International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Erection Trusted and Effective Request Services in the Cloud with RASP Data Perturbation

Rashmi Kadu[1], Sonali Patil[2]

[1]ME CSE- Final Year, BSIOTR Wagholi (Pune), Maharashtra, India

[2]Assistant Professor, CSE Dept, BSIOTR Wagholi (Pune), Maharashtra, India

Abstract: ***Cloud computing infrastructure madeinformation accessible to public which has become an appealing solution for the advantages on scalability and cost-saving. However, some data is so sensitive that the data owner does not want to move to the cloud unless the data confidentiality and query privacy are guaranteed. The RASP data perturbation method provides secure and efficient range query and kNN query services for protected data in the cloud. The kNN-R algorithm works with the RASP range query algorithm to process the kNN queries. The nearest neighbours concept involves interpreting each entry in the database as a point in space. k Nearest Neighbours (kNN) algorithm selects k entries which are closest to the new point. However kNN algorithm performs slowly on large databases since each new entry has to be compared to every other entry. There is a alternative method proposed which is fit to the large sized databases. This method is FCNN i.e. fast condensed nearest neighbour data reduction method. In this method the database is summarized by finding only the important data points. The main purpose of this method is to approximate the nearest neighbour algorithm, 1NN, with a smaller, more representative set of data points.***

Keywords: Cloud Computing, RASP, security,FCNN algorithm

1. Introduction

While using the exiting services of cloud computing, the growing concern is how to store, manage, and analyze a large volume of data while preserving the privacy [6]. The RAndom Space Perturbation (RASP) approach is used to construct practical range query and k-nearest-neighbor (kNN) query services in the cloud. The RASP perturbation is a unique combination of OPE i.e. order preserving encryption, dimensionality expansion, random noise injection, and random projection, which provides strong confidentiality guarantee [1] [5]. In FCNN several training set condensation algorithms have been introduced, also known as instance-based, lazy, memory- based, and case-based learners. These methods can be grouped into three categories depending on the objectives that they want to achieve competence preservation, competence enhancement, and hybrid approaches. The goal of competence preservation methods is to compute a training set consistent subset removing superfluous instances that will not affect the classification accuracy of the training set. Competence enhancement methods aim at removing noisy instances in order to increase classifier accuracy. Finally, hybrid methods search for a small subset of the training set that, simultaneously, achieves both noisy and superfluous instances elimination. Competence enhancement and preservation methods are combined in order to achieve the same objectives of hybrid methods [3].

2. Literature Survey

Title 1: RASP-QS: Efficient and Confident Query Services in the Cloud

Author of this paper Zohreh Alavi, Lu Zhou, James Powers, Keke Chen studied that data perturbation allows user to select one of the datasets. The perturbation parameters have to be generated OPE parameters are dataset-specific and the size of matrix A is subject to the dimensionality of the dataset. The perturbated data is sent to the server. The server then conducts multidimensional indexing on the perturbated data space. This demonstration shows a prototype for efficient and confidential range/kNN query services built on top of the random space perturbation (RASP) method. The RASP approach provides a privacy guarantee practical to the setting of cloud based computing, while enabling much faster query processing compared to the encryption-based approach. This demonstration will allow users to more intuitively understand the technical merits of the RASP approach via interactive exploration of the visual interface [4]. The main purpose of this demonstration is to show the key ideas of the RASP-based query processing approach for efficiently and confidentiality hosting query services in the cloud.

Title: Fats Condensed Nearest Neighbor Rule

Author of this paper Fabrizio Angiulli presents a novel algorithm for computing a training set consistent subset for the nearest neighbor decision rule. The algorithm, called FCNN rule, has some desirable properties. Indeed, it is order independent and has sub- quadratic worst case time complexity, while it requires few iterations to converge, and it is likely to select points very close to the decision boundary. [3] The comparison took place between the FCNN rule with state of the art competence preservation algorithms on large multidimensional training sets, showing that it outperforms existing methods in terms of learning speed and learning scaling behavior, and in terms of size of the model, while it guarantees comparable prediction accuracy. This paper presented a novel order independent method for computing a training set consistent subset for NN rule and compared it with existing state of the art competence preservation method. The observed superior learning speed of the new method is substantiated by the learning behavior comparison. This work can be extended in

several ways, e.g. studying the impact of different metrics on the FCNN rule and the behavior of FCNN- based hybrid method.

3. Existing System

Random space perturbation (RASP) approach to constructing practical range query and knearest-neighbor (kNN) query services in the cloud. The RASP kNN query service (kNN-R) uses the RASP range query service to process kNN queries. The nearest neighbours approach involves interpreting each entry in the database as a point in the neighbours are taken from a set of objects for which the class or the object property value is known. The nearest neighbours approach involves interpreting each entry in the database as a point in space. Then, the similarity of two points is measured by the distance between them. The nearest neighbours approach then classifies the new sample by looking at the classifications of those closest to it. In the k Nearest Neighbours (kNN), this is achieved by selecting the k entries which are closest to the new point. The best choice of k depends upon the data. Generally, large values of k reduce the effect of noise on the classification [7]. The accuracy of the kNN algorithm can be degraded by the presence of noisy or irrelevant the features, or if the feature scales are not consistent with their importance. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant (e.g. the same measurement in both feet and meters) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called feature eta in order to extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this ion reduced representation instead of the full size input. Feature extraction is performed on raw data prior to applying k-NN algorithm on the transformed data in feature space [8].

4. Proposed System

A. Fast Condensed Nearest Neighbours (FCNN)Data Reduction

Data Reduction: The database is summarized by finding only the important data points. Data points in the training set are divided into three types [3], Outliers, Prototypes, Absorbed. The purpose of this method is to be able to approximate the nearest neighbour algorithm, 1NN, with a smaller, more representative set of data points.

- Outliers are points whose k nearest points are not of the same class.
- X={x1, x2,..., xn} (without outliers) P={x1}
- We scan all elements of X and move individual elements to P if their nearest prototype (their nearest element from P) has a different class label
- Repeat until no more new prototypes are found.

Properties of CNN

- Absorbed points are the points which are not prototypes.
- CNN reduces the amount of data necessary for classification.
- Points are labelled as either prototypes, outliers or absorbed points.
- Absorbed points and outliers are then no longer used in classification tasks, validation tests or maps of the data set.

B. Algorithm

Algorithm FCNN (T: training set)

1) Initialize the set S to the empty set
2) Initialize the set ΔS to the set Centroids (T)
3) While the set ΔS is not empty:
4) Augment the set S with the set ΔS
5) Initialize the set ΔS to the empty set
6) For each object y in the set S, insert into ΔS the representative object of the Voronoi enemies of y in the T w.r.t S
7) Return the set S

5. System Architecture

Each record x in the outsourced database contains two parts: the RASP-processed attributes

D′ =F(D,K) and the encrypted original records, Z =E(D,K′), where K and K′ are keys for perturbation and encryption, respectively. The RASP-perturbed data D′ are for indexing and query processing [2]. There are a number of basic procedures in this framework [5]:

1) F(D) is the RASP perturbation that transforms the original data D to the perturbed data D';
2) Q(q)transforms the original query q to the protected form q′ that can be processed on the perturbed data;
3) H(q′,D′) is the query processing algorithm that returns the result R′.

Figure 1 shows the system architecture for both RASP-base range query service and kNN service.

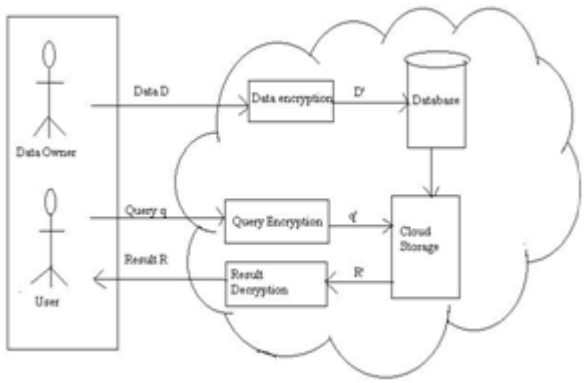

Figure 1: System Architecture of RASP

6. Modules

A. User Module

In this module, Users are having authentication and security to access the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

B. Dimensional Reduction

Dimensional are derived from R-tree like algorithms, where the axis-aligned minimum bounding region (MBR) is the

construction block for indexing the multidimensional data. For data, an MBR is a rectangle. For higher dimensions, the shape of MBR is extended to hyper-cube. The MBRs in the R-tree for a dimensional dataset, where each node is bounded by a node MBR. The R-tree range query algorithm compares the MBR and the queried range to find the answers.

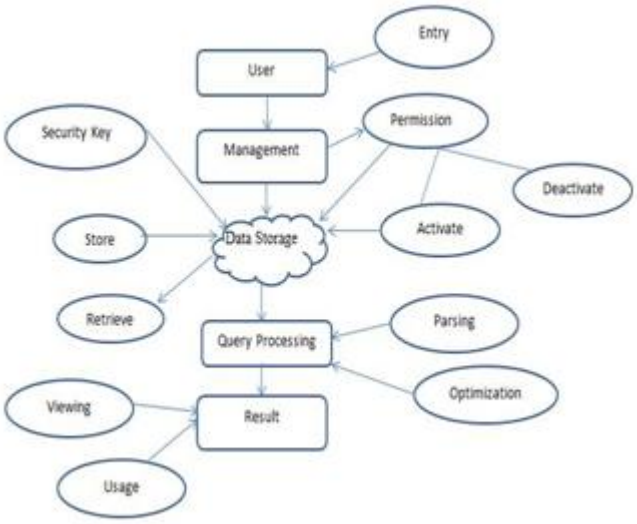

Figure 2: Dataflow Diagram

C. Performance of FCNN Rule Processing

In this set of experiments, we investigate several aspects of CNN query processing.

1) We will study the cost of (n, δ)-Range algorithm, which mainly contributes to the server-side cost.
2) We will show the overall cost distribution over the cloud side and the proxy server.
3) We will show the advantages of FCNN over another popular approach: the Casper approach for privacy-preserving FCNN search.

D. Preserving Query Privacy

Private information retrieval (PIR) tries to fully preserve the privacy of access pattern, while the data may not be encrypted. PIR schemes are normally very costly. Focusing on the efficiency side of PIR, Williams et al. use a pyramid hash index to implement efficient privacy preserving data-block operations based on the idea of Oblivious RAM. It is different from our setting of high throughput range query processing. Hu et al. addresses the query privacy problem and requires the authorized query users, the data owner, and the cloud to collaboratively process FCNN rules. However, most computing tasks are done in the user's local system with heavy interactions with the cloud server. The cloud server only aids query processing, which does not meet the principle of moving computing to the cloud.

7. Results

A. Input Given

Problem Definition

A training database which trains us to know what the different types of things look like.We are having a database of the characteristics land admin, state, country, landmark, facility. Then we will take a new sample and want to know what classification it should be. The classification is based on the items of the training database, the new sample is similar to Aim is to use this database to give a new person a perfect landmark and facility. Again, we want to classify it with the type of landadmin it is most similar to.

Requirement:

1) Database Creation:

We start with a database of objects who's classification we already know.

The datasets, "Hospitals.csv, Rest.csv, Toll_Transaction.csv" are taken from UCI repository. The website of UCI Repository is http://archive.ics.uci.edu/ml/

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. This repository currently maintain 308 data sets as a service to the machine learning community.

2) Origin:

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University .

3) Creator:

Harrison, D. and Rubinfeld, D.L.

'Hedonic prices and the demand for clean air', J. Environ. Economics& Management, vol.5, 81-102, 1978.

Data Set Characteristics:	Multivariate	Number of Instances:	506	Area:	N/A
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	14	Date Donated	1993-07-07
Associated Tasks:	Regression	Missing Values?	No	Number of Web Hits:	128761

Figure 3: Database specification

3) Data Set Information:

The database used in this project contains concerns housing values in suburbs of Boston

Software Requirement:

1) WampServer:

WampServer is a Windows web development environment. It allow to create web applications with Apache2, PHP and a MySQL database. Alongside, PhpMyAdmin allows to manage database easily.

2) MySQL:

The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. Free-software-open source projects that require a full-featured database management system often use MySQL. MySQL is also used in many high-profile, large-scale World Wide Web products, including Wikipedia, Google (though not for searches), Facebook, and Twitter.

Figure 4: Viewing a Dataset

B. Output Observed

Table 1: Comparison of Nearest Neighbour Techniques

Sr No	Technique	Key Idea	Advantages	Disadvantages	Target Data
1	K nearest Neighbor (kNN) [8]	Uses nearest neighbor rule	1. training is very fast 2. Simple and easy to learn 3. Robust to noisy training data 4.Effective if training data is large	1. Biased by value of k 2.Computation Complexity 3.Memory limitation 4. Being a supervised learning lazy algorithm i.e. runs slowly 5. Easily fooled by irrelevant attributes	large data samples
2	Condensed nearest neighbor (CNN) [9,10,11]	Eliminate data sets which show similarity and do not add extra information	1. Reduce size of training data 2. Improve query time and memory requirements 3.Reduce the recognition rate	1. CNN is order dependent; it is unlikely to pick up points on boundary. 2. Computation Complexity	Data set where memory requirement is main concern

Performance of KNN:

- In order to investigate how good the initial training set is, a procedure called cross validation gets used.
- This involves running the kNN algorithm on each of the points in the training set in order to determine whether they would be recognized as the correct type
- It can clearly be seen that including more random noise points in the training set increases the number of cross validation errors
- As the number of random noise points becomes very large, the percentage of points which
- fail the cross validation tends to 50%

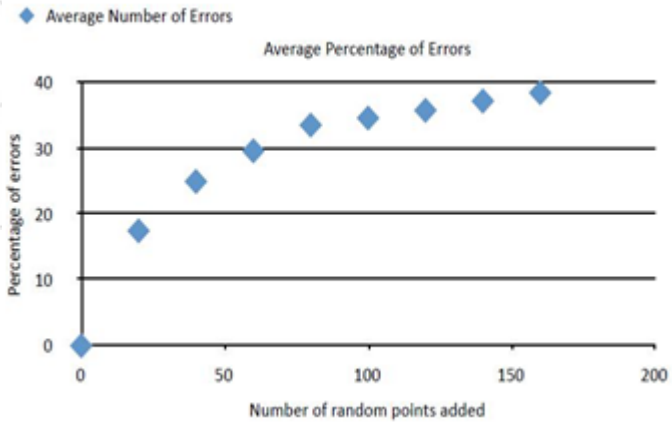

Figure 5: Performance of kNN algorithm

Problems with kNN:

- kNN algorithm is difficult to implement for some datatypes. This is because it relies on being able to get a quantitative result from comparing two items
- Slow for large databases: Since each new entry has to be compared to every other entry

Performance of FCNN algorithm:

- Classifying a new sample point is now faster, since we don't have to compare it to so many other points.
- This is the trade of we have to make, between speed and accuracy is better than kNN algorithm. Percentage of points classed as outliers increased dramatically.
- Percentage of points classed as absorbed decreased.
- Percentage of points classed as prototypes increased slightly

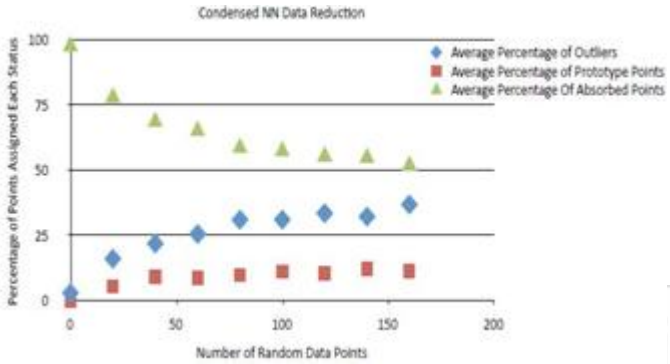

Figure 7: Performance of FCNN algorithm

Figure 6: Application Homepage

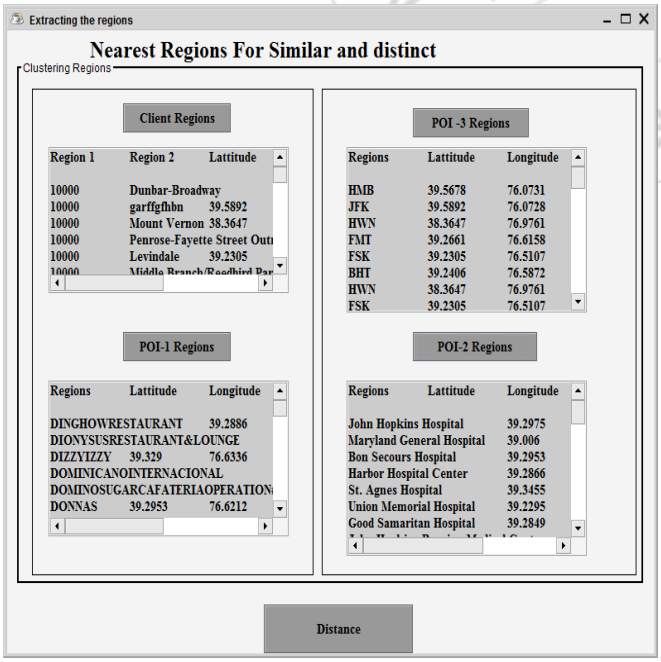

Figure 8: Extracting the Regions

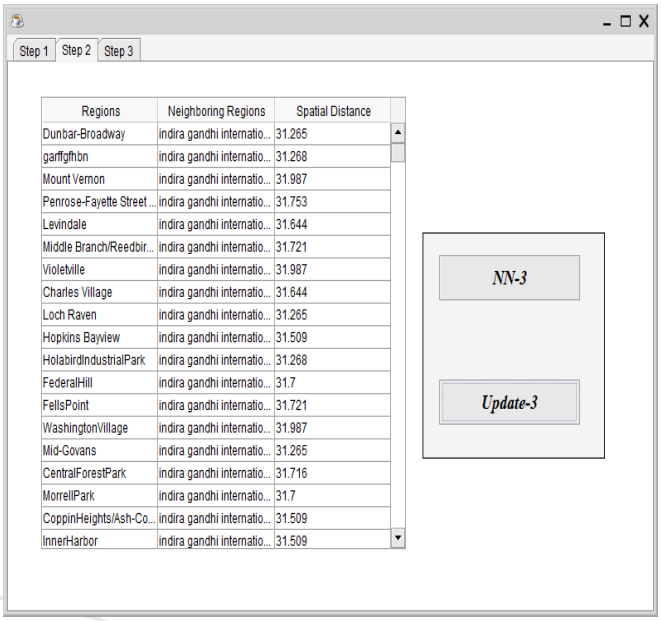

Figure 9: Calculate Distance Using NN-3 Algorithm

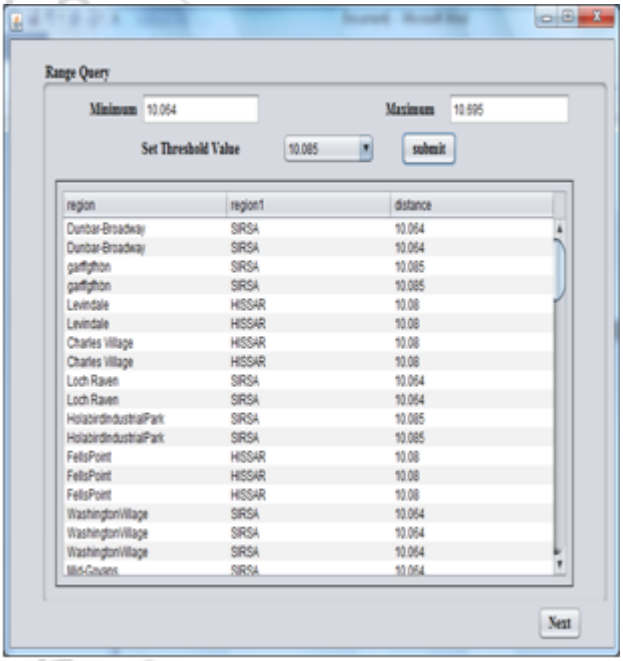

Figure 10: Calculate Distance Using Range Query

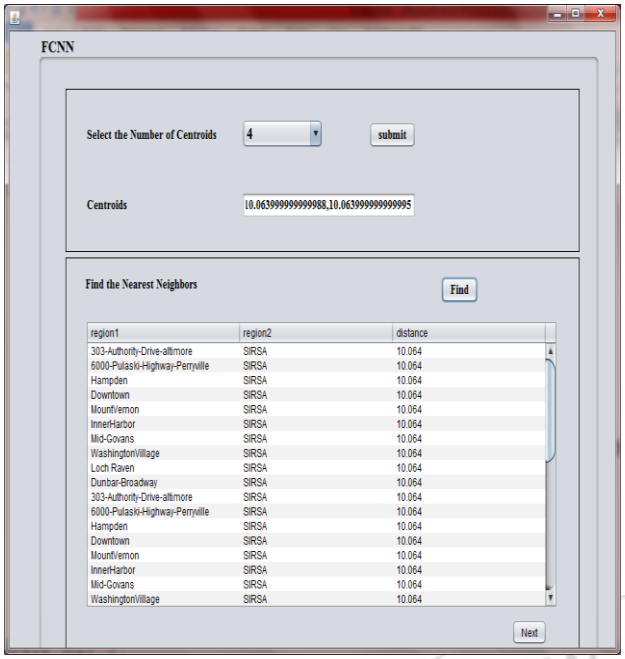

Figure 11: Distance calculation Using FCNN Algorithm

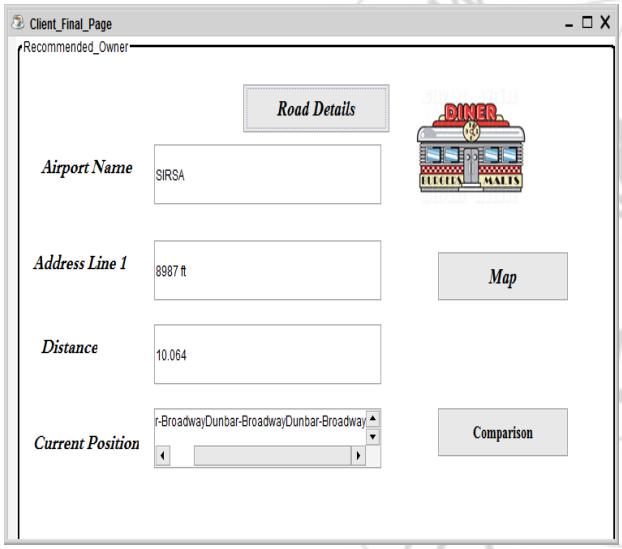

Figure 12: Entering Road Details

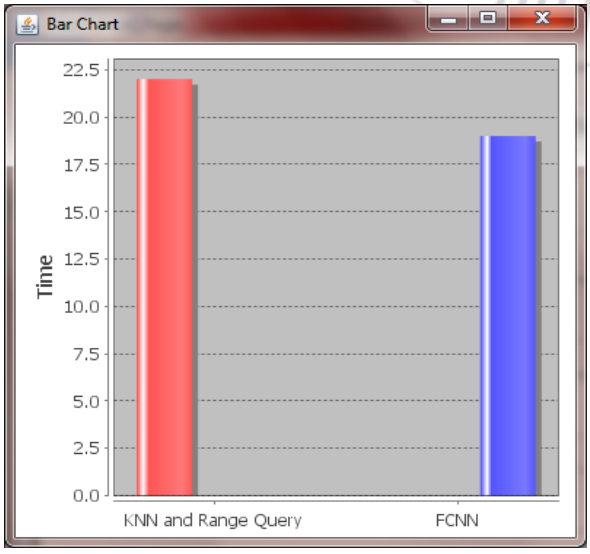

Figure 13: Comparison between KNN, Range Query and FCNN

8. Conclusion

RASP aims to preserve the topology of the queried vary in the rattled area, and permits to use indices for efficient vary question process. With the topology-preserving features, one can develop economical vary question services to realize sub. kNN classifies unknown instances based on a majority vote of the k nearest examples from the training set. The most frequent class label amongst these k nearest examples is then assigned to our unknown instance. Where as in CNN, the condensed nearest neighbour rule for data reduction, The database is summarized by finding only the important data points. It reduces the data set to a condensed data set. It labels points as either prototypes, outliers or absorbed points. Only prototypes are used in classification tasks, validation tests and maps, as they are considered to be more or less representative of all of the points in the initial data set. The performance of FCNN is better than kNN for large dataset as it find out prototypes first and then perform kNN algorithm later.

References

[1] B. Wang, B. Li, and H. Li, "Public Auditing for Shared Data with Efficient User Revoation in the Cloud," *in the Proceedings of IEEE INFOCOM2013, 2013, pp. 2904–2912.*

[2] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing," *Communications of the ACM,vol. 53, no. 4, pp. 50–58, Apirl 2010.*

[3] F. Anguilli, Fast Condensed Nearest Neighbor Rule. *Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 2005.*

[4] Zohreh Alavi, Lu Zhou, James Power, Keke Chen, "RASP-QS: Efficient and Confidential Query Services in the Cloud", *Proceedings of theVLDB Endowment, Vol. 7, No. 13*

[5] C.Wang,Q.Wang,K.Ren,andW.Lou, "Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing"

[6] Huiqi Xu, Shumin Guo, Keke Chen, "Building Confidential and Efficient Query Services in the Cloud with RASP Data Perturbation" *Knowledgeand Data Engineering, IEEE Transactions on (Volume:26, Issue: 2)*

[7] Everitt,B.S.,Landau,S.,Leese,M. and Stahl, D. (2011), "Miscellaneous Clustering Methods, in Cluster Analysis", *5th Edition, John Wiley & Sons,Ltd, Chichester, UK*

[8] Cover TM, Hart PE (1967). "Nearest neighbor pattern classification". *IEEE Transactions onInformation Theory.*

[9] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification*", IEEE Trans. Inform.Theory, Vol. IT-13, pp 21-27, Jan 1967.*

[10] K. Chidananda and G. Krishna, "The condensed nearest neighbor rule using the concept of mutual nearest neighbor", *IEEE Trans. InformationTheory, Vol IT- 25 pp. 488-490, 1979.*

[11] F Angiulli, "Fast Condensed Nearest Neighbor", *ACM International ConferenceProceedings, Vol 119, pp 25-32*