



K-Nearest Neighbour Search Algorithm by Using Random Projection Forests

KALIGITHI RAJANI^{#1}, V.SARALA^{#2}

^{#1} MSC Student, Master of Computer Science,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

^{#2} Assistant Professor, Master of Computer Applications,

D.N.R. College, P.G.Courses & Research Center, Bhimavaram, AP, India.

ABSTRACT

Data mining is the process of extracting valuable information from a large data source and this is the primary step in the process of knowledge discovery. One among the best classification algorithm for classification is K-nearest neighbors (kNN) for the process of knowledge discovery. Inspired by the huge positive reviews about the tree-based methodology over the last decades, we try to develop a new method for data search under random projection forests (rpForests). Here the random forest projection method is mainly used to search the specified document randomly from the several groups and try to find out the document which is found or not. As we are using K-NN algorithm if the file is not found in one group the same file can be immediately search in opposite group in random manner and if the file is found the search will be immediately stopped at that point and the result will be displayed immediately. If the file is not found then the search continues till the end and the process terminates at the final stage if the file is not present in the entire groups. By conducting various experiments on our proposed model we finally came to a conclusion that our proposed approach is best in finding the solution to search the documents in random projection manner by using K-NN in an accurate manner.

Key Words: Data Mining, Random Projection, Knowledge Discovery, Classification Algorithm, Positive Reviews.

1. Introduction

In general K-Nearest Neighborhood (KNN) Algorithm can be used to solve the two main problems like classification and regression. This is widely used by almost all MNC and large scale companies to solve the classification and regression problems which arise in their company. The following are the main aspects in MNC companies for using this K-NN algorithm:

1. The K-NN is mostly used because it is very easy to retrieve the desired output.
2. The K-NN gives accurate results
3. The K-NN algorithm optimizes the time complexity for solving the task.
4. The K-NN is best suited for prediction that compared with several other primitive classification methods.

Now let us elaborate some example to use KNN in overall scale:

	Logistic Regression	CART	Random Forest	KNN
1. Ease to interpret output	2	3	1	3
2. Calculation time	3	2	1	3
3. Predictive Power	2	2	3	2

Table 1: Represent the Overall Scale of Various Algorithms

From the table 1, we can see several algorithms are represented and they all are compared with the several parameters for identifying the best algorithm. One thing we can see is the proposed K-NN algorithm is having good reviews for all the different parameters which we considered to test the efficiency of classification. Hence we are really motivated with this K-NN algorithm for classification of desired input in our current application. The K-NN algorithm is assumed to be as one of the best instance-based learning, which is mainly used for finding the approximate distance from one object group to a new object which is found to be added recently. This will try to find out the distance between the new object and for the existing group and then check in which group this new object should be assigned. As we all know that this algorithm has very positive impact compared to many primitive algorithms of classification, this may increase its accuracy for some more level by normalizing the data.

2. LITERATURE SURVEY

Literature survey is that the most vital step in software development process. Before developing the tool, it's necessary to work out the time factor, economy and company strength. Once this stuff is satisfied, ten next steps are to work out which OS and language used for developing the tool. This literature survey is mainly used for identifying the list of resources to construct this proposed application.

MOTIVATION

Data Mining (DM) is the process of extracting the useful information from a large data source in order to extract the useful information from that raw data. In general the data mining process will require some integration of techniques from multiple area's such as statistics, machine learning, database technology, and spatial data analysis. The process of DM requires a very keen observation by taking a set of algorithms and also the task which is accomplished for that user. Almost several type of algorithms are used to fit the model in best way. In this proposed application we try to use deep collaborative filter for mining the text reviews given on set of products and also try to identify the rating present for that products or post.

In general the data mining algorithms can be classified based on the following ways like:

1. Model Based Approach:

This approach is mainly used to identify the purpose of the algorithm which is required to fit the data.

2. Preference Based Approach:

This is another form of approach in which we try to identify the preference or criteria that is used to execute the task.

3. Search Based Approach:

This is the third model in which we try to identify the data based on search time complexity.

So based on the above methods we can able to classify the algorithms and then perform the process of information extraction

In current day's one of the emerging application and demand for accurate K-NN search is random forest. For example, if we look at the robotic route planning [25], if there is any minute inaccuracy in a kNN search algorithm can able to possibly re-route the methodology in a wrong way and may cause wrong results. Another example of K-NN search classification for face identification in surveillance systems, the surveillance cameras try to identify a person based on the facial expression by comparing a set of target images which are already registered in the system. The accuracy of those observed images will mainly depend on the query image which was collected during the test phase and this is mainly used for identification of a person. If the image is not properly captured or classified then the target will not able to identify properly and this may leave wrong outcome for the end users. Till now there is lot of works [1],

[14], [16], [27], [47] processed on the fast computation of K-Nearest Neighborhood Search but no work is completely satisfied in providing the accurate results.

3. K-NEAREST NEIGHBOR SEARCH ALGORITHM BY USING RANDOM PROJECTION FORESTS

In this section we will mainly discuss about the proposed K-Nearest Neighbor Search Algorithm by Using Random Projection Forests. Now let us discuss about this proposed model in detail as follows:

MOTIVATION

Here the random projection classifier mainly contains the following attributes like

d : This is always small and for our application we assume that $d \leq 10$.

Let us assume that $d < p$

Where p is the random projection variable

We try to assume $C_{n,d}$ as the classifier which is used to classify the input data from a set of training samples.

Now the algorithm used for random projection using Gaussian method :

ALGORITHM FOR RANDOM PROJECTION

Result:

Data: and the test point $x \in \mathbb{R}^p$

Input: $\alpha \in [0, 1]$, B_1 , B_2 , $d \in \mathbb{N}$, a projected data base classifier $C_{n,d}$

for $b_1 = 1, \dots, B_1$

for $b_2 = 1, \dots, B_2$ do

Generate a Gaussian projection

Project the training data to give

Estimate by

End

Set , where

End

Let .

Here from the above algorithm we can able to search the documents from a set of clusters using random projection method and once the file is found the data search will be terminated and the result will be displayed as the out come.

4. IMPLEMENTATION PHASE

We have implemented the proposed concept on Java programming language with JSE as the chosen language in order to show the performance this proposed model. The front end of the application takes AWT, Swings and Socket Programming and as a Back-End Data base we try to use the MY-SQL as the database for storing the details about the nodes. The proposed application is divided into mainly 5 modules; now let us look about them in detail as follows:

1) Forest Network User Module

In this module, Forest network user Register to a particular Group in Router and Login by using his Username, Password and Group Name(Group1, Group2 and Group3).Then he will upload a file to associated node in the Router, Based on its minimal cost of node in group.

2) Forest Network Router Module

The Router is responsible to issue the query to the group as depicted by an arrow to the associated nodes, when the query request is transmitted from one node to another node in the group. The Router Can View the files in the router with their tags such as File Name, Digital Signature, Private Key, Secret Key, Username and Group Name. View the Registered Users in Router with their tags Username, Group Name, Password and Access Permission, Assign the cost for the Users in groups and also can view the distance details of all the users, if a user tries to download a file which is not available with any users then he will be revoked. He also responsible for unrevoked the user.

3) State of Nodes Module

A User state can be in three status, i.e., active, negatively terminated, or positively terminated, as defined below.

A) ACTIVE ELEMENT:

An active state is a state that allows further transitions. The query request will be further transmitted under an active state. Which is also called an active element.

B) POSITIVELY TERMINATED STATE:

An active state becomes a positively terminated state if the query request is answered by the active element. Upon receiving a query request in Categories.

C) NEGATIVELY TERMINATED:

An active state becomes a negatively terminated state if the query request is dropped by the active element.

4) FOREST NETWORK MANAGER

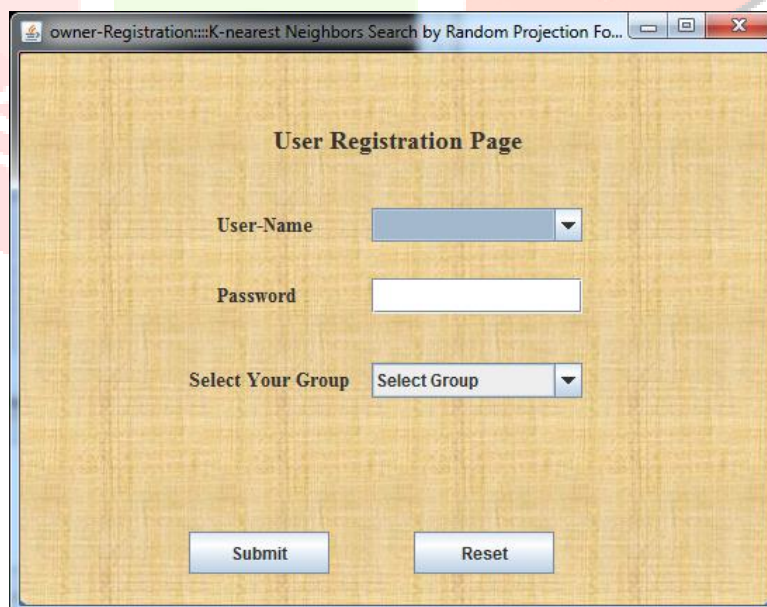
The Network Manager is responsible to view the Query Transactions with their tags Requested user, File Name, Secret key, Group Name, Responded user for that file details and he can also specify the access permission for searching the file in different group.

5) FOREST DATA CONSUMER

In this module Data Consumer is only responsible to download the file by specifying their file name and Secret key to router, it will search the availability of files within the same group. If file is not available within the same group then it should predict access permission with Social Network Manager for search in other groups. If file or secret key hadn't been matched in any group then he will be considered as an attacker.

5. EXPERIMENTAL RESULTS

1) Data user Registration Window



owner-Registration:::K-nearest Neighbors Search by Random Projection Fo...

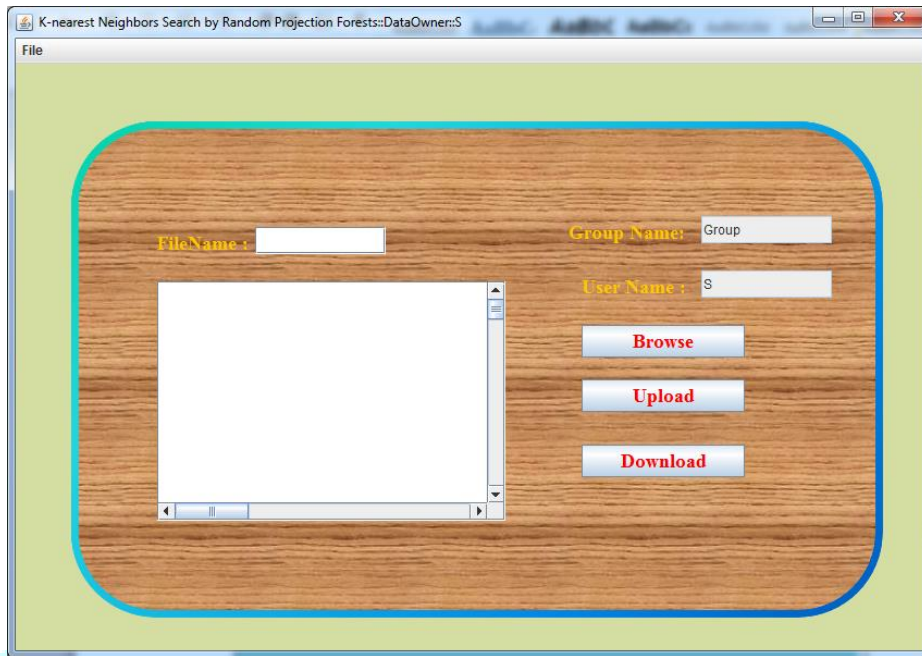
User Registration Page

User-Name

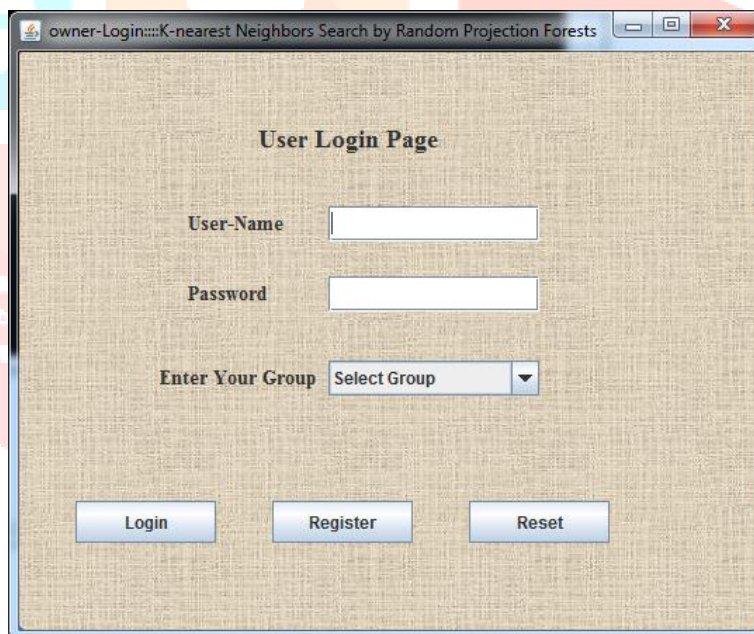
Password

Select Your Group

2) User Login page:



3) User Login Window



6. CONCLUSION

In this paper, we for the first time designed a model like random forest projection which is used to search the specified document randomly from the several groups and try to find out the document which is found or not. As we are using K-NN algorithm if the file is not found in one group the same file can be immediately search in opposite group in random manner and if the file is found the search will be immediately stopped at that point and the result will be displayed immediately. If the file is not found then the search continues till the end and the process terminates at the final stage if the file is not present in the entire groups. By conducting various experiments on our proposed model we finally came to a conclusion that our proposed approach is best in finding the solution to search the documents in random projection manner by using K-NN in an accurate manner.

7. REFERENCES

- [1] S. Arya, D. Mount, N. Netanyahu, R. Silverman, and A. Wu. An optimal algorithm for approximate nearest neighbor searching. *Journal of the ACM*, 45:891–923, 1998.
- [2] J. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.
- [3] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.
- [4] P. J. Bickel and L. Breiman. Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *The Annals of Probability*, 11(1):185–214, 1983.
- [5] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *Technical Report*, University of Minnesota, 2007.
- [6] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. In *Fortieth ACM Symposium on Theory of Computing (STOC)*, 2008.
- [7] P. J. Bickel and D. Yan. Sparsity and the possibility of inference. *Sankhya: The Indian Journal of Statistics, Series A* (2008-), 70(1):1–24, 2008.
- [8] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [9] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.
- [10] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.
- [11] S. Dasgupta and K. Sinha. Randomized partition trees for nearest neighbor search. *Journal Algorithmica*, 72(1):237–263, 2015.