APPLICATION OF MACHINE LEARNING IN OPTIMIZING SOCIO-ECONOMIC PLANNING

Vijay Maurya¹, Rashmi Ranjan Rout², Anuj Verma³

¹Assistant Prof., Electronics Engineering Department, Institute of Engineering & Technology, U.P., India
²M.Tech. Scholar, Systems Engineering Department, IIT BHU Varanasi, U.P., India
³Assistant Prof., Electrical Engineering Department, International Maritime Institute, U.P., India

ABSTRACT

Effective use of land is a very crucial parameter in laying foundation of development process of a state or a country. Government Agencies want effective use of land to ensure rapid growth. State Government conducts surveys on Land Utilization Patterns of each district every year. The important attributes of Land Utilization are total geographic area, forest area, miscellaneous trees and grooves, permanent pastures, cultivable waste, land put to non Agricultural use, barren and uncultivable land, current fallows, other fallows, net area sown, total irrigated land etc. Based on these attributes the different districts can be clustered in to desire number of clusters. Again determining districts having closer values of the given demographic attributes with each district is also important. Clustering of districts can be done precisely by K-means clustering approach and K-Nearest Neighborhood approach can be implemented to find districts having very closer data.


1. INTRODUCTION

1.1 Data Mining in Brief

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to cluster data points having multiple dimensions. Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics. The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows, have allowed the development of new techniques based on a brute-force exploration of possible solution. This paper uses two techniques: one for clustering (K-means Clustering) and the other for possible replacements of each data point (KNN classification).

1.2 K-means Clustering:

Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in data sets. It is a process of grouping data objects into disjointed clusters so that the data in the same cluster are similar, yet data belonging to different cluster differ. The demand for organizing the sharp increasing data and learning valuable information from data, makes clustering techniques widely popular in many application areas such as artificial intelligence, biology, customer relationship management, data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, pattern recognition, psychology, statistics and so on.

1.3 K-means Clustering:

KNN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The KNN is the fundamental and simplest classification technique when there is little or no prior knowledge about the distribution of the data. This rule simply retains the entire training set
during learning and assigns to each query a class represented by the majority label of its k-nearest neighbours in the training set. The Nearest Neighbour rule (NN) is the simplest form of KNN when K = 1.

2. K-MEANS CLUSTERING ALGORITHM
Step 1: Given m objects of n dimensions, initialize k cluster centres.
Step 2: Assign each object to its closest cluster centre.
Step 3: Update the centre of each cluster i.e. calculates the mean value of objects in each cluster.
Step 4: Repeat Step 2 and Step 3 until no change in each cluster centre.

Programming Approach
Step 1: Load data matrix Z having dimension m-by-n
Step 2: Find k cluster centres using MATLAB inbuilt function [IDX, C] = kmeans (Z, k); Where C is a matrix of dimension k-by-n and contains k cluster centres
Step 3: Find Euclidean distance of each row in C from each row in store the distances in matrix X, which is a k-by-m matrix
Step 4: Find the minimum element in each column of matrix X Display the index number of the minimum element

3. KNN ALGORITHM
Step 1: Load data matrix Z of dimension m-by-n
Step 2: Calculate Euclidean distance of each data node from all other data nodes
\[ d = \sqrt{\sum_{i=1}^{k} (X_i - Y_j)^2} \]
Here k = n
Step 3: Construct distance matrix X of dimension m-by-m taking the Euclidean distance of each node from other node.
Step 4: Specify the value of K
Step 5: Find K smallest value for each row other than the diagonal elements of X

4. EXPERIMENTAL RESULTS
This paper uses the land utilization pattern of the 30 districts of the state Odisha for the financial year 2010-2011 having 11 attributes.
area in ‘000 hect.

Fig-1 Snapshot of data set of 30 districts having 11 attributes

This data is further divided into 4 datasheets:
1. totaldata.xlsx
2. agrobased.xlsx
3. industries.xlsx
4. agriculture.xlsx

The sole motive of splitting of data is to meet different objectives.
1. When totaldata.xlsx is chosen all the districts are clustered into k clusters based on all the 11 attributes.
2. When agrobased.xlsx is chosen all the districts are clustered based on 4 attributes i.e. geographical area, forest cover, miscellaneous trees and grooves and permanent pastures.
3. When agriculture.xlsx is chosen clustering is done based on 5 attributes i.e. cultivable waste, current fallow, other fallows, net area sown and total irrigated area.
4. When industries.xlsx is chosen clustering is done based on 5 attributes i.e. cultivable waste, current fallow, other fallows, permanent pastures and barren lands & uncultivable land.
4.1. Experiment 1:
If for clustering the file totaldata.xlsx is chosen and number of clusters is given to be 7. Then the 30 districts will be clustered into Cluster Number - 1 to Cluster Number- 7 based on all 11 attributes.
Clustered districts of state ODISHA into 7 clusters:

Fig-2 the normalized data
4.2. Experiment 2:
The second experiment is to search for k nearest neighbours or k replacements for any data node. If for KNN, the file industries.xlsx is chosen and given value of k is 6. We seek 6 nearest neighbours of node 7 (Jajpur) then the following is the output of the program.
5. CONCLUSION

From the output it is clear that for setting up a manufacturing Industry based on land utilization pattern for JAJPUR: DEOGARH, NAYAGARH, SAMBALPUR, PHULBANI, GANJAM and NABARANGPUR are the 6 nearest neighbours in terms of the given set of attributes. The closest neighbour of JAJPUR can be obtained by setting K=1, which is found to be DEOGARH.

6. REFERENCES

[1]. Directorate of Agriculture, Govt. of Odisha-(www.agriodisha.nic.in)
[4]. Shi Na, Guan Yong and Liu Xumin, “Research on K-means clustering algorithm”
[5]. O. Beucher and M. Weeks, “Introduction to matlab and simulink-a project approach”
[6]. Amos Gilat, “MATLAB: An introduction with applications.”