ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

# ANALYSIS OF KNN QUERY USING CLUSTERING ALGORITHM IN DATA MINING

A.Roslin Deepa ,Ph.D Scholar,
Christhu Raj College, Bharathidasan University, Trichy, Tamilnadu - India.
E-Mail: roslindeepa@googlemail.com

Dr. Ramalingam Sugumar,Professor,
Christhu Raj College,Trichy, Tamilnadu - India.

*Abstract*— **Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis.[1] From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. (KNN) k-Nearest Neighbor method to find the hidden patterns inside the large dataset so as to transfer the retrieved information into usable knowledge for classification and prediction. In this paper we analysis of knn query using clustering algorithm in data mining.**
**Key Words: Data Mining, Clustering, (KNN), K-Nearest Neighbor, Clusters.**

## I. Introduction

The goal of this survey is to provide a comprehensive review of different clustering techniques in data mining. Clustering is a division of data into groups of similar objects. [3]Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. [2] It represents many data objects by few clusters, and hence, it models data by its clusters.

Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept. Data mining deals with large databases that impose on clustering analysis additional severe computational requirements. These challenges led to the emergence of powerful broadly applicable data mining clustering methods surveyed below.

## II. CLUSTER ANALYSIS

### ❖ Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers.

**Example:**

Use in finding Fraud usage of credit cards. Outlier Analysis may uncover Fraud usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the location and type of purchase or the purchase frequency.

### ❖ Clustering

Clustering analyses data objects without consulting a known class label. [2] The unsupervised learning technique of clustering is a useful method for ascertaining trends and patterns in data, when there are no pre-defined classes. There are two main types of clustering, hierarchical and partition. In hierarchical clustering, each data point is initially in its own cluster and then clusters are successively joined to create a clustering structure. [4]This is known as the agglomerative method. In partition clustering, the number of clusters must be known a priori. The partitioning is done by minimizing a measure of dissimilarity within each cluster and maximizing the dissimilarity between different clusters.

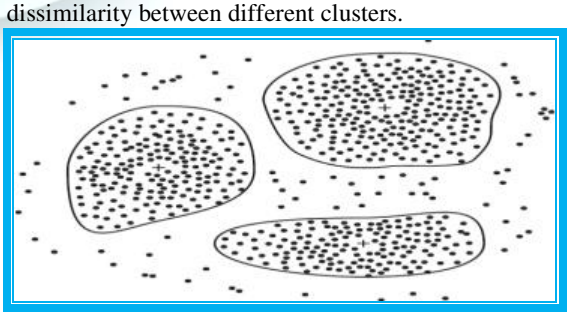

**Fig. 1 Cluster Analysis**

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

## III. CLASSIFICATION AND PREDICTION

Classification is the process of finding a model that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label .

Classification model can be represented in various forms such as

1) IF-THEN Rules

2) A decision tree

3) Neural network.

4) K-Nearest Neighbor etc.

Age(X, "Youth") AND income(X, "High") → Class(X, "A")
Age(X, "Youth") AND income(X, "Low") → Class(X, "B")
Age(X, "Middle_Aged") → Class(X, "C")
Age(X, "Senior") → Class(X, "C")

**Fig. 2 Classification model (Sample 1)**

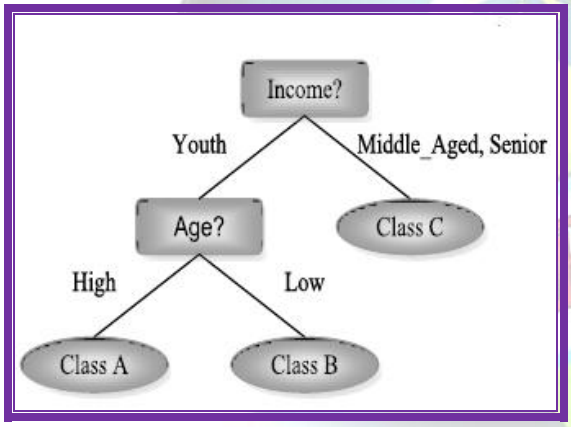


**Fig. 3 Classification model (Sample 2)**

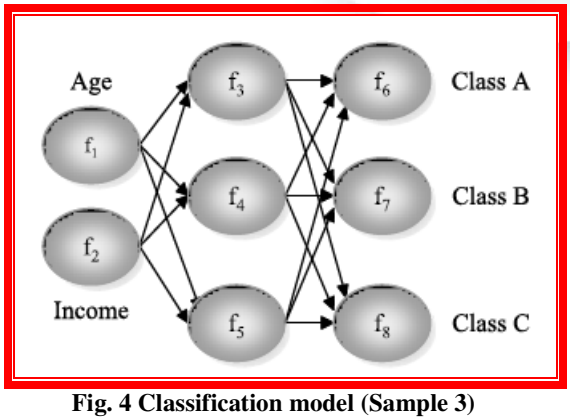


**Fig. 4 Classification model (Sample 3)**

## IV. HUB-BASED CLUSTERING

Hubness is an aspect of the curse of dimensionality pertaining to nearest neighbors which has only recently come to attention, unlike the much discussed distance concentration phenomenon. Let D

IRd be a set of data points and let Nk(x) denote the number of k-occurrences of point x D,

**i.e.,** the number of times x occurs in k-nearest neighbor lists of other points from D. As the dimensionality of data increases, the distribution of k-occurrences becomes considerably skewed . As a consequence, some data points, which we will refer to as hubs, are included in many more k-nearest-neighbor lists than other points. [4] In the rest of the text, we will refer to the number of k-occurrences of point x D as its hubness score. It has been shown that hubness, as a phenomenon, appears in high-dimensional data as an inherent property of high dimensionality, and is neither an artifact of finite samples nor a peculiarity of some specific data sets. Naturally, the exact degree of hubness may still vary and is not uniquely determined by dimensionality.

**Emergence of Hubs**

Hubs also exist in clustered data, tending to
be situated in the proximity of cluster centers. [4]In addition, the degree of hubness does not depend on the embedding dimensionality, but rather on the intrinsic data dimensionality, which is viewed as the
minimal number of variables needed to account for all pair wise distances in the data.

**Hubs and Data Clusters**

There has been previous work on how well
high-hubness elements cluster, as well as the general impact of hubness on clustering algorithms. A correlation between lowhubness elements and outliers was also observed. A low-hubness score indicates that a point is on average far from the rest of the points and hence probably an outlier. In highdimensional spaces, however, low-hubness elements are expected to occur by the very nature of these spaces and data distributions. [5] These data points will lead to an average increase in intracluster distance. It was also shown for several clustering algorithms that hubs do not cluster well compared to the rest of the points. This is due to the fact that some hubs are actually close to points in different clusters. Hence,
they lead to a decrease in intercluster distance. [6]This has been observed on real data sets clustered using state-of-the-art prototypebased methods and was identified as a possible area for performance improvement.

**Deterministic Approach**

A simple way to employ hubs for clustering
is to use them as one would normally use centroids.
In addition, this allows us to make a direct
comparison with the K-means method.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 20, April 2016

### Probabilistic Approach

Even though points with highest hubness scores are without doubt the prime candidates for cluster centers, there is no need to disregard the information about hubness scores of other points in the data. In the algorithm described below, we implemented a squared hubness-proportional stochastic scheme based on the widely used simulated annealing approach to optimization. The temperature factor was introduced to the algorithm, so that it may start as being entirely probabilistic and eventually end by executing deterministic K-hubs iterations.

### A Hybrid Approach

The algorithms do not require knowledge of
data/object representation, so all that is required is a distance/similarity measure defined for each pair of data objects. [7]

We will refer to this algorithm as hubness-proportional K-means (HPKM). It is nearly identical to HPC, the only difference being in the deterministic phase of the iteration, as the configuration cools down during the annealing procedure: instead of reverting to K-hubs

## V. CONCLUSION

Clustering techniques are used to partition
the transactions based on the relationships. Hubness measures are used to select the cluster centrality data points. Hubness-proportional K-means (HPKM) algorithm performs the clustering process with predefined cluster count. Kernel mapping model and shared neighbor clustering algorithm are enhanced with hubness mechanism. In this paper we analyzed of KNN query using clustering algorithm in data mining.

## VI. REFERENCES

1) **Analysing the Performance of Apriori Algorithm for by Combining with Clustering Techniques, Nisha Rani,Yamini Chouhan,Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org Volume 2, Issue 3, July (2015) ISSN:2395-5317 ©EverScience Publications.**

2) **APPROXIMATE K-NEAREST NEIGHBOUR BASED SPATIAL CLUSTERING USING K-D TREE Dr. Mohammed Otair Department of Computer Information Systems, Amman Arab University, Amman, Jordan Otair@aau.edu.jo, International Journal of Database Management Systems ( IJDMS ) Vol.5, No.1, February 2013. DOI: 10.5121/ijdms.2013.5108.**

3) **Optimal Clustering with Nearest Neighbor Relationships Ms. N. Narmatha,Dr. T. Senthil Prakash,Ms.S. Kokila,International Journal On Engineering Technology and Sciences – IJETS™ ISSN (P): 2349-3968, ISSN (O): 2349-3976 Volume 1 Issue 6, October 2014.**

4) **Survey of Clustering Data Mining Techniques, Pavel Berkhin, Accrue Software, Inc.Author s address: Pavel Berkhin, Accrue Software, 1045 Forest Knoll Dr., San Jose, CA, 95129; e-mail: pavelb@accrue.com.**

5) **Temperature and Humidity Data Analysis for Future Value Prediction using Clustering Technique: An Approach Badhiye S. S. 1, Dr. Chatur P. N.2, Wakode B. V.3.International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 1, January 2012).**

6) **Supporting KDD Applications by the k-Nearest Neighbor Join Christian Böhm and Florian Krebs University for Health Informatics and Technology, Innrain 98, 6020 Innsbruck, Austria Christian.Boehm@umit.at, Krebs.Florian@symplex.de**

7) **High Performance Data Mining Using the Nearest Neighbor Join Christian Böhm Florian Krebs University for Health Informatics and Technology University of Munich Christian.Boehm@umit.at krebs@dbs.informatik.uni-muenchen.de**