PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
31st March 2011. Vol. 25 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFICATION OF HERBAL GARDENS IN INDIA USING DATA MINING

[1*]NUKELLA SRINIVASA RAO, [2]SUSANTA KUMAR DAS

[1]Scientist (Computer Application in Agriculture), Directorate of Medicinal and Aromatic Plants Research, Boriavi, Anand, India, Pin- 387 310.

[2]Reader, Department of Computer Science, Berhampur University, Berhampur, India, Pin – 760 007.

ABSTRACT

There are a large number of herbal gardens in India and most of them are not linked and classified the availability of medicinal plant species of these gardens. In this study, the hierarchical clustering technique of data mining was applied on the herbal gardens of India and classified in order to discover meaningful patterns such as what type of habit of medicinal plant species is present in which location of India.

Keywords: *Herbal gardens, Medicinal plant, Clustering, Data mining*

1. INTRODUCTION

Medicinal plants are nature's wonderful gift to the mankind. The relationship between the human beings and the plants is as old as the history of the mankind itself. Since time immemorial, human beings have been utilizing plants for basic curative health care. The use of plants as a source of medicine is, in fact, based on the experience of many generations of traditional physicians and herbal practitioners of different ethnic societies, prevailing all over the world. The World Health Organization (WHO) has listed over 21,000 species (including synonyms) that have been reported for medicinal uses around the world [2]. Moreover, it estimates that more than 80 percent of the world's population relies on traditional health care. It was also reported that 25% of the drugs were derived from plant sources and many others were actually synthetic analogues of the drugs isolated from plant species in modern pharmacopoeia [10].

Many of the medicinal plant species are facing threats of extinction due to over and improper exploitation, habitat loss, degradation of land, urbanization, *etc*. for which there is an urgent need of conservation, documentation and classification. On the other hand, the increasing global demand for the medicinal plants necessitates an accelerated cultivation and conservation of them. However, before the widespread domestication of such plant species is implemented, it would be important to determine their genetic diversity so that the useful genotypes could be effectively used as cultivars by farmers or breeders and it would, in turn, facilitate the efficient conservation, management and utilization of the species [6].

To overcome these problems, a hierarchical clustering analysis on herbal gardens was carried out to classify the data and to determine the clusters best suit the data such as the habit of herb available in the different gardens and their locations.

2. REVIEW OF LITERAURE

2.1 Knowledge Driven Databases In Agriculture:

The knowledge sector of modern economies has grown extremely rapidly and the value of knowledge is now reckoned to be a major economic force. There is a need for experts, agriculturists, horticulturists and others in this field to think of ways to integrate, network and classify the available data. The knowledge acquisition process in expert system design is the most valuable asset in output accuracy. However, much of this asset is either hidden in databases as information that has not yet been tested out and made explicit or locked up in individual principals

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and employees. An emerging field: Knowledge Discovery in Databases (KDD) extends the scope of knowledge engineering research to extracting knowledge from data records collected for routine use.

2.2 Data Clustering:

Data analysis underlies many computing applications, either in a design phase or as part of their on-line operations. Data analysis procedures can be dichotomized as either exploratory or confirmatory, based on the availability of appropriate models for the data source, but a key element in both types of procedures (whether for hypothesis formation or decision-making) is the grouping, or classification of measurements based on either

(i) goodness-of-fit to a postulated model, or
(ii) Natural groupings (clustering) revealed through analysis.

Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity [7].

2.3 Clustering And Data Mining:

With so much data available, it is necessary to develop algorithms which can extract meaningful information from the vast stores. Searching for useful nuggets of information among huge amounts of data has become known as the field of data mining [5].

2.4 Data Mining:

Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns of data. Data mining is the core step in the process which results in the discovery of knowledge. Data mining is a high-level application technique used to present and analyze data for decision-makers. There is an enormous wealth of information embedded in huge databases belonging to enterprises and this has spurred tremendous interest in areas of knowledge discovery and data mining.

Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. [8].

2.5 The Role Of Botanical Gardens In Horticultural Science:

Since the first modern botanical garden was founded in Padua, Italy in 1545, botanical gardens have tremendous social and economic impact for both horticultural science and industry. There are 2479 botanical gardens in 167 countries in the world, which are responsible for living collection, conservation, and research. Moreover, many new horticultural cultivars have been bred in the botanical gardens and distributed to horticultural institutes and industries. The germplasm in botanical gardens is the essential resource for breeding work. The collections and selections from both wild (species, subspecies, varieties, and forms) and cultivated (cultivars) populations in the botanical gardens have been widely utilized by horticultural professionals. Beijing Botanical Garden (BBG), for example, has collected 61 ornamental peach cultivars, which can be grouped into six groups and 12 forms.

Both morphological and molecular features have been applied for studying genetic relationships among ornamental peach cultivars. Two early blooming cultivars and one variegated leaf cultivar were bred and selected by BBG. They have been widely cultivated in Chinese gardens. BBG has been celebrating the Ornamental Peach Festival since 1989. There are more than 5000 ornamental peach trees cultivated and this offers an opportunity to educate the public about plants and their horticultural applications. Obviously, botanical gardens are not only the resources and institutes for horticultural field, but also the promoters of horticultural science to the public through plant collections and educational programs [4].

2.6 Medicinal Plant Resources *ex situ*:

There have been a few efforts to collect and conserve medicinal plant species. Herbal gardens are one of the main repositories of medicinal plants and good examples are set by world renowned gardens at Kew, New York Botanic Garden, Missouri Botanic Garden, *etc*. Most of them are over 100 years old and there are other newly established botanic gardens which focus on medicinal plant maintenance and conservation in various countries like Thailand

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and India. Due to constraint of space very few plants of a any given species are cultivated in such gardens either on ground or in pots. The objective of such a collection is to establish species diversity. The genetic diversity of the useful species needs to be well studied to select superior plants for sustainable conservation, or cultivation and use. Most of the basic research on medicinal plants in Asian developing countries is carried out in universities and in some specific institutes [9].

3. MATERIALS AND METHODS

3.1 Location Of Study:

The study was conducted at Agricultural Research Information Systems Cell (ARIS cell) & Computer Cell of Directorate of Medicinal and Aromatic Plants Research (DMAPR), Boriavi, Anand, Gujarat, India.

3.2 Data Collection:

The different herbal gardens available in India were identified. The information on different medicinal plant gardens existing in India were collected, maintained and compiled through the help of different state forest departments. The information thus compiled was utilized for the data analysis in this study.

3.3 Data Mining:

The data mining technique employed here enables decomposition of the garden system into data cohesive subsystems (sub classification based on location, type of habit *etc.*) mining the different types of by detecting associations between programs sharing the same files. Clustering analysis used here is useful for Similarity/Dissimilarity analysis. It helps in the analysis of what data points in a given dataset are close to each other. In our case, mutually exclusive groups of gardens based on habit, location *etc.* is created according to their similarities.

A hierarchical cluster analysis has been performed using SPSS software. Hierarchical clustering allows users to select a definition of distance, then select a linking method for forming clusters, then determine how many clusters best suit the data. Hierarchical clustering generates representation of clusters in icicle plots and dendograms. Here a hierarchical cluster analysis was carried out between the habit of herb available in the different gardens and their location.

The following steps are carried out in the hierarchical cluster analysis using SPSS:

1. *In SPSS* the hierarchical clustering is called as the Cluster procedure. The desired variables (i.e. the habit type and the location) are moved to the variable list box.
2. The agglomeration schedule and the proximity matrix were created and the maximum number of clusters required was also set.
3. The similarity/distance measures were selected in the Measure area of the Method subdialog obtained by pressing the Method button in the Classify dialog. The *proximity matrix* table in the output shows the actual distances or similarities computed for any pair of cases.
4. The cluster (linkage) method and the distance measure to be used were selected. The distance measure choices will depend on the level of measurement specified: interval, count, or binary.
5. The output was obtained and is discussed in the results section.

4. RESULTS

To classify the herbal gardens information, the data was collected from different herbal gardens, developed a database and uploaded online at www.herbalgardenindia.org. This website at present indicates the presence of a total of 71 herbal gardens from all over the country that are registered as members in this network. A total of 1024 species are available in these herbal gardens from which majority of the species present in the garden are herbs of plant type. It is noted that the total number of quantity of planting material available in the entire registered herbal garden is 9,06,426 cuttings.

From the above large quantity of datasets, the data mining experiments was carried out to discover meaningful patterns and rules. Initiative behind this experiment was the question that what type of habit of species is present in which location.

We modeled these questions as to find the relationship between location of species and their habit.

4.1 Agglomeration Measures:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The agglomeration schedule was created using SPSS are given in Table 1 shows the following features. The rows are stages of clustering, numbered from 1 to (n-1), since there are 13 states in our analysis the row number stands at 12. The 12th stage includes all the cases in one cluster. There are two "Cluster Combined" columns which give the cluster numbers for combination at each stage. Here the clustering has been carried out using a distance measure like Euclidean distance, stage 1 combines the two cases which have lowest proximity score.

The proximity agglomeration coefficient in the "Coefficients" column is an indicator of how far the agglomeration algorithm has to reach to combine an existing cluster with the next closest cluster. It can be seen above that there is a large jump in the values all through the cluster analysis. A large agglomeration coefficient will correspond with a long distance in the dendogram (Figure 2).

4.2 Icicle Plot Measures:

The icicle plot measure for the data mining experiment carried out has been illustrated in the figure (Figure 1). Icicle plots are usually horizontal, showing cases as rows and number of clusters in the solution as columns. A vertical icicle plot has been carried out where the variable values have been read from the last row bottom to top. This is a clear indication of how agglomeration proceeded. This is a visual way of representing information on the agglomeration schedule, but without the proximity coefficient information.

In the figure (Figure 1), from hierarchical cluster analysis,
12- Uttarkhand and Meghalaya are present in one cluster, all others in their own cluster.
10- Maharashtra and Karnataka are present in one cluster, Uttarkhand and Meghalaya are present in one cluster all others in their own cluster
8- Orissa falls with Maharashtra and Karnataka in one cluster, Uttarkhand and Meghalaya are present in one cluster ,Madhya Pradesh and Kerala are present in one cluster all others in their own cluster
4- Uttaranchal joins Madhya Pradesh and Kerala in a cluster
2- Himachal Pradesh joins the Uttar Pradesh and Uttaranchal Cluster.
1- Rajasthan is in a cluster of its own.

4.3 Dendrogram Using Average Linkage Between Groups

The Dendrogram (Figure 2) represents the relative size of the proximity coefficients at which clusters were combined. The bigger the distance coefficient the more clustering involved combining unlike entities, which is undesirable. Those showing low distance are close, with a line linking them a short distance from the left of the dendogram, indicating that they are agglomerated into a cluster at a low distance coefficient, indicating alikeness. It can be observed that Meghalaya, Uttarkhand, Assam, Karnataka and Orissa form one of the first clusters.

This methodology indicates that interlinking between species of different locations can be modelled through this data mining technique and answers may be obtained to enable making the database which would be more easily accessible. The strength of this method of data mining lies in clustering the evidence scattered in the data and hidden from the naked human intelligence

5. DISCUSSION

Agricultural and biological research studies have used various techniques of data analysis including, natural trees, statistical machine learning and other analysis methods" [3]. Therefore, incorporation of a data mining technique to the database system on herbal gardens developed has potentials to become an important source on cultivated medicinal species all over the country.

The 'mined' information is typically represented as a model of the semantic structure of the dataset, where the model may be used on new data for prediction or classification and this technology is on rise in the fields of agriculture and related research [1]. The benefit of this data mining method in herbal garden database is that there is a considerable ease in the access of required information with regard the different search modules developed.

6. CONCLUSIONS

The advent of new computing technologies has brought about an increase in the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

demand for novel methods of collecting information and data especially in the field of agriculture. This method of information on medicinal plant species can be a stepping stone to ensure that knowledge of therapeutic plants which has started to decline and become obsolete due to the lack of recognition by younger generations will be passed on due to involvement of techniques like data mining. Due to important uses of medicinal plants, it is essentially needed to retrieve the valuable information knowledge with the expectation of developing the medicinal plants sector. Particular to this study on herbal gardens, the mining techniques involved makes access of information to the user much easier and are of extremely important.

REFRENCES:

[1]. Abdullah, A., S. Brobst, and M. Umer. 2004. The Case study for an Agri Data Warehouse: Enabling Analytical Exploration Integrated Agricultural Data. Proceeding of the IASTED International Conference on Databases and Applications, Innsbruck, Austria.

[2]. Chandel, K.P.S., G. Shukla and N. Sharma. 1996. Biodiversity in Medicinal and Aromatic Plants in India: Conservation and Utilization, 239.

[3]. Cunningham, S. J., and G. Holmes. 1999. Developing innovative applications in agriculture using data mining. Proceeding of Southeast Asia Regional Computer Confederation Conference.

[4]. Dongyan, H., and Z. Zuoshuang. 2008. The role of botanical gardens in horticultural science. Acta Hort. (ISHS), 769:493-496.

[5]. Everitt, B. S., 1993. Cluster Analysis. Edward Arnold Ltd., London, UK.

[6]. Gupta, V.K., 2000. An approach for establishing Traditional Knowledge digital library. J Intellectual Property Rights, 5(6): 307-319.

[7]. Lee, R.C.T., 1981. Cluster analysis and its applications. In Advances in Information Systems Science, J. T. Tou, Ed. Plenum Press, New York, NY.

[8]. Michalski, R.S. and K.A. Kaufman. 1998. Data Mining and Knowledge Discovery: A Review of Issues and a Multistrategy Approach, in Michalski, R.S., Bratko, I. and Kubat, M. (Eds.), Machine Learning and Data Mining: Methods and Applications, John Wiley & Sons London, pp. 71-112.

[9]. Natesh, S., 2000. Biotechnology in the conservation of medicinal and aromatic plants: Biotechnology in horticultural and plantation crops. Malhotra publishing house, new Delhi , Pp 548-561

[10]. Rao V.R. and R.K. Arora. 2004. Rationale for conservation of medicinal plants, Medicinal Plant Research in Asia- volume I: The framework and the project. pp 7-22.

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
31st March 2011. Vol. 25 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Average Linkage (Between Groups)

Agglomeration Schedule						
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	10	15	145158.000	0	0	2
2	1	10	810408.000	0	1	4
3	6	9	1.451E7	0	0	4
4	1	6	2.874E7	2	3	6
5	7	8	9.123E7	0	0	7
6	1	11	1.668E8	4	0	10
7	3	7	5.038E8	0	5	10
8	16	17	7.089E8	0	0	11
9	5	13	8.404E8	0	0	12
10	1	3	2.585E9	6	7	11
11	1	16	7.192E9	10	8	12
12	1	5	2.501E10	11	9	0

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
31st March 2011. Vol. 25 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

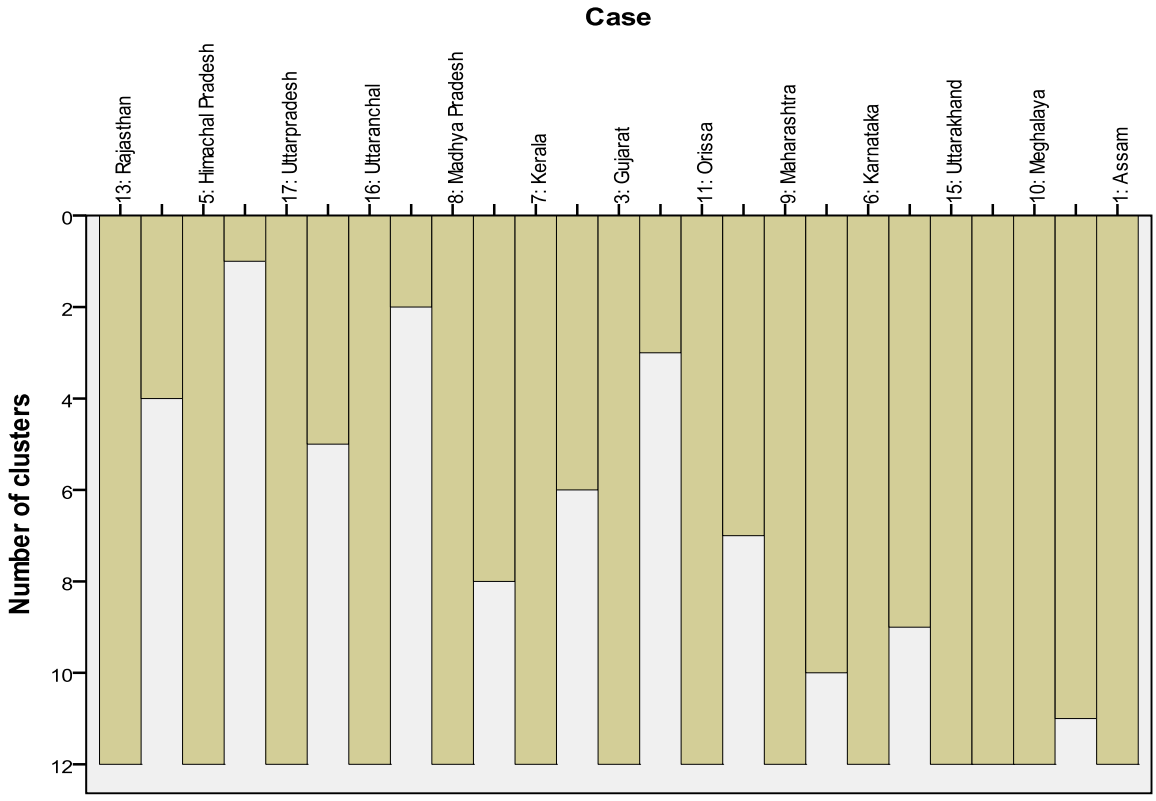

Figure 1. Icicle plot in Data Mining Analysis

PUBLICATION OF Little Lion Scientific R&D, Islamabad PAKISTAN
Journal of Theoretical and Applied Information Technology
31st March 2011. Vol. 25 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

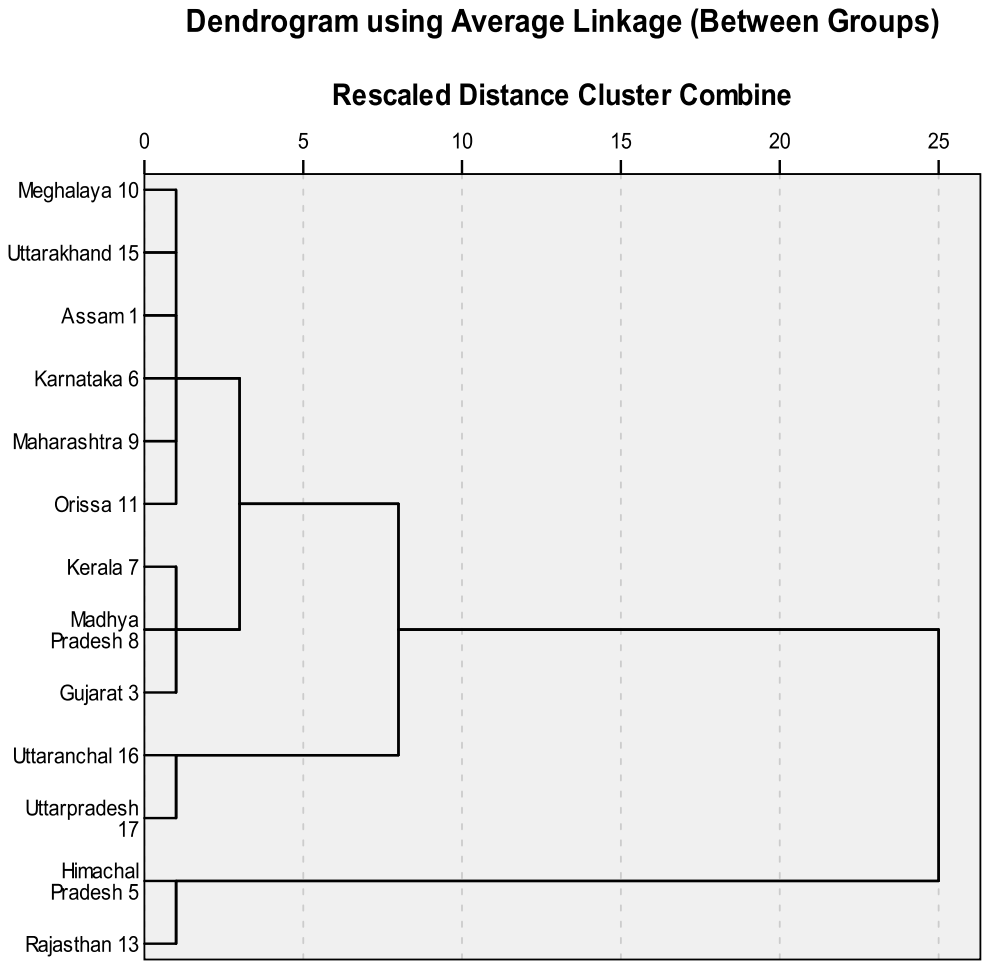

Figure 2. Dendrogram using Average Linkage