International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

Analysis of K-Means Clustering Algorithm

Rishabh Sharma, Dr. Shipra Arora
Dronacharya College Of Engineering,
Farrukh Nagar, Gurgaon - 123506

Abstract:- As we know, Clustering analysis is a very important feature of machine learning. This process divides the data points on the basis of their characteristics. Similar data points come into the same group and unsimilar data points in different groups. So basically it makes a group of similar data points study further.
This paper analysis K-means clustering algorithm by taking an example of cricket stats in which player capability is determined using their stats like wickets taken and run scored in their career. This capability can be used for the team selection purpose by selecting appropriate batsman, allrounders and bowler.

Keywords: Analysis, K-means, clustering, pattern matching

1. INTRODUCTION

Clustering algorithms perform an action on given data sets or populations to divide data points into various groups in such a manner that similar data points come into the same group and unsimilar data points in different groups. Cluster analysis is used for both understanding the relationship and difference between data. Nowadays it has various applications in machine learning. Cluster analysis may disclose not only the underlying relationship and difference between data, but it can also serve as a critical foundation for subsequent data analysis and knowledge discovery.

There are multiple stages are involved in clustering process. First Step is “Feature Selection and Extraction”, which is used to identify the data input or pattern which it received and perform dimensionality reduction which is key to reducing model complexity and overfitting [6] . Then “Interpattern Similarities” is done to determine the similarity between dataset by finding pattern or trend. Then “Grouping” is done to form cluster.

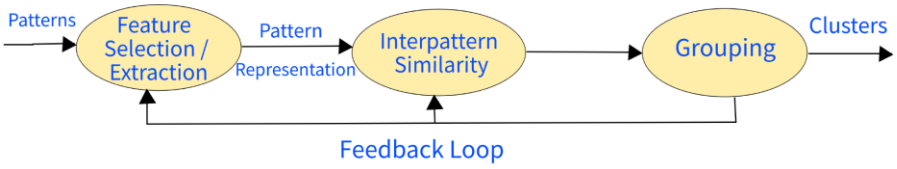

Figure 1: Stages in Clustering

2. RELATED WORK

G. Kesavaraj and S. Sukumaran [1] perform a study on different classification techniques. The main purpose of the classification of data is to identify patterns in a given dataset and divide data on basis of classification. It discusses various examples of classification, for example, the segmentation of customers on the basis of credit risk to provide loans. It helps to create a model to classify the dataset on given attributes and datasets. It describes the commonly used methods for data mining classification i.e. Decision tree induction methods, Rule-based methods, Memory-based learning, Neural networks, Bayesian networks, and Support vector machines [1]. And finally concluded that no technique is superior to other techniques. Not all problems can be solved by a single technique with the optimum solution. For each domain, there is a specific algorithm.

Bini B. S. & Mathew T. [2] made a price prediction using clustering and regression techniques. The regression technique is used to predict the price and clustering to find the pattern between data. First of all, Clustering is done on the available stocks validation index and the K-means algorithm appears to be more efficient among all clustering algorithms. Then cluster is passed to multiple regression to find a predicted price.

For this purpose, data is collected for National Stock Market for various stocks like WIPRO, TCS, ROLTA, POLARIS, PERSISTENT, NIITTECH, NAUKRI, MINDTREE, INFY, and HCL TECH [2]. On the selected stocks, various clustering technique is performed to get the best stocks. To get the best technique among all clustering algorithms comparison is done using the c-index, the Rand index, Jaccard index and silhouette index [2]. After comparing all clustering algorithms K-means algorithm and EM algorithm shows better performance than density-based clustering and hierarchical based clustering algorithm. After cleaning data by removing unwanted stocks, price prediction is done using multiple regression techniques. After predicting the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

price of all stocks it compares the price for TCS stocks to the predicted and actual prices and it comes out to be approximately accurate. So it helps investors to select the stocks from a given bunch of stocks to choose wisely

Li, Y., & Wu, H [3] discussed the improved k-means clustering algorithm. The new algorithm is based on selecting the initial focal point. So it improves the K-means clustering by not randomly selecting centroid for the first run. It first finds the minimum and maximum distance between data points, to understand the centroid value more accurately. Centroid value not to be adjacent and should be with minimum distance to nearest data points. The new algorithm basically solves two issues of the standard K-means algorithm a). Choosing the beginning focal point dependency. b) Being stuck in a local minimum.

"So in this paper, Li, Y., & Wu, H uses the largest minimum distance algorithm to determine K's initial cluster point. The largest minimum distance algorithm is used in order to find hidden information in the field of pattern recognition. It is done to choose the pattern in which pairwise distances are farther apart as much as possible to be the cluster focal point. Thus, it helps to determine the best initial cluster focal point and also increases the efficiency of dividing the initial data congregation. So it makes the initial focal points to be more representative and decentralized" [3].

3. METHODOLOGY

The "K-Means Clustering" is an "unsupervised learning" approach used to solve clustering challenges in "data science and machine learning". It's an iterative technique that divides the input dataset into "K" unique clusters, each of which contains only 1 dataset. The "K-means algorithm" locates "K" centroids before assigning each data point to the cluster with the closest centroids. Basically as explained in [6], "K-means algorithm name consists of two words, *K* - represents the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the centre of the cluster. *Means* - refers to averaging of the data, i.e finding the centroid" [6].
The algorithm works by continuously finding the best value for "K centre point" or "Centroid" using iterative methods. The K-means algorithm begins with the initial group of randomly chosen centroids, which serve as the starting points for each cluster, and then performs iterative (repetitive) calculations to optimise the centroids' placements [6]. It stops forming and optimising clusters when either the centroids have stabilised or the specified number of iterations was completed [6].

Selecting The Right Number Of Clusters.
We already knew the "K value" in the situation presented before. However, knowing the "K value" before to computation is impossible in every scenario. As a result, the best value for "K" must be chosen. We shall not choose the no. of clusters in the approach at arbitrary. Each cluster is created by calculating and comparing the average distances between its centroid.

To calculate the correct number of clusters, the "Within-Cluster-Sum-of-Squares" (WCSS) method can be utilised. WCSS stands for the sum of the squares of the data points' distances from the cluster's centroid in each cluster [6].

The "Elbow Method" is a popular method for determining the correct value of "K".

4. RESULTS

To analyse the K-means clustering algorithm, we used the cricket problem to understand the K-means clustering. For the team selection process in cricket, it is necessary to understand the capability of the players. For this purpose, we cluster three groups of players on the basis of runs scored and wickets taken to classify them on basis of category of Batsman, Bowler and All-rounder.

Firstly we extracted the data of known cricket players and cleaned data on basis of matches played so that there is no outlier in the results. The player data is shown in Figure 2. Secondly, we perform a clustering algorithm on runs scored and wickets taken to make 3 clusters using sklearn library (Refer to Figure 4). The plot is drawn to represent the data in user-interactive form (Refer to Figure 5.).

As a result, we can conclude that K-means clustering performs very well as Bowlers like Anil Kumble, James Anderson and Shane Warne are classified in a single group and batsmen like Sachin Tendulkar, Ricky Pointing, Rahul Dravid, Brian Lara are represent in another group.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

	Player Name	Runs	Wickets
0	Sachin Tendulkar	15921	46
1	Ricky Ponting	13378	5
2	Rahul Dravid	13288	1
3	Brian Lara	11953	0
4	Anil Kumble	2506	619
5	Harbhajan Singh	2225	417
6	Glen Mcgrath	563	641
7	Sunil Gavaskar	10122	1
8	Kapil dev	5248	434
9	Imran Khan	3807	362
10	Sir Garry Sobers	8032	236
11	James Anderson	1262	640
12	Shane Warne	3154	708
13	Shaun Pollock	3781	421

Figure 2: Player Stats

	Player Name	Runs	Wickets	Y-axis	X-axis
0	Sachin Tendulkar	15921	46	1.000000	0.064972
1	Ricky Ponting	13378	5	0.834419	0.007062
2	Rahul Dravid	13288	1	0.828558	0.001412
3	Brian Lara	11953	0	0.741633	0.000000
4	Anil Kumble	2506	619	0.126514	0.874294
5	Harbhajan Singh	2225	417	0.108217	0.588983
6	Glen Mcgrath	563	641	0.000000	0.905367
7	Sunil Gavaskar	10122	1	0.622412	0.001412
8	Kapil dev	5248	434	0.305053	0.612994
9	Imran Khan	3807	362	0.211225	0.511299
10	Sir Garry Sobers	8032	236	0.486326	0.333333
11	James Anderson	1262	640	0.045514	0.903955
12	Shane Warne	3154	708	0.168707	1.000000
13	Shaun Pollock	3781	421	0.209532	0.594633

Figure 3: Player Stats after adding x and y axis value

	Player Name	Runs	Wickets	Y-axis	X-axis	cluster
0	Sachin Tendulkar	15921	46	1.000000	0.064972	0
1	Ricky Ponting	13378	5	0.834419	0.007062	0
2	Rahul Dravid	13288	1	0.828558	0.001412	0
3	Brian Lara	11953	0	0.741633	0.000000	0
4	Anil Kumble	2506	619	0.126514	0.874294	2
5	Harbhajan Singh	2225	417	0.108217	0.588983	1
6	Glen Mcgrath	563	641	0.000000	0.905367	2
7	Sunil Gavaskar	10122	1	0.622412	0.001412	0
8	Kapil dev	5248	434	0.305053	0.612994	1
9	Imran Khan	3807	362	0.211225	0.511299	1
10	Sir Garry Sobers	8032	236	0.486326	0.333333	1
11	James Anderson	1262	640	0.045514	0.903955	2
12	Shane Warne	3154	708	0.168707	1.000000	2
13	Shaun Pollock	3781	421	0.209532	0.594633	1

Figure 4: Player Stats with cluster value
(0 represent batsman, 1 represent Allrounder and 2 represent bowler)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

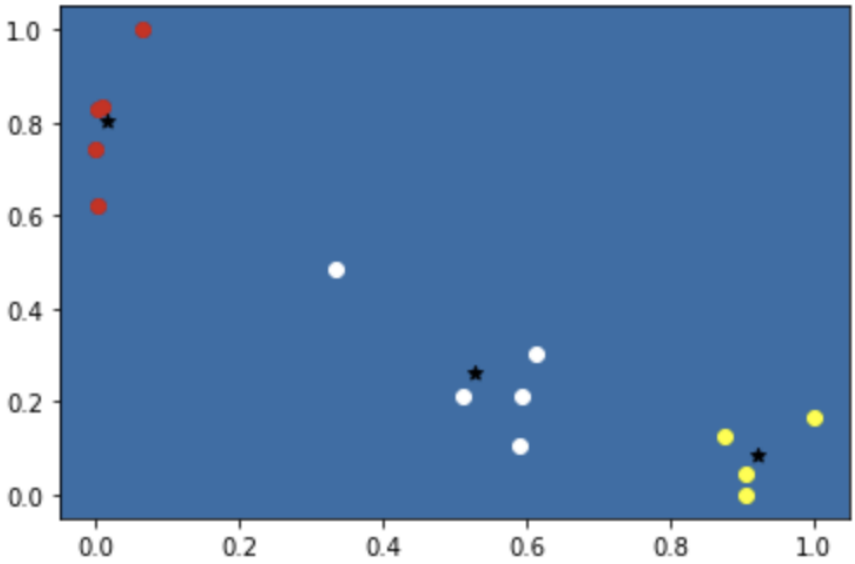

Figure 5: Plotting K-means graph
(red color represents batsman, white color Allrounder, yellow color as a bowler, black color represent centroid of each cluster)

5. CONCLUSION AND FUTURE WORK

The technique of discovering similarities between any set of data is known as clustering. Its applications include data mining, pattern identification, web document categorization, and grouping of geographic information such as seismic and so on. In this paper, k-means the popular clustering algorithm has been discussed. This paper discusses the cricket problem to analyse their performance and capability using the K-means algorithm. And it successfully categories players on basis of their into Batsman, Bowler and All-rounder. The same categorisation can be used to select the team for the match by finding data of the player for the given match venue or similar condition venue. This may help to select the right combination of batsman, bowler and allrounder for the match.

6. REFERENCES

[1] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining", 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1-7, 2013.
[2] Bini, B. S., & Mathew, T., " Clustering and Regression Techniques for Stock Prediction", *Procedia Technology*, pp. 1248–1255, 2016.
[3] Li, Y., & Wu, H., "A Clustering Method Based on K-Means Algorithm", *Physics Procedia*, *pp.* 1104–1109. 2012
[4] Mohebi, A., Aghabozorgi, S., Ying Wah, T., Herawan, T., & Yahyapour, R., " Iterative big data clustering algorithms: A review", *Software - Practice and Experience*, *pp.* 107–129, 2016
[5] Rahman, M. A., Chowdhury, A. K. M. R., Rahman, D. M. J., & Kamal, A. R. M, "Density based clustering technique for efficient data mining", Proceedings of 11th International Conference on Computer and Information Technology, ICCIT 2008, (Iccit), pp. 248–252, 2008
[6] Sharma, R., " ANALYSIS OF K-MEANS CLUSTERING ALGORITHM", IRJMETS, 2022
[7] Dr. Jhansi Rani P, Aditya Vidyadhar Kamath , Aadith Menon, Prajwal Dhatwalia , "SELECTION OF PLAYERS AND TEAM FOR AN INDIAN PREMIER LEAGUE CRICKET MATCH USING ENSEMBLES OF CLASSIFIERS ", Department of Computer Science & Engineering, CMR Institute of Technology, Bangalore, India, 2020
[8] Shah, S., Hazarika, P. J., & Hazarika, J., " A Study on Performance of Cricket Players using Factor Analysis Approach", International Journal of Advanced Research, pp. 656 - 660, 2019
[9] S, M., & E, M. "An Analysis on Clustering Algorithms in Data Mining", *International Journal of Computer Science and Mobile Computing*, pp. 334–340, 2014
[10] Deng, D., "DBSCAN Clustering Algorithm Based on Density." Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA, pp. 949–953, 2020