A Novel Datamining Approach to Determine the Vanished Agricultural Land in Tamilnadu

S.Megala
Research Scholars in Computer Science,
Karpagam University, Coimbatore.

Dr M.Hemalatha
Head, Department of Software Systems,
Karpagam University, Coimbatore.

ABSTRACT

The presence of wide heterogeneity in the investigational material that is often used in agricultural research, led to the application of data mining tools and as a result many refinements and newer developments in statistics followed. Data mining, in fact, provides scientific tools for representative data collection, appropriate analysis and summarization of data and inferential procedures for drawing conclusions in the face of uncertainty. There is a need to provide remunerative prices for farmers in order to maintain food security and increase income of framers. Farmer finds himself thriftily poor and the most of the grains of rich agriculture have been appropriated by other section of the community. The difference between the engineering industry profit and agriculture sector produce profit tend them to leap over to the other sector. For this they need a capital investment and they find better way to make money is selling away the valuable and cultivable land to non agriculture purpose. So the food producing land is lost to the non-food producing sector, year by year the population of cultivable farmer and crop cultivable land is diminishing in a large chunk. By using clustering techniques this paper examines the current usage and details of agriculture land vanished in the past seven years.

Keywords

Data mining, Clustering, Agriculture, Food Security

1. INTRODUCTION

Data mining software relevance, using diverse methodologies, have been residential by both viable and research centre. These techniques have been used for engineering, commercial and scientific purposes. For example, data mining has been used to analyze large data sets and establish useful classification and patterns in the data sets. “Agriculture and biological research studies have been used various techniques of data analysis including, nature tree, statistical machine learning and other analysis methods”(Cunningham and Holmes,1999)[1]. Data Mining or "the efficient discovery of valuable, on-obvious information from a large collection of data"[2] has a goal to discover knowledge out of data and present it in a form that is easily comprehensible to human.

This research determined whether data mining techniques can also be used to improve pattern recognition and analysis of large land datasets. The research aimed to establish if data mining techniques can be used to assist in the clustering methods by determining whether meaningful patterns exist across various land profiles at various research site across Tamil Nadu in India.

Clustering analysis is a kind of statistical analysis and used to identify clusters embedded in the data, where a cluster is a collection of data objects that are ‘similar’ to one another [3]. It can be expressed by distinct functions, specified by users or experts. A superior clustering process produces high eminence clusters to make certain that the inter-cluster resemblance is low and the intra-cluster similarity is high[4]. For example, one may cluster the parcels according to their land use, cover, agriculture type, ownership and geographical locations. The remainder of the paper is organized as follows. Section 2 gives an overview of agriculture research using mining techniques. Section 3 presents the clustering Analysis .Section 4 present the refinement Experiment and Result . Section 5 Discuss some Further analysis and concludes the present work.

2. MINING TECHNIQUES IN AGRICULTURE

Agriculture plays a very important task in the profitable development of a country. Its is fundamentally different from an industry. The marketing of farm product generally tends to be a complex process. The type of agriculture commodities produced in our country are various and varied. Agriculture price plays an important role in achieving growth and equity in Indian economy[5]. The price is one of the important instruments in achieving food security by improving production, employment and incomes of the farmers. There is a need to provide remunerative price for farmer in order to maintain food security and increase income of farmers[6]. It has come under serious attack in recent period on the ground of higher support prices than the cost of production warrant and supposed distortion of the market leading to food deprivation.

The major challenges are to transfer income to farmers whose farm size is not adequate to provide them income above poverty line or provide good living standard[7]. The rural sector provides primary commodities for use in the urban sector for processing or direct consumption. But expenditure of urban sector does not reach rural sector as commodities needed are not produced in the rural areas. Rural economy remains as small centers of producing primary agriculture commodities that hardly provide good living standard.

The resources flow is largely from the government investment in rural areas. But due to corruption and other leakages not much

reaches in rural areas. The gross capital formation in farm sectors is merely 1.3 percent of GDP. There are very meager alternate employment opportunities in rural area on which the marginal and small farmer can fall back. In the absence of this, their economic condition hardly improves[8]. It due to economic desperation and financial crisis farmer sell the land, he is reduced to land-less labor or make exit to urban area in search of new job. This is the scenario of the farm sector and the marginal and small farmer in particular[9]. So cultivable land is the source of food both for human being and animals. If the amount of land is insufficient or the land available for cultivation is not properly managed in that case we will not get a sufficient food and it will be end in food demand.

The overall aim of the research was to determine the land utilization for agriculture and non-agriculture areas for the past ten year. The study is monitored in Six area in Tamil Nadu. This research adopted action research methodology, where improvement and changes may have to be undertaken to provide the outcome that meets their required specification for the development. The research used Software to conduct qualitative analyses and to create a benchmark for the analysis of the dataset. The dataset was then analyzed using a clustering process within the data mining software. The process of data mining in the modern agriculture decision is shown as below.

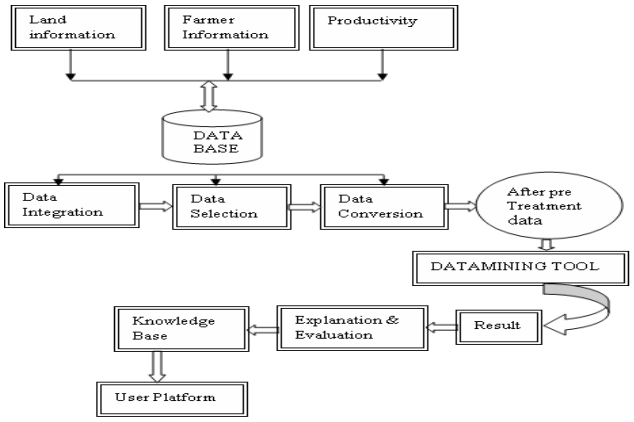

Figure 1: Procedure of Data mining

A. Data Source:

The data source that the data excavation needs come from the database of the logistics information system, the main data source has two mostly. They are cultivable land information, uncultivable land information, total cropped area.

B. The Data Preparation:

The data preparation is important function to data mining which reprocess to the agriculture logistics system data, check the integrity and consistency of data, process to the inaccuracy and worthless data. The method of data organize includes data integration, the data choice, and the data conversion. The data integration mainly withdraws and integrates data from several differences operability databases, documents, or remnants system. The data choice means choosing data from the database, and recognizing data set need to be analyzed to narrow the scope of data mining, avoid blind search, and exalt data mining speed and quality. The data conversion is scouring treatment to the selected data before mining and guiding mining direction interactive mode by experts inputting interesting knowledge.

C. The Choice of Data Mining Tool:

The hidden novel mode withdraw from the data set which through iterating and searching again and again with the appropriate data mining technology and algorithm based on data mining task determined which is the knowledge that the customer need. Clustering are used in classification problem to build the model using training data and classify the land lost stages from agricultural data sets and compared with the algorithms.

D. The Explanation and Evaluation of the Knowledge:

The data mining gain may be not satisfied to the users, since the different data mining process get also different result. Carry on analysis to the information withdraw according to the resolution purpose of the end customer and distinction the most worthy information, finally hand over decision maker through the decision support system. The process of explanation and evaluation is to percolation processing of information, which is to decide whether to deposit gained rules to knowledge base. If it can't produce meaning to the decision of users, that the data mining process repeated.

E. E.The Customer Interface and Knowledge Base:

The usage of visualization technology and choice of suitable visualization tools make users confirm the reliability of discovered knowledge. The knowledge in the knowledge base of the saving related realm and the summary detective knowledge through the data mining method and the other science method, which are likely to be varied on the manifestation, such as chart or the rule representation, provide a strong decision support for the decision maker. The decision makers can understand the knowledge by visualization interface and do a decision.

3. CLUSTERING ANALYSIS

Clustering analysis is the searching for group (clusters) in the data, in such a way that samples belonging to the same cluster resemble each other, whereas sampling in different cluster are dissimilar. Generally speaking, there are two categories of

methods for clustering(Wulan 2002;Zhang Guojiang, 2002;Theodore P.Beauchaine,2002)[10].

3.1 Based on Partitioning Algorithms

A partitioning algorithm describes a method that divides the data set into k clusters, where the integer k needs to be specified by the user[11]. Typically, the user runs the algorithms for a range of k-values. Algorithms of this type include k-means, partition around medoids, fuzzy clustering etc. k-mean algorithm is one of the partitioning algorithms. Among all partition algorithms, k-means algorithm is most widely used[12]. It is applied in distinguishing relative homogeneous sample set. the land data are geographical features and they have closely characterized. The k-mean algorithm is applied as (a) it partitions N data points into K disjoint subsets(cluster)R_i, containing N_j data points so as to minimize the sum of squares criterion. Where X_j is vector representing the j^{th} point W_i is the geometric centroid of the data point in R_i.

$$J = \sum_{i=1}^{k} \sum_{\substack{X \\ j \in Ri}} \left\| \bar{X}_j - \bar{W}_i \right\|^2$$

This algorithm initially takes the number of components of the agriculture land used equal to the final required the total number of the land equal to the final number of clusters. In this stride itself the final essential number of clusters is select such that the points are reciprocally farthest apart. Next, it examines each component in the Land and assigns it to one of the clusters depending on the minimum distance. The geometric centroid's position is recalculated every time a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.

Table 1: Closing Cluster Centers

	Cluster		
	1	2	3
Agree	.9652	.4673	.8737
Percentage			

Table 2: Distances among Closing Cluster Centers

Cluster	1	2	3
1		.919	.524
2	.919		.492
3	.524	.524	

Cluster	1	36
	2	51
	3	51
Valid		139

Table 3: No of Cases in Each Cluster

Machine language is used to employ here for the analysis. Table 1 shows the centers after clustering process. The values of centers represent the means of the samples in every class. According to the values of centers, we can easily find that cluster 1, 3, 2 correspond to high, medium, low requirement degree respectively Table 1 illustrate the Euclidean distance between cluster centers. The distance between cluster 1 and cluster 2 is bigger than the one between cluster 1 and cluster 3, because cluster 1 and cluster 2 represent high and low requirement degree respectively. And from the semantic view, 'high' and 'low' have a big difference. Table 4 tells the number of cases (samples) in each final cluster.

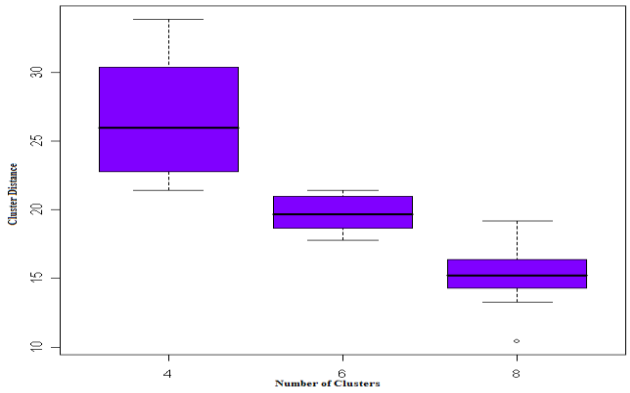

Figure 2: Categorization the Cluster Detachment

We need further analysis to see whether cluster number 3 is suitable. If not, we would have to cluster again. Through analyzing distances of cases from its classification cluster centre, we can tell whether the cluster number 3 is suitable. If many cases were seriously out of centre, then the cluster number is not suitable. Figure2 is a box diagram: the black bold line in the centre represents the mean; the rectangle box is the bound of interquarti1e range. We can find that all cases are not out of centre seriously. To some degree, this indicates that cluster numbe3 is suitable

3.2 Based on Hierarchical Algorithms*:*

The basic process of hierarchical clustering:

Step 1: Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) among the clusters equal the distances (similarities) among the bits and pieces they contain.

Step 2: Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.

Step 3: Compute distances (similarities) between the new cluster and each of the old clusters. Repeat steps 2 and 3 until all items are clustered into a single cluster can be done in different ways, which is what distinguishes average-link clustering from single-link and complete-link[13]. Stuck between the clusters equal the distances (similarities) between the items they contain.

The following is the details of average-link, single-link and complete-link clustering.

$$d(R,Q) = \frac{1}{\left|R\right\|Q\right|} \sum_{i \in R,\, j \in ()} d(i.j)$$

We adopt average-link clustering and get the dendrogram, see Figure3. From it, the land information of estimated area for cultivation is show above and we can easily find that if the clustering number is 5, then there would be one single sample (alley) in a class. This is complicated to deal with. So clustering number should no more 5. If it were 2, then there would be 2 types of semantics: need and not need, which is also unreasonable. So clustering number should be 3

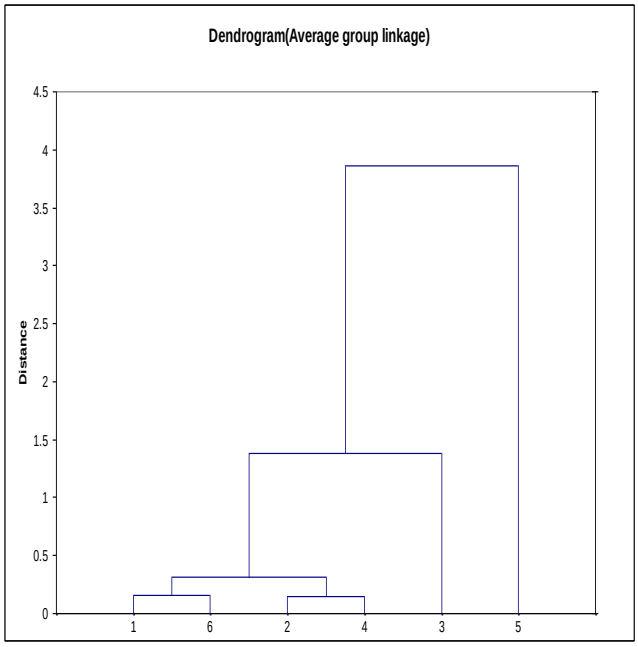

Figure 3: Dendrogram by Average Linkage among Land group

4. RESEARCH AND OUTCOME

The collection of information and data has increased with the advent of new computing technology, but establishing patterns within this data has become more difficult and requires new approaches and tools if it is to be undertaken. The advent of this problem has provided an opportunity from which data analysis has started to take over from current methods. Furthermore, this technology has reduced the time taken to undertake data analysis and has increased automation of the process. The integrity of the data is critical to ensure that results are not affected by outliers and null values in the data set, or other adverse factors.

The establishment of clusters in the data required a large amount effort by the researchers when using current methods. Furthermore, the current methods still required some post Excel analysis because the platform was limited in the interpretation of the graphs generated. The use of land classification maps have been shown in order to understand the land lost with in the past seven years. This has allowed a greater understanding of biophysical and environmental management. Figure 4 shows the classification of land used details of six district (Coimbatore, Erode, Dindigul, Salem&Namakkal, Dharmapuri) around Tamil Nadu in India.

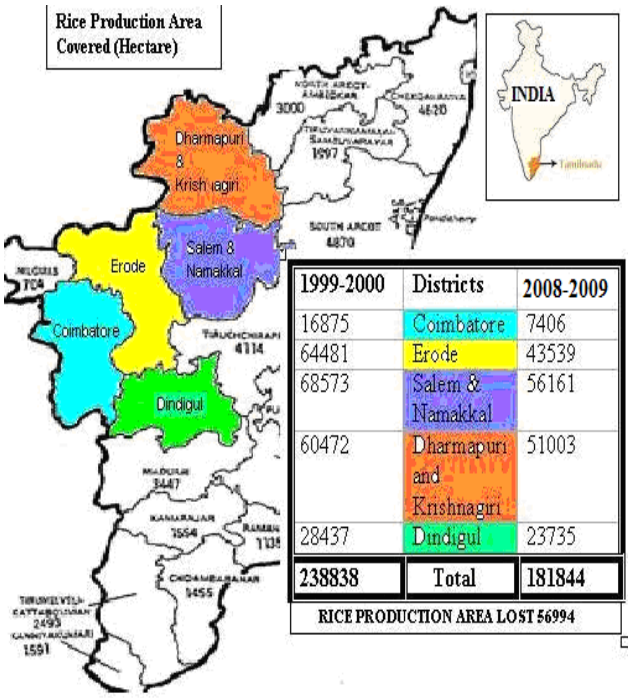

Figure 4 : Rice Production Area.

Table 4: Rice Production in Tonnes

Rice Production In Tonnes		
1999-00	District	2005-06
60494	**Coimbatore**	**20665**
270611	**Erode**	**179564**
186402	**Salem,**	**137721**
105619	**Namakkal**	**68849**
201243	**Dharmapuri& Krishnagiri**	**85390**
		57797
118998	**Dindigul**	**75304**
943367	Total	**625290**
7 Year Production loss = 318077 Tonnes		

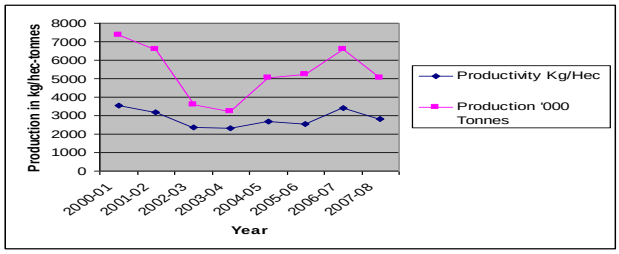

Figure 6: Analysis of Land & Productivity Details

In the above Table and Figure 6 shows the impact of the food gain cultivatable land loss on food production which ultimately reflect heavily on food grain supply and demand.

5. CONCLUSION AND FUTURE WORK

The analysis of these agriculture data sets with various data mining techniques may yield useful to research in the agriculture. It is envisaged that the information gained from this research will contribute to the improvement of land and the usages of food security and Zero hunger programs. Due to land loss, the farmers who were the land owners became land laborers. They tend to Migrate to other industries and sectors which is more remunerative to them than the agricultural labor. My research leads to mitigate the predicament of the farmers and find the ways and means retain them in the agricultural sector itself.

6. REFERENCES

[1] Cunningham, S.J. and Holmes, G., 1999.The Proceedings of the Southeast Asia regional computer confederation conference.

[2] Agrawal, R., Imielinski, T., and Swami, A., 1993. "Mining association rules between sets of items in large databases." Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, D.C., 207-216.

[3] Wang, K., Xu, C., & Liu, B., 1999. Clustering transactions using large items. International Conference on Information and Knowledge Managmen, CIKM'99 Kansas city, Missouri United States, 483–490.,

[4] Guha, S., Mishra, N., Motwani, R., & O'Callaghan, L.,2000. Clustering data streams Symposium on Foundation of Computing Science, 359-366.

[5] N.Sakthivel, Dr.A.Selvaraj .Farmer Perception Towards Regulated Markets. A Case Analysis, Financing Agriculture-Anature Journal of Agriculture &Rural Development, January-February-2009,pg:6-12.

[6] Chadha G. K. et al., 2004. Land Resources, State of the Indian Farmer: A Millennium Study, Ministry of Agriculture, Department of Agriculture and Cooperation, New Delhi.

[7] A. Abdullah & A. Hussain, , 2006. "Data Mining a New Pilot Agriculture Extension Data Warehouse", Journal of

Research and Practice in Information Technology, vol. 38, no. 3, page. 229-249.

[8] Shiva, Vandana, Jalees, Kunwar.'Farmers Suicides in India'Research Foundation for Science, Technology and Ecology, New Delhi, India.

[9] Gupta, R.P. and S.K. Tewari., 1985. Factors Affecting Crop Diversification: An Empirical Analysis, Indian Journal of Agril. Economics, Vol. XL, No. 3, July-Sept, pp. 304-305.

[10] Gonzalez, T.F. 1985. Clustering to minimize the maximum inter cluster distance Theoretical Computer Science, 38, 293-306.

[11] Hartigan, J. and Wong, M., 1979. Algorithm AS136: A k-means clustering algorithm. Applied Statistics, 28, 100-108.

[12] K. Alsabti, S. Ranka, and V. Singh, Mar. 1998. "An Efficient k- meansClustering Algorithm," Proc. First Workshop High Performance Data Mining,.

[13] 13)Day, W. and Edelsbrunner, H., 1984. Efficient Algorithms For Agglomerative Hierarchical Clustering Methods. Journal of Classification, 1, 7, 7-24.

[14] Mark A. Friedl, Carla E. Brodley, and Alan H. Strahler. 1999. Maximizing Land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales. IEEE Transactions on Geoscience and Remote Sensing, 37:969–977.

[15] Anil K. Jain and Richard C. Dubes., 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ-07632,

BIBLIOGRAPHY

S.Megala received her MCA, M.Phil., from Bharathiar University, Coimbatore. Currently she is pursuing her Ph.D Full- time in Karpagam University. Her research interest in the areas of Data Mining and Knowledge Discovery, Pattern Recognition and Machine Learning.

Dr.M.Hemaltha completed MCA MPhil., PhD in Computer Science and Currently working as a Asst . Professor and Head , Dept of Software systems in Karpagam University. Ten years of Experience in teaching and published Twenty seven papers in International Journals and also presented seventy papers in various National conferences and one international conference. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. Also reviewer in several National and International journals.