APPLICATION OF MACHINE LEARNING IN OPTIMIZING SOCIO-ECONOMIC PLANNING

Vijay Maurya[1], Rashmi Ranjan Rout[2], Anuj Verma [3]

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

Keyword: - *K-means clustering, Machine Learning K-Nearest Neighbor, land utilization pattern*

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

Step1. Given m objects of n dimensions, initialize k cluster centres.
Step2. Assign each object to its closest cluster centre.
Step3. Update the centre of each cluster i.e. calculates the mean value of objects in each cluster.
Step4. Repeat Step2 and Step3 until no change in each cluster centre.

Programming Approach

Step1. Load data matrix Z having dimension m-by-n
Step2. 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
Step3. 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
Step4. 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

$$d = \sqrt{\sum_{i=1}^{k} (Xi - Yi)^2}$$ 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.

Districts	geographical area	Forest area	Misc trees &grooves	Permnt pastures	cultivable waste	Land under nonagricultural use	Barren Land	Current Fallow	Other fallows	net area sown	Net area irrigated
Balasore	381	33	25	16	9	33	10	34	5	216	102.19
Bhadrak	250	10	3	11	11	33	1	16	5	160	114.5
Balangir	657	154	4	46	18	53	23	6	13	340	62.6
Sonepur	234	41	3	13	8	22	12	5	7	123	58.91
Cuttack	393	79	11	11	10	83	10	44	1	144	98.5
Jagatsingpur	167	13	4	7	6	13	13	13	7	91	63.12
Jajpur	290	72	4	4	4	51	5	5	5	140	62.38
Kendrapara	264	25	5	8	6	49	5	11	14	141	74.19
Dhenkanal	445	174	6	8	4	42	5	43	20	143	59.64
Angul	638	272	23	36	19	48	7	41	17	175	39.77
Ganjam	821	315	22	20	11	21	20	45	6	361	221.58
Gajapati	433	247	8	12	4	12	68	1	6	75	26.72
Kalahandi	792	254	8	23	21	35	57	39	16	339	124.29
Nawapara	385	185	1	2	2	3	2	22	1	167	46.59
Keonjhar	830	310	6	20	26	77	93	35	0	163	48.12
Koraput	881	188	17	45	44	54	210	36	19	268	78.2
Malkangiri	579	335	1	21	4	23	38	2	15	140	45.18
Nabarangpur	529	246	13	8	15	44	9	1	8	185	36.82
Rayagada	707	281	18	26	22	124	38	36	5	157	49
Mayurbhanj	1042	439	41	28	10	58	16	98	13	339	108.31
Phulbani	802	571	34	10	14	9	30	19	6	109	21.95
Boudh	310	128	19	17	20	21	12	8	4	81	51.48
Puri	348	14	9	9	3	115	8	54	1	135	102.86
Khordha	281	62	10	5	8	46	15	24	6	105	53.26
Nayagarh	389	208	6	4	5	25	6	11	1	123	43.48
Sambalpur	666	363	4	13	19	38	18	13	17	181	62.51
Bargarh	584	122	5	20	15	47	20	51	6	298	132.14
Deogarh	294	156	1	5	6	51	6	3	2	64	18.38
Jharsuguda	208	20	6	20	15	39	17	19	3	69	12.31
Sundargarh	971	496	25	26	16	29	66	24	0	289	66.26

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.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Balasore	1	0.086614173	0.065616798	0.041994751	0.023622047	0.086614173	0.016146719	0.088138845	0.02312336	0.566819134	0.268215223
2	Bhadrak	1	0.04	0.012	0.044	0.044	0.132	0.004	0.064	0.02	0.64	0.458
3	Balangir	1	0.234398782	0.00608828	0.070015221	0.02739726	0.080669711	0.03500761	0.00913242	0.01978691	0.517503805	0.095281583
4	Sonepur	1	0.175213675	0.012820513	0.055555556	0.034188034	0.084017094	0.051182051	0.021367521	0.02891453	0.525641026	0.251752137
5	Cuttack	1	0.101017812	0.027988822	0.027988822	0.025445293	0.211195929	0.015445293	0.111959288	0.00254529	0.366412214	0.250636132
6	Jagatsingpur	1	0.077844311	0.023952096	0.041916168	0.035928144	0.077844311	0.077844311	0.077844311	0.041916168	0.54491018	0.377964072
7	Jajpur	1	0.248275862	0.013793103	0.013793103	0.013793103	0.175862069	0.017241379	0.017241379	0.017241379	0.482758621	0.215103448
8	Kendrapara	1	0.09469697	0.018939394	0.03030303	0.022727273	0.185606061	0.018939394	0.041666667	0.053030303	0.534090909	0.281002727
9	Dhenkanal	1	0.391001136	0.013483146	0.017977528	0.008988764	0.094382022	0.011235955	0.096629213	0.04494382	0.321348315	0.134022472
10	Angul	1	0.416332288	0.036050157	0.056426332	0.029780564	0.07523511	0.010971787	0.064263323	0.026645768	0.274284671	0.062335423
11	Ganjam	1	0.383678441	0.01579659	0.014360536	0.013398295	0.015578663	0.014360536	0.054811206	0.007308161	0.439707674	0.269883078
12	Gajapati	1	0.570438799	0.018475751	0.027713626	0.009237875	0.027713626	0.15704388	0.001309469	0.013856813	0.173210162	0.061708007
13	Kalahandi	1	0.320707071	0.00101101	0.02904404	0.026515152	0.044191919	0.071969697	0.049242424	0.02020202	0.428030303	0.156931818
14	Nawapara	1	0.480519481	0.001597403	0.005194805	0.005194805	0.007792208	0.005194805	0.057142857	0.001597403	0.493766234	0.121012987
15	Keonjhar	1	0.373493976	0.007228916	0.024096386	0.031325301	0.092771084	0.112048193	0.042168675	0	0.31686747	0.057975904
16	Koraput	1	0.213393871	0.019196154	0.05107832	0.048943046	0.061295984	0.138365484	0.040863656	0.021666402	0.304199773	0.08876277
17	Malkangiri	1	0.578583765	0.00172711 6	0.03626943	0.006908463	0.039723661	0.065630397	0.003454231	0.025906736	0.2417962	0.078031088
18	Nabarangpur	1	0.465018355	0.014574689	0.015122873	0.028355388	0.083175803	0.017013133	0.001890359	0.015122873	0.349716446	0.059603025
19	Rayagada	1	0.397454031	0.025459689	0.036775106	0.031117397	0.175388967	0.053748232	0.050919378	0.007072136	0.222065064	0.069306931
20	Mayurbhanj	1	0.421305182	0.039347409	0.00687401	0.009596929	0.055662188	0.015355086	0.094049904	0.001474008	0.315335893	0.103944338
21	Phulbani	1	0.711570075	0.042394015	0.012468828	0.017456359	0.011221945	0.037406484	0.023690773	0.007481297	0.135910224	0.027369077
22	Boudh	1	0.412903226	0.051290323	0.05483871	0.064516129	0.067741935	0.038709677	0.015806452	0.011903226	0.161290323	0.166064516
23	Puri	1	0.040025885	0.025862069	0.025862069	0.00862069	0.33045977	0.022988506	0.155172414	0.001873563	0.387931034	0.295574713
24	Khordha	1	0.220640569	0.035587189	0.017793594	0.028469751	0.163701068	0.053380783	0.085409253	0.021351313	0.37366548	0.189537367
25	Nayagarh	1	0.53470437	0.015424165	0.010282776	0.01285347	0.064167352	0.015424165	0.028277635	0.002570694	0.316195373	0.111773779
26	Sambalpur	1	0.545045045	0.006006006	0.01951952	0.028528529	0.057057057	0.027027027	0.01951952	0.025525526	0.271771772	0.093858859
27	Bargarh	1	0.20880411	0.008561644	0.034246575	0.025684932	0.08047945 2	0.034246575	0.087328767	0.010273973	0.510273973	0.226267123
28	Deogarh	1	0.530612245	0.003401361	0.017006803	0.020408163	0.173469388	0.020408163	0.010204082	0.006802721	0.217687075	0.065317007
29	Jharsuguda	1	0.096153846	0.038846154	0.096153846	0.072115385	0.1875	0.081730769	0.091346154	0.014423077	0.331730769	0.059182692
30	Sundargarh	1	0.510813594	0.025746653	0.025776519	0.016477858	0.029866117	0.067971164	0.024716787	0	0.197631308	0.068238929

Fig-2 the normalized data

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:

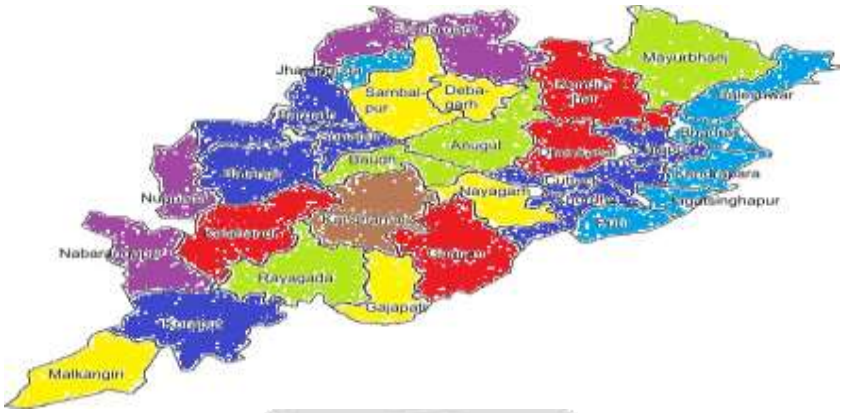

Fig-3 clustered districts of ODISHA state

```
>> kmeansd
Enter choice of option :
enter 1 for TOTAL,2 FOR AGRO-FOREST BASED ,3 FOR AGRICULTURE,4 FOR OTHER MANUFACTURING
INDUSTRY
1
enter the number of clusters :

7
1  belongs to cluster no. 6
2  belongs to cluster no. 6
3  belongs to cluster no. 5
4  belongs to cluster no. 5
5  belongs to cluster no. 5
6  belongs to cluster no. 6
7  belongs to cluster no. 5
8  belongs to cluster no. 6
9  belongs to cluster no. 2
10  belongs to cluster no. 2
11  belongs to cluster no. 1
12  belongs to cluster no. 7
13  belongs to cluster no. 1
14  belongs to cluster no. 1
15  belongs to cluster no. 2
16  belongs to cluster no. 4
17  belongs to cluster no. 7
18  belongs to cluster no. 2
19  belongs to cluster no. 3
20  belongs to cluster no. 2
21  belongs to cluster no. 7
22  belongs to cluster no. 2
23  belongs to cluster no. 6
24  belongs to cluster no. 5
25  belongs to cluster no. 2
26  belongs to cluster no. 2
27  belongs to cluster no. 5
28  belongs to cluster no. 3
29  belongs to cluster no. 4
30  belongs to cluster no. 2
```

Fig-4 program interface

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.

```
>> knn
Enter choice of option :
enter 1 for TOTAL,2 FOR AGRO-FOREST BASED ,3 FOR AGRICULTURE,4 FOR OTHER MANUFACTURING↙
INDUSTRY
4
Enter the value of k :
6
Enter if selected district OR total districts
1 for selected , 2 for total :
1
Enter the district INDEX
7
    28

    25

    26

    18

    21

    11

>>
```

```
>> knn
Enter choice of option :
enter 1 for TOTAL,2 FOR AGRO-FOREST BASED ,3 FOR AGRICULTURE,4 FOR OTHER MANUFACTURING↙
INDUSTRY
4
Enter the value of k :
1
Enter if selected district OR total districts
1 for selected , 2 for total :
1
Enter the district INDEX
7
    28

>>
```

Fig-5 Output of Experiment 2:

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 Agriculure, Govt. of Odisha-(www.agriodisha.nic.in)
[2]. Junjie Wu, "Advances in K-means Clustering-a data mining thinking"
[3]. Sadegh Bafandeh Imandoust And Mohammad Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for predicting economic events-Theoretical background"

[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.