An Optimized Hybrid Techniques of Training set reduction for Performance Improvement of k- Nearest Neighbour Classifier to apply it on Agricultural Soil health card dataset

Bhagirath Parshuram Prajapati[1*], Priyanka Puvar[2]

Department of Computer Engineering, A. D. Patel Institute of Technology, New V. V. Nagar, India
*Corresponding Author

Received:- 19 December 2024/ Revised:- 25 December 2024/ Accepted:- 28 December 2024/ Published: 31-12-2024

Abstract— *In non-parametric algorithms such as k-nearest neighbour the fundamental predicaments are the larger storage and computational requirements. Moreover, the effectiveness of classification task affected significantly due to uneven distribution of training data. To overcome the drawbacks of lazy learner like k-nearest neighbour classifier, the scope of training set reduction by editing and condensing the training set is explored in this research work. Additionally, the reduction of training set is carried out by hybrid techniques of training set reduction namely TSR-FkNN (Elbow method) and TRS-FkNN (Silhouette value) in optimized way to achieve improvement of classification performance*.

Keywords— Machine Learning, k-NN, Hybrid method.

I. INTRODUCTION

In Machine learning (ML) algorithm like the simple k-nearest neighbour (k-NN) classifier the input training set consists of vectors and associated class labels [1]. This training set is used in training phase of ML task and size of the input training set is not changed while taken as input [2]. The ML algorithm calculates the distance between a new input test vector and each vector of the stored training set then assigns a class label to the test vector [3]. Hence, the *k*-NN classifier requires a large amount of memory to store the training dataset and a large amount of time required to execute this algorithm, because in contrast to parametric classification algorithms where parameters are learned from training set and algorithm uses these parameters to compute similarity measure, the non-parametric classifiers stores all training instance [4].

Since non-parametric classifiers stores all training instances, it motivated us to find the solution to reduce time and space of k-NN classifier. There are a few solutions to this problem which are feature selection, training set reduction by removing noisy and unimportant training instances [5]. In this research work, we have evaluated hybrid training set reduction techniques with optimization. These techniques are Training set reduction Fast k-NN by applying SSE (TSR-FkNN, Elbow method) and Training set reduction Fast k-NN by applying silhouette value (TSR-FkNN, silhouette value).

The evaluation of above approaches is carried out on agriculture soil health card dataset. And results suggest that the effectiveness of above approaches is significant to existing methods. This paper is organized as follows. Section 2 presents the background about the research topic. Section 3 is covering proposed research work. Section 4 is about comparison of all methods and analysis. Section 5 is concluding this research work.

II. BACKGROUND

The ML field is divided into three major areas namely supervised, unsupervised and semi-supervised. In supervised approach the labeled data is used to supervise the algorithms, for example classification [6]. In unsupervised ML approach, learning algorithms learn from the data itself, for example clustering. While in semi-supervised learning approach, the mixture of labeled and unlabeled records are provided as an input to the ML algorithm.

2.1 Soil health card data set (SHCDS):

This research work is concentrated on exploring the applicability of Machine learning techniques on Agricultural Dataset of Soil health card and to propose improved efficient Machine learning algorithm to classify soil sample into the categories of the deficiencies of micro and macro nutrients.

TABLE 1
SOIL HEALTH CARD DATA SET [5]

Sr. No	SHC_POTASS	SHC_SULPHUR	SHC_MG	SHC_PHOSPHORUS	SHC_IRON	SHC_MANGANESE	SHC_ZINC	SHC_CU	Label
1	454	6	6	99	5	6	5	6	MaMi 179
2	429	12.2	1	31	8.36	9.8	0.46	0.98	MaMi 145
3	479	9.7	2	17	0.56	8.46	7.32	0.22	MaMi 130
4	369	8.2	1.5	21	8.44	10.4	0.86	0.32	MaMi 148
5	370	9.3	1	35	7.38	8.4	0.8	0.92	MaMi 145
6	351	12.2	2.5	17	0.98	7.5	7.56	0.48	MaMi 163
7	242	11.6	2	31	8.06	4.12	0.8	0.65	MaMi 145
8	360	14	2	20	9.6	11.44	0.45	0.92	MaMi 133
9	237	14.3	2.5	35	7.12	8.1	0.74	0.23	MaMi 176
10	356	13.5	1	21	8.44	7.56	0.44	0.52	MaMi 145
11	438	10.8	2	31	8.9	7.6	0.46	0.58	MaMi 145
12	315	18.2	2.5	12	7.58	9.2	0.74	0.44	MaMi 161
13	310	16.5	1	33	6.12	8.06	0.74	0.55	MaMi 145
14	233	12.2	2.5	33	9.8	7.52	0.56	0.88	MaMi 177
15	378	18.5	2	26	8.9	7.36	0.56	0.42	MaMi 145
16	397	6.2	1.5	14	11.4	8.9	0.8	0.2	MaMi 136
17	283	9.2	1	35	8.44	9.38	0.5	0.62	MaMi 145

2.2 Applying *k*-NN on SHCDS:

ALGORITHM 1: *k*-Nearest Neighbour (*k*-NN) classifier

Input: A set of Agriculture records R = {R1, R2. . . Rn}, where n is the total number of SHCDS records.
Procedure:

- **Step 1:** Divide the record data into one training set and test set as 50-50 split.
- **Step 2:** For each test record, calculate similarity with each training record.
- **Step 3:** Sort the training records in the descending order of the maximum cosine similarity and select the top k training records.
- **Step 4:** Assign a class to test record which occurs maximum times in the top k training records.
- **Step 5:** Construct a confusion matrix.
- **Step 6:** Calculate performance measures from the confusion matrix.

2.3 Selection of prototype:

k-NN is having a high computational cost requirement and it is a major and severe drawback in spite of various advantages. To achieve two major advantages of the low computational cost and improved storage need to store the subset(a small set from training set) the selecting prototypes is applied for similar of sometimes even an improves classification performance. Different ways of taking an optimized and proper set of representatives have been studied so far. There are two methods which lead to the reduction of the training set size are editing and condensing, they are giving optimized set and referred as Prototype Selection (PS) methods [6].

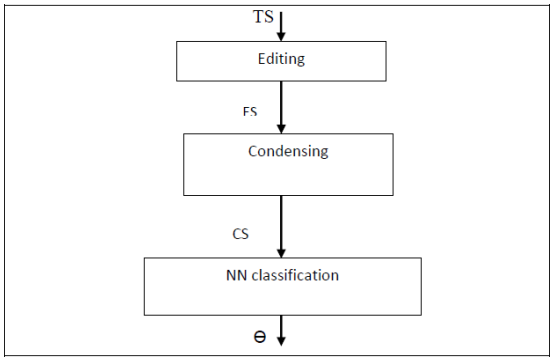

FIGURE 1: Selection of prototype

The learning process is consists of two steps to be finished see Fig 1, editing and condensing in the case when the classifier uses the NN rule. The main focus of editing is to remove noisy instances, and the condensing maintains only the representative instances means it generates prototypes, see Fig. 1. Here, the training set (TS) is the input to the editing, the output of editing is edited set (ES), which in sequence give as input to condensing, whose output is condensed set (CS). Finally, the unknown sample x is classified using the resulting condensed set as input to the classifier. The result: the class ϴ to which the sample x belongs.

III. PROPOSED WORK

In previous section we have discussed training set editing and condensing techniques respectively. In this section, we have proposed a novel techniques which uses both editing and condensing both. These hybrid techniques can be understood from Fig. 2. It takes Initial training set (TS), and then it will condense it followed by applying editing algorithm. Finally, the edited set is applied to the *k*-NN classifier.

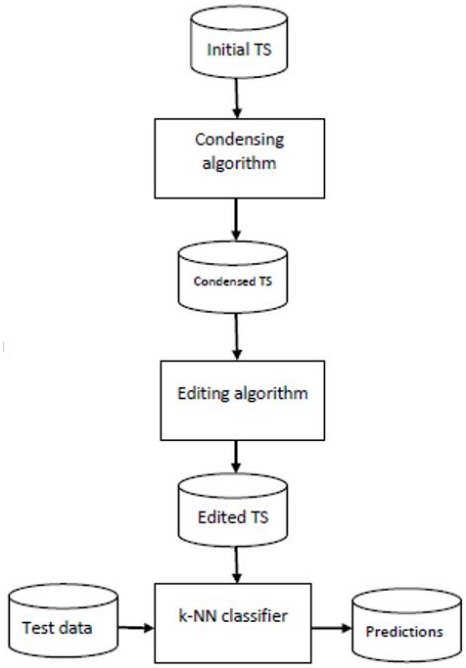

FIGURE 2: Data reduction by hybrid method

Fig. 3 provided an overview of a hybrid approach based on the previous two methods of training set reduction and then clustering. This method is a hybrid method, where we have combined features of both fast k-NN and training set reduction. In hybrid method, the training set reduction techniques are applied on training set feature vector and the technique reduces training set. The reduced training set is given as an input to clustering algorithm and a set of clusters are given as an input to Machine learning algorithm and classifier model is learned. The classifier model assigns a class to a new test instance.

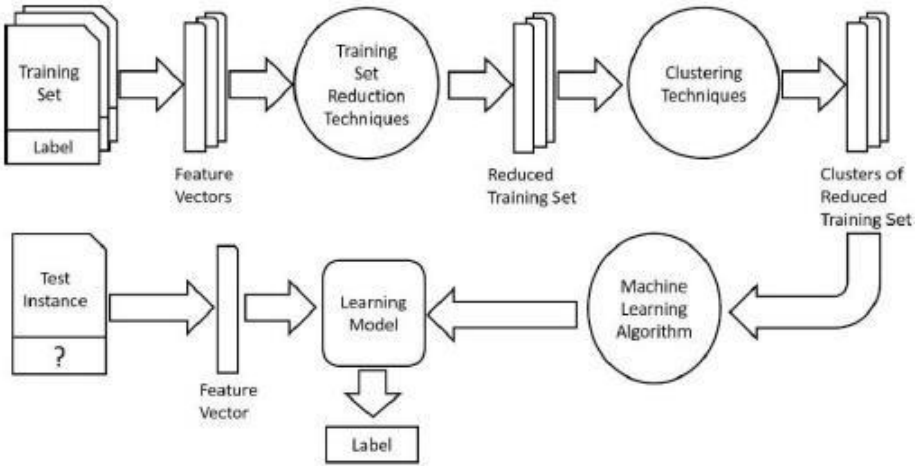

FIGURE 3: Overview of hybrid machine learning technique TSR-FkNN

3.1 Training set reduction fast *k*-nearest neighbour (TRS-FkNN), Elbow method:

This hybrid approach is implemented as per algorithm 2. As shown in step 2, the training set reduction method is applied which reduces the training instances. It is condensing technique the shrink subtractive method which reduces the input training size considerably. The reduced training set is now taken as input to the step 3 where editing method of clustering is applied. To find optimal k value of k-Means clustering algorithm. Here elbow method is applied. It can be noticed that in step 2 and step 3 we are applying editing and condensing techniques consequently. Hence it is a hybrid approach of combining methods for training set reduction.

ALGORITHM 2: Training set reduction Fast k-NN (TRS-FkNN), optimum k-Means by SSE

Input: A set of Agriculture records R = {R1, R2,… Rn}, where n is the total number of agriculture records reduced training record set D.

Procedure:

Step 1: Divide the record data into one training set and test set as 50-50 split. Step 2: Shrink (subtractive) algorithm applied on training set,

- **Step 2.1:** Assign all the training documents into S.
- **Step 2.2:** Select randomly an instance P from S.
- **Step 2.3:** Classify the instance P using remaining instances from S.
- **Step. 2.4:** Remove the instance P if it is correctly classified.
- **Step 2.5:** Repeat step 2.2 to 2.4 till no such instance left in S. Step 2.6: Take the new reduced set S as a training set for step 3.

Step 3: Construct k clusters using k-Means clustering algorithm and validate k value for k-Means clustering by Elbow method step 3.1 to 3.4 and assign a class label to each cluster centroids based on maximum occurrences of a particular class in that cluster.

- **Step 3.1:** Initialize k = 1.
- **Step 3.2:** Increment the value of k.
- **Step 3.3:** Measure the value of SSE for the optimal solution.
- **Step 3.4:** At some point, the effective cost of the solution reaches significantly, then take that value of k and stop. If not then repeat steps 3.2- 3.4.

Step 4: For each test record, calculate similarity with each cluster's centroid.

Step 5: Sort the training records in the descending order of the maximum cosine similarity and select the top k training records.

Step 6: Assign a class to test record which occurs maximum times in the top k training records.

Step 7: Construct a confusion matrix.

Step 8: Calculate performance measures from the confusion matrix.

3.2 Training set reduction fast *k*- nearest neighbour (TRS-FkNN), Silhouette value:

The second proposed hybrid approach is implemented as per algorithm 3. As shown in step 2, the training set reduction method is applied which reduces the training instances. It is condensing technique the shrink subtractive method which reduces the input training size considerably. The reduced training set is now taken as input to the step 3 where editing method of clustering is applied. To find optimal k value of k- Means clustering algorithm, here silhouette value is computed. It can be noticed that in step 2 and step 3 we are applying editing and condensing techniques consequently. Hence it is a hybrid approach of combining methods for training set reduction.

ALGORITHM 3: Training set reduction fast k-NN (TRS-FkNN), optimum k-Means by silhouette value

Input: A set of Agriculture records R = {R1, R2,… Rn}, where n is the total number of agriculture records reduced training record set D.
Procedure:
Step 1: Divide the record data into one training set and test set as 50-50 split.
Step 2: Shrink (subtractive) algorithm applied on training set,
- **Step 2.1:** Assign all the training documents into S.
- **Step 2.2:** Select randomly an instance P from S.
- **Step 2.3:** Classify the instance P using remaining instances from S.
- **Step 2.4:** Remove the instance P if it is correctly classified.
- **Step 2.5:** Repeat step 2.2 to 2.4 till no such instance left in S.
- **Step 2.6:** Take the new reduced set S as a training set for step 3.

Step 3: Construct k clusters using k-Means clustering algorithm and validate k value for k-Means clustering by calculating silhouette value and assign a class label to each cluster centroids based on maximum occurrences of a particular class in that cluster.
Step 4: For each test record, calculate similarity with each cluster's centroid.
Step 5: Sort the training records in the descending order of the maximum cosine similarity and select the top k training records.
Step 6: Assign a class to test record which occurs maximum times in the top k training records.
Step 7: Construct a confusion matrix.
Step 8: Calculate performance measures from the confusion matrix.

IV. COMPARISONS OF RESULTS OF PROPOSED CLASSIFIERS

In this section comparison between different proposed classification techniques is carried out in terms of performance measures and time of classification in milliseconds. For the experiment, we have taken Kutch district data set from SHCDS, having total 14000 entries. These results are performed on a computer with Intel i5 processor and 4GB Ram, the software IDE is NetBeans 8.2. Depends on hardware some of the results may vary. The observed results are on average of five times run.

4.1 Comparison of accuracy of various classifiers:

TABLE 2
COMPARISON OF ACCURACY FOR ALL *k*-NN CLASSIFIERS

	Value *k* for *k*-NN	*k* Nearest Neighbor	Hybrid method	
Sr. No	*k*	*k*-NN	TSR-FkNN (Elbow method)	TSR- FkNN (Silhouette value)
1	31	90.21	88.92	92.64
2	33	88.85	90.42	92.14
3	35	90.41	90.85	93.85
4	37	90	88.85	91.85
5	39	90.35	89.28	92.35
6	41	89.51	90.75	93.34
7	43	90.55	88.75	92.46
8	45	90.27	90.35	91.75

In table 2, the accuracy of different k-NN classifiers is compared. The accuracy presents the ratio between a number of predictions those are correctly classified to the total number of predictions (the number of test data points) [7-10]. It is observed that accuracy of proposed TSR-FkNN (Silhouette value) is high in comparison with other classifiers.

4.2 Training set comparison:

In table 3, different classifiers training instances are compared. In our experimental setup, the simple k-NN classifier trains on 7000 instances which is the benchmark to compare with other reduction techniques. All classifiers start with 7000 instances in training set, then we are applying our proposed prototype selection techniques to reduce the size of the training set and resulting reduced training set it indicated. In our research TSR-FkNN (Elbow method) has lowest training instances when the value of k is 33 and 35 respectively, all other classifiers have higher training instances while k- NN has highest training instances. Here, training instances of all classifiers other than k-NN are reduced by applying novel techniques designed for this research.

TABLE 3
COMPARISON OF TRAINING SET FOR ALL *k*-NN CLASSIFIERS

Sr. No	Value k for k-NN	*k* Nearest Neighbor	Hybrid method	
	k	***k*-NN**	**TSR-FkNN (Elbow method)**	**TSR- FkNN (Silhouette value)**
1	31	7000	141	131
2	33	7000	61	151
3	35	7000	71	131
4	37	7000	111	141
5	39	7000	151	161
6	41	7000	131	141
7	43	7000	**91**	151
8	45	7000	101	161

4.3 Classification time comparison:

In table 4, comparison of all classifier is done in terms of classification time in millisecond [11-12]. In our research, it is observed that TSR-FkNN (applying SSE) is having lowest classification time when the value of k is 33 and 45 respectively.

TABLE 4
COMPARISON OF CLASSIFICATION TIME FOR ALL *k*-NN CLASSIFIERS

Sr. No	Value k for k-NN	*k* Nearest Neighbor	Hybrid method	
	k	***k*-NN**	**TSR-FkNN (Elbow method)**	**TSR- FkNN (Silhouette value)**
1	31	5766	248	261
2	33	5779	143	219
3	35	5777	180	192
4	37	5746	217	217
5	39	5749	221	245
6	41	5753	175	217
7	43	5776	185	213
8	45	5745	157	224

V. CONCLUSION

5.1 Storage reduction:

Storage requirement in *k*-NN is very high in comparison to other algorithms.

For TSR- FkNN (Elbow method) storage requirement is lowest when the value of *k* is 33 and 35 respectively followed TSR-FkNN (Silhouette value). Hence, in terms of storage TSR-FkNN is efficient.

5.2 Execution time:

- Execution time is highest in *k*-NN as it store more number of instances for training purpose.
- Execution time is lowest in TSR-FkNN (Elbow method) as it store less number of training instances.

5.3 Generalization accuracy, precision, recall and F1 measure:

- Generalize accuracy of TSR-FkNN (Silhouette value) is highest compared to other algorithms hence in terms of accuracy TSR-FkNN (Silhouette value) is recommended.

In terms of Time, Space and Accuracy comparisons. The proposed novel hybrid algorithm TSR-FkNN (Silhouette value) is the best algorithm hence it can be recommended for classifying soil samples in respective nutrients deficiencies category.

REFERENCES

[1] Tomek, I., 1976. A generalization of the k-NN rule. IEEE Transactions on Systems, Man, and Cybernetics, (2), pp.121-126.

[2] Tan, P.N., 2006. Introduction to data mining. Pearson Education India.

[3] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[4] Rokach, L. and Maimon, O., 2005. Data mining and knowledge discovery handbook. Springer US.

[5] Bhagirath, P. and Dhaval, K., 2017, A Hybrid Machine Learning Technique for fusing fast kNN and training Set Reduction: Combining both Improves the Effectiveness of Classification, 2017. ICACIE 2017.

[6] Cios, K.J., Pedrycz, W., Swiniarski, R.W. and Kurgan, L.A., 2007. Data mining: a knowledge discovery approach. Springer Science & Business Media.

[7] Raicharoen, T. and Lursinsap, C., 2005. A divide-and-conquer approach to the pairwise opposite class-nearest neighbor (POC-NN) algorithm. Pattern recognition letters, 26(10), pp.1554-1567.

[8] Singh, B. and Ryan, J., 2015. Managing fertilizers to enhance soil health. International Fertilizer Industry Association, Paris, France, pp.1-24.

[9] Armstrong, L.J., Diepeveen, D. and Maddern, R., 2007, December. The application of data mining techniques to characterize agricultural soil profiles. In Proceedings of the sixth Australasian conference on Data mining and analytics- Volume 70 (pp. 85-100). Australian Computer Society, Inc.

[10] Khedr, A.E., Kadry, M. and Walid, G., 2015. Proposed Framework for Implementing Data Mining Techniques to Enhance Decisions in Agriculture Sector Applied Case on Food Security Information Center Ministry of Agriculture, Egypt. Procedia Computer Science, 65, pp.633-642.

[11] Williams, N., Zander, S. and Armitage, G., 2006. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. ACM SIGCOMM Computer Communication Review, 36(5), pp.5-16.

[12] Bost, R., Popa, R.A., Tu, S. and Goldwasser, S., 2015, February. Machine Learning Classification over Encrypted Data. In NDSS.