

A K- Nearest Neighbors' based on Clustering for High Performances and High Volumes of the Data

Vijayalakshmi. K Associate Prof. Priya .M

Dept. of Computer Science Engg.
Bharathiyar College of Engineering and Technology
Karaikal, Puducherry, India

Abstract- A new classification method is presented which uses clustering techniques to augment the performance of K-Nearest Neighbour algorithm. This new method is called Nearest Cluster approach, [NC]. In this algorithm the neighbour samples are automatically determined using clustering techniques. After partitioning the train set, the labels of cluster centers are determined. For specifying the class label of a new test sample, the class label of the nearest cluster prototype is used. Computationally, the NC method of KNN is faster than the k-means with 2 times. Also the clustering techniques leads to find the best number of neighbours based on the nature of feature space. The proposed method is evaluated standard sales dataset. Experimental results show the excellent improvement both in accuracy and the time complexity in comparison with the KNN method. Many clustering algorithms have been developed to address the problem of very large data size as well as for high dimensional data. This algorithm have their own disadvantages but they are often impractical when the data is large in both aspects. The proposed system uses the KNN algorithm based clustering so the KNN queries can be evaluated with minimum seconds.

Keyword- KNN algorithm, cluster, Sales dataset, automatically determined.

I. INTRODUCTION

Many efficient algorithm for K-nearest neighbour queries that uses clustering, a pruning of the search space, and caching to improve performance. We call our algorithm ckSearch. The main goal of this work is to improve performance of queries in a k nearest neighbor (KNN) system [2]. In this paper ,a new search algorithm for supporting the nearest neighbour queries called ck search algorithm is proposed. we provide an overview of the KNN algorithm, and brief coverage of the performance challenges facing KNN implementations. This paper proposed with experimental domain of our application is described with standard analysis and then provide details on our approach. The KNN method simultaneously overcome the 'curse of dimensionality' problem and using the fast clustering of FensiVAT with nine state-of-the-art approaches. Its used to large sample size or high-dimensional data clustering to improving the performance level and reducing the time strategy in a large volumes of high-dimensional datasets in a minimum seconds.

II. RELATED WORK

1. Data clustering algorithms and application.

Data clustering Algorithm [1] describes the proposed solution uses a univariate time series model. It takes the price of a product as a parameter that influences systematically the prediction. The price influence is

computed based on historical sales data using correlation analysis and adjustable price ranges to identify products with comparable history. Compared to other techniques this novel approach is easy to compute and allows to preset the price parameter for predictions and simulations. Tests with data from the Data Mining Cup 2012 demonstrate better results than established sophisticated time series methods.

2. A vision architectural elements and future direction.

This paper [2] mainly explores when the agricultural industry faces grain crop price fluctuations and natural climate changes, it will take which level of price of grain crops and what probability of climate changes for developing a dynamic grain crop rotation model. In the existing system , Lin introduce the mixed strategy of game theory to construct a 2-player game. In consideration of the pursuit of the maximization of their own interests, the decision-making of dynamic grain crop rotation is the main focus of the previous paper, and it will be extended to a multiple stable dynamic grain crop rotation strategy cycle. And now the authors [2] develop a stationary Markov process as the basis for a final decision. Markov chain is a method frequently used in decision-making and is a model simple to be discussed.

3. Discretized streams: An Efficient and fault tolerant

Food production in India is largely dependent on cereal crops including rice, wheat and various pulses. The sustainability and productivity of rice growing areas is dependent on suitable climatic conditions. Variability in seasonal climate conditions can have detrimental effect, with incidents of drought reducing production. Developing better techniques to predict crop productivity in different climatic conditions can assist farmer and other stakeholders in better decision making in terms of agronomy and crop choice. Machine learning techniques can be used to improve prediction of crop yield under different climatic scenarios. This paper [3] presents the review on use of such machine learning technique for Indian rice cropping areas. This paper discusses the experimental results obtained by applying SMO classifier using the WEKA tool on the dataset of 27 districts of Maharashtra state, India. The dataset considered for the rice crop yield prediction was sourced from publicly available Indian Government records.

4. A Resilient distributed graph system spark

The term Big Data, refers to sizably voluminous data whose volume, variability, and velocity make it very arduous to manage, process or analyzed. To analyze this sizably voluminous kind of data Hadoop will be utilized. However, Processing is very time-consuming. To resolve this quandary & to decrement replication time one solution is to executing the job partially, where an approximate, early result becomes available to the utilize, afore completion of job. The Proposed system [4] gives a more incipient Map Reduce architecture that sanctions data to be divided for easier & early processing. This is not time consuming and amends system utilization for batch jobs as well. Proposed system presents a more incipient version of the Hadoop Map Reduce framework that fortifies on-Process aggregation, which sanctions & avails users to get early results of a job as it is computing. It will evaluate this technique utilizing authentic world datasets and applications and endeavor to amend the systems performance in terms of precision and time.

5. Machine learning in Apache spark.

The apparent of internet has lead to the data explosion which results in the emergence of Data mining. Extraction of useful knowledge based content and recognizing the patterns in the dataset are comprehended in the recent decade. Analyzing meaningful data and applying knowledge to various disciplines for monitoring is the important features in agriculture domain. India's economy is agriculture based, where majority of Indian population have agriculture and farming as main occupation. Analysis of large datasets in effective way requires understanding of appropriate techniques in data mining. The focus of this paper [5] is

to provide and build agricultural based information system for Customer and Farmer interaction where scalability, reliability and integrity of information can be access through cloud based technology. This paper aims to analyze and use data mining techniques specially Regression analysis to forecast the crop production. The forecasting of respective crops analyzes patterns in knowledge lie information of certain parameters and historical data.

6. A platform for fine grained resource sharing in the data centre.

Random Forest is an ensemble of classification algorithm widely used in much application especially with larger datasets because of its outstanding features like Variable Importance measure, OOB error detection, Proximity among the feature and handling of imbalanced datasets. This paper discusses many applications which use Random Forest to classify the dataset like Network intrusion detection, Email spam detection, gene classification, Credit card fraud detection, and Text classification. In this paper focus [7] each application is briefly introduced and then the dataset used for implementation is discussed and finally the real implementation of Random Forest algorithm with steps wise procedure and also the results are discussed. Actual Random Forest Algorithm and its features are also discussed to highlight the main features of Random Forest Algorithm more clearly.

III. PROBLEM STATEMENT AND PROPOSED SYSTEM

1. Proposed work

The K-nearest neighbor algorithm (KNN) is a well-known statistical search or learning method used in a wide range of problem solving domains: e.g., robotics navigation, data mining, and image processing]. In robotic navigation KNN is used to select an appropriate action of a robot by evaluating similar (K) instances from the 'nearest neighbor feature set' in training data. In forestry KNN is used to map satellite image data to inventory forest resources is used to classify, here the feature space include alcohol level, hue and wine opacity.

More formally, KNN finds the K closest (or most similar) points to a query point among N points in a d-2 dimensional attribute (or feature) space. K is the number of neighbors that are considered from a training data set and typically ranges from 1 to 20. Advantages of KNN algorithm include that it is fairly simple to implement and it is well suited for multi-modal classes. A key idea of our ckSearch algorithm is to improve performance by avoiding costly distance computations for the KNN search [10]. We use a divide-and-conquer approach. First, we divide the training data into clusters based on

similarity between the data points in terms of Euclidean distance. Next we perform a linearization of data points in each cluster for faster lookup. The data points in a cluster can be sorted based on their similarity to the center of the cluster [8]. Our data linearization process takes advantage of this similarity and produces metric indexes for each data point in a cluster.

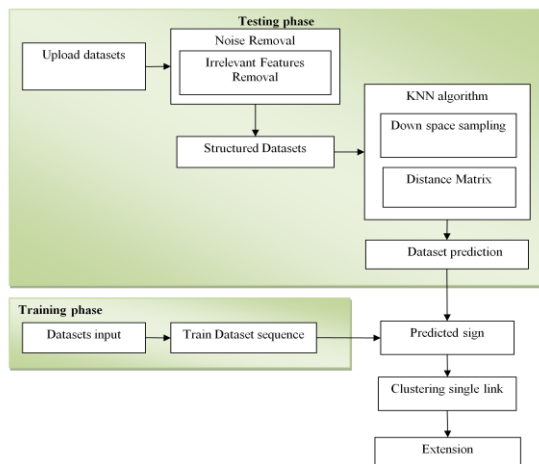


Fig.1 Architecture diagram for KNN algorithm using clustering process.

The Product will perform the following functions

- Dataset Acquisition
- Pre processing
- Clustering
- Feature Selection
- Classification

1. Dataset Acquisition

In this module is used to upload the sales prediction details. It contains the 'Year', 'Stock', 'Area of Sowing', 'Product', 'Quantity and Quality' and 'Production'.

2. Pre processing

Data pre-processing is an important step in the data mining process. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. In this module we remove noise words such as and stemming words.

3. Clustering

Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps the users to understand the natural grouping or structure in a data set. In this module, we can implement clustering algorithm to group the features.

4. Feature Selection

Feature selection is the process of selecting a subset of relevant, useful features for use in building an analytical model. The Feature selection helps narrow the field of

data to just the most valuable inputs, reducing noise and improving training performance

5. Classification

In this module, implement classification algorithm to classify the data, finally predict the yield production.

6. Algorithm

- Generally MapReduce paradigm is based on sending the computer to where the data resides!
- Map Reduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

7. Map stage- The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

8. Reduce stage- This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

- During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
- The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the
- cluster between the nodes.
- Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
- After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.

IV. CLUSTERING

Clustering is a technique in data mining to find interesting patterns in a given dataset. The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters information's into k groups, where k is considered as an input parameter [9]. It then assigns each information's to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then more computed and the process begins again. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical data and related fields. K-Means algorithm is a divisive, unordered method of defining clusters.

1. Feature Selection

In this module is used to select the features of the given dataset. Attribute selection was performed to determine the subset of features that were highly correlated with the class while having low inter correlation.

2. Classification

The KNN Classification Algorithm represents a statistical method as well as supervised learning method

for classification. It assumes a probabilistic model which allows us to solve the diagnostic and predictive problems. Bayes classification has been proposed which is based on Bayes rule of conditional probability [6]. Naïve Bayesian rule is a technique used to estimate the likelihood of a property from the given data set. The approach is called “naïve” because it assumes the independence between the various attribute values. Bayesian classification can be seen as both a descriptive and a predictive type of algorithm. The probabilities are descriptive and used to predict the class membership for a target tuple.

V. EXPERIMENTAL RESULT

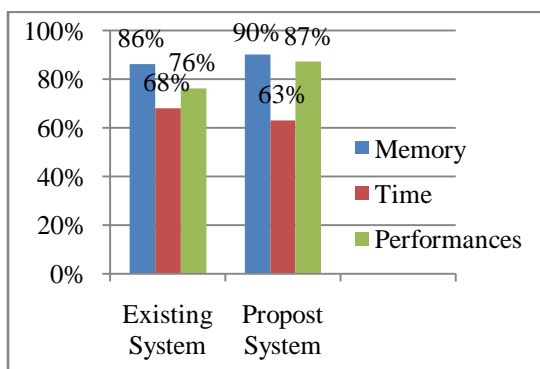


Fig. 2 Bar graph of existing and proposed system.

In this above graph is representing the comparison of the existing and proposed system classification algorithms. In KNN Algorithm is providing the accuracy is higher than the other classification method.

VI. CONCLUSION AND FUTURE WORK

Building a productive arrangement demonstrates for order issues with various dimensionality and diverse example measure is essential. The principle assignments are the determination of the highlights and the choice of the order strategy [11]. In this paper, we utilized PSO to perform include choice and after that assessed wellness esteems with a SVM, which was joined with the one-versus-rest technique, for five characterization profiles. Experimental outcomes demonstrate that our strategy disentangled element determination and the aggregate number of parameters required successfully, along these lines acquiring a higher order exactness contrasted with other element choice strategies. The proposed technique can fill in as a perfect pre-handling instrument to help streamline the component determination process, since it expands the characterization precision and, in the meantime, keeps computational assets expected to a base. It could likewise be connected to issues in different zones later on.

REFERENCE

- [1]. C. C. Aggarwal and C. K. Reddy, Data clustering: algorithms and applications. CRC Press, 2013.
- [2]. J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of things (iot): A vision, architectural elements, and future directions,” *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.
- [3]. M. Zaharia, T. Das, H. Li, S. Shenker, and I. Stoica, “Discretized streams: an efficient and fault-tolerant
- [4]. R. S. Xin, J. E. Gonzalez, M. J. Franklin, and I. Stoica, “Graphx: a resilient distributed graph system on spark,” in *First Int. Workshop on Graph Data Management Experiences and Systems*. June 2013.
- [5]. X. Meng, J. Bradley, B. Yuvaz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, and D. Xin, “Mllib: machine learning in apache spark,” *Journal of Machine Learning Research*, vol. 17, no. 34, pp. 1-7, 2016.
- [6]. R. S. Xin, J. Rosen, M. Zaharia, M. J. Franklin, S. Shenker, and I. Stoica, “Shark: SQL and rich analytics at scale,” in *Proc. of the 2013 ACM Int. Conf. on Management of Data (SIGMOD)*, June 2013.
- [7]. B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. H. Katz, S. Shenker, and I. Stoica, “Mesos: a platform for fine-grained resource sharing in the data center,” in the *14th USENIX Symp. on Networked Systems Design and Implementation (NSDI)*, March 2011.
- [8]. E. Bingham and H. Mannila, “Random paperion in dimensionality reduction: applications to image and text data,” in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 245–250.
- [9]. A. McCallum, K. Nigam, and L. H. Ungar, “Efficient clustering of high-dimensional data sets with application to reference matching,” in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2000, pp. 169–178.
- [10]. A. S. Shirshorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, “Big data clustering: a review,” in *International Conference on Computational Science and Its Applications*. Springer, 2014, pp. 707–720.