DATA CLASSIFICATION ON USER LOCATION WITH SPATIAL DATA

[1] K. MAHALAKSHMI, [2] P.S. HAMSHAVARTHINI

[1] Assistant Professor, [2] UG Research Scholar,
[1] Department of B. Com (Business Analytics),
[1] PGSR Krishnammal College for Women Coimbatore, Tamilnadu, India.

Abstract: The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. K-nearest neighbors (KNN) algorithm may be a sort of supervised machine learning algorithm which may be used for both classifications also as regression predictive problems. However, it is mainly used for classification predictive problems in many industries. It is widely used for large datasets to achieve the best result. The main objective of KNN algorithm is to classify the dataset and categorize the nearest neighbors. KNN is a classification algorithm which is used in this paper to classify the dataset and cluster the nearest location. Classifying the spatial dataset of Coimbatore using k-nearest neighbors algorithm and cluster the closest common location into a cluster. The clustered result will used to find the Euclidean distance to find the nearest group query.

Keywords: **Algorithm, Classify, Cluster, Location, nearest, group query.**

I. INTRODUCTION

In datamining and statistics k-nearest neighbor algorithm is a classification algorithm used to cluster the data with nearest or homogeneous data. This classification algorithm is used in many fields like machine learning, data compression, computer graphics, image analysis etc., there are several clustering algorithms and here we are using KNN algorithm to classify and cluster the spatial dataset of Coimbatore. KNN is a machine learning algorithm which is mainly used for classification predictive problems in many industries. Clustering is that the process of grouping of things which have similar characteristics into subsets where each subset will have their characteristic different from one another. In this study by loading the data into the dataset, classify the dataset as a systematic arrangement of groups or category and cluster the nearest and homogenous location into a group which is done by k-nearest neighbor algorithm.

II. OBJECTIVES

The present study has framed the following objectives. They are
1. To analyse the spatial dataset of Coimbatore
2. To classify and cluster the dataset using KNN algorithm
3. To offer the suggestion based on study

III. RELATED WORK

Spatial data is defined by spatial indexing structures and access methods that pose challenges to extract information from spatial data are wont to access spatial data. Space data mining is an analysis of indirect information and other patterns that are not directly contained in space databases [2].

In spatial data mining, data access methods are distinct from accessing methods in relational database and handling complex spatial objects using traditional data mining methods are difficult. Spatial data processing algorithms lacks in refining discovery patterns. The error patterns are increasing the search space of algorithm and there is a need to design an effective knowledge discovery algorithm for removing unnecessary data [5].

Location based service (LBS) is broadly defined as offering value added and interesting service or information to the user, like the user geo location coordinate where the user is present. rapid increase in smart phones plotted with the GPS device gives the way for this new technology [4].

Location analytics is a segment of business analytics that is used for the purpose of producing geographic intelligence. Arc News (2012) defined location analytics as focused on "thematic mapping [1].

SDIs can be defined as a means of employing technologies, policies, standards and stakeholders to assist in providing access to spatial data and enabling them to be disseminated. SDIs are centralized from the standpoint of knowledge management. They also support multi-participant involvement in long-term environmental projects [3].

Data Classification is that the conscious choice allocates a level of sensitivity to data because it's being created, amended, enhanced, stored, or transmitted. The classification of any property should be determined by the extent which the data must be controlled and secured and it is additionally supported by its value in terms of worth as a business asset. The classification of all property (including data and documents) is indispensable of a corporation and differentiate between that which is of little (if any) value, which is extremely sensitive and confidential [12].
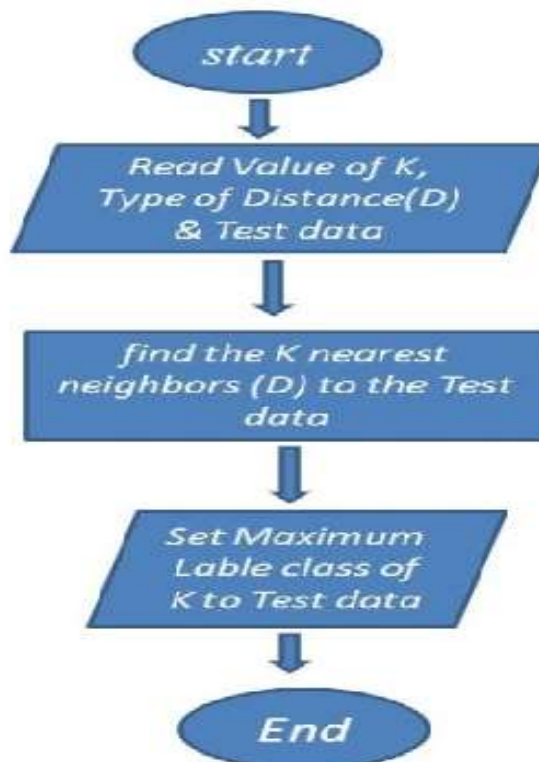
IV. METHODOLOGY

A. K-NEAREST NEIGHBOUR ALGORITHM

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.[6]

- The K-NN working can be explained on the basis of the below algorithm:

  - **Step-1:** Select the number K of the neighbors
  - **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
  - **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
  - **Step-4:** Among these k neighbors, count the number of the data points in each category.
  - **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
  - **Step-6:** Our model is ready.[6]

FLOWCHART OF THE PROCESS

start

Read Value of K, Type of Distance(D) & Test data

find the K nearest neighbors (D) to the Test data

Set Maximum Lable class of K to Test data

End

V.   IMPLEMENTATION AND RESULT

DATASET

Fig: 1.1 (dataset)

In the above Fig:1.1 shows the dataset of the Coimbatore spatial data set which is used to classify the data into category using KNN classification algorithm and cluster the nearest group location. This dataset incorporates the attributes namely feature, category, place, latitude and longitude of the spatial data. The feature gives information about the location of the selected place, the latitude and longitude give the exact geographically location of the Coimbatore location. The dataset is collected and used for the process of KNN classification which we are going to use in this paper
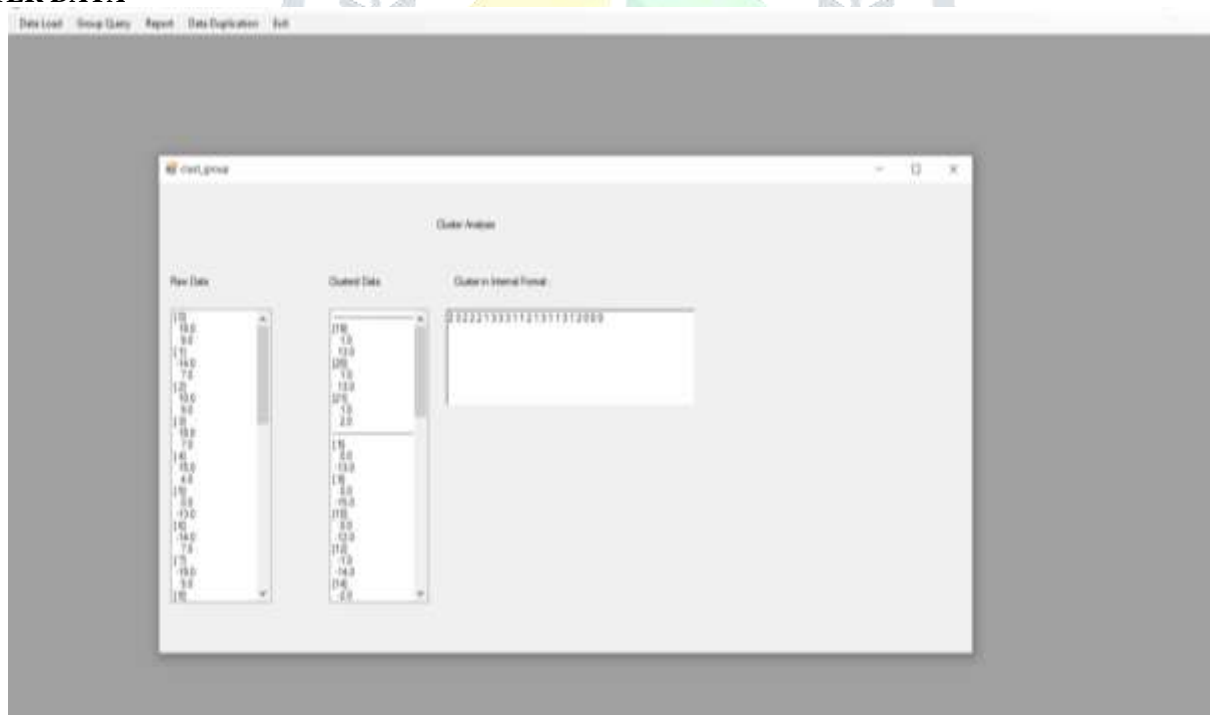
CLUSTER DATA

Fig: 1.2 (Clustered Attributes)

In the above Figure 1.2 dataset are classified using K Nearest Neighbor algorithm. Raw data is transferred to clustered data by using the KNN algorithm. The process of algorithm is to classify the entire dataset and categorize the data and cluster the nearest neighbor data. KNN algorithm assumes the similarity between the new case/data and available cases and the algorithm will cluster the nearest group attributes. On the application of KNN, the raw data has been converted into clusters for the further analysis.

VI.   CONCLUSION AND FURTHER WORK

In this paper, Microsoft visual studio is the tool which was used to analyse the spatial data to find the nearest group query with minimum distance and the main objective is to classify the dataset using KNN algorithm which means classifying the nearest data into category and cluster the data into groups. To analyse and classify the dataset apply the KNN algorithm to classify the Coimbatore dataset and categorize the data according to the result and cluster the nearest location, the raw data has been converted into clusters for the further analysis.

REFERENCES:

[1] Chua, A, L Servillo, E Marcheggiani and AV Moere (2016). **Mapping Cilento: Using geotagged social media data to characterize tourist flows in southern Italy.** Tourism Management, 57, 295-310.

[2] Lu.W, Han. J, and Ooi B.C 1993, **"Discovery of General Knowledge in Large Spatial Databases"**, In Proc. Far East Workshop on Geographic Information Systems pp. 275-289, Singapore, June.

[3] Maguire, D.J.; Longley, P.A. **The emergence of geoportals and their role in spatial data infrastructures.** *Comput. Environ. Urban Syst.* **2005**, *29*, 3–14.

[4] PeiZ,Lionel M, Ni N**. Spotlight: the rise of smartphone**. IEEE Distributed Systems Online. 2006; 7(3):1–5

 [5] Sumathi.N, Geetha.S and Dr. S. SathiyaBama 2008**, "Spatial Data Mining Techniques Trends and its Applications"**, Journal of Computer Applications, Vol.1, No.4, Oct-Dec.

 [6] https://towardsdatascience.com/

 [7] Umesh, R. M. (1988). A technique for cluster formation. Pattern Recognition, 21(4), 393–400. doi:10.1016/0031-3203(88)90052-0.

 [8] Yue. P, H. Zhou, J. Gong, L. Hu. **Geoprocessing in cloud computing platforms–a comparative analysis**. International Journal of Digital Earth, 6 (2013), pp. 404-425

[9] https://www.researchgate.net/

[10] M.F. Goodchild Citizens as sensors: **the world of volunteered geography** GeoJournal, 69 (2007), pp. 211-221.

[11] Mennis J., and Liu, J. W. 2005. Mining Association Rules in Spatio Temporal data: An Analysis of Urban Socioeconomic and Land Cover Change. Transactions in GIS, 9(1), pp.5-17.

[12] Umesh, R. M. (1988). A technique for cluster formation . Pattern Recognition, 21(4), 393–400. doi:10.1016/0031-3203(88)90052-0.

[13]https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm