International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

Use of Class Dependent Features in K-NN Classifier for the Classification of Encrypted Data

Nayana Jangale[1], Sandip Patil[2]
[1]Computer Department, Research Scholar, SSBTCOET Jalgaon, Maharashtra, India
[2]Computer Department, Associate Professor, SSBTCOET Jalgaon, Maharashtra, India

Article Info
Volume 9, Issue 2
Page Number : 44-50
Publication Issue :
March-April-2022
Article History
Accepted : 05 March 2022
Published: 15 March 2022

ABSTRACT

Cloud Computing stores the data in encrypted form. Classification of the data is required in many machine learning applications, so in the field of cloud computing, classification of the encrypted data is one of the major challenges. Extracting the class dependent features from the encrypted data and using these features in a well-known K-NN classifier can be used to classify the encrypted data at the cloud. In the proposed work we have encrypted and data extracted the correlation coefficient between the two or more variables and feeding it into the K-NN classifier and classifying the data. We have calculated the precision, recall, F1 score and accuracy of the proposed system and evaluated the performance of it with SVM and Naïve Byes classifiers.

Keywords: Cloud Computing, Features Selection, Privacy-Preserving, Data Encryption, Machine Learning

I. INTRODUCTION

The cloud computing (CC) paradigm is transforming the way organisations interact with data, notably how they store, access, and analyse data. CC is gaining popularity as a processing paradigm because of its low cost, flexibility, and lack of regulatory burden. Frequently, associations outsource their computations and data to the cloud. Despite the immense benefits the cloud provides, concerns about cloud security prevent enterprises from taking use of those benefits. When sending sensitive data to the cloud, it must be encrypted first. However, when data is jumbled, regardless of the underlying encryption scheme, executing data mining tasks becomes exceedingly difficult. With the growing popularity of cloud computing, consumers may now outsource their data management chores to the cloud. Large-scale services may now be provisioned using CC platforms like Google App Engine, Amazon EC2, and Microsoft Azure. The CC concept has changed how businesses interact with information for decades. Almost every corporation uses the cloud to better their data. Despite the great benefits of the cloud, privacy and security concerns prevent enterprises from taking full use of these benefits. Data that is private or extremely sensitive must be encrypted before being sent to the cloud. However, encrypting data makes data mining difficult without decrypting it.

A safe KNN classifier over semantically secure encrypted data is proposed in this work. Our protocol

selects relevant characteristics first, then encrypts them. No calculations occur after the encrypted data are sent to the cloud. So Alice gets no information. Our protocol also passes the following privacy standards: (1) The cloud should not be shown Data or any intermediate outcomes. (II) Only Bob should know q. Also, don't tell Bob anything else. (III) Data access patterns, such as k-nearest neighbours of q, should be kept secret from Bob and the cloud (to prevent any inference attacks). 4 Ensure privacy and security (v) Data privacy.

Section 1 introduces cloud computing, machine learning, and encryption, Section 2 reviews the literature, Section 3 shows the suggested system model, Section 4 explains the results and discussion, and Section 5 closes the article.

II. LITERATURE REVIEW

On the other hand, De Capitani et al. [2] outlined the major privacy concerns associated with data outsourcing and cloud Various methods were shown to handle these issues such as privacy in the cloud, privacy for users, privacy for stored data, and privacy for data access. This was achieved via data transmission, anonymized communication, collaborative querying, external data storage, and digital interactions. These methods protected the user's risk privacy. Attribute-based access control and user privacy choices reduced user privacy risk. Using selective access ensured confidentiality and integrity of stored data. Data access privacy was protected by ensuring query integrity.

Hu et al. [3] developed an efficient method based on privacy homomorphism and a safe traversal scheme. Using an index-based strategy, this system could handle large datasets. Authors devised secure protocol for KNN queries on R-tree index. Various optimization strategies for query processing protocols were also provided.

D. Bogdanov et al. This framework focuses on safe multi-party computing but also brings new concepts for enhancing application and development efficiency. The framework's major theoretical contribution is a set of computing procedures that use a ring of 32-bit integers instead of a finite field. Thus, writers may create simple and efficient protocols. They have developed a fully working SHAREMIND prototype that outperforms previous comparable frameworks.

Privileged data mining is addressed by R. Agrawal et al [10]. So, two parties control sensitive databases and want to run a data mining algorithm on the combined databases without releasing unneeded information. To secure privileged information while allowing its usage for research or other objectives drives their work. As such, it can be solved using well-known generic protocols. Authors concentrate on decision tree learning using the famous ID3 method to overcome these issues. Their protocol is much more efficient than generic methods, using fewer communication rounds and less bandwidth.

P. Zhang et al. [11] describe a privacy-preserving Naive Bayes Classifier for horizontally partitioned data. Data is often separated across companies. These firms may desire to use all data to improve prediction models while without disclosing their training data or cases to be categorised. NB Classifier is a basic yet effective baseline classifier. The authors use a Naive Bayes classifier to classify a dispersed dataset.

[12] provide a framework for mining association rules from transactions with categorical categories when the data is randomised to protect individual transaction privacy. While simple randomization may recover association rules while protecting privacy, the found rules can be used to detect privacy violations. In addition, the authors suggest a family of randomization operators that are much more successful than uniform randomization in reducing privacy intrusions. In this paper, the authors establish equations for an unbiased

support estimator and its variance, which enable recovering itemset supports from randomized datasets. A final set of findings verify the approach on actual datasets.

Using a flexible and highly combinatorial model of attribute value generalization, R. J. Bayardo et al. [13] They approached the challenge by using a tree-search technique to examine anonymizations from the most general to the most specialized. The approach uses cost lower bounding for node trimming and tail value reordering for tree rearranging. Authors also used data management tactics to avoid repeating sorting the full dataset, resulting in faster node assessment. They presented an iterated 2-phase greedy algorithm that outperforms current incomplete approaches for k-anonymization.

M. Kantarcioglu provided a safe approach for performing KNN classification from remote inputs in article [14]. Many privacy advocates see privacy vs data mining as an either/or scenario. Ensuring privacy and measuring the costs of obtaining information allows for more rational argument.

III. SYSTEM ARCHITECTURE

A. Proposed System

All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified. In this system, a novel approach is developed to successfully address the problem of privacy and security on the basis
of encrypted data exchanged with a cloud. Our primary emphasis is on the categorization problem.

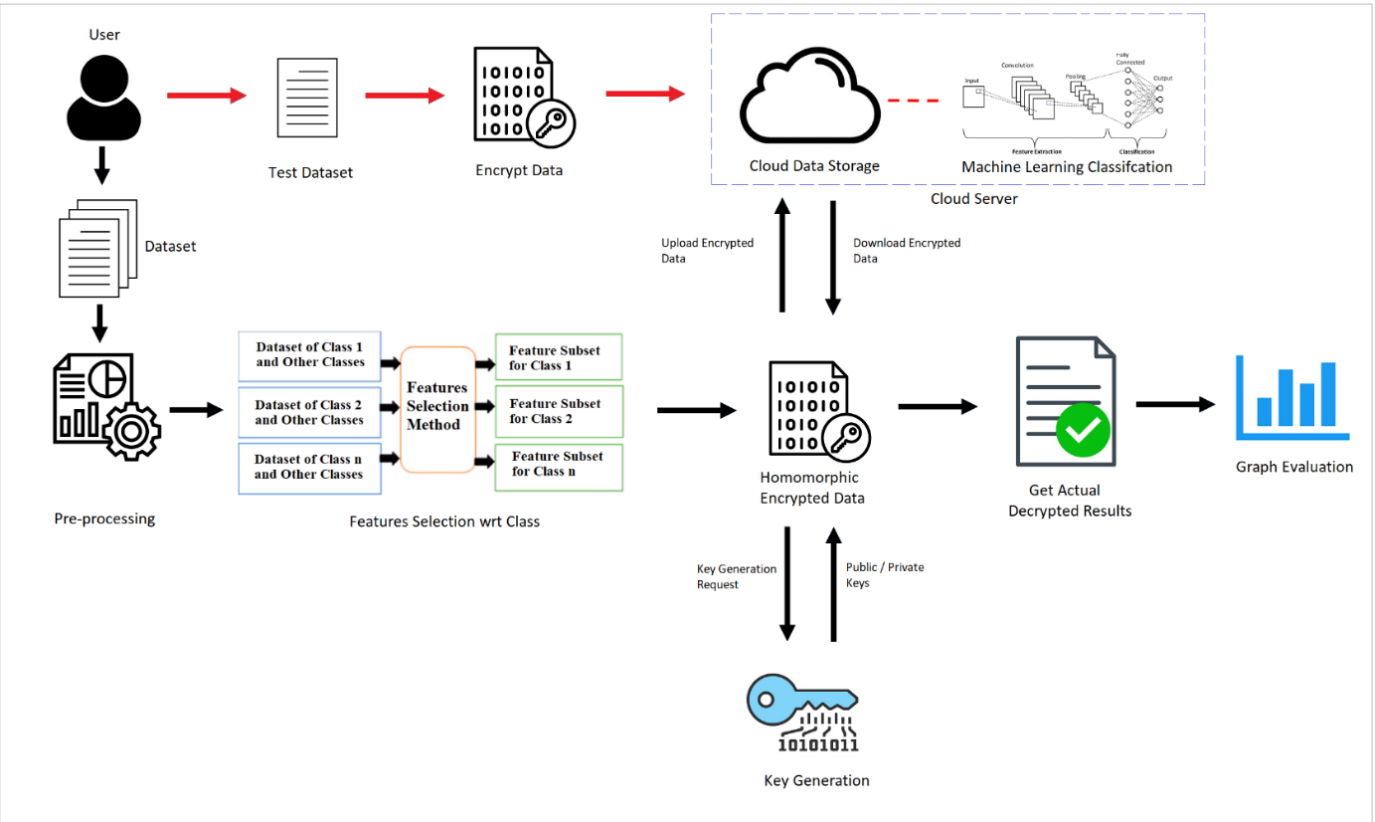

Fig. 1 System Architecture

When compared to other data mining jobs, classification is a familiar task. This method is used on encrypted data inside cloud computing. In our suggested framework, the SVM, KNN, and NB algorithms are employed for classification as well as to avoid the problem of classification on encrypted data.

To choose features from a dataset, class dependent algorithms are also utilised. The proposed technique saves memory and time when
compared to current algorithms, and the whole information system is run on the cloud. The proposed system includes numerous clouds, a user, and a data source, with the data source sending information to the cloud to be stored in encrypted format. When saving the data, the data source sends the encrypted private key to the cloud. When a customer requests information from the cloud, we use machine learning categorization. A client or a user sends an encrypted query to the cloud. On the encrypted data, both clouds undertake training and testing activities. The result is returned to the user in encrypted form, which the client decrypts.

B. Class Dependent Feature Selection

The class-dependent (CD) strategy selects a unique feature subset for the input dataset of a multiclass issue at each step (class). Because each class has discrimination in terms of categorising, the qualities that rely on classes may have distinct feature subsets. In a many-class classification issue, CD features are features whose discriminating performance varies greatly depending on the classes. As a result, as compared to the Class Independent feature selection approach, CD feature selection may enhance prediction accuracy while lowering feature measurement costs. Figure 2 is an illustration of the CD method.

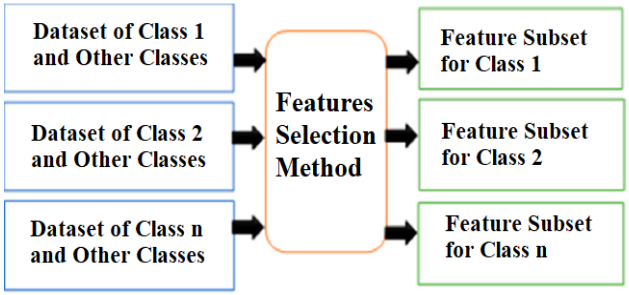

Fig 2. Class-dependent feature selection technique

Correlation Coefficient: The correlation coefficient is a measure of the linear connection between two or more variables. We can predict one variable based on the other using correlation. The reasoning for utilizing correlation to pick features is that the excellent variables are significantly associated with the aim. Furthermore, variables should be linked with the aim but uncorrelated with one another. We can anticipate one variable from the other if they are associated. As a result, if two characteristics are associated, the model only requires one of them since the second one adds no more information. After selection features from all independent classes all features are merge and duplicate features are removed and finally training file is generated which is pass to the hybrid classification model. Here for Features selection are consider feature importance and Correlation Coefficient.

The classification pipeline is made up of the following components:

1. Training text: This is the material from which our supervised learning model learns and predicts the needed class.
2. Feature Vector: A feature vector is a vector that includes information characterising the input data's qualities.
3. Labels: These are the preset categories/classes that will be predicted by our model.
4. ML Algo: This is the algorithm that allows our model to cope with text categorization (In our case: SVM, NB and KNN)
5. Predictive Model: A model that has been trained on previous data and can predict label values.

C. Algorithms

1. Encryption by Paillier [4]

The greatest common divisor of x and y is returned by the helper function gcd(x,y).
The least common multiple of x and y is returned by lcm(x,y).

Key generation

The following is how key generation works:

1. Select two huge prime numbers, p and q, at random and independently. Ascertain that gcd(pq,(p-1)(q-1)) = 1. If not, start again.
2. Determine n=pq.
3. Create the function L(x)=x1n.
4. Calculate as lcm (p-1, q-1).
5. Pick a random integer g in the set Z∗n2 (integers between 1 and n2).
6. Find the modular multiplicative inverse, μ=(L(gλmodn2)) – 1modn. If μ does not exist, go back to step 1.
7. What is the public key (n,g). Use this to encrypt data.
8. The secret key is λ. This is for decryption.

Encryption

Encryption applied to any m in the range 0≤m<n:

1. Choose a random number r from the range 0<r<n.
2. Determine the ciphertext c=gm.rnmodn2.

Decryption

Decryption requires a cipher-text generated by the preceding encryption procedure, with c in the range 0<c<n2.

1. m=L(cmodn2)modn is the plaintext.

2. KNN

Step-1: Determine the neighbours' K-numbers
Step-2: Compute the Euclidean distance between K neighbours.
Step-3: Determine the K closest neighbours based on the Euclidean distance determined.
Step-4: Count the number of data points in every category among these k neighbours.
Step-5: Allocate the new data points to the category that has the greatest number of neighbours.
Step-6: The KNM model is complete.

3. *Naive Bayes (NB)*

NB is Based on Bayes' theorem; a NB classifier is a basic probabilistic classifier. This classifier presupposes that the classes are conditionally independent of one another for a given class. The presence or absence of a single class characteristic (i.e. feature) is unrelated to the presence or absence of any other feature, according to the NB classifier. The mathematical method for determining whether a review will be positive or negative is

$$P(s|E) = \frac{P(s) * P(E|s)}{P(E)}$$

Where s denotes anticipated class output and E denotes test data for which the class is being predicted. P(s) and P(E|s) are acquired throughout the course of training. The NB classifier calculates the mean and variance of the variables required for classification using less training data [16].

4. SVM

SVM is a ML technique for classification and regression analysis that studies data and recognises patterns. Given a set of training samples, each of which is allocated to one of n classes. A SVM learning approach generates a model that divides new data into n categories. In an SVM learning algorithm model, the samples are represented as points in space. These points are mapped such that samples from distinct classes are separated by a greater gap. New samples are also mapped into the same space and given to a class depending on their location on either side of the split. [17]

IV. EXPERIMENTATION

A. Dataset Description

Car dataset to predict the with car is suitable to buy is predicted. Dataset is downloaded from https://archive.ics.uci.edu/ml/datasets/car+evaluation. It contains car acceptability, price, the size of luggage boot, buying, number of doors, capacity in terms of persons to carry, estimated safety of the car features.

B. Experimental Setup

All the experiments are performed on environment of Intel i5, 3.2 GHz, 8 GBs of RAM, 1 TB of hard disk, 512 GB of SSD and Windows 10 operating system.

NetBeans IDE is used and JAVA (JDK 1.8) technology is used.

C. Performance Parameters

The results of the proposed system are compared to those of Naive Bayes, SVM and KNN. In a 10-fold cross-validation test, the car dataset is categorized using these classifiers and the proposed multiclass model. The accuracy of chosen classifiers is assessed using a variety of performance indicators. The classification accuracy, precision, f1 measure and recall are used to confirm the results of the proposed model with other traditional classifiers. The formula for calculating these measures is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

V. RESULT AND DISCUSSION

Figures 3 and Figure 4 show the accuracy, precision, and recall scores comparison graphs of all the classifiers. The results of the performance evaluation of several categorization techniques are shown in Table 1. And Table 2. Naïve bayes had the lowest performance when compared to the others making it inappropriate for learning complicated structures for the subject data. All other classification models were surpassed by the proposed system class dependent feature selection with fine tune KNN gives the best accuracy, precision, recall, and F1 Score compare to other ML classifiers techniques.

TABLE I

PERFORMANCE PARAMETERS COMPARISON OF ML ALGORITHM (WITHOUT FEATURES SELECTION)

	Precision	Recall	F1 Score	Accuracy
Naïve Bayes	71.1	74.2	70.9	74.24
SVM	79	80.05	78.7	80.49
KNN	93.1	92.9	93.5	93.50

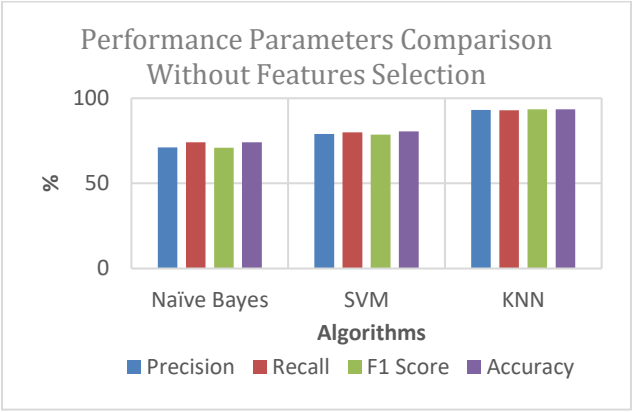

Figure 3. Class-dependent feature selection technique Performance Parameters Comparison Graph Without Features Selection.

TABLE III

PERFORMANCE PARAMETERS COMPARISON OF ML ALGORITHM (WITH FEATURES SELECTION)

	Precision	Recall	F1 Score	Accuracy
Naïve Bayes	71.5	74.4	71.2	74.42
SVM	79.4	80.7	79	80.67
KNN	94.2	94.3	94	94.5

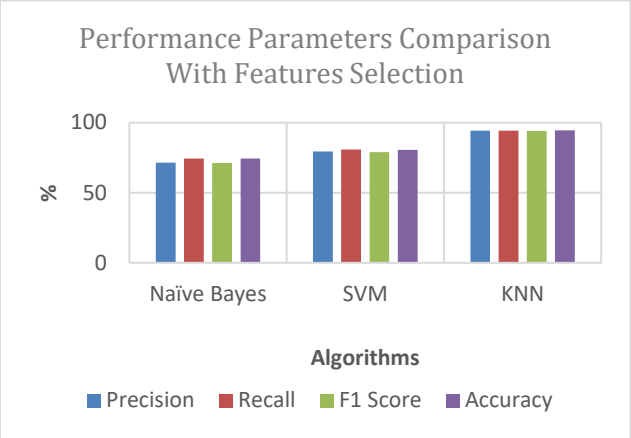

Figure 4. Class-dependent feature selection technique Performance Parameters Comparison Graph with Features Selection

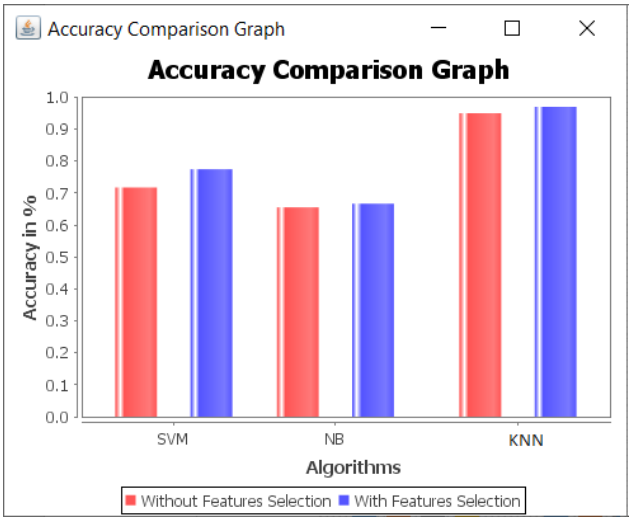

Figure 5. Class-dependent feature selection technique Performance Parameters Comparison Graph with Features Selection

Fig 5 shows the Accuracy Comparison Graph of Algorithms with and without features selection technique. KNN with class dependent feature selection methods outperforms all the other classification techniques and achieve accuracy of 94.5%.

VI. RESULT AND DISCUSSION

We look at three ways to classify things in this paper: Random Forest (RF), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). We combine those classifier methods with the class-dependent features selection method to get the best classifier technique for each classifier based on its accuracy. The experiment shows how important it is to pick the right features for classification data analysis. Table 2 also shows that the KNN algorithm is very accurate across all of the experiment groups. In terms of how well our experiment worked, it is clear that KNN with the class dependent features selection method is the best classifier. Along with these all classification algorithms are work on encrypted dataset which increases the security of our system. In addition, a public dataset of cars is used in all tests. As a result, even though traditional feature selection techniques are very useful and effective, we should not use every single feature in the dataset. As a result, it will have an effect on how quickly and accurately the data is processed, as well as how many different types of data there are and how big they are. 94.5 percent of the time, our suggested method with four characteristics works well.

VII. REFERENCES

[1] P. Song, C. Geng and Z. Li, "Research on Text Classification Based on Convolutional Neural Network," 2019 International Conference on Computer Network, Electronic and Automation (ICCNEA), Xi'an, China, 2019, pp. 229-232, doi: 10.1109/ICCNEA.2019.00052.

[2] S. De Capitani di Vimercati, S. Foresti, and P. Samarati, “Managing and accessing data in the cloud: Privacy risks and approaches,” in CRiSIS, pp. 1 –9, 2012.

[3] H. Hu, J. Xu, C. Ren, and B. Choi, “Processing private queries over untrusted data cloud through privacy homomorphism,” in Proc. IEEE 27th Int. Conf. Data Eng., 2011, pp. 601–612.

[4] P. Paillier, “Public key cryptosystems based on composite degree residuosity classes,” in Encrypt, pp. 223–238, 1999.

[5] B. K. Samanthula, Y. Elmehdwi, and W. Jiang, “k-nearest neighbor classification over

semantically secure encrypted relational data." E-print arXiv:1403.5001, 2014.

[6] C. Gentry, "Fully homomorphic encryption using ideal lattices," in ACM STOC, pp. 169–178, 2009.

[7] C. Gentry and S. Halevi, "Implementing gentry's fully- homomorphic encryption scheme," in EUROCRYPT, pp. 129– 148, Springer, 2011.

[8] A. Shamir, "How to share a secret," Commun. ACM, vol. 22, pp. 612–613, 1979.

[9] D. Bogdanov, S. Laur, and J. Willemson, "Sharemind: A framework for fast privacy-preserving computations, "in Proc. 13th Eur.Symp. Res. Computer Security, 2008, pp. 192–206.

[10] R. Agrawal and R. Srikant, "Privacy-preserving data mining," ACM Sigmod Rec., vol. 29, pp. 439–450, 2000.

[11] P. Zhang, Y. Tong, S. Tang, and D. Yang, "Privacy preserving Naive Bayes classification," in Proc. 1st Int. Conf. Adv. Data Mining Appl., 2005, pp. 744-752.

[12] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," Inf. Syst., vol. 29, no. 4, pp. 343-364, 2004.

[13] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in Proc. IEEE 21st Int. Conf. Data Eng., 2005, pp. 217-228.

[14] M. Kantarcioglu and C. Clifton, "Privately computing a distributed k-nn classifier," in Proc. 8th Eur. Conf. Principles Practice Knowl. Discovery Databases, 2004, pp. 279-290.

[15] Y. Elmehdwi, B. K. Samanthula, and W. Jiang, "Secure k-nearest neighbor query over encrypted data in outsourced environments," in Proc. IEEE 30th Int. Conf. Data Eng., 2014, pp. 664-675.

[16] Manek, A.S., Shenoy, P.D., Mohan, M.C. et al. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World Wide Web 20, 135–154 (2017).

[17] K. Huang, H. Jiang and X. Zhang, "Field Support Vector Machines," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 1, no. 6, pp. 454-463, Dec. 2017, doi: 10.1109/TETCI.2017.2751062.

Cite this article as :

Nayana Jangale, Sandip Patil, "Use of Class Dependent Features in K-NN Classifier for the Classification of Encrypted Data", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 2, pp. 44-50, March-April 2022. Available at

Journal URL : https://ijsrset.com/IJSRSET22926