
verySmall	broad	Alpine
Small	broad	Temperate
none	verylowArea	Unknown
Small	broad	Alpine

Continued on next page

Table 4.10 – continued from previous page

ExtentOfOccurrance.eoo.	AreaOfOccupancy.aoo.	zones.or.habitats
Average	broad	Temperate
none	verylowArea	Subtropical,Tropical
none	verylowArea	Temperate
Average	broad	Alpine
none	lowArea	Temperate
Small	broad	Alpine
none	verylowArea	Alpine
verySmall	broad	Alpine
none	verylowArea	Subtropical
verySmall	midArea	Temperate
none	verylowArea	Temperate
Small	broad	Alpine
NA	NA	Temperate
Small	broad	Subtropical
NA	NA	Temperate
none	lowArea	Alpine
none	lowArea	Subtropical
none	lowArea	Alpine
veryLarge	broad	Subtropical,Tropical
verySmall	midArea	Temperate
none	verylowArea	Temperate
veryLarge	broad	subalpine
verySmall	midArea	Alpine
verySmall	midArea	subalpine
NA	NA	Alpine
none	lowArea	Temperate
verySmall	midArea	Temperate
none	lowArea	Subtropical,Tropical

Continued on next page

Table 4.10 – continued from previous page

ExtentOfOccurrance.eoo.	AreaOfOccupancy.aoo.	zones.or.habitats
veryLarge	broad	Temperate

Table 4.11: Species Data Table from column 13 to column 15

Propagation	Major Threats	Habit
Seed, tuber	lh, uc	herb
Seed	lh, oh	herb
Seed	lh, uc	herb
Rhizome, Seed	lh, uc	herb
Seed	oe	tree
Seed	lh	tree
Seed	lh, oe, g	herb
Seed	oe, lh	herb
Rhizome, Seed	oe, lh	herb
Rhizome, Seed	oe, lh	herb
Rhizome, Seed	-	herb
Seed	oe	tree
Seed	lh, oh	climber
Seed	lh, uc	herb
Seed, tuber	oe, lh	orchid
Seed, tuber	oe	herb
Seed	oe	herb
Seed	uc	herb
Seed	lh, oh	herb
Seed	oe	tree
Seed, tuber	lh	herb
Rhizome, Seed	oe	herb

Continued on next page

Table 4.11 – continued from previous page

Propagation	Major Threats	Habit
Rhizome, Seed	oh	herb
Seed	lh, uc	tree
Seed	lh, uc	herb
Bulb	lh, uc	herb
Rhizome, Seed	lh, uc	herb
Seed	oe	herb
Seed	oe	herb
Seed	lh, uc	Shrub
Rhizome, Seed	oe	herb
Rhizome, Seed	oe, lh	herb
Rhizome, Seed	oe, lh	herb
Rhizome, Seed	oe	herb
Rhizome, Seed	lh, uc, oe	herb
Rhizome, Seed	uc	herb
Seed	oe	tree
Seed	oe	herb
Stem, Seed	oe	herb
Rhizome, Seed	oe	herb

Chapter 5

Experiments and Result Discussion

5.1 System setup

Downloading and Installing R

R version 3.5.0 for Windows (62 megabytes, 32/64 bit) was downloaded. R installation was done at the system location C:\Program Files \ files\R\R3.5.0

In figure 5.1, figure 5.2 and figure 5.3 downloading and installing steps of R have been shown.

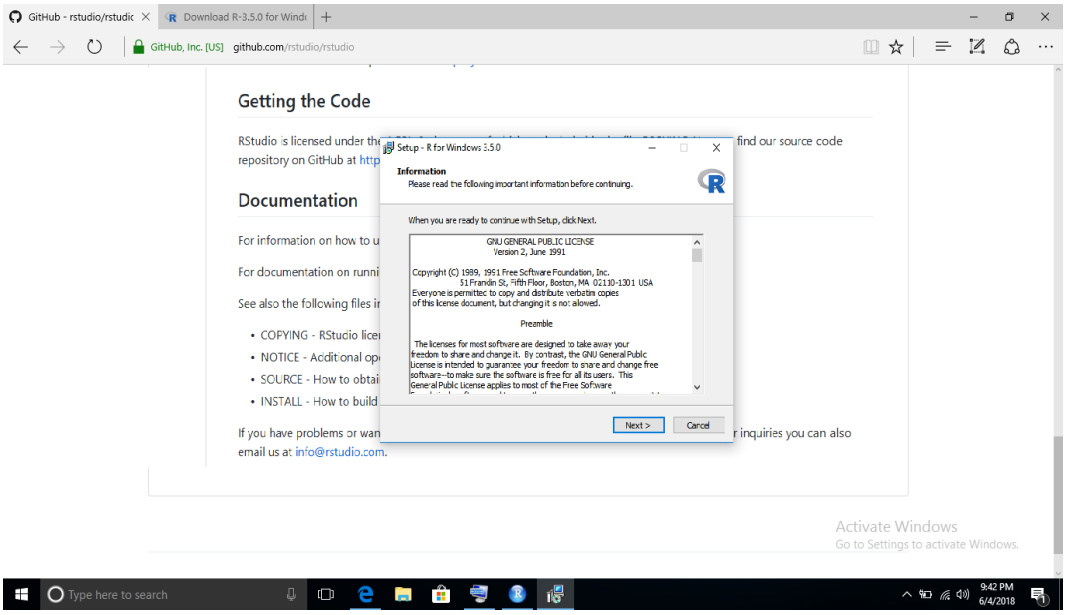


Figure 5.1: first step of downloading R

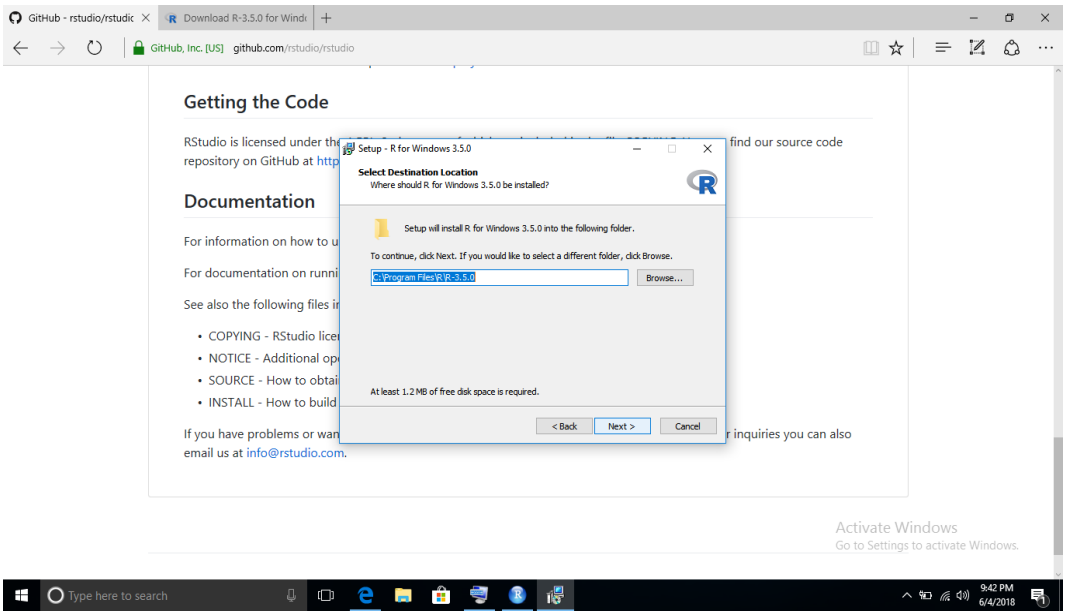


Figure 5.2: Second step of downloading R with path selection

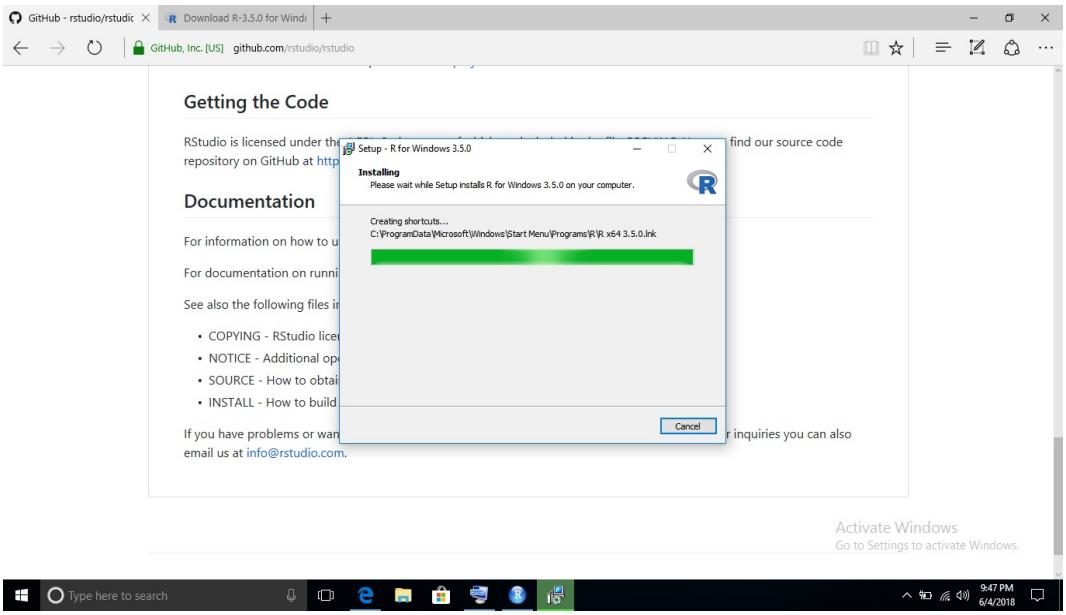


Figure 5.3: Installing R

R studio downloading and installation

RStudio desktop for Open Source License was selected to download RStudio. Installation of R studio was done at the same path location of R, so that R can be connected with R studio.

5.2 Experiments on the Data set

Applying apriori algorithm

The species data frame had been converted into transaction object and finally arules provided Apriori algorithm had been applied on it. By changing the parameters value, tried to get some valuable results. When apriori method with minimum support value .1 and minimum confidence value .8 had been applied on the species transaction object, total 1253 rules had been generated. A scatter plot representation of those rules with parameters support and confidence value and with another parameter lift value is shown in Figure 5.4. In this plot we can see maximum rules with higher lift value has support value from .1 to .125. Rules had been selected by considering which were providing relevant, useful information of conservation of threatened Medicinal plants species. In this respective, to generate the limited and relevant rules duplicate or redundant rules were removed and for appearing with specific attributes value, 'appearance' parameter was used. Each rule can be read as if the lhs attributes value-set before the symbol '⇒' occurs then rhs will happen and the the lhs attributes value-set before the symbol '⇒' describes the support value, it is the number of species out of total species which supports the lhs attributes value-set and rhs attribute value describes confidence percentage of validation of the rule. Figure 5.5 shows the top 25 frequent items or elements in my data set and their respective individual absolute frequency.

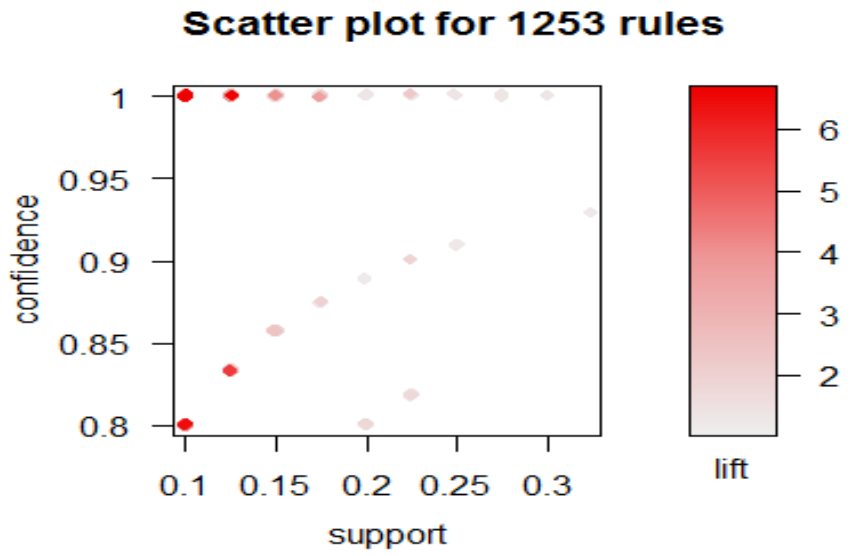


Figure 5.4: scatter plot of total 1253 rules

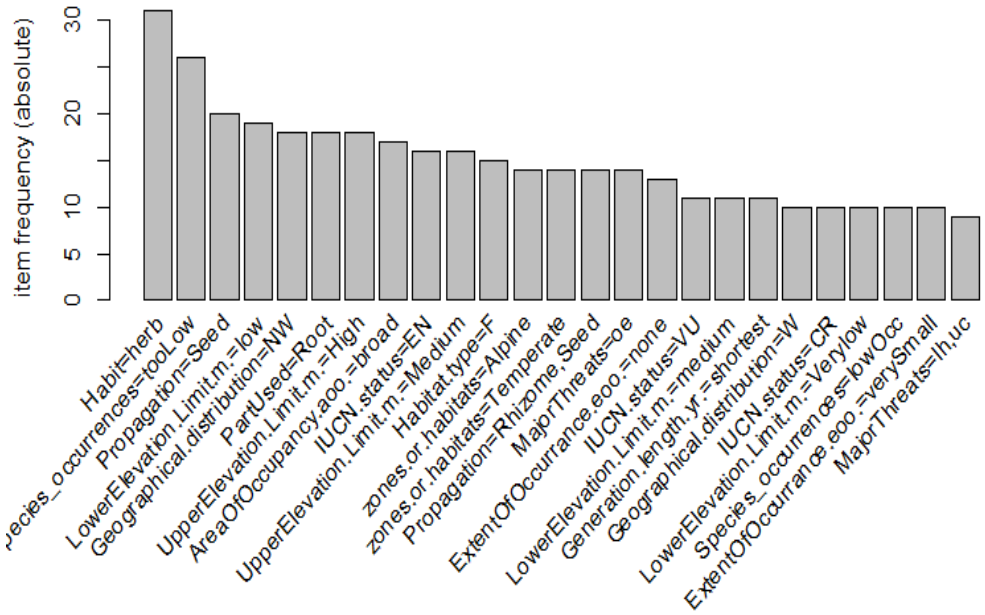


Figure 5.5: Top 25 frequent elements

5.3 Result Presentation

5.3.1 Presenting summary of result

Rule which gives 'IUCN.status= CR'

There were 3 rules generated with consequent value 'IUCN.status= CR'. After removing the redundant rule, I got 2 rules. In Figure 5.6 inspect(CR) shows the generated all 3 rules and inspect(new.CR) generates the rules after removing the redundant one, where support value is .1 and confidence value is.8, from which one can say there are 10 percent species in my data table supports the rules and 80 percent chances that their threat status according to IUCN is CR(Critically endangered). Those 2 rules say if species are found in low lower elevation limit to medium upper elevation limit, species occurrences is toolow and seed is used for propagation or if species are found in low lower elevation limit to medium upper elevation limit, species occurrences is toolow and habit is herbal then their threat or conservation status according to IUCN is Critically endangered. A graph representation of these two rules was shown in Figure 5.7.

Rules having consequent element IUCN.status=EN

Total 9 rules had been generated with the consequent value IUCN.stastus=EN. shown in figure 5.8. From the total rules 3 rules had been selected shown in Figure 5.9, for each Rule confidence value is 1 can be defined as, there are 100 percent chances of happening EN(Endangered) for the attributes value set appearing in the left hand side of the rule.

In Figure 5.8 rule no. 1 says if species belongs to high (range 3733.333 ≤ to ≤ 5300) upper elevation limit, F(forest) habitat type and herb(herbal) habits then their conservation status according to IUCN is EN. Total 10 percent Species out of total species in data set supports this rule. Thus, rule number 2 and rule number 3 also can be defined.

Rules where consequent element is IUCN.status=VU

Two rules had been selected after removing the duplicate rules. In Figure 5.11 and Figure 5.12 the visualized representation of those rules have been shown.

Rules which give MajorThreats= lh,uc as a consequent element

Total 4 rules had been generated, Species which are found in high upper elevation limit and have a shortest generation length their major threats are lost of habitat and unsustainable collection(lh,uc). Graph visualization of those rules was shown in Figure 5.13.

Rules which give 'MajorThreats= oe' as a consequent element

Total 14 rules out of 42 rules had been generated by removing the redundant rules, shown in Figure 5.15. Five selected rules where element MajorThreats = oe had been appeared as a consequent shown in Figure 5.17.

From rule no. 10 in figure 5.17 we can say if species distribution is in NW(North Western) Himalayan region in temperate zone then there are 80 percent chances of occurring over exploitation(oe) as a major threat.

Rules which give Propagation= Rhizome,seed as a consequent element

Total 6 rules out of 10 rules had been generated by removing the redundant rules, shown in figure 5.18 From which first 3 rules had been selected, shown in figure 5.20.

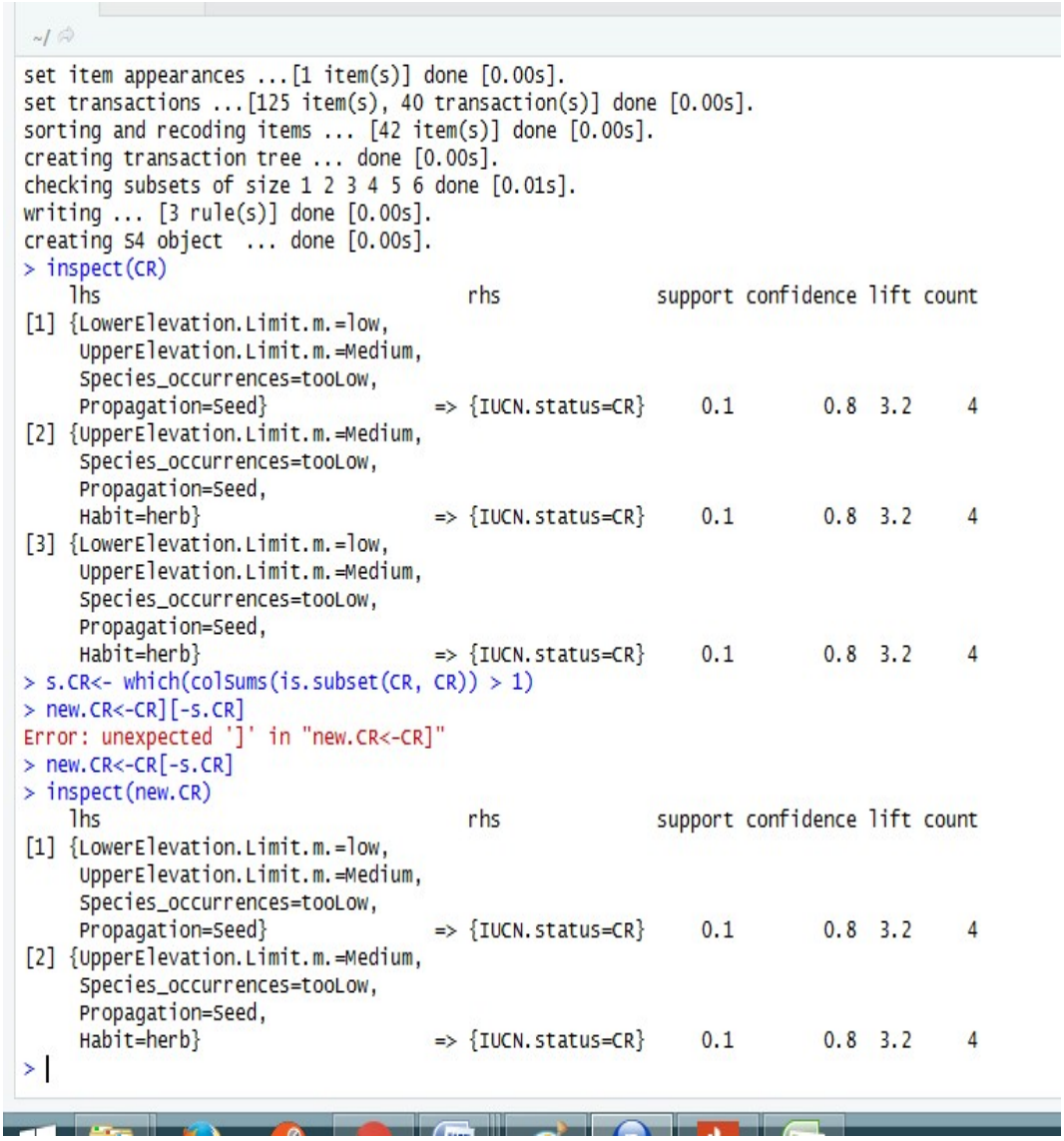


Figure 5.6: Rules having CR as consequent

In Figure 5.6 inspect(CR) shows the generated all 3 rules and inspect(new.CR) generates the rules after removing the redundant one, where support value is .1 and confidence value is .8. From which one can say there are 10 percent species in my data table supports the rules and 80 percent chances that their threat status according to IUCN will be CR(Critically endangered).

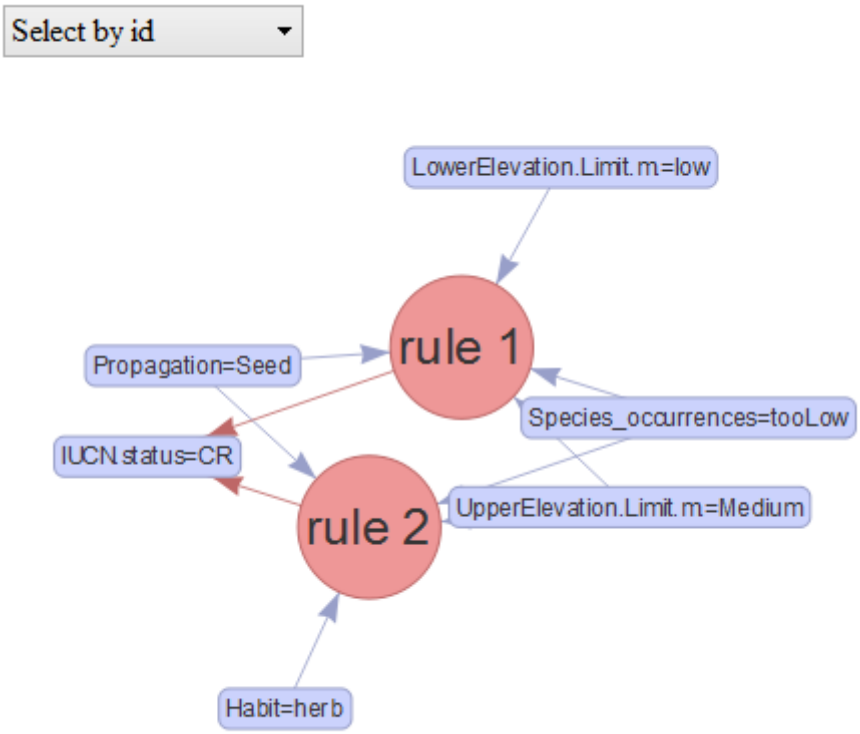


Figure 5.7: Graph based plot of 2 rules

In the above Figure 5.7 rule 1 and rule 2 represents IUCN.status=CR as their consequent by connected with the red coloured arrows and the blue coloured arrows connected with each rule represents which elements are antecedent.

```
To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
> inspect(subEN)
    lhs                                  rhs              support confidence lift count
[1] {AreaOfOccupancy.aoo.=lowArea,
     MajorThreats=oe}                 => {IUCN.status=EN}   0.125          1  2.5     5
[2] {Habitat.type=F,
     AreaOfOccupancy.aoo.=lowArea}    => {IUCN.status=EN}   0.125          1  2.5     5
[3] {Species_occurrences=tooLow,
     AreaOfOccupancy.aoo.=lowArea}    => {IUCN.status=EN}   0.100          1  2.5     4
[4] {Habitat.type=F,
     ExtentOfOccurrance.eoo.=none,
     Habit=herb}                      => {IUCN.status=EN}   0.125          1  2.5     5
[5] {ExtentOfOccurrance.eoo.=none,
     Propagation=Seed,
     Habit=herb}                      => {IUCN.status=EN}   0.100          1  2.5     4
[6] {Habitat.type=F,
     MajorThreats=oe,
     Habit=herb}                      => {IUCN.status=EN}   0.125          1  2.5     5
[7] {UpperElevation.Limit.m.=High,
     MajorThreats=oe,
     Habit=herb}                      => {IUCN.status=EN}   0.125          1  2.5     5
[8] {Species_occurrences=tooLow,
     MajorThreats=oe,
     Habit=herb}                      => {IUCN.status=EN}   0.100          1  2.5     4
[9] {UpperElevation.Limit.m.=High,
     Habitat.type=F,
     Habit=herb}                      => {IUCN.status=EN}   0.100          1  2.5     4
>
```

Figure 5.8: All generate Rules where consequent element is 'IUCN.status=EN'

The above Figure 5.8 shows the selected rules where species threat status is endangered.

Rule No.	Rules	Support	confidence	Lift	Count	Species
1	{UpperElevation.Limit.m.=High, Habitat.type=F,Habit=herb} => {IUCN.status=EN}	.1	1	2.5	4	Fritillaria cirrhosa Paris polyphylla, Picrorhiza kurrooa, Podophyllum hexandrum
2	{UpperElevation.Limit.m.=High, MajorThreats=oe,Habit=herb} => {IUCN.status=EN}	.125	1	2.5	5	Ephedra gerardiana, Paris polyphylla, Picrorhiza kurrooa, Podophyllum hexandrum, Rheum australe
3	{Habitat.type=F, MajorThreats=oe,Habit=herb} => {IUCN.status=EN}	.125	1	2.5	5	Dioscorea deltoidea, Paris polyphylla, Picrorhiza kurrooa, Podophyllum hexandrum, Tinospora cordifolia

Figure 5.9: Selected Rules where consequent element is 'IUCN.status=EN'

Figure 5.9 shows all the selected rules with confidence value 1 and the group of species which support those rules.

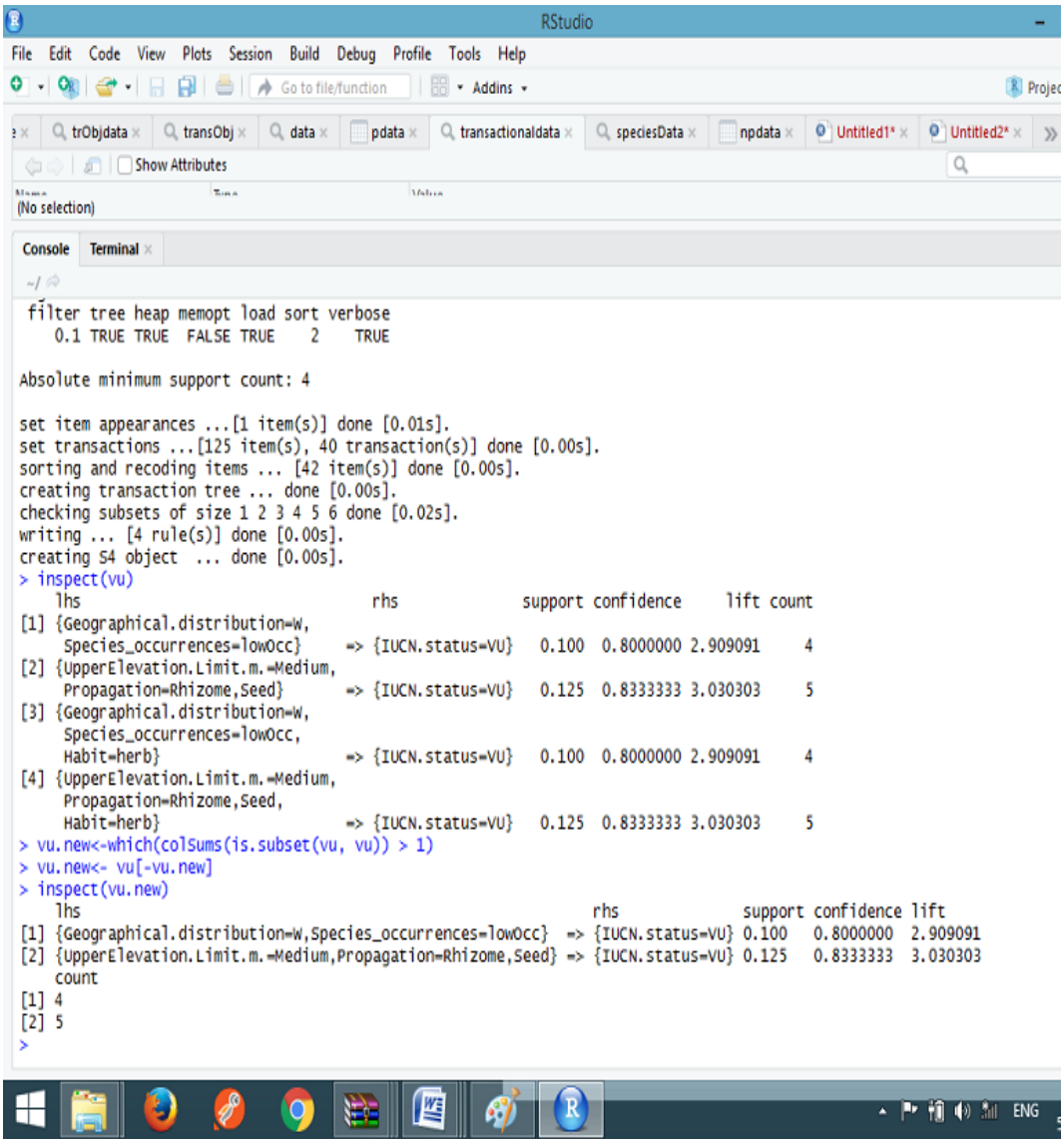


Figure 5.10: Rules where consequent element is IUCN.stastus=VU

First 4 rules in this figure represents the rules without removing the redundant rules, After removing redundant rules when applied inspect(vu.new), it generated 2 rules, where the first rule says if species are found in western Himalayan region and their occurrences is between 43 to 86 then their threat status is vulnerable and the second rule say if species upper elevation altitude range is from 2166.667 m to less than 3733.333 m and propagation material is rhizome and seed then those species are vulnerable.

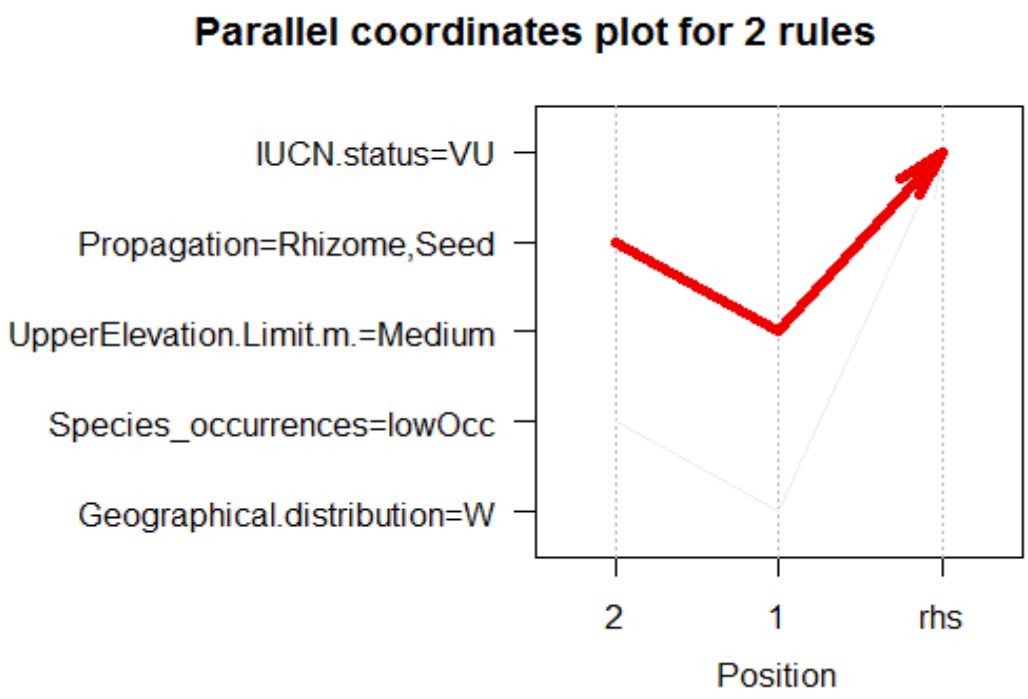


Figure 5.11: Rules representation in a parallel coordinate plot where consequent element is 'IUCN.status=VU'

A parallel coordinate plot shown in the above figure visualize the 2 rules, where the red arrow specifies species which upper elevation limit is medium and propagation material is rhizome and seed are likely to be vulnerable (VU) and the second arrow specifies species found in Western (W) Himalayan region and lowOcc occurrences are likely to be vulnerable.

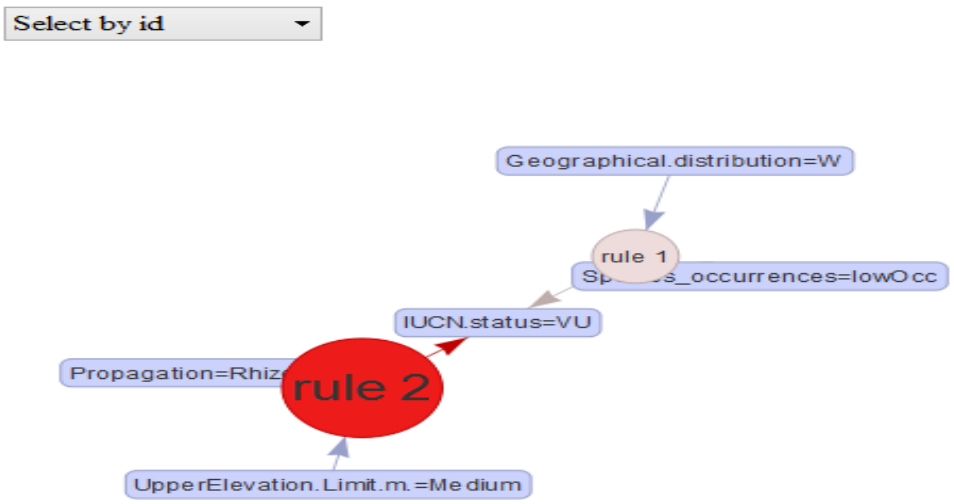


Figure 5.12: Graph based visualization of rules where consequent value is vu

Figure 5.12 shows both rules represent IUCN.status=VU as a consequent element and rule 2 gives greater lift value that means elements connected with rule2 are more likely to be happened.

	lhs	rhs	support	confidence	lift	count
[1]	{UpperElevation.Limit.m.=High, Generation.length.yr.=shortest}	=> {MajorThreats=lh,uc}	0.1	0.8	3.555556	4
[2]	{PartUsed=Root, LowerElevation.Limit.m.=low, Generation.length.yr.=shortest}	=> {MajorThreats=lh,uc}	0.1	0.8	3.555556	4

Figure 5.13: Selected rules where MajorThreats= lh,uc represents as a consequent

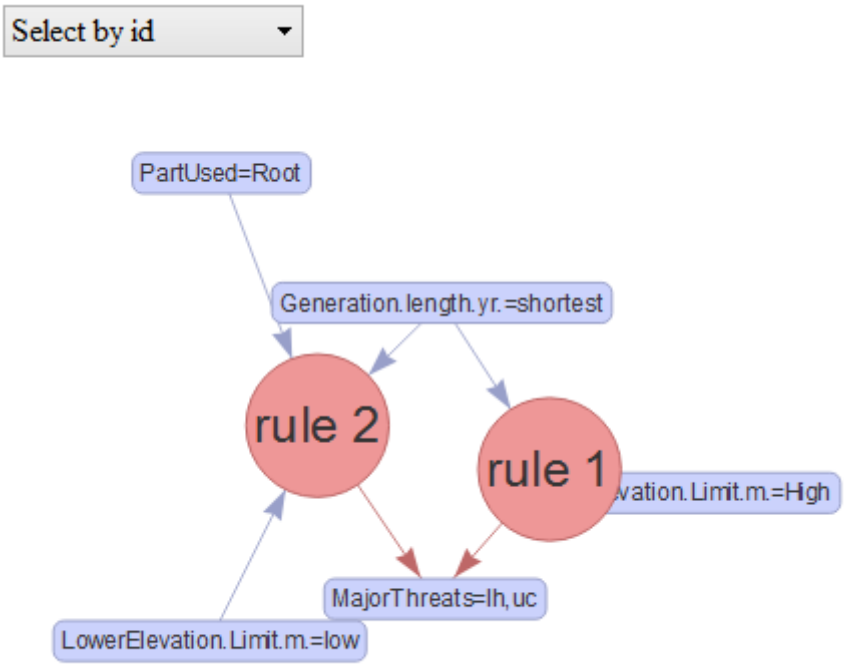


Figure 5.14: Graph based visualization of rules where consequent element is MajorThreat=lh,uc

In the above Figure 5.14 the elements which are connected with each individual rule represents the elements set contained by that rule. Thus for rule 1 can say, species with shortest generation length and high upper elevation limits are majorly threatened due to loss of habitat and unsustainable collection.

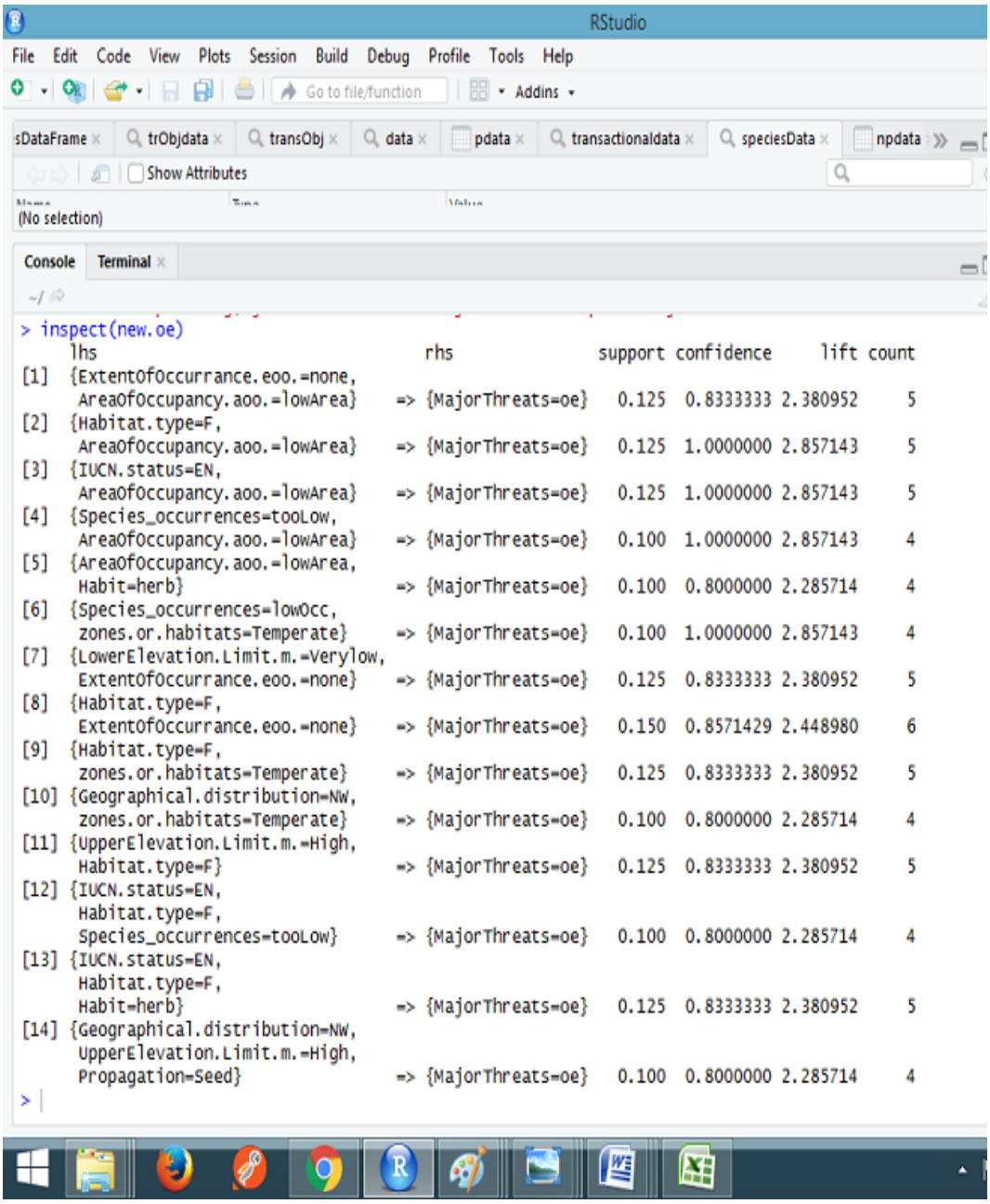


Figure 5.15: Total generated 14 rules where consequent element is MajorThreat=oe

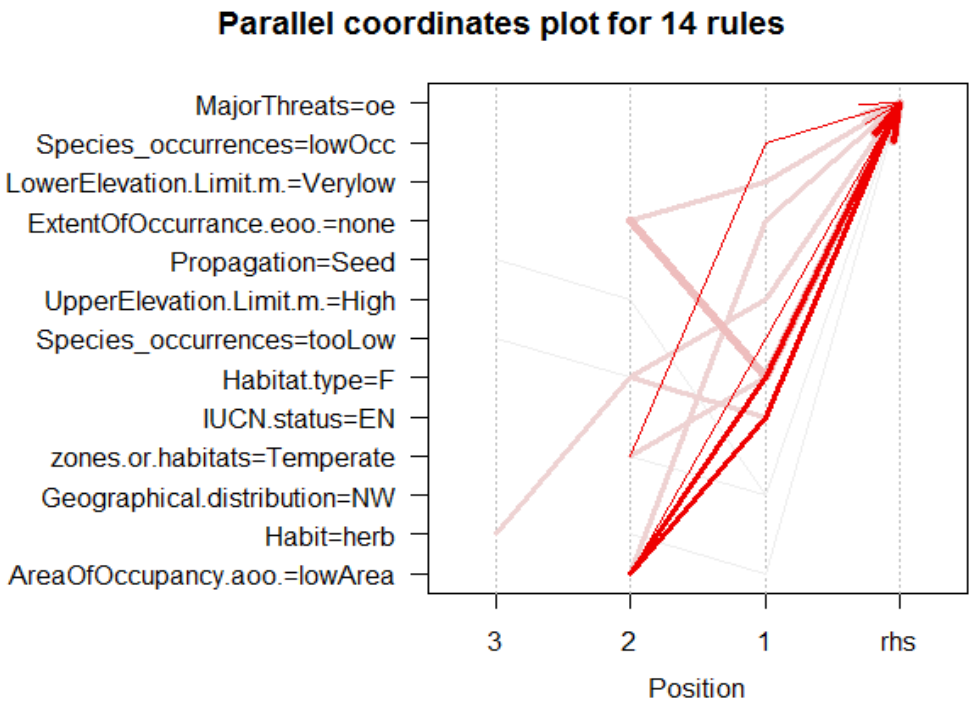
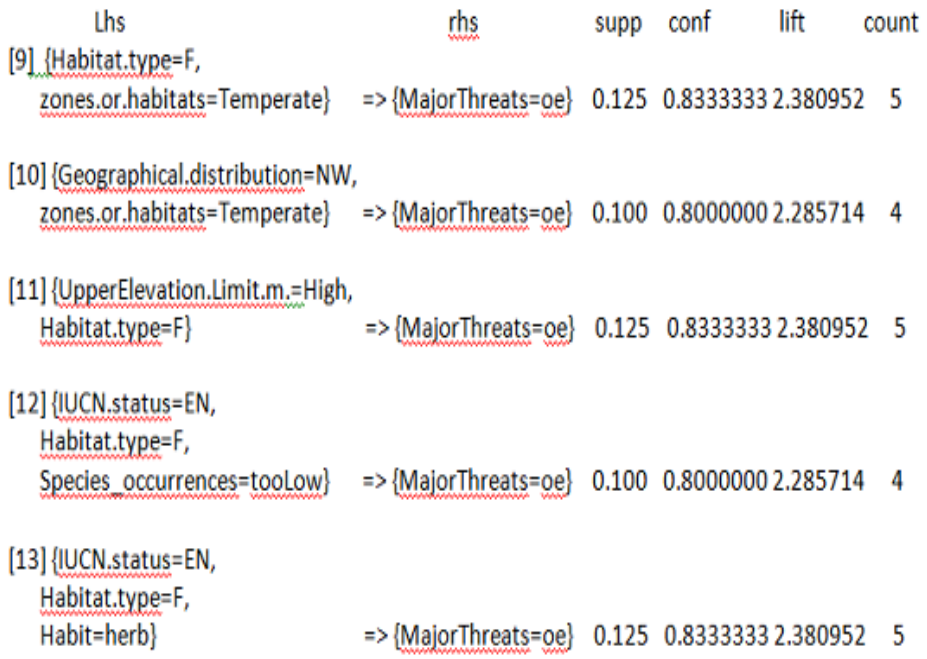


Figure 5.16: Parallel coordinate plot of 14 rules where consequent element is 'MajorThreat=oe'

The lowest dark red arrow in the above plot in Figure 5.16 specifies if species IUCN status is Endangered and area of occupancy is lowArea then for those species major threat is over exploitation(oe). Chances of happening of over exploitation is more in case of dark red arrow rather than the others arrows.



	Lhs	rhs	supp	conf	lift	count
[9]	{Habitat.type=F, zones.or.habitats=Temperate}	=> {MajorThreats=oe}	0.125	0.8333333	2.380952	5
[10]	{Geographical.distribution=NW, zones.or.habitats=Temperate}	=> {MajorThreats=oe}	0.100	0.8000000	2.285714	4
[11]	{UpperElevation.Limit.m.=High, Habitat.type=F}	=> {MajorThreats=oe}	0.125	0.8333333	2.380952	5
[12]	{IUCN.status=EN, Habitat.type=F, Species_occurrences=tooLow}	=> {MajorThreats=oe}	0.100	0.8000000	2.285714	4
[13]	{IUCN.status=EN, Habitat.type=F, Habit=herb}	=> {MajorThreats=oe}	0.125	0.8333333	2.380952	5

Figure 5.17: Selected 5 rules where consequent item is MajorThreat=oe

In Figure 5.17, rule number 9 says, there are 12.5 percent species objects or total 5 species in my data set which supports the lhs attributes value set and around 83.3 percent chances among them that the consequent element will be 'MajorThreat=oe'. Thus the remaining rules can be defined. From rule no.11 we can say if species are found in high upper elevation limit and in F(Forest) habitat type then 83.33 percent chances of their major threat will be oe (Over Exploitation). Rule no.12 can be defined as if species threat status is EN(Endangered), habitat type F and their occurrences toolow then there are 80 per cent chances of their major threat will be oe. Thus rule number 10 and rule number 13 can be defined according to their confidence value.

```
> plot(new.propagation, method= paracoord )
> inspect(new.propagation)
    lhs                                  rhs                          support confidence     lift count
[1] {Species_occurrences=tooLow,
     MajorThreats=oe,lh}              => {Propagation=Rhizome,Seed}    0.100  1.0000000 2.857143     4
[2] {MajorThreats=oe,lh,
     Habit=herb}                      => {Propagation=Rhizome,Seed}    0.100  0.8000000 2.285714     4
[3] {IUCN.status=VU,
     UpperElevation.Limit.m.=Medium}  => {Propagation=Rhizome,Seed}    0.125  0.8333333 2.380952     5
[4] {IUCN.status=VU,
     AreaOfOccupancy.aoo.=broad,
     Habit=herb}                      => {Propagation=Rhizome,Seed}    0.100  0.8000000 2.285714     4
[5] {PartUsed=Root,
     LowerElevation.Limit.m.=low,
     UpperElevation.Limit.m.=High}    => {Propagation=Rhizome,Seed}    0.100  0.8000000 2.285714     4
[6] {PartUsed=Root,
     UpperElevation.Limit.m.=High,
     Species_occurrences=tooLow}      => {Propagation=Rhizome,Seed}    0.100  0.8000000 2.285714     4
>
```

Figure 5.18: Total generated 6 rules which give 'Propagation=Rhizome,Seed' as consequent element

In the above figure rule number 1 gives confidence value 1, with support value .10, that means there are 100 percent chances for species with tooLow occurrences and over exploitation and loss of habitat as major threats can be propagated using rhizome and seed.

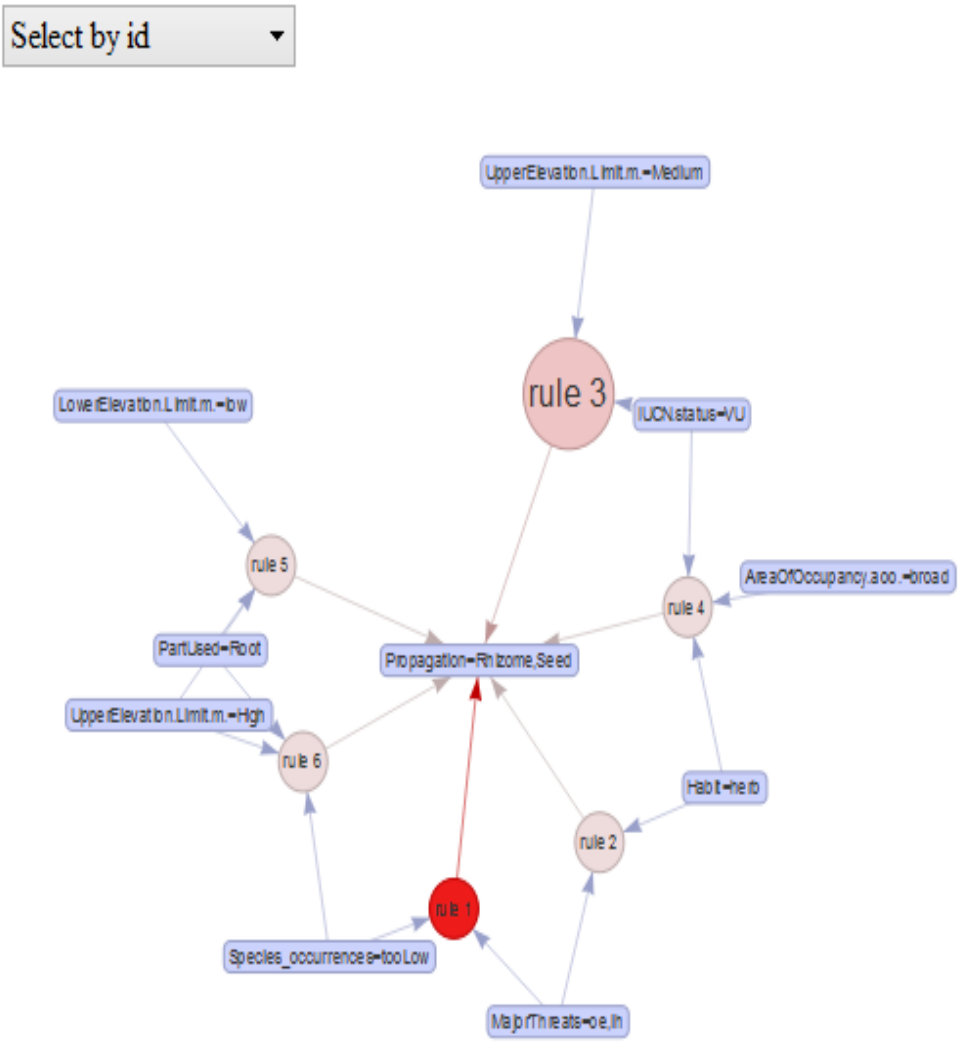


Figure 5.19: Graph based visualization of rules where consequent element is Propagation=Rhizome,Seed

In Figure 5.19, all 6 rules point the consequent element Propagation=Rhizome,Seed by connecting with the the out going arrow. Here, rule 1 contains species_occurrences= tooLow and MajorThreats=oe,lh as lhs elements set and gives consequent element Propagation=Rhizome,Seed and rule 1 represents more likely happening elements set rather than others 5 rules.

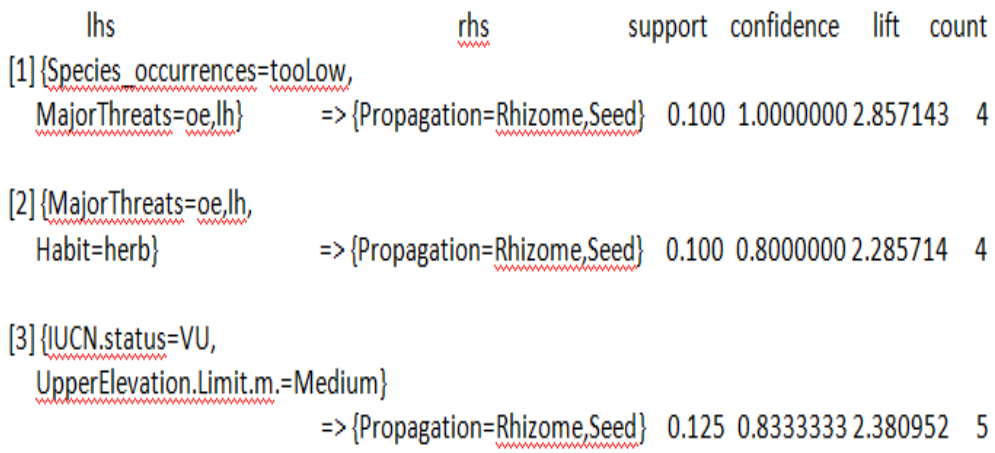


Figure 5.20: Selected 3 rules which give 'Propagation=Rhizome,Seed' as consequent element

In Figure 5.20 the first rule give confidence value 1 and lift value 2.857143 and the rule 1 can be described as if species occurrences is less than 43 and major threat is over exploitation and loss of habitat then for their propagation rhizome cutting and seed are used.

5.4 Discussion

By analyzing the results, some new useful correlation has been identified between the elements set of species data base. For a group of species relationship in species major threats with respect to species threat status, distribution, occurrence had been identified. For a group of species relationship in species propagation material with respect to their occurrences and distribution, threat status has been identified. Relationship in species threat status with respect to species altitude range and habitat type, altitude range and major threats, habitat type

and major threats had been identified.Then analysis the rule got that species Aconitum chasmanthum, Arnebia benthami, Gentiana kurroo, Lilium polyphyllum are found in low (1233.333 ≤ to < 2466.667) lower elevation limit to medium (2166.667 ≤ to < 3733.333) upper elevation limit, those species occurrences is less than 43 and seed is used for propagation and their threat or conservation status according to IUCN is Critically Endangered, shown in Figure 5.21.

Rule	sup	conf	lift	count	species
{LowerElevation.Limit.m.=low,UpperElevation.Limit.m.=Medium,Species_occurrences=tooLow,Propagation=Seed,Habit=herb}=> {IUCN.status=CR}	.1	.8	3.2	4	Aconitum chasmanthum, Arnebia benthami, Gentiana kurroo, Lilium polyphyllum

Figure 5.21: Group of species which supports the rule which are critically endangered

Relationship in species threat status with respect to species altitude range, habit and habitat had been identified. Fritillaria cirrhosa, Paris polyphylla, Picrorhiza kurrooa, Podophyllum hexandrum are found in high(range 3733.333 ≤ to ≤ 5300) upper elevation limit, F(forest) habitat type and herb(herbal) habits then their conservation status according to IUCN is Enangered.
Aconitum violaceum, Bergenia stracheyi, Hedychium spicatum,Thalictrum foliolosum species are vulnerable where they are found in western himalayan region with occurrences value from 43 to 86 and those species habit is herb.

Species in Forest habitat type and temperate zone are majorly threatened due to over exploitation. Acer caesium, Betula utilis, Dioscorea deltoidea, Podophyllum hexandrum, Taxus wallichiana belong to this categoty. Species in North western Himalayan region in temperate zone are threatened due to their major threat over exploitation. Acer caesium, Betula utilis, Dioscorea deltoidea, Valeriana jatamansi belong to this category. wspecies Acer caesium, Betula utilis, Paris polyphylla, Picrorhiza kurrooa, Podophyllum hexandrum are also threatened due to major threats over exploitation and they are found in high upper elevation limit in forest habitat type. where Species Dioscorea deltoidea, Paris polyphylla, Picrorhiza kurrooa, Tinospora cordifolia, Podophyllum hexandrum are over exploited, where their threat status is endangered and those species are Forest habitat type. So it can be said that species Paris polyphylla, Picrorhiza kurrooa, Podophyllum hexandrum are endangered, found in high upper elevation limit, their habitat type is forest.

Rules	Sup	Conf	Lift	Count	Species Name
{Habitat.type=F, zones.or.habitats=Temperate} =>{MajorThreats=oe}	.125	0.833	2.380952	5	Acer caesium, Betula utilis, Dioscorea deltoidea, Podophyllum hexandrum, Taxus wallichiana
{Geographical.distribution=NW, zones.or.habitats=Temperate} =>{MajorThreats=oe}	.1	0.800	2.285714	4	Acer caesium, Betula utilis, Dioscorea deltoidea, Valeriana jatamansi
{UpperElevation.Limit.m.=High, Habitat.type=F} => {MajorThreats=oe}	.125	0.833	2..380952	5	Acer caesium, Betula utilis, Paris polyphylla, Picrorhiza kurrooa, Podophyllum hexandrum
{IUCN.status=EN, Habitat.type=F, Habit=herb} => {MajorThreats=oe}	.125	0.833	2.380952	5	Dioscorea deltoidea, Paris polyphylla, Picrorhiza kurrooa, Tinospora cordifolia, Podophyllum hexandrum

Figure 5.22: Group of species which support the rules which give over exploitation as major threat

If species occurrences is tooLow(0≤ to < 43) and major threat is over exploitation and loss of habitat then those species can be cultivated using rhizome cutting and seed production. Species name Figure 5.23 shows species Bergenia ciliate, Bergenia ligulata, Polygonatum cirrhifolium and Polygonatum verticillatum belong to this category.

If species threat status is vulnerable and are found in medium upper elevation limit(range 2166.667 ≤ and < 3733.333) then those species more likely(80 percent)can be cultivated using rhizome and seed. Bergenia ciliate, Bergenia

ligulata, Hedychium spicatum, Polygonatum verticillatum, Valeriana jatamansi belong to this category, shown in Figure 5.23.

If species major threat is over exploitation and loss of habitat and they are herbal habit then for their cultivation rhizome cutting and seed can be used, Bergenia ciliate, Bergenia ligulata, Polygonatum cirrhifolium and Polygonatum verticillatum are belong to this category, shown in Figure 5.23. Thus observe that tool of association rule mining in r identified relation by rule generation which is useful to generate information between a large number of data elements set.

Rules	Sup	Conf	Lift	Count	Species
{Species_occurrences=tooLow, MajorThreats=oe,lh} => {Propagation=Rhizome,Seed}	.1	1.0	2.56	4	Bergenia ciliate, Bergenia ligulata, Polygonatum cirrhifolium, Polygonatum verticillatum
{MajorThreats=oe,lh, Habit=herb} =>{Propagation=Rhizome,Seed}	.1	.87	2.28	4	Bergenia ciliate, Bergenia ligulata, Polygonatum cirrhifolium, Polygonatum verticillatum
{IUCN.status=VU, UpperElevation.Limit.m.=Medium} => {Propagation=Rhizome,Seed}	.125	.83	2.38	5	Bergenia ciliate, Bergenia ligulata, Hedychium spicatum, Polygonatum verticillatum, Valeriana jatamansi

Figure 5.23: Group of species which support the rule for which propagation material are used rhizome and seed

Chapter 6

Conclusion

Ecologists need a tool which can easily extract information from a large biodiversity data base and can generate the knowledge and then can use this information in their research. In this respect, application of data mining had been considered to extract information from a large data set and to identify the relationship among elements by generating rules or pattern. In this study data mining is used to obtain information about species threat status, major threats and propagation material.

Application of data mining technique on biodiversity data successfully identified new association rules providing new information of threatened species data. It had shown from the generated rules some relationships in different elements set had been identified. Relationships in species major threat with respect to their threat status and habitat type; distribution and habitats; altitude range and habitat type had been identified. Relationship in different threat status with respect to their habitat type and major threats, distribution and occurrences, altitude range and major threats had been identified. Another relationship in species propagation with respect to their occurrences and major threats; threat status and altitude range had been identified. Thus, applying association rule mining in r ecologist can extract and identify new information. I had used a small species data set and then applied association rule mining tool on species data, but one can apply this tool for a large biodiversity data base to get valuable knowledge.

Chapter 7

Species data references

Table 7.1: Species Data collection references

Species Number	References
1	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.]
2	
3	[Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi]
4	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.]
5	
6	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill]
7	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi and Maikhuri(2017)], [Ganaie et al.(2010)Ganaie, Aslam, and Nawchoo]
8	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.]
9	[KUMARI et al.(2012)KUMARI, Chandra JOSHI, and Mohan TEWARI]
10	[Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Negi and Maikhuri(2017)], [Butola et al.(2008)Butola, Badola et al.]
11	[Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Butola et al.(2008)Butola, Badola et al.]
12	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi]
13	

Continued on next page

Table 7.1 – continued from previous page

Species Number	References
14	[Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.]
15	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.], [Negi and Maikhuri(2017)]
16	[Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Srivastava et al.(2016)Srivastava, Srivastava, and Dangwal]
17	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.], [Barik et al.(2018)Barik, Rao, Haridasan, Adhikari, Singh, Tiwary et al.]
18	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill]
19	[Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.]
20	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.], [Barik et al.(2018)Barik, Rao, Haridasan, Adhikari, Singh, Tiwary et al.]
21	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi]
22	[Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Negi and Maikhuri(2017)], [Phondani et al.(2016)Phondani, Bhatt, Negi, Kothyari, Bhatt, and Maikhuri], [Rawat et al.(2018)Rawat, Jugran, Bhatt, and Rawal]
23	[Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Butola et al.(2010)Butola, Vashistha, Malik, and Samant], [Butola et al.(2008)Butola, Badola et al.], [Jan et al.(2018)Jan, Singh, Maqbool, and Nawchoo]
24	[Angami et al.(2017)Angami, Bhagawati, Touthang, Ronya et al.]
25	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.], [Dhyani et al.(2018)Dhyani, Baskin, Nautiyal, and Nautiyal]

Continued on next page

Table 7.1 – continued from previous page

Species Number	References
26	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi]
27	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Purohit et al.(2012)Purohit, Chauhan, Andola, Prasad, Nautiyal, and Nautiyal]
28	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.], [PAUL et al.(2015)PAUL, GAJUREL, and DAS]
29	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.], [Negi and Maikhuri(2017)]
30	
31	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.], [Negi and Maikhuri(2017)], [Butola et al.(2008)Butola, Badola et al.]
32	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [KUMARI et al.(2012)KUMARI, Chandra JOSHI, and Mohan TEWARI], [Butola et al.(2008)Butola, Badola et al.]
33	[Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [KUMARI et al.(2012)KUMARI, Chandra JOSHI, and Mohan TEWARI], [Butola et al.(2008)Butola, Badola et al.]
34	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Haridasan et al.(2018)Haridasan, Mao, Janarthanam, Pandey, Barik, Srivastava, Panda, Geetha, Borthakur, Datta et al.], [Negi and Maikhuri(2017)]
35	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill], [Butola et al.(2008)Butola, Badola et al.]

Continued on next page

Table 7.1 – continued from previous page

Species Number	References
36	[Butola et al.(2008)Butola, Badola et al.]
37	[Siwach et al.(2013)Siwach, Siwach, Solanki, and Gill]
38	[Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [KUMARI et al.(2012)KUMARI, Chandra JOSHI, and Mohan TEWARI]
39	
40	[Negi et al.(2018)Negi, Kewlani, Pathak, Bhatt, Bhatt, Rawal, Sundriyal, and Nandi], [KUMARI et al.(2012)KUMARI, Chandra JOSHI, and Mohan TEWARI], [Phondani et al.(2016)Phondani, Bhatt, Negi, Kothyari, Bhatt, and Maikhuri]

Bibliography

[0]

V. S. Negi, P. Kewlani, R. Pathak, D. Bhatt, I. D. Bhatt, R. S. Rawal, R. Sundriyal, S. Nandi, Criteria and indicators for promoting cultivation and conservation of medicinal and aromatic plants in western himalaya, india, Ecological indicators 93 (2018) 434–446.

M. Siwach, P. Siwach, P. Solanki, A. R. Gill, Biodiversity conservation of himalayan medicinal plants in india: A retrospective analysis for a better vision, International Journal of Biodiversity and Conservation 5 (2013) 529–540.

A. Pandey, K. Chandra Sekar, B. Joshi, R. Rawal, Threat assessment of high-value medicinal plants of cold desert areas in johar valley, kailash sacred landscape, india, Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology 153 (2019) 39–47.

P. KUMARI, G. Chandra JOSHI, L. Mohan TEWARI, Biodiversity status, distribution and use pattern of some ethno-medicinal plants, International journal of conservation science 3 (2012).

N. Sharma, C. P. Kala, Harvesting and management of medicinal and aromatic plants in the himalaya, Journal of applied research on medicinal and aromatic plants 8 (2018) 1–9.

T. Djatna, I. M. Alitu, An application of association rule mining in total productive maintenance strategy: an analysis and modelling in wooden door manufacturing industry, Procedia Manufacturing 4 (2015) 336–343.

A. Rahman, S. Das, Data mining for studentâĂŹs trends analysis using apriori algorithm, Int. J. Control Theory Appl 10 (2017) 107–115.

K. Haridasan, A. Mao, M. Janarthanam, A. Pandey, S. Barik, S. Srivastava, P. Panda, S. Geetha, S. Borthakur, B. Datta, et al., Contributions of plant taxonomy, herbarium and field germplasm bank to conservation of threatened plants: case studies from the himalayas and eastern and western ghats (2018).

V. S. Negi, R. Maikhuri, Forest resources consumption pattern in govind wildlife sanctuary, western himalaya, india, Journal of environmental planning and management 60 (2017) 1235–1252.

K. Ganaie, S. Aslam, I. Nawchoo, Development of agrotechnology for the cultivation and conservation of arnebia benthamii-a critically endangered medicinal plant of north west himalaya, Journal of American Science 6 (2010) 1133–1141.

J. S. Butola, H. K. Badola, et al., Threatened himalayan medicinal plants and their conservation in himachal pradesh, J Trop Med Plants 9 (2008) 125–142.

A. Srivastava, S. Srivastava, L. Dangwal, Specific habitat requirement and ex-situ conservation of some threatened plant species of western himalaya, 2016.

S. Barik, B. Rao, K. Haridasan, D. Adhikari, P. Singh, R. Tiwary, et al., Classifying threatened species of india using iucn criteria, Current Science 114 (2018) 588–595.

P. C. Phondani, I. D. Bhatt, V. S. Negi, B. P. Kothyari, A. Bhatt, R. K. Maikhuri, Promoting medicinal plants cultivation as a tool for biodiversity conservation and livelihood enhancement in indian himalaya, Journal of Asia-Pacific Biodiversity 9 (2016) 39–46.

S. Rawat, A. K. Jugran, I. D. Bhatt, R. S. Rawal, Hedychium spicatum: a systematic review on traditional uses, phytochemistry, pharmacology and future prospectus, Journal of Pharmacy and Pharmacology 70 (2018) 687–712.

J. S. Butola, R. K. Vashistha, A. Malik, S. Samant, Assessment of inter-population

variability in heracleum candicans wall with emphasis on seed characteristics and germination behavior, Journal of Medicinal Plants Research 4 (2010) 1523–1534.

M. Jan, S. Singh, F. Maqbool, I. A. Nawchoo, Direct shoot regeneration from hypocotyl explants of heracleum candicans wall: A vulnerable high value medicinal herb of kashmir himalaya, African Journal of Agricultural Research 13 (2018) 1419–1424.

T. Angami, R. Bhagawati, L. Touthang, T. Ronya, et al., Star anise (illicium griffithii hook. f. and thoms.): A socially important tree species from high altitude region of arunachal pradesh, Indian Forester 143 (2017) 390–391.

A. Dhyani, C. C. Baskin, B. P. Nautiyal, M. C. Nautiyal, Overcoming root dormancy and identifying the storage behaviour of lilium polyphyllum seeds, Botany 97 (2018) 161–166.

V. K. Purohit, R. Chauhan, H. C. Andola, P. Prasad, M. Nautiyal, A. Nautiyal, Nardostachys jatamansi dc. is at risk in the himalayan region, Current Science 103 (2012) 251–252.

A. PAUL, P. R. GAJUREL, A. K. DAS, Threats and conservation of paris polyphylla an endangered, highly exploited medicinal plant in the indian himalayan region, Biodiversitas Journal of Biological Diversity 16 (2015).

Appendix A

User guide for Tools used

R and R studio provide open access licence, any one can download and install it.

R Downloading and installation

Steps of downloading and installing R were shown below

1. Go to the site https://cran.r-project.org/bin/windows/base
2. Click on the link Download R 3.5.0 for Windows (62 megabytes, 32/64 bit)
3. To install R 3.5.0 click run
4. Select the path where want to save it and finish the installation.

R studio Downloading and installation

Here the steps of downloading and installing R studio were shown below

1. To download go to the site https://www.rstudio.com/products/rstudio/download
2. Choose the option RStudio desktop for Open Source License to download RStudio
3. Choose the installer according to your OS and click on it to download
4. Finish the Installation
5. Open R studio to make sure it can properly connect with R.

Installing Tools from R library

In r studio for installing each tool from the library install.packages() method is used. Installation of Spocc package install.packages('spocc') loading Spocc package library(spocc) To use apriori algorithm R provides "arules" package. To install and load "arules" use code

install.packages("arules")

library(arules)

Then apply apriori algorithm as:

rules⇐ apriori(transactional.speciesdata,parameter = list(supp = 0.1,conf=.8, minlen=3))

, where parameter value can be change by the users, by default it takes .1 as minimum support and .8 as minimum confidence value

To show the generated rules inspect(rules) is used.

To write the rules on data.csv file use write method as:

write(rules, file = "data.csv", quote=TRUE, sep = ",")

.