INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 01, January 2024, pp : 651-653

IMPROVING THE K VALUE IN CLUSTERING WITH THE K-NN ALGORITHM BY INCORPORATING THE EXPECTATION MAXIMIZATION ALGORITHM

M Poornima[1]

[1]Dept. of CS, Fatima College, India.

ABSTRACT

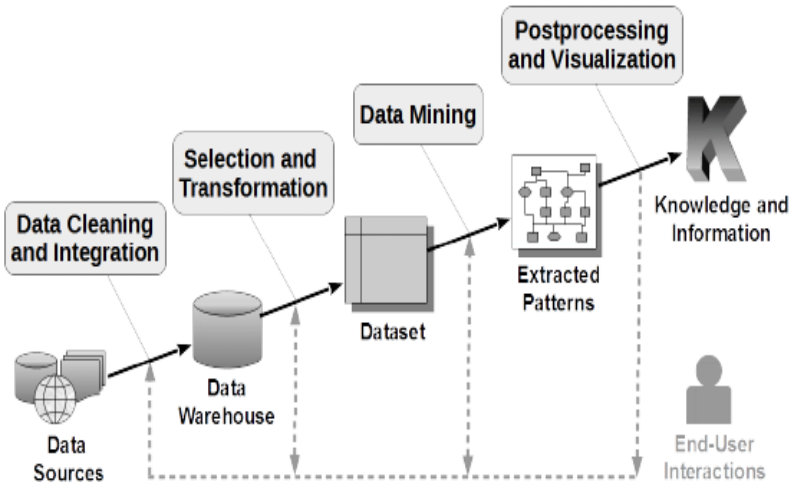
Data stands as the cornerstone of any study, and the research outcomes directly correlate with the data quality employed. The presence of missing data, denoting the absence of a value for a specific attribute in the dataset, poses a significant challenge. Researchers commonly turn to the k-nearest Neighbor (KNN) method to address this issue. However, KNN has drawbacks, particularly in selecting an appropriate value for k, which can impact classification performance.The accuracy of classification results in KNN is influenced by parameters such as the choice of k. Using more than one k parameter involves employing majority voting to determine the classification outcomes. A k value of 1 in KNN leads to tightly bound results, relying solely on the nearest neighbor for classification. Conversely, a higher k value in KNN results in more diffuse classification outcomes.This study aims to optimize the k parameters in UN tax clustering using the Expectation Maximization (EM) algorithm. The research delivers clustering information, considering optimized k values and those without optimization. Subsequent analysis of the clustered data reveals that the EM algorithm's k optimization enhances cluster results, reducing the error rate from 66% to 64%. Although not reaching the pinnacle of measurement accuracy, this improvement signifies a notable advancement in cluster outcome quality.

1. INTRODUCTION:

One approach to categorizing data involves the use of clustering, a method designed to organize data into groups or clusters based on similarities, ensuring that related data is grouped together. Various clustering algorithms exist, including partitional algorithms like Expectation-maximization and K-Means, hierarchical algorithms such as Centroid Linkage and Single Linkage, overlapping algorithms like Fuzzy C-Means, and hybrid approaches.Partitional algorithms address the challenge of arbitrary grouping, allowing a document to initially belong to one cluster in the process and then be reassigned to another cluster in subsequent processes. This algorithm aims to determine the Maximum Likelihood estimation of parameters in a probabilistic model. Its distinctive feature lies in its ability to classify unlabeled or untagged data, and its classification results consistently converge. The Expectation-Maximization algorithm comprises two phases: the Expectation phase and the Maximization phase. Prior to initiating the essential data grouping process, pre-processing steps, including cleansing, tokenizing, and parsing, are undertaken. The labeling of a cluster involves identifying the most prevalent actual label within a cluster and adopting it as the cluster's label. Through the utilization of the EM algorithm in the budget clustering process, classification occurs, and the optimal number of clusters is determined.

2. STAGES OF DATA MINING

Data mining is inherently a component of the Knowledge Discovery in Databases (KDD) process, rather than an autonomous and standalone technology. Within the KDD framework, data mining plays a pivotal role in the extraction and computation of patterns from data, as illustrated in Figure 1 below.



INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 01, January 2024, pp : 651-653

A. Data cleaning:

To eliminate the data noise (irrelevant data / dealing directly with the ultimate goal of data mining process, eg data mining that aims to analyze the results of the sale, then the data in the collection as "employee name", "age", and so on can -ignore) and inconsistent.

B. Data integration :

To combine multiple data sources.

C. Data selection :

To retrieve the appropriate data for analysis.

D. Data transformation:

To transform data into a form more suitable for mining. Data mining is the most important process in which a particular method is applied to generate the data pattern.

3. METHODOLOGY

The fundamental principle behind the K-Nearest Neighbor (KNN) involves identifying the shortest distance between the data under evaluation and K neighbors, which are the closest points in the training data set. This method falls within the realm of nonparametric classification, where the distribution of the data to be grouped is not a primary consideration. The simplicity and ease of implementation make this technique noteworthy.Similar to clustering methods, KNN involves classifying new data based on its proximity to multiple nearby data points or neighbors. The primary objective of the KNN algorithm is to categorize new objects by considering their attributes in comparison to a training sample. The classifier does not rely on any specific model but is memory-based.When given a query point, the algorithm identifies a specified number of objects (k training points) that are closest to the query point. The classification is determined through majority voting among the classifications of the k objects. The KNN classification algorithm predicts the value of the new query instance based on its proximity to neighboring points. The method is straightforward, operating by determining the K-nearest neighbors through the shortest distance from the query instance to the training sample.

The special case where the classification is based on the trainingdata diprekdisikan closest (in other words, k = 1) is called Nearest Neighbor algorithm.excess KNN (K-Nearest Neighbor):

1. Resilient to training data that has a lot of noise.

2. Effective if training data is huge.

The weakness of KNN (K-Nearest Neighbor):

1. KNN need to determine the value of the parameter k (the number of nearest neighbors).

2. Training based on distance is not clear on what kind of distance that must be used.

3. Which attributes should be used to get the best results.

4. The computational cost is high because the necessary calculation of the distance of each

query instance in the whole training sample.

Optimization Rated K :

The k-Nearest Neighbor (KNN) method employs supervised algorithms to classify the outcome of a new query instance based on the majority of categories within its KNN. The objective of this algorithm is to categorize new objects by considering their attributes and comparing them to the training sample. Unlike traditional models, the classifier does not rely on a specific model but is memory-based. When provided with a query point, the algorithm identifies a specified number of objects or K (training points) closest to the query point.

Expectation Maximization Clustering:

The Expectation Maximization algorithm is an unsupervised learning algorithm with the capability to explore a set of data that lacks labels or specific class targets. It achieves this by assessing the instances' values distributed into a Gaussian distribution, specifically a Gaussian mixture. The algorithm iteratively ascends to find the highest likelihood value for each instance, effectively determining its proximity to each cluster within the Gaussian mixture. The EM algorithm utilizes the mixture of Gaussian components in its process.

The EM algorithm fundamentally comprises two steps: expectation and maximization. In the expectation step, likelihood probability values are calculated. In the subsequent maximization step, these probabilities are refined by adjusting parameters within the Gaussian mixture, aiming to attain maximum likelihood. Several key aspects of the EM algorithm include:

1.Maximum Likelihood Estimation (MLE)

2.Mixtures of Gaussians

3.Estimation-Maximization (EM)

The EM algorithm employs Gaussian mixtures, emphasizing the exploration and refinement of the obtained distribution. Its primary tasks involve identifying each Gaussian within the mixture and optimizing each discovered Gaussian under optimal conditions (ensuring a more fitting model), a process known as maximization. This contributes to the clustering process.

Interpretation / Evaluation:

In the assessment and interpretation phase, the patterns derived from clustering data through the EM-cluster method are carefully evaluated. If the obtained results are deemed unsatisfactory, the process is iterated back to the clustering stage. This phase represents the concluding step in the Knowledge Discovery in Databases (KDD) process, enabling an examination to ascertain if identified patterns or information conflict with factual data. The information gleaned from the data mining process, in the form of patterns

Influence Selection of Parameter Values k:

During the test, the impact of optimizing the parameter k value on the success rate with the Expectation Maximization clustering algorithm will be analyzed. The k value represents the number of nearest neighbors considered when determining the cluster decision. Two distance parameters, namely Euclidean distance and Hamming distance, are used for optimization in simulation data, with k set to 13.Upon processing the data, the initial results without additional parameters revealed an incorrect identification of 11 clusters, accounting for 66%. However, through the optimization of parameters, the number of clusters was refined to 9, leading to a reduction in incorrect cluster identification to 64%.

4. CONCULSION

Utilizing clustering algorithms enables the identification of the EM-attainment status and budget plans for the upcoming year. The K-NN algorithm, specifically with k set to 13, is employed in this clustering process, proving effective for high-dimensional data types. The determination of the parameter k in the K-NN algorithm has a significant impact, contributing to the enhancement of the number of clusters in the forecast.

5. REFERENCES

[1] Gunadi, G., Sensuse, D., I., 2012, Application of Data Mining Methods Market Basket Analysis to book the product sales data by using algorithms Apriori and Frequent Pattern Growth .

[2] Ian H. Witten, f. E. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). (ASBurlington, Ed.) United States of America: Morgan Kaufmann.

[3] Yanto. R, Khoiriah. R., 2015, Implementation of Data Mining with Apriori Algorithm Method in Determining Drug Purchasing Patterns.

[4] Syaifullah. (2010). Implementation of Data Mining Algorithm Apriori Sales System, Amikom, Yogyakart a.

[5] Mardiani, 2014, Comparison Algorithm K-Means and EM for Clusterisasi Value Based Home School Students, CITEC Journal, Vol. 1, No. 4, 316-325.

[6] Yanto. R, Khoiriah. R., 2015, Implementation of Data Mining with Apriori Algorithm Method in Determining Drug Purchasing Patterns, CITEC Journal, Vol. 2, No. 2, 101- 113