JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-8, Issue-1, Jan 2020 E-ISSN: 2347-2693

A Hybrid Classification Algorithm Using Landmark Based Spectral Clustering

Shivani Walia[1*], P S Mann[2]

[1]Dept. of Computer Science and Engineering, DAVIET, PTU, Jalandhar, India
[2]Dept. of Information Technology, DAVIET, PTU, Jalandhar, India

**Corresponding Author: shivaniwalia02@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v8i1.3039 | Available online at: www.ijcseonline.org

Accepted: 08/Jan/2020, Published: 31/Jan/2020

Abstract— Landmark-based Spectral Clustering (LSC) is used for large scale spectral clustering. The basic idea of the our approach is designing an efficient way for graph construction. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure like distance functions. k-NN is a type of instance-based learning, or lazy learning. In this field, the CRF approach is relatively novel and considered a prominent choice as compared to other methods. However, a lot of scope for further enhancement of the CRF(Conditional Random Field) with Knn optimization problem. The Performance of CRF-Knn has shown quite significant resultsand using different datasets in this paper.The proposed technique improves the selection process using KNN algorithm. The results obtained show that the CRF found to be better than that of LSC in terms of Accuracy, time, recall and precision.

Keywords— Landmark-based Spectral clustering, K nearest neighbors, CRF, Accuracy, time, precision and recall.

I. INTRODUCTION

Data mining is a procedure that finds and summarized the link between potential data and laws in a large number of irregular, disorderly, chaotic, structured data, as well as the process of learning and prediction data.[1,2] Extract knowledge through data mining can be used in statistics and processing, information inquiry, decision analysis, formulas summarizing and process control. Data mining is the process of withdrawal of unrevealed predictive information from a large number of databases; it is an extremely powerful technology with the prominent potential to help organizations that focus on the more considerable information in their data warehouses. Data mining is the process to study the data from various perspectives and encapsulate it into information that is useful for future use -information that can be used to increase revenue, to cut the costs or both. Data Mining is a very important analytic method designed to explore knowledge.[3,4] It allows users or customers in analyzing the data from different dimensions or angles, categorize it, and summarize the relationships identified. Nowadays, data mining is primarily used by many companies with a strong consumer focus such as in communication, retail, financial and marketing organizations. Clustering is one of the fundamental problems in data mining, pattern recognition,and many other research fields. A series of methods have been proposed over the past decades (Jain, Murty, and Flynn 1999). Among them, spectral clustering, a class of methods that are based on the eigendecomposition of matrices, often yields more enhanced superior experimental performance comparing to other algorithms (Shi and Malik 2000). While many clustering algorithms are based on Euclidean geometry and consequently place limitations on the shape of the clusters, spectral clustering can adapt to a wider range of geometries and detect non-convex patterns and linearly non-separable clusters (Ng, Jordan, and Weiss 2001; Filippone et al. 2008). Despite its good performance, spectral clustering is limited in its applicability to large-scale problems due to its high computational complexity. Inspired by the recent progress on sparse coding (Lee et al. 2006) and scalable semi-supervised learning (Liu, He, and Chang 2010), we propose a scalable spectral clustering method termed Landmark-based Spectral Clustering (LSC) in this paper. Specifically, LSC selects p (n) representative data points as the landmarks and represents the remaining data points as the linear combinations of these landmarks. The spectral embedding of the data can then be efficiently computed with the landmark-based representation.[5,6,7]

Organization of the paper

The other sections of the document are structured as follows: Section 2 outlines the related work. Section 3 highlights the proposed model. Section 4 elaborates on the simulation parameters. Section 5 comprises the results and analysis. At last, the conclusion with a future scope is mentioned in Section 6.

II. RELATED WORK

Jasmina D.Novakovic, 2018 et al., [8] discussed presents the filter methods which use K-Nearest Neighbor and these filter methods are applied to numerous dataset. K-Nearest Neighbor is an algorithm which is non-parametric nature. It doesn't use any training data points for generalizing the data. For enhancing the accuracy, feature selection methods are used. The challenging task for the development of feature selection is to select an optimal subset from a large number of features that are possible. The k should be chosen optimally, its size should neither be too long nor too small. Distance metric and the k is used to determine k-nearest Neighbor. The experiment was performed in which 18 datasets were taken with the k-Nearest Neighbor Algorithm and classification accuracy is compared. Filter Methods are used for dimensionality reduction. The paired t-test is used here so that to indicate which one is having better accuracy as compared to others. Attributes are ranked according to the filter methods such as Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SU), Relief-F (RF), One-R (OR) and Chi-Squared (CS). The accuracy of the classification algorithm using IBk is checked for each dataset, filter methods, and attributes. The result shows the removal of noisy data and redundancy in data and contributes towards accuracy.

Sadegh Bafandeh , 2013 et al., [9] discussed steps in data mining where the proper description of data and its attribute is explained and is pictorially represented by graph and chart where the link between variables are observed. Different type of prediction methods such as Classification and regression are discussed wherein Classification groups are made and Case belong to which group is identified and in Regression Unknown values are predicted based on the values which already exists. K-Nearest Neighbor is the technique that is used in this paper. In KNN for Classification if training set and an unknown sample are given then the distance between them can be computed by KNN. The smallest distance is one which is closest to unknown sample. The main challenge is to select optimal size of k , moreover when points are not distributed uniformly. In KNN for Regression independent variables are used to predict dependent variables. To measure distance Euclidian is used. After the selection of k , the prediction is made based on voting methods.

Amir Ali, 2017 et al., [10] discussed the working of K-Nearest Neighbor. Iris data set is taken where 3 attributes are taken into consideration(Sepal length, Sepal width and target attribute). Euclidean distance between actual and observed sepal length and sepal width is calculated. Rank is assigned based on distance and then k-Nearest Neighbor is computed from the rank that is calculated. In this paper practical implementation of K-Nearest Neighbor in Scikit learn is explained. The data set that is imported is divided into testing and training set and then Feature Scaling is done that is an integral part of data preprocessing. KNN Classifiers are modeled using Scikit Libraries. The result is predicted and the accuracy score is measured based on visualized result and the result that was predicted.

Farrukh Arslan, 2018 et al., [11] discussed KNN that is used to determine SOP (Standard Operating Procedure) based on implicit feedback in which information is without feedback of users and there is no negative feedback and is complicated. SOP documents were used as data set and value of each user was determined whether it was suitable or not after observing for 5 weeks. Then classification parameter was determined and ranking and Euclidean distance was computed.

Shufeng Chen, 2018 et al., [12] discussed optimized K-Nearest Algorithm for text classification by reducing the complexity of calculating similarity and improving the speed of algorithm. The CURE clustering algorithm is used which combines hierarchical method and partitioning method. In this paper sample taken is sorted and K nearest neighbor is applied to increase the speed. 6500 News Essays were taken to test the algorithm in which 5500 of them are used as training samples and other 1000 were taken as testing sample. The improved algorithm's speed was much faster than the traditional one whereas the accuracy remained the same.

Yun-lei Cai, 2010 et al., [13] discussed KNN text categorization method in which BM25 similarity calculation method is used. Verification of this method is done in NTCIR-Patent Classification method. Classification of research papers written in Japanese or English is done into IPC at subclass, main group and subgroup levels. Similarity Calculation is important part of KNN algorithm. BM25 similarity method is used which is used to retrieve documents related to query keywords. Similarity Summing Algorithm based on Nearest Neighbor is used as common category decision making method. A Experiment based on comparison of shared Nearest Neighbor + KNN and KNN was conducted on Japanese corpus and English corpus.

Khalid Alkhatib 2013 et al., [14] discussed the stock price prediction using KNN. 6 major companies were taken as a sample on which k nearest algorithm and regression method was applied to predict stock prices. There were 200 records of each company and 3 attributes were taken into consideration where closing price was the main attribute which was used to predict the price and KNN algorithm was applied to 1000 records. The results show that predicted value and the actual value were approximately the same. Total squared errors , RMS errors and actual errors were calculated which was very small that indicated proper performance of the model.

III. PROPOSED MODEL

The methodlogy of proposed model thate is ,the literature review will be conducted on the statistical classification techniques for the data classification. The techniques will be shortlisted for the creation of the new classification model for the data classification based on the LSC, k-NN with CRF techniques. The shortcomings of the existing model would be overcome using the proposed model design based upon important feature with appropriate classification. The problem formulation would be formed after finding the research gaps in the existing data classification models. The proposed model would be framed and finalized using the repeated rounds of algorithm improvement and theoretical debugging.

A. Landmark-based Spectral Clustering (LSC):- Landmark-based Spectral Clustering (LSC) is used for large scale spectral clustering. The basic idea of our approach is designing an efficient way for graph construction and Laplacian matrix Eigen decomposition. The rationale of LSC is to select p («n) representative sampleas landmarks and represents the original samples as the linear combinations of these landmarks. Different from that traditional spectral clustering method use the entire samples to represent each sample, the LSC significant reduces the complexity of affinity matrix. At the same time, the complexity of solving the eigen value down to scales linearly. LSC(Landmark-based Spectral Clustering) tries to compress the original samples by finding a set of basis vector and the representation for the bases for each sample,i.e., searching for p representative samples. kNN(k nearest neighbors) method is not training process, we propose to introduce a new training process for kNN, which blocks training dataset by a clustering algorithm with linear complexity.

B. Conditional Random Fields:-
Conditional Random Fields are a probabilistic framework for labeling and segmenting structured data, such as sequences, trees and lattices. This is especially useful in modeling time-series data where the temporal dependency can manifest itself in various forms. CRF are statistical modeling methods, often applied to pattern recognition and machine learning problems. CRFs have seen wide application in many areas, including natural language processing, computer vision, and bioinformatics. CRFs also used for feature selection and kNN used for classification purposes. Landmark spectral based clustering also using for the clustering method. [17]

C. K-nearest-neighbor: K -nearest neighbor techniques will be shortlisted for the creation of the new classification model for the data classification based on the LSC, k-NN with CRF techniques. The shortcomings of the existing model would be overcome using the proposed model design based upon important feature with appropriate classification. The problem formulation would be formed after finding the research gaps in the existing data classification models. The proposed model would be framed and finalized using the repeated rounds of algorithm improvement and theoretical debugging. The proposed and existing models would be implemented in the Python programming in Anaconda environment and the results would be collected in the form of various performance parameters. The collected results would be analyzed in-depth and the conclusion would be prepared to project the final results of the proposed model.[18]

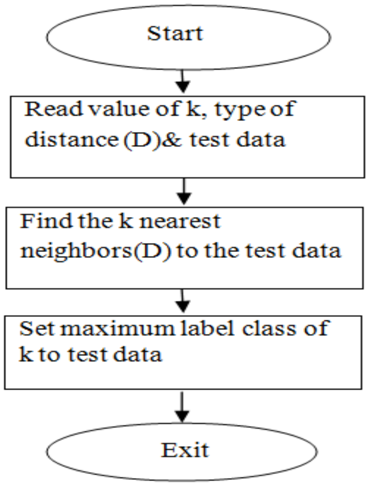
Start
Read value of k, type of distance (D)& test data
Find the k nearest neighbors(D) to the test data
Set maximum label class of k to test data
Exit

Figure 1: Flowchart of K-nearest neighbor

The flowchart of K- nearest neighbour is explained stepwise

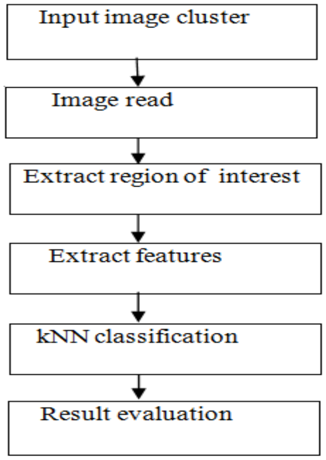

Figure 2: Steps of k-nearest neighbor

Firstly It works based upon on minimum distance from the query instance to the training samples to determine the K-nearest neighbors.[19,20] The data for KNN algorithm consist of several multivariate attributes name that will be used to classify. K-nearest neighbors is simply used for the prediction of query instance. The K-nearest-neighbor (KNN) algorithm measures the distance between a query scenario and a set of scenarios in the data set. [21,22]

IV. SIMULATION ENVIRONMENT

The simulation can be done using the accuracy and time parameter.

Parameters for evaluation

As different parameters are been considered and evaluated for comparing the proposed framework. To compare the existing and proposed algorithm based on following parameters. Four parameters can be achieved, these are following below:-

A. Accuracy
B. Time
C. Recall
D. Precision

A. Accuracy

In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated. % of testing set examples correctly classified by the classifier.
To calculate the accuracy, the following formula has been used:

$$\text{Accuracy} = \frac{\text{Correctly Obtained Results}}{\text{Total Results}} \text{ X } 100$$

- True positive: forged images correctly identified as forged.
- False positive: Authentic images incorrectly identified as forged.
- True negative: Authentic Images correctly identified as Authentic
- False negative: Forged Images incorrectly identified as Authentic.

Table 1 : Comparison of five datasets using with accuracy parameter

DATASETS	LC-kNN	CRF-kNN
USPS	93.55	95.38
MNIST	83.89	98.98
GISEETTE	95.26	99.6
LETTER	94.95	98
PENDIGITS	97.21	97.3

In this table to shown the value of m using the different datasets to find the accuracy through of existing (landmark based spectral clustering with knn and the proposed model(conditional random field with knn)and the value of m=10 in this table

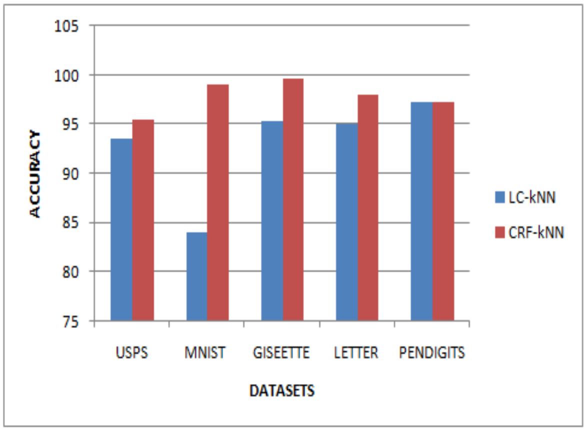

Figure 3: Graph show the accuracy through of using five different datasets

Table 2 : Comparison of fivedatasets using with accuracy parameter

DATASETS	LC-kNN	CRF-kNN
USPS	93.38	95.5
MNIST	83.64	98.34
GISEETTE	94.94	99.19
LETTER	94.69	97.89
PENDIGITS	97.11	97.24

In this table to shown the value of m using the different datasets to find the accuracy through of existing (landmark based spectral clustering with knn and the proposed model(conditional random field with knn)and the value of m=15 in this table.

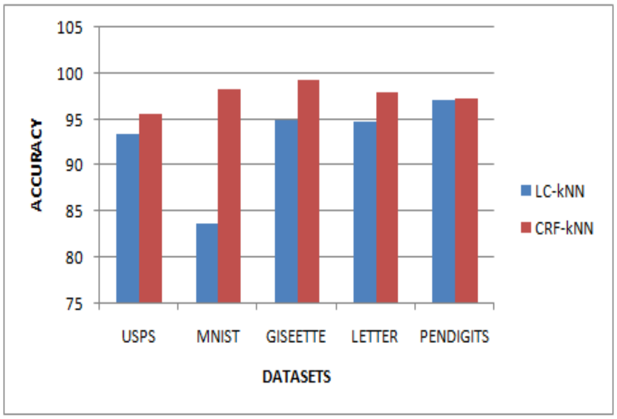

Figure 4: Graph show the accuracy through of using five different datasets

Table 3: Comparison of five datasets using with accuracy parameter

DATASETS	LC-kNN	CRF-kNN
USPS	93	95.14
MNIST	83.53	97.89
GISEETTE	94.11	98.76
LETTER	94.51	97.5
PENDIGITS	97	96.54

In this table to shown the value of m using the different datasets to find the accuracy through of existing (landmark based spectral clustering with knn and the proposed model (conditional random field with knn)and the value of m=20

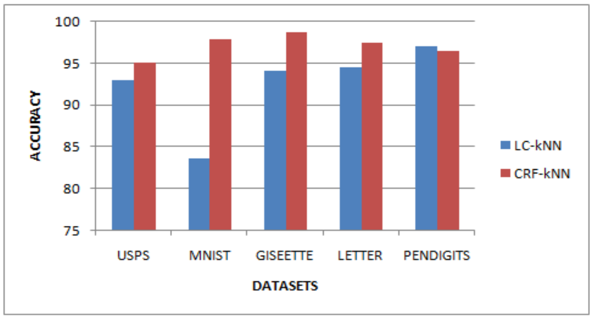

Figure 5: Graph show the accuracy through of using five different datasets

Table 4 :Comparison of five datasets using with accuracy parameter

DATASETS	LC-kNN	CRF-kNN
USPS	92.84	94.86
MNIST	83.38	97.43
GISEETTE	93.21	98.2
LETTER	94.23	97.32
PENDIGITS	96.87	95.17

In this table to shown the value of m using the different datasets to find the accuracy through of existing (landmark based spectral clustering with knn and the proposed model(conditional random field with knn)and the value of m=25

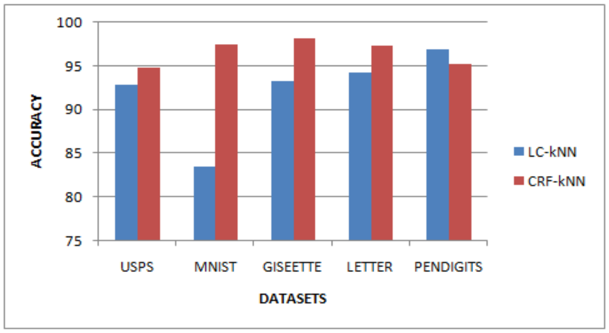

Figure 6: Graph show the accuracy through of using five different datasets

Table 5 : Comparison of five datasets using with accuracy parameter

DATASETS	LC-kNN	CRF-kNN
USPS	92.75	94.65
MNIST	83.13	97
GISEETTE	91.92	97.65
LETTER	94.03	97
PENDIGiTS	96.83	95.15

In this table to shown the value of m using the different datasets to find the accuracy through of existing (landmark based spectral clustering with knn and the proposed model(conditional random field with knn)with value of m=30

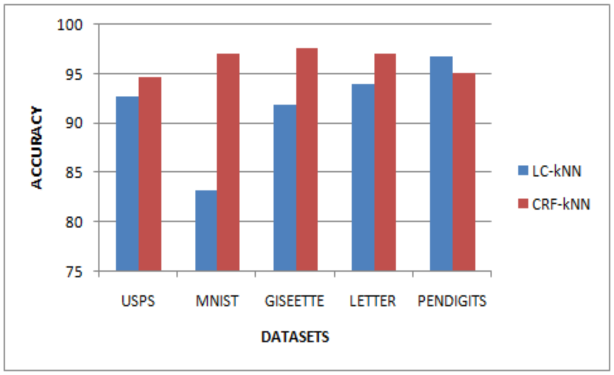

Figure 7: Graph show the accuracy through of using five different datasets

B. TIME:-

The thing that is measured as seconds, minutes, hours, days, years, etc. In the sproposed model time is calculated in seconds.

Elapsed time= stop_time -start_time.

Table no. 6 : Comparison of 5 dataset with Time parameter

DATASETS	LC-kNN	CRF-kNN
USPS	3.76	3.5
MNIST	3.55	4.27
GISEETTE	28.59	4.27
LETTER	3	11
PENDIGITS	2.4	2.2

This table to shown the value of m using the different datasets to find the time through of existing (landmark based spectral clustering with knn and the proposed model(conditional random field with knn and the time is measured in seconds. Every table shown the different value of m=10

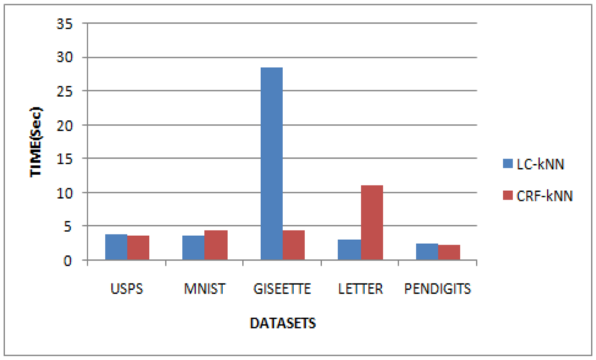

Figure 8: Graph show the time through of using five different datasets

Table no. 7 : Comparison of 5 dataset with Time parameter

DATASETS	LC-kNN	CRF-kNN
USPS	2.72	3.25
MNIST	3.12	4
GISEETTE	23.19	4
LETTER	3.43	10.4
PENDIGITS	2.57	2.1

This table to shown the value of m using the different datasets to find the time through of existing (landmark based spectral clustering with knn and the proposed model(conditional random field with knn and the time is measured in seconds.

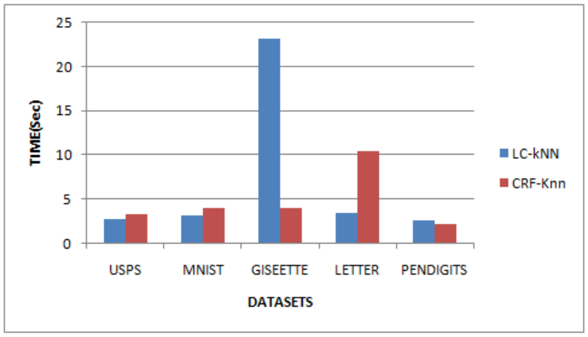

Figure 9: Graph show the time through of using five different datasets

Table 8 : Comparison of 5 dataset with Time parameter

DATASETS	LC-kNN	CRF-kNN
USPS	2.51	3
MNIST	2.14	3.53
GISEETTE	16.2	3.92
LETTER	3.12	10
PENDIGITS	2.25	2.3

This table to shown the value of m using the different datasets to find the time through of existing LSC knn andproposed model(conditional random field with knn and the time is measured in seconds.

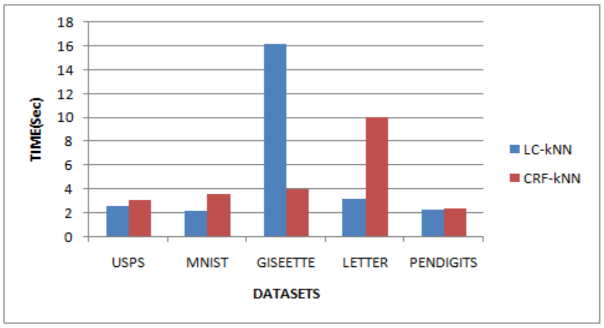

Figure 10:Graph show the time through of using five different datasets

Table 9 : Comparison of 5 dataset with Time parameter

DATASETS	LC-kNN	CRF-kNN
USPS	1.99	2
MNIST	2.11	3.14
GISEETTE	13.86	3.76
LETTER	3.48	9.4
PENDIGITS	2.54	2.13

This table to shown the value of m using the different datasets to find the time through of existing (landmark based spectral clustering with knn and the proposed model(conditional random field with knn and the time is measured in seconds. Every table shown the different value of m.

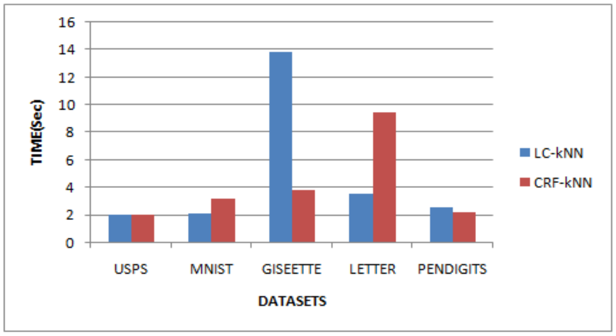

Figure 11: Graph show the time through of using five different datasets

Table 10 : Comparison of 5 dataset with Time parameter

DATASETS	LC-kNN	CRF-kNN
USPS	1.92	2.45
MNIST	1.72	2.45
GISEETTE	11.39	3.54
LETTER	3.11	9.3
PENDIGITS	2.2	2.1

This table to shown the value of m using the different datasets to find the time through of existing (landmark based spectral clustering with knn and the proposed model(conditional random field with knn and the time is measured in seconds.

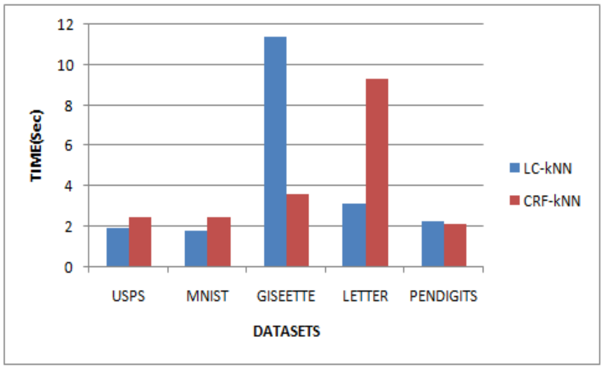

Figure 12:Graph show the time through of using five different datasets

C. PRECISION:-

Precision can be seen as a measure of exactness or quality. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant ones.

Table 11 : Comparison of 5 dataset with Precision parameter

DATASETS	LC-kNN	CRF-kNN
USPS	91.68	93.90
MNIST	97	100
GISEETTE	94.82	99.9
LETTER	82.35	100
PENDIGITS	92.10	99.8

This table to shown the value of m using the different datasets to find the time through of existing LSC with knn and the proposed model(conditional random field with knn Every table shown the different value of m=10

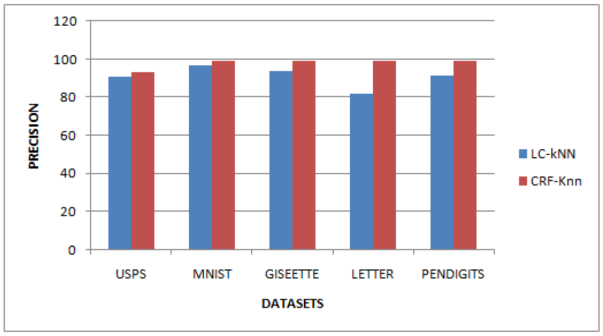

Figure 13 :Graph show the precision through of using five different datasets

Table 12 : Comparison of 5 dataset with Precision parameter

DATASETS	LC-kNN	CRF-kNN
USPS	91.55	93.89
MNIST	96.9	99.8
GISEETTE	94.94	99.89
LETTER	82.25	99.8
PENDIGITS	91.94	99.87

This table to shown the value of m using the different datasets to find the time through of existing LSC with knn and the proposed model(conditional random field) with knn Every table shown the different value of m

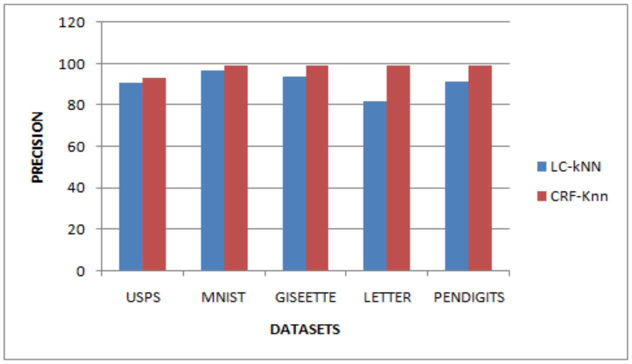

Figure14 :Graph show the precision through of using five different datasets

Table 13:Comparison of 5 dataset with Precision parameter

DATASETS	LC-kNN	CRF-kNN
USPS	91.43	93.8
MNIST	96.8	99.7
GISEETTE	94.11	99
LETTER	82.07	99.8
PENDIGITS	91.8	99.8

This table to shown the value of m using the different datasets to find the time through of existing LSC with knn and the proposed model(conditional random field) with knn Every table shown the different value of m=20

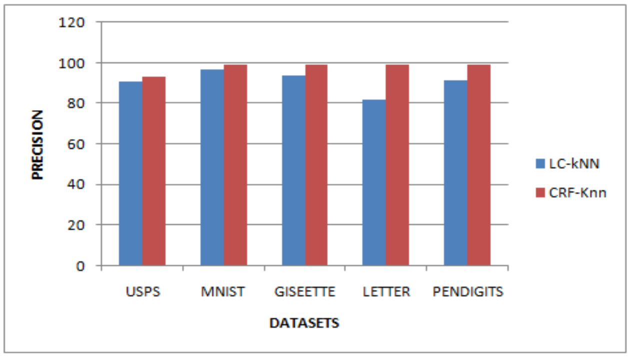

Figure15 :Graph show the precision through of using five different datasets

Table 14 : Comparison of 5 dataset with Precision parameter

DATASETS	LC-kNN	CRF-kNN
USPS	91.34	93.65
MNIST	96.69	99.65
GISEETTE	93.81	99.6
LETTER	81.97	99.65
PENDIGITS	91.69	99.6

This table to shown the value of m using the different datasets to find the time through of existing LSC with knn and the proposed model(conditional random field) with knn Every table shown the different value of m=25

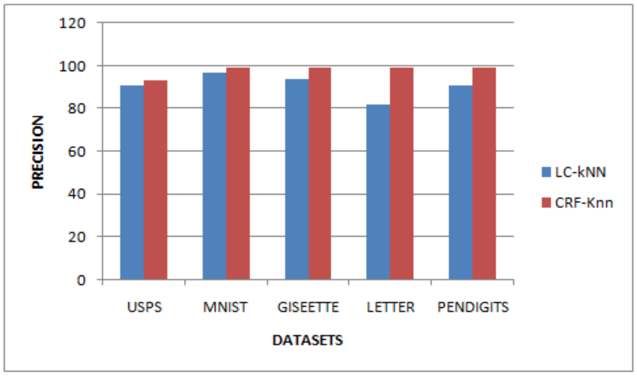

Figure16:Graph show the precision through of using five different datasets

Table 15: Comparison of 5 dataset with Precision parameter

DATASETS	LC-kNN	CRF-KNN
USPS	91	93
MNIST	96.5	99
GISEETTE	93.7	99
LETTER	81.8	99
PENDIGITS	91.5	99

This table to shown the value of m using the different datasets to find the time through of existing LSC with knn and the proposed model(conditional random field) with knn Every table shown the different value of m

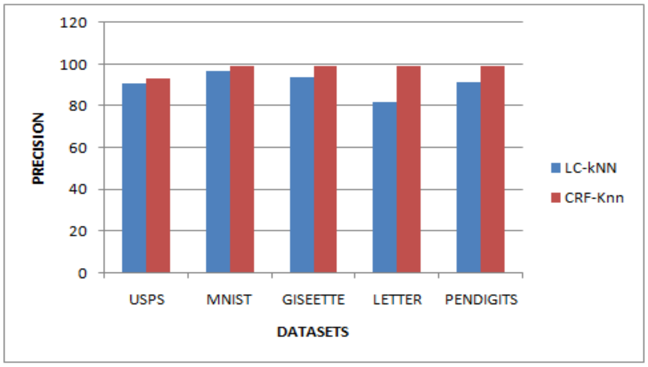

Figure17:Graph show the precision through of using five different datasets

D. Recall:-

Recall is a measure of completeness or quantity. In simple terms, high recall means that an algorithm returned most of the relevant results. This table is shown the value with different value of m. Here, m=10.

Table 16: Comparison of 5 dataset with Recall parameter

DATASETS	LC-kNN	CRF-kNN
USPS	87.33	95.06
MNIST	95.28	98.98
GISEETTE	97.23	99.6
LETTER	97.22	99.3
PENDIGITS	87.93	95.17

This table is shown the value with different value of m. Here, m=15.

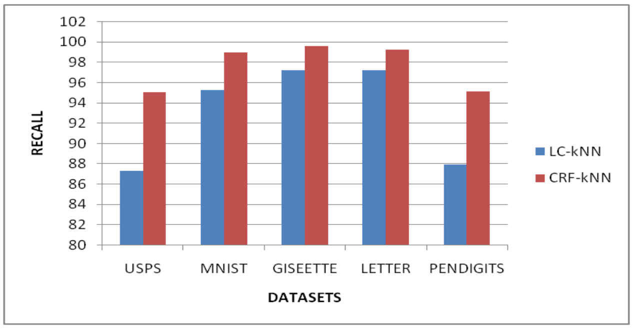

Figure 18: Graph show the recall through of using five different datasets

Table 17 :Comparison of 5 dataset with Recall parameter

DATASETS	LC-kNN	CRF-kNN
USPS	87	94.52
MNIST	95	98.52
GISEETTE	97	99.52
LETTER	97.12	99
PENDIGITS	87.68	94.52

This table is shown the value with different value of m. Here, m=15.

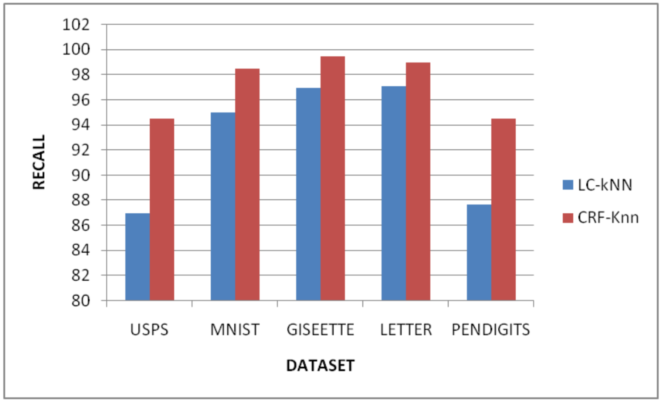

Figure 19: Graph show the recall through of using five different datasets

Table no.18:Comparison of 5 dataset with Recall parameter

DATASETS	LC-kNN	CRF-kNN
USPS	86.87	94.38
MNIST	94.9	98.38
GISEETTE	96.98	99
LETTER	97	99.38
PENDIGITS	87.5	94

This table is shown the value with different value of m. Here, m=20.

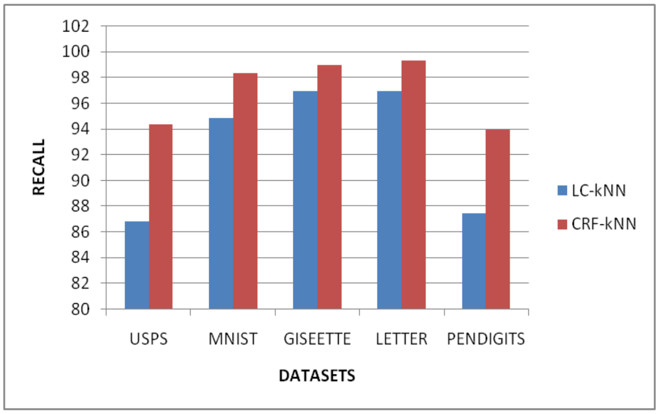

Figure 20:Graph show the recall through of using five different datasets

Table 19: Comparison of 5 dataset with Recall parameter

DATASETS	LC-kNN	CRF-kNN
USPS	86.6	94.2
MNIST	94.78	98.2
GISEETTE	96.90	99.2
LETTER	96.88	99
PENDIGITS	87.3	94.2

This table to shown the value of m using the different datasets to find the time through of existing LSC with knn and the proposed model(conditional random field) with knn Every table shown the different value of m

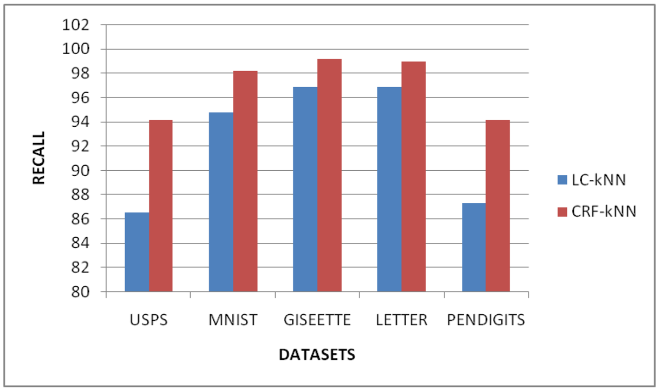

Figure 20:Graph show the recall through of using five different datasets

Table 20: Comparison of 5 dataset with Recall parameter

DATASETS	LC-kNN	CRF-kNN
USPS	86.34	94
MNIST	94.6	97.9
GISEETTE	96.7	98.9
LETTER	96.76	98
PENDIGITS	87	93.9

This table is shown the value with different value of m. Here, m=30.

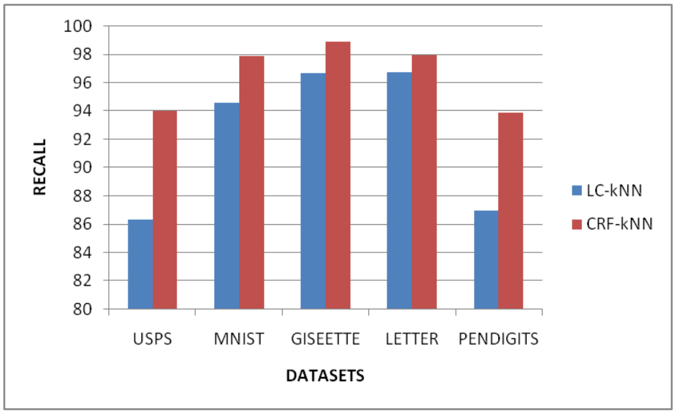

Figure 22:Graph show the recall through of using five different datasets

Conclusion and future Scope

In this paper, five datasets are used with the accuracy parameter and with the time parameter,recall and precision The five datasets which are used in the thesis are USPS, MNIST, GISEETTE, LETTER and PENDIGITS. The value of m parameter is changed every time as it varies from m=10 to m=30. In this paper,we propose to first conduct a clustering to separate the whole dataset into several parts,each of which is then conducted kNN classification. The accuracy of LC-Knn is lower than that of CRF-Knn. Overall ,the results of CRF-kNN better than the LC-kNN, because of landmark clustering only picked up the data through of main destination but CRF-knn ,they can data divided through of segement. The result of recall and and precision better than the previous work.While observing all the readings, the Performance of CRF-Knn has shown quite significant results. The proposed technique improves the selection process using KNN algorithm. Future work can be further extended by using the more parameters which would provide more significant improvement over accuracy and time. There are still a lot of additional problems to be addressed around this research. Some of our ideas that we would like to see further investigation like, Although the proposed methods are easy to implement, the four parameters are used and other parameters can also be considered to speed up the algorithm. The other optimization techniques have also be considered for evaluation.

REFERENCES

[1] Smita,Priti Sharma,”*Use of Data Mining in Various Field: A Survey Paper*”,IOSR Journal of Computer Engineering (IOSR-JCE)7Volume **16**,pp.18-21,**2014**.

[2] Dhara Patel, Ruchi Modi , Ketan Sarvakar *“A Comparative Study of Clustering Data Mining: Techniques and Research Challenges*”, IJLTEMAS, Volume **3**, **2014**.

[3] Amit Tate,Bajrangsingh Rajpurohit, Jayanand Pawar,UjwalaGavhance,”*Comparative Analysis used for Disease Prediction in Data Mining*”,International Journalof Engineering and Techniques, Volume **2**, **201**6.

[4] Thair Nu Phyu ,”*Survey of Classification Techniques in Data Mining* “,International MultiConference of Engineers and Computer Scientists,Volume **1**, **2009**.

[5] Neha Midha,Vikram Singh,“*A Survey on Classification Techniques in Data Mining*”, International Journal of Computer Science & Management Studies, Volume **16**, **2015**

[6] Deng, Zhenyun, Xiaoshu Zhu, Debo Cheng, Ming Zong,Shichao Zhang. "*Efficient kNN classification algorithm for big data*.",proceedings Neurocomputing 195 ,**2016**.

[7] Iyer, S. Jeyalatha,R. Sumbaly, “*Diagnosis of Diabetes using Classification Mining Techniques*”, IJDKP, Volume **5**, pp. 1-14, **2015**.

[8] Jasmina novakovic, “*Experimental Study Of Using The K-Nearest Neighbour Classifier With Filter Methods*,” in computer science and technology at varna, Bulgaria.

[9] Imandoust, Sadegh Bafandeh, Mohammad Bolandraftar. *"Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background*." International Journal of Engineering Research and Applications 3.5 **2013**.

[10] Amir ali, “*An Intuitive Guide of K-Nearest Neighbor with Practical*”, Wavy AI Research Foundation in k-Nearest Neighbor.

[11] Arslan, Farrukh. "*An Efficient K-Nearest Neighbor Algorithm to Determine SOP File System*." ,**2018**.

[12] Shufeng chen , “*K-Nearest Neighbor Algorithm Optimization in Text Categorization*” IOP conference series, earth and environment sciences.

[13] Yun-lei cui , *“A KNN Research Paper Classification Method Based on Shared Nearest Neighbor*” , Proceedings of NTCIR-8 Workshop Meeting,Tokyo, Japan,**2010**.

[14] Khalid Alkhatib, “*Stock price prediction using KNN algorithm*” in International Journal of Business, Humanities and Technology Volume **3**,**2013**.

[15] H.P. Channe ,Sayali.D.Jadhav ,“ *Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques*”, International Journal of Science and Research (IJSR),Volume **5**, **2016**

[16] Ramana, Bendi Venkata, M. Surendra Prasad Babu, N. B. Venkateswarlu. "*A critical study of selected classification algorithms for liver disease diagnosis*." International Journal of Database Management Systems , Volume **3**,**2011**.

[17] Rajkumar ,G. S. Reena, “*Diagnosis of Heart Disease Using Datamining Algorithm*,” Global Journal of Computer Science and Technology, Volume **10**, **2010**.

[18] Fadl Mutaher Ba-Alwi ,Houzifa M. Hintaya,” *Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach*”, International Journal of Scientific & Enginerring Research, Volume **4**, **2013**..

[19] Rohit Arora,Suman“ *comparative Analysis of Classification Algorithms on Different using WEKA*”, International Journal of Computer Applications ,Volume **54**,**2012**.

[20] Samir Kumar Sarangi , Dr. Vivek Jaglan, Yajnaseni Dash ,“ *A Review of Clustering and Classification Techniques in Data Mining*”,International Journal of Engineering, Business and Enterprise Applications,**2015**.

[21] Marie Fernandes*,” Data Mining: A Comparative Study of its Various Techniques and its Process*”, International Journal of Scientific Research in Computer Science and Engineering Volume-**5**, Issue-**1**, pp.19-23, **2017**(E-ISSN: 2320-7639).

[22] Himanshi , Komal Kumar Bhatia*,” Prediction Model for Under-Graduating Student’s Salary Using Data Mining Techniques*”, International Journal Scentific Research in Network Security and Communication, Volume-**6**, Issue-**2**, April **2018**(ISSN: 2321-3256).

AUTHOR PROFILE

Shivani walia has pursed Bachelor of Technology from Rayat Bahra College of Hoshiarpur in year 2015 and i have pursing my Master of Technology in Department of Computer Science and Engineering, DAVIET under Punjab technical university, jalandhar, India. Her main research work focuses on data mining.

Dr. PS.Mann has pursed Ph.d and he is currently working as Assistant Professor in Department of Information technology, DAVIET,Jalandhar, India His main research work focuses on Computational Intelligence, WSNs, Image Processing. He has 13 year experience in teaching field and also gold medalist in B.tech.