Finding Similarities and Patterns from District-Level Data of India (Census 2011)

Abhinav Gupta
MT18031

Mohit Choudhary
MT18111

Suraj Pandey
MT18025

Udit Pant
MT18049

Abstract

This project aims to study the vast information from census 2011 and to find out the similarities among districts. After analysis of the techniques, applied to our created dataset, we found out that in case of clustering, districts belonging to the same clusters are similar in most aspects. Also, in case of association patterns, some of the association rules which we obtained were insightful and expressed important facts about the living conditions of the people at the district level.

1 Introduction

Census provides plenty of data to study the demographics and its changing patterns. Analysis of such data yields useful information which is put to use by the government for formulating policies for growth and development and demarcating constituencies to ensure optimal management. However, the sheer size of data makes it difficult to analyse and extract useful information out of it. The challenge remains to isolate sets of features which have higher degrees of correlation. The benefit of such a scheme is to observe the changes in a particular feature or a set of features relative to the other group of features. For example, good hygiene practices followed in an area (determined by observing the number of closed drainages, pipelines, waste diposal, etc.) leads to curbing of diseases caused otherwise.

2 Literature Review

Preliminary results for census 1991 have been listed out in a simplified manner by taking the census methodology into account [1]. Intercensal rate of population growth and number of males has been reported to fall significantly [2]. The growth of urbanisation, specifically in the intercensal periods, led to the emergence of census towns but these are still governed under rural administrative framework [3]. Pandey et al determined the impact of spreading awareness among the rural population about entitled services [4]. The idea of electrification of rural areas has less likelihood of solving the energy access sitaution of India [5]. Patterns such as high tobacco and alcohol use was seen prevalent among rural population as compared to urban population [6].

3 Dataset

In our attempt to study census 2011 data, we have extracted various features from the main source. We have a total of 639 data points. Each data point represents a district of India. Features include the following categories - population count, household conditions, availability of water and electricity resources, facilities like education, banking, medical, etc., condition of roads, cooking fuels, quality of drainages and assets owned. To ensure complete coverage, we've included the counts for each feature in urban, rural and total population separately. Data is mostly present in the form of .xls files scattered across links throughout the official website. Features existed as independent files but had numeric values, mostly percentages. Also, data has been categorised by type. The challenge here was to explore as many links and possible and chalk out relevant data. For the same, we built a web crawler to extract relevant features. These features existed as independent files. So the next task was to assemble these features into a single file.

3.1 Identifying Relevant Features

Since our project aims to cluster districts based on developmental patterns, identifying features was pretty intuitive. Features like literacy rate, education facilities, medical facilities, drainage systems, water supplies, electrification, etc. made to our dataset based on the assumption that higher scores in these categories play a crucial factor in better developmental growth. On the other hand, features like population count of scheduled caste, scheduled tribes were disregarded.

3.2 Data Cleaning

Slum population percentage for each district has been wrongly computed in the original data. So, it was re-calculated for each district. Similarly, while storing data points into separate tables, where each table represented an entire feature, led to some formatting errors. For example, due to redundant blank columns in some tables, while computing aggregate percentages, the values increased beyond hundred. These formatting issues were identified and removed.

3.3 Handling Errors

Some tables had misspelt names of districts. While combining features from different tables, same districts appeared as different data points due to typographical errors. For example, Rajasthan was misspelt as Rsjasthan. So, another cumbersome task was to eliminate these inconsistencies which was done by correcting these spelling errors manually.

3.4 Data Imputation

There were a few instances where some features had missing values. For example, slum population for some urban areas was left blank. Since, this field was well-populated for other areas and on cross-verification with ground truth, we concluded that these urban areas would not contain slums and thereby filled the missing values with zeroes. For example, Central Delhi was filled with a zero as no slums exists there in reality.

4 Methodology

After the entire data processing pipeline, the final dataset consisted of a total of eighty one features. The next step was to apply the following two strategies:

- Apply clustering to obtain similarities among districts.
- Apply association pattern mining to obtain association rules.

Clustering applied on any set of data points groups them into clusters each of which represents a unique class. This fact justifies our use of clustering as we wanted to group districts which have similar developmental patterns. We tried clustering with various algorithms by providing the entire dataset as input. Also, we varied our inputs by picking out specific feature vectors to be part of the sample dataset. This helped us determine how clustering was affected by varying input datasets. For example, we excluded the feature vectors which were representatives for total population and instead kept features for urban and rural poulation only. To evaluate the clustering performance, we used silhouette coefficient and compared the quality of clusters. Also, we observed the shifting of certain data points (districts) when we varied the number of clusters.

Association pattern mining is a useful technique when there is a need to identify interesting patterns exhibited by our dataset. The motivation for applying pattern mining was to able to visualize what kind of dependencies exist between the chosen feature vectors. Since our dataset had features for total, urban and rural population individaully, we were able to apply rule mining on all these features colletively as well as individually. We applied one of the most popular algorithms (Apriori algorithm) for obtaining association rules. One daunting task was to convert our numerical data to nominal data as we can't provide numerical data as input to our algorithm. For this conversion, we broke down our percentage figures into three equi-width intervals - low, medium and high. The width of each interval was approximately thirty-three percent, i.e., low interval ranged from zero percent to thirty-three percent, medium interval from thirty-four percent to sixty-six percent and the rest belonged to the high interval. With this kind of grouping, we were able to obtain association rules for our dataset and analyse them as well.

4.1 K-Means Clustering

Initially, we applied the algorithm on the entire dataset by simply varying the number of clusters and observing the relative changes appearing in the clusters. Changes appearing in successive iterations were tracked for analysing the potential membership of any data point in the next iteration. Next, we selectively picked features from the original dataset and applied clustering to these new set of features to see if these group of features were clusteriing well together. Sihouette coefficients were calculated for each run of the algorithm with varying number of clusters.

4.2 K-Medians Clustering

The same approach of varying the number of clusters was applied here as well. Sihouette coefficients were computed and compared against each other for varying number of clusters.

4.3 DBScan

Instead of using distance functions, we tried using the density characteristics of individual data points and then using it to form clusters. After applying the algorithm, the final clusters were also evaluated using silhouette coefficient.

4.4 Apriori Algorithm

For obtaining association rules, we applied the apriori algorithm on the nominal features which we derived from our original dataset. We also obtained frequent itemsets appearing in our dataset.

5 Analysis

Applying clustering and association pattern mining was essential for our project. Both of the techniques yielded useful results, which for the most part, were not obvious when observed directly from raw data.

5.1 Tools Used

i) WEKA: This tool was used primarily for association pattern mining. Weka allowed for easy selection of constraints which we wanted to impose on our dataset for obtaining rules. The support values were varied from 0.1 to 0.8 to look at the change in number of rules for higher confidence values. Besides obtaining association rules, we were also able to get the sets of frequent itemsets as they changed successively in each of the iterations.

ii) NCSS: This tool was used for analysing the clusters. NCSS generates detailed reports on clustering. Specifically, it lists out the computed values for each step which helps visualise how the clusters are taking shape. For K-Means clustering, we varied the number of clusters to be formed and observed how data points shifted from cluster to cluster.

5.2 Clustering

K-means clustering and K-medians clustering was applied on the dataset. Clustering was carried out for two, three, four and five clusters. Silhouette coefficients were calculated for each of these number of clusters. The silhouette coefficients for each clustering algorithm against the number of specified clusters are given in Table 1.

Algorithm	SC(2)	SC(3)	SC(4)	SC(5)
K-Means	0.25	0.17	0.17	0.15
K-Medians	0.25	0.16	0.13	0.10

Table 1: Sihouette Coefficients (SC) for 2, 3, 4 & 5 clusters

For analysis, we compared some of the districts which were in the same cluster and had high distance from the cluster representative. Some of them are listed below:

- Kolkata (6.2469) and Dehradun (1.9786)
- Lakshadweep (5.07) and Pune (2.44)
- Surat (3.50) and Gurgaon (1.61)

The values (in brackets) represent the distance of the data point from the cluster representative. For the above-mentioned districts, the values of almost all features is similar except features like slum population ratio and fuel used in cooking but other features are almost in the same range. Since these districts belong to the same cluster, we conclude that these are similar to each other.

Change in the number of clusters from 3 to 4:
Since, the distance function to representative 3 for some data points are very large as compared to others . So, when there are four clusters, it will go to another cluster due to some features difference. Some of them are listed below:

- Gujarat: Ahmedabad and Bhavnagar
  Rural and Urban ratio - (For Rural Ahmedabad (15.96) and Bhavnagar (58.95) and Urban Population ratio Ahmedabad (84.04) and Bhavnagar (41.04))
  Roof Concrete is also a feature which has large difference Ahmedabad (63.4) and Bhavnagar (44.6)
  Tap Water from treated source Ahmedabad (67) and Bhavnagar (42.8)
  Location of drinking water source 86.1 and 66.6
  Latrine Facility within Premises 83 and 53.9
  Bathroom within Premises 86.6 and 69.1
  Total waste water outlet connect to closed

drainage 76.6 and 40
Total waste water outlet connect to open drainage 3 and 9
Total Fuel Used LPG Biogas 65.5 and 32.1
Total cooking inside house 92.2 and 77.8
Total households availing banking services 67.3 and 50.7

- Haryana: Karnal and Sirsa
  Roof Concrete is also a feature which has large difference Karnal (43.4) and Sirsa (27)
  Total waste water outlet connect to closed drainage 20.5 and 22.4
  Total waste water outlet connect to open drainage 70.5 and 47.6
  Total Fuel Used LPG Biogas 50.7 and 29.6

- Himanchal Pradesh: Shimla and Kullu
  Total waste water outlet connect to closed drainage 52.4 and 25.8
  Total waste water outlet connect to open drainage 23 and 34.5
  Latrine Facility within Premises 68.3 and 58.4
  Total Fuel Used LPG Biogas 60.3 and 41.9

- Karnataka: Bangalore and Udupi
  Roof Concrete is also a feature which has large difference (67.8) and (24.5)
  Tap Water from treated source (66.6) and (11.7)
  Total Fuel Used LPG Biogas 76.2 and 34.9

- Kerela: Thiruvananthapuram and Palakkad
  Roof Concrete is also a feature which has large difference (53.8) and (30.1)
  Total waste water outlet connect to closed drainage 21.4 and 11.6
  Total waste water outlet connect to open drainage 11.6 and 21.8
  Total Fuel Used LPG Biogas 42.2 and 28.8

- Maharastra :Pune and Satara
  Roof Concrete is also a feature which has large difference (45.9) and (16.1)
  Total waste water outlet connect to closed drainage 62.8 and 18.3
  Total waste water outlet connect to open drainage 13.5 and 40.4
  Total Fuel Used LPG Biogas 68.3 and 43.4

Inference: We started with k=5 and found out that one of the cluster represents districts which mostly have higher values for most features. As we started decreasing k, the data points nearly remained same for this cluster except for a very little number of data points which got added. They were also found to be of good quality in most features but had lower values for a few features like quality of household roofs, quality of drainage systems and fuel used for cooking. This indicates that the features which had lower values in this context have more dominance as compared to others.

DBScan for the fixed values of ϵ, resulted in :

ϵ	**SC**	**Clusters**
130	0.31	3
100	0.01	4

Table 2: Sihouette Coefficients & No. of Clusters for ϵ

We observed that the silhouette coefficients for DBScan were very low. Therefore, DBScan doesn't seem to produce good clusters for our dataset.

5.3 Association Pattern Mining

We applied pattern mining individually on urban population, rural population and total population. We found out the following association rules:

- Total livable condition households (%)=M 480 ==> Total cooking inside house (%)=L 450$< conf : (0.94) >$

- Total Workers (%)=M Total households with no assets (%)=S 432 ==> Total cooking inside house (%)=L 405 $< conf : (0.94) >$

- Total Male Workers (%)=M Total Female Workers (%)=S 407 ==> Total cooking inside house (%)=L 383 $< conf : (0.94) >$

- Total Literacy Rate Male (%)=L Total livable condition households (%)=M Total waste water outlet connect to closed drainage (%)=S 426 ==> Total Male Workers (%)=M Total cooking inside house (%)=L 401 $< conf : (0.94) >$

- Total Literacy Rate (%)=L 457 ==> Total households with no assets (%)=S 412 $< conf : (0.9) >$

- Rural Population (%)=L Rural Fuel Used LPG+Biogas (%)=S 450 ==> Rural waste water outlet connect to closed drainage (%)=S 449 $< conf : (1) >$

- Urban Fuel Used LPG+Biogas (%)=M 362 $==>$ Urban households with no assets (%)=S 362 $< conf : (1) >$
- Total Literacy Rate Male Urban (%)=L Urban Fuel Used LPG+Biogas (%)=M 358 $==>$ Urban households with no assets (%)=S 358 $< conf : (1) >$
- Urban Fuel Used LPG+Biogas (%)=L Urban households availing banking services (%)=L 158 $==>$ Urban Latrine facility within premises (%)=L Urban Bathroom with roof within premises (%)=L Urban cooking inside house (%)=L Urban households with no assets (%)=S 151 $< conf : (0.96) >$

6 Conclusion

After application of clustering and association pattern mining techniques, we obtained important inferences about our dataset. Districts belonging to the same clusters are similar in most aspects (except some features like slum population and fuel used in cooking). Also, some of the association rules which we obtained were insightful and expressed important facts about the living conditions of the people at the district level.

Acknowledgements

We would like to express our heartfelt gratitude to our instructor Dr. Sumit Bhatia for providing us with the opportunity to work on this project. His continuous support and guidance was crucial in helping this project take shape. We would also like to thank the teaching assistants for their suggestions and help.

References

[1] Bhagat, Ram B. ”Emerging pattern of urbanisation in India.” Economic and political weekly (2011): 10-12.

[2] Dyson, Tim. ”The preliminary demography of the 2001 census of India.” Population and Development Review 27.2 (2001): 341-356.

[3] Pradhan, Kanhu. ”Unacknowledged urbanisation: New census towns of India.” (2013).

[4] Pandey, Priyanka, et al. ”Informing resource-poor populations and the delivery of entitled health and social services in rural India: a cluster randomized controlled trial.” Jama 298.16 (2007): 1867-1875.

[5] Bhattacharyya, Subhes C. ”Energy access problem of the poor in India: Is rural electrification a remedy?.” Energy policy 34.18 (2006): 3387-3397.

[6] Neufeld, K. J., et al. ”Regular use of alcohol and tobacco in India and its association with age, gender, and poverty.” Drug and alcohol dependence 77.3 (2005): 283-291.