Application of K-Means Algorithm to Mapping Poverty Outline by Province in India

Pushpendra Kumar Verma, Preety

Abstract*: India has a second largest population and seventh largest country in the world, the UN data in 2018 recorded that there were 1,368,681,134 more people scattered throughout the Indian provinces. In addition, India also has a variety of social problems, one of which is poverty. The poverty line number in Indonesia needs to be improved. Data utilization techniques become new information called data mining. One of the most popular data mining methods is clustering using the k-means algorithm. K-means can process data without being notified in advance of the class label. This study will produce three provincial groups according to very low, low and sufficient income figures. Data processing of poverty line numbers in India using the k-means algorithm to get the results of the Davies Bouldin index of 0.271. These results are considered well enough because the closer the results obtained with zeros, the better the data similarity between members of the cluster.

Keywords: Poverty Line, K-means, cluster analysis Data Mining.

I. INTRODUCTION

Poverty is one of the social problems and also become a challenge for many communities around the world to always find a solution. At the global level, data on poverty regarding the number of poor people is dominated by developing countries. However, in developed countries like the United States as well, there are still poor people. So poverty is everywhere universally to be a problem with the community and the world. The experience of poverty reduction in the past have shown many weaknesses, such as: (1) macro growth orientation without considering aspects of equity, (2) centralized policy, (3) more caricature than transformative, (4) positioning communities as objects rather than subjects, (5) orientation poverty alleviation tends to caricature and instantaneous than sustainable productivity, and (6) perspectives and solutions that are generic to the problems of poverty that exist regardless of diversity which exists. Because it is so diverse nature of the challenges that exist, the handling of the problem of poverty must touch the bottom of the source and root of the real problem, either directly or indirectly.

1.1 Problem Formulation

Based on the background of the problem, the problem formulation taken in this research is "How to implement the algorithm k-means for mapping poverty line numbers in India from data that has been collected for the past four years".

1.2 Research purposes

This research is intended to process data on poverty line numbers in Indonesia sourced from the website The Central Statistics Agency uses the method clustering as a field of science viz data mining. In addition to getting the results of poverty line mapping by applying an algorithm k-means, so that the right solution can be taken for each different region in India

1.3 The objectives of this research are:

1. Conduct further review related to poverty line numbers in Indonesia compiled by the Central Statistics Agency over the past four years.
2. Grouping provinces in India into three groups according to the income figures for each region uses an algorithm k-means

1.4 Research benefits

The benefits of this thesis research are the results of mapping poverty line numbers which are expected to take appropriate solution steps according to the respective regional level. The other benefits of this study are:

For Authors- Can add knowledge and insight and can apply the theory that has been obtained during the lecture.

For Students-Can be a reference for students who study the algorithm k-mean in the future.

II. LITERATURE REVIEW

1. Clustering the determination of the potential for regional crime in the city of Banjarbaru by the method k-means (Rahayu, S., Nugrahadi, D.T., Indriani, F. 2014). The research discusses the application of algorithmsk-means in determining potential crimes based on crime data owned by the Republic of Indonesia National Police in the South Kalimantan area of Banjarbaru Resort. The variables used inclustering Determination of the potential for regional crime in the city of Banjarbaru is punishment, month and report. After getting the data the next step is to prepare the data viz. data selection, data preprocessing, transformation until the method is applied k-means. The conclusion of this study clustering the potential for regional crime in the city of Banjarbaru is processed based on alignment. This is done so that the results of grouping the potential for regional crime in the city of Banjarbaru are more specific.

Revised Manuscript Received on February 01, 2020.

***** Correspondence Author

Dr. Pushpendra Kumar Verma, Associate Professor, School of Computer Science and Application, IIMT University, Meerut, UP, India.
E-mail : dr.pkverma81@gmail.com

Dr. Preety, Assistant Professor, Faculty of Management Swami Vivekanad subharti University Meerut UP India. Email: mailpreity81@gmail.com

2. Analysis of the method hierarchical clustering and k-means with the LRFMP model in customer segmentation (Muhidin, A. 2017). The research discusses the application of algorithms k-means in determining potential consumer segmentation. Variables used in clustering customers based on the LRFMP model (Length, Recency, Frequency, Monetary, Payment).

In this study the process of customer segmentation begins with the process preprocessing, analytic hierarchy process (AHP), the search for the best K value of all methods hierarchical clustering by comparing values index. Then the selected k value is made as the initial value on k-means clustering. Results clustering It is used to segment using the RFM model to get the consumer class. Results clustering can be used as a marketing reference in determining the treatment of customers.

3. Utilization of methods k-means clustering in determining the direction of high school students (Aziz, A., Purmaningsih, C., Saptono, R. 2014). The research discusses the application of algorithms k-means. The research discusses the application of algorithms k-means in grouping fruit producing regions. The variables used are based on harvested area (Ha), production (tons) and year of harvest. After getting the data the next step is to prepare the data viz data cleaning,

4. Application of the method k-means clustering to classify fruit production potential in the province of the Special Region of Yogyakarta (Murti, M.A.W.K. 2017). Transformation until the method is applied k-means. The conclusions of the study provide a mapping of areas with high, medium and low fruit production.

2.1 Definition

2.1.1 Algorithm

According to Munir (2012: 176) algorithm is a logical sequence of problem solving steps that are arranged systematically. The ordering method is described in a number of limited steps that lead to the solution of the problem.

2.1.2 K-means

According to Vulandari (2017) k-means is an algorithm that sets values the cluster (k) in a manner random, for a while the value becomes the center of the cluster commonly called centroid. Then calculate the distance of each existing data against each of them centroid using a formula 9las an9n until you find the closest distance from each data with centroid up to value centroid unchanged (stable).

According to Suyanto (2017) k-means is a clustering algorithm that has a simple basic idea by minimizing Sum of Squared Error (SSE) between data objects and a number of k centroid.

Algorithm: k-means. The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

1) Arbitrarily choose k objects from D as the initial cluster centers;

2) Repeat

3) (re)assign each object to the cluster to which the object is the most similar,

4) based on the mean value of the objects in the cluster;

5) Update the cluster means, that is, calculate the mean value of the objects for

6) Each cluster;

7) Until no changes;

2.1.3 Clustering

According to Suyanto (2017) Clustering is the process of grouping a set of data objects (into multiple groups) or the cluster so the objects in one group have a high similarity, but are very different from the objects in other groups. According to Han, et al (2012: 445) the clustering is the process of partitioning a set of data objects (observations) into a subset that can be used to organize search results into groups and present results in a concise and easily accessible way. Clustering widely used in various fields with a variety of very important applications including market research, recommendation systems, security systems and search engines.

III. DATA MINING

3.1 Definition data mining

The development of information technology has contributed to the rapid growth in the amount of data collected and stored in large databases (big data). Big data is a term that describes a large volume of data, both structured data and unstructured data. Big data has high potential to gather key insights from business information. Big data can be analyzed for insights that lead to better business decision making and strategy. A method or technique is needed to be able to transform the data into valuable information or knowledge that is useful to support decision making. A technology that can be used to make it happen is data mining. Recently data mining has been implemented into various fields, including business or trade, education and telecommunications.

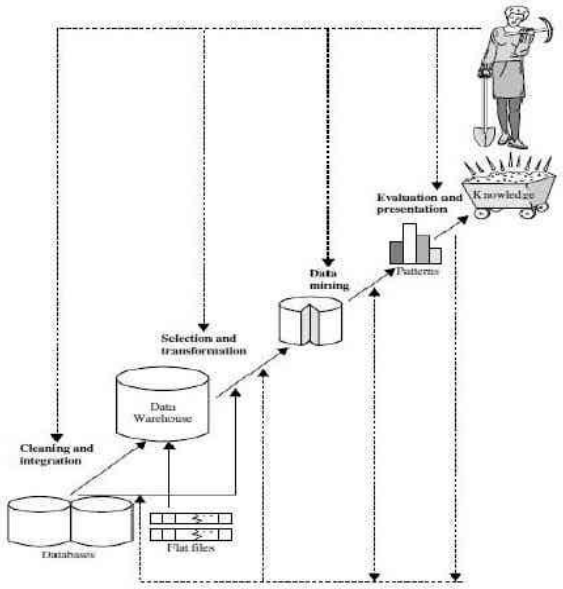

Fig.1.0-The discovery of new knowledge. Source: Han, et al, 2012.

3.2 Function data mining

In general, usability data mining divided into two namely descriptive and predictive. Descriptive means to look for patterns that can be understood by humans that explain the characteristics of the data while predictive is used to form a knowledge model to make predictions. Based on its functionality, tasks data mining can be grouped into six groups namely:

The detailed description of the six groups is as follows:

1. Classification (classification)-The process of generalizing structures that are known to be applied to new data.
2. Clustering (clustering)-Grouping data of unknown class labels into a number of specific groups according to the size of the similarity.
3. Regression (regression)-Find a function that models data with minimum prediction error.
4. Anomaly detection (anomaly detection)- Identifying unusual data, in the form of outlier, changes or deviations that may be very important and need further investigation.
5. Dependency modeling-Looking for relations between tables.
6. Summary (summarization)-Provides simpler data representation, including visualization and report generation.

3.3 Learning techniques data mining

The technique used in data mining closely related to discovery and learning which is divided into three main methods of learning, namely:

A. Supervised learning

A technique that involves a training phase where historical training data whose characters are mapped to known outcomes and is processed in an algorithm data mining. This process trains algorithms to recognize variables and key values which will later be used as a basis for making estimates when new data is provided.

B. Unsupervised learning

Learning techniques that do not involve training phases such as supervised learning that is, it depends on the use of an algorithm that detects all patterns that emerge from specific important criteria in the input data. This approach leads to the making of many rules 16 characterizes the invention associations, clusters and segment which is then analyzed to find important things.

C. Reinforcement learning

Techniques that have applications that are continuously optimized over time and have adaptive control. Resembles real life which is like "on job training “Where a worker is given a group a task that requires decisions. Reinforcement learning very appropriate to be used to solve difficult problems that depend on time.

IV. RESULT AND DISCUSSION

The following is a systematic system general descriptive model built:

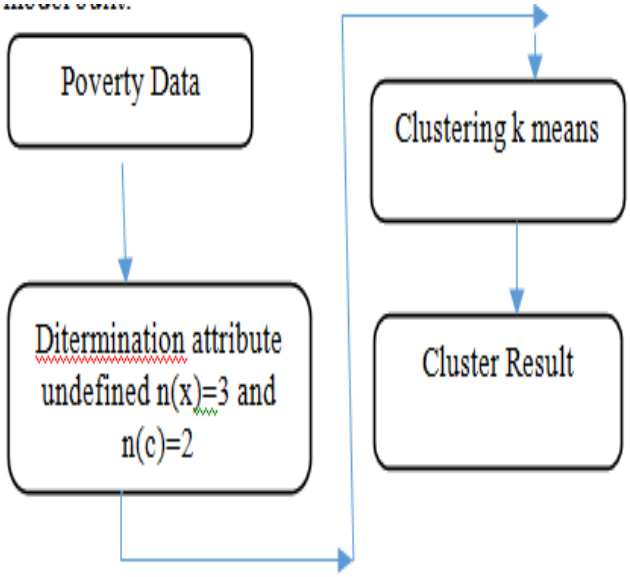

Fig.2.0: system flow in general

Figure 1 explains the descriptive a model where the data are processed is the data poverty in 12 cities or districts in the province Riau. Variable used is percentage data poverty (x1) and lines poverty (x2) as well the number of the population is poor (x3) to determine clusters of areas with high poverty rates (c3) and cluster with the region with poverty level normal (c2) as well region with level poverty low (c1). Data the processed with Muse k-means clustering to find patterns that are provide an overview of the clusters of each region. Following is the process of clustering k- means

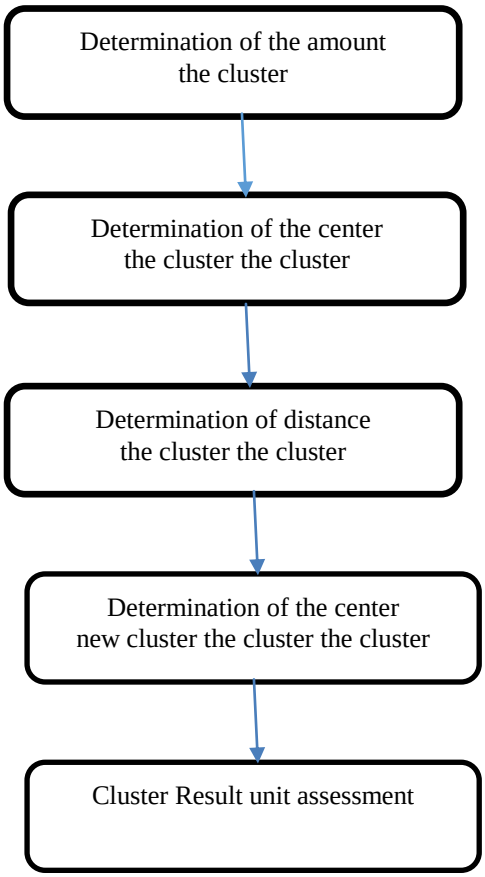

Fig. 3.0: The K-Means clustering process

Starting with determining the number of clusters, in terms of this cluster who used three cluster, Wiley cluster with Nagaland level poor normal and cluster with poverty level height Odisha region with them level real high. Determination center random cluster of values found in variable l processed. Next looking for Uttar Pradesh and value and cluster determination new. Until ah iteration is complete done. The following is an analysis and has il from Descriptive models are carried out

Table 1: Indian poverty estimates (% below poverty line) (1993– 2012)

Year	Rural	Urban	Total
1993 – 94	50.1	31.8	45.3
2004-05	41.8	25.7	37.2
2009-10	33.8	20.9	29.8
2011-12	25.7	13.7	13.7

Table2: State-wise poverty estimates (% below poverty line) (2004-05, 2011-12)

State	2004-05	2011-12	Decrease
Andhra Pradesh	29.9	9.2	20.7
Arunachal Pradesh	31.1	34.7	-3.6
Assam	34.4	32	2.4
Bihar	54.4	33.7	20.7
Chhattisgarh	49.4	39.9	9.5
Delhi	13.1	9.9	3.2
Goa	25	5.1	19.9
Gujarat	31.8	16.6	15.2
Haryana	24.1	11.2	12.9
Himachal Pradesh	22.9	8.1	14.8
Jammu and Kashmir	13.2	10.4	2.8
Jharkhand	45.3	37	8.3
Karnataka	33.4	20.9	12.5
Kerala	19.7	7.1	12.6
Madhya Pradesh	48.6	31.7	16.9
Maharashtra	38.1	17.4	20.7
Manipur	38	36.9	1.1
Meghalaya	16.1	11.9	4.2
Mizoram	15.3	20.4	-5.1
Nagaland	9	18.9	-9.9
Odisha	57.2	32.6	24.6
Puducherry	14.1	9.7	4.4
Punjab	20.9	8.3	12.6
Rajasthan	34.4	14.7	19.7
Sikkim	31.1	8.2	22.9
Tamil Nadu	28.9	11.3	17.6
Tripura	40.6	14.1	26.5
Uttar Pradesh	40.9	29.4	11.5
Uttarakhand	32.7	11.3	21.4
West Bengal	34.3	20	14.3
All Inda	37.2	21.9	15.3

Table 3: Indian poverty lines (in Rs. per capital per month) A c value k was taken from score each respectively variable that use. Take score random b influence on the number of iterations in k-means clustering.

Year	Rural	Urban	Total
1993 – 94	50.1	31.8	45.3
2004-05	41.8	25.7	37.2
2009-10	33.8	20.9	29.8
2011-12	25.7	13.7	13.7
	X1	X2	X3
C1	13.7	3.27	282,361
C2	23.2	23.2	296.77
C3	29.6	5.54	328,158

Table 4: Random Cluster center

Next determine center the cluster new. On the table 5 seen that cluster center only focus on C2 and C3.

	X1	X2	X3
C1	-	-	-
C2	40.85	7.79	290.07
C3	40.99	10.78	371.58

Table 4: New Cluster center

Descriptive model Where data yang processed is poverty data at in India where variables du use are: poverty percentage data (x1) and gar is poverty (x2) and amount residents e.g. k in (x3) to determine cluster territory with levels normal poverty and cluster with w region with high poverty rates and regions with level to poor . The data is processed by using k- means clustering for find different patterns Build a picture of cluster each the state.

V. CONCLUSION

Based on the results of research conducted by the author, the following conclusions can be drawn:

1. The application of the k-means algorithm divides the dataset into three groups: very low, low and quite in accordance with the similarity of income / capital / month levels.
2. The test results get a value Davies Bouldin index amounting to 0.288 which means similarity between members the cluster which is quite good.

5.1 Suggestions

Considering there are still many things that cannot be implemented from this research, the authors consider several suggestions, namely:

1. Results clustering formed can be developed into a knowledge base for the provincial mapping decision support system with the average income of each region in accordance with its similarity.
2. Combine with other methods or approaches to get better research results.

REFERENCES

1. G. O. Young, “Synthetic structure of industrial plastics (Book style with Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Computing Surveys. 1999;31(3):264–323.
2. Celebi ME, Kingravi HA, Vela PA. A comparative study of efficient initialization methods for the K-means clustering algorithm. Expert Systems with Applications. 2013;40(1):200–210.
3. Bordogna G, Pasi G. A quality driven hierarchical data divisive soft clustering for information retrieval. Knowledge-Based Systems. 2012;26:9–19.
4. Luo C, Pang W, Wang Z. Semi-supervised clustering on heterogeneous information networks. In: Advances in Knowledge Discovery and Data Mining; 2014. p. 548–559.
5. Yang HJ, Kim YE, Yun JY, Kim HJ, Jeon BS. Identifying the clusters within nonmotor manifestations in early Parkinson’s disease by using unsupervised cluster analysis. PLoS One. 2014;9(3):e91906. pmid:24643014
6. Bogner C, Trancón y Widemann B, Lange H. Characterising flow patterns in soils by feature extraction and multiple consensus clustering. Ecological Informatics. 2013;15:44–52.
7. E. Prasetyo, "Data Mining Processing Data into Information withMatlab " , Yogyakarta: Andi, 2016.
8. W. Nengsih, "DESCRIPTIVE MODELING USING K-MEANSFOR CLUSTERING OF POVERTY LEVELS IN RIAU PROVINCE, "2016.
9. F. Fajrianti, MN Bustan and MA Tiro, "USE OF CLUSTER ANALYSISK-MEANS AND DISCRIMINANT ANALYSIS IN VILLAGE GROUPINGPOOR IN PANGKEP DISTRICT, 2018.
10. AN Ulfah and S. Uyun, "Performance Analysis of Fuzzy C-Means and K-Means on Poverty Data " Jatisi, vol. Vol. 1 No. 2, pp. 139 -148, 2015.
11. J. Han and M. Kamber, " Data Mining Concepts and Techniques, SanFrancisco: Morgan Kaufmann " , 2011.
12. C. Vercellis, Business Intelligence: " Data Mining and Optimization forDecision Making ", Italy: Wiley, 2009.
13. DT Larose, " Discovering Knowledge in Data" , America: Wiley, 2005.
14. AE Putri and M. Budiharto, “Understanding JKN (National Health Insurance)" ,Jakarta: Friedrich-Ebert-Stiftung, 2014.
15. Agusta and Yudi, K-Means, “Web Blog for Data Mining and Clustering ", Yogyakarta: Andi, 2001.

AUTHORS PROFILE

Dr. Pushpendra Kmar Verma, is an Associate Professor in the SoCSA, IIMT University, Meerut, UP India. He has a MCA, M.Tech.(CSE).MPhil (CS) and has completed Ph.D. from S.V. Subharti University, Meerut, UP. His area of interest is in CNS. He has written original research articles in various International journals and interested in academics and research.

Dr. Preety is an Assistant Professor in the Faculty of Management, Swami Vivekanand Subharti University, Meerut, UP India. She have MSc, MBA and has completed Ph.D. from S.V. Subharti University, Meerut, UP. Her area of interest is in Finance and HR. She has written original research articles in various International journals and interested in academics and research.