Efficient Classification for Dataset Using PSO and KNN Implementation

[1]Bhagwati Galande
Department of Computer Engineering,
BSCOER ,
Pune, India

[2]Prof. M. K. Kodmelwar
Department of Computer Engineering,
BSCOER,
Pune, India

***Abstract* -- Nearest neighbour (kNN) methodology could be a widespread classification methodology in data processing, classification and statistics due to its easy implementation and significant classification performance. However, it's impractical for ancient kNN strategies to assign a k to any or samples. Previous solutions of classification assign completely different k values to different samples by the cross validation methodology however area unit sometimes time intense or time consuming. This paper proposes a kTree methodology to be told completely different best k values for various test/new samples, by involving a coaching stage within the kNN classification and Particle Swarm improvement (PSO). Specifically, within the coaching stage, PSO improvement is employed to search out best k values and discarded week samples then kTree methodology 1st learns best k values for all coaching samples by a brand new distributed reconstruction model, then constructs a tree namely, kTree. kTree quick outputs the best k worth for every sample, and then, the kNN classification is conducted. As a result, the planned kTree methodology features a similar running value however higher classification accuracy, compared with ancient kNN strategies, that assign a fixed k to any or all samples.**

I. INTRODUCTION

K-means cluster analysis is an algorithm that groups similar objects into groups called clusters. The endpoint of cluster analysis is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Typically, k-means cluster analysis is performed on a table of information, wherever every row represents associate degree object and also the columns represent quantitative characteristic of the objects. These quantitative characteristics area unit referred to as agglomeration variables.

k-means agglomeration could be a technique of vector division, originally from signal process, that's fashionable for cluster analysis in data processing. k-means agglomeration aims to partition n observations into k clusters within which every observation belongs to the cluster with the closest mean, serving as a example of the cluster. This leads to a partitioning of the info area into Voronoi cells.

II. LITERATURE SURVEY

Different from model-based strategies that first learn models from coaching samples and so predict take a look at samples with the learned model [1]–[6], the model-free k nearest neighbors (kNNs) methodology doesn't have coaching stage and conducts classification tasks by first conniving the gap between the take a look at sample and every one coaching samples to get its nearest neighbors and so conducting kNN classification1 (which assigns the take a look at samples with labels by the bulk rule on the labels of elite nearest neighbors). due to its easy implementation and significant classification performance and result, kNN methodology could be a very fashionable methodology in data processing, classification and statistics and so was voted in concert of prime data processing algorithms [7]–[13].

Previous kNN strategies include: 1) assignment Associate in Nursing optimum k price with a fixed expert-predefined price for all take a look at samples [14]–[19] and 2) assignment of completely different optimum k values for training, take a look at samples [18], [20], [21]. For instance, Lall and Sharma [19] indicated that the fixed optimal-k-value for all take a look at samples ought to be k =√n (where n > a hundred and n is that the variety of coaching samples), whereas Zhu et al.

[21] Projected to pick completely different optimum k values for training, take a look at samples via multiple cross validation methodology. However, the standard kNN methodology or algorithm, that assigns fixed k to any or all take a look at samples (fixed kNN strategies for short), has been shown to be impractical in real applications. As a consequence, setting Associate in Nursing or assigning optimal-k-value for every take a look at sample to conduct kNN classification (varied kNN strategies for short) has been changing into a awfully fascinating analysis topic in data processing and machine learning [22]–[29].

plenty of efforts are centered on the various kNN strategies, that efficiently set completely different optimal-k-values to different samples [20], [30], [31]. For instance, Li et al. [32]

projected to use completely different numbers of nearest neighbours for various labels or classes and Sahigara et al. [33] projected to use the town validation methodology to pick Associate in Nursing optimum smoothing parameter k for every take a look at sample. Recently, Cheng et al.

[20] projected a sparse-based kNN methodology to find out Associate in Nursing optimal-k-value for every take a look at sample and Zhang et al. [30] studied the kNN methodology by learning an acceptable k price for every take a look at sample supported a reconstruction framework [34].

Previous varied kNN strategies typically first learn a personal optimal-k-value for every take a look at sample and so use the standard kNN classification (i.e., school of thought on k coaching samples) to predict take a look at samples by the learned optimal-k-value. However, either the method of learning Associate in nursing optimal-k-value for every take a look at sample or the method of scanning all coaching samples for finding nearest neighbors of every take a look at sample is long. Therefore, it's difficult for at the same time addressing these problems with kNN methodology, i.e., optimal-k-values learning for various samples, time value reduction, and performance improvement

While kNN methodology allows to output outstanding performance and has been proved to around accomplish to the error rate of mathematician optimisation beneath terribly gentle conditions, it's wide been applied to several types of applications, like regression, classification, and missing price imputation. The performance of kNN methodology will be full of plenty of problems, like the choice of the k price and therefore the selection of distance measures.

III. METHODOLOGY

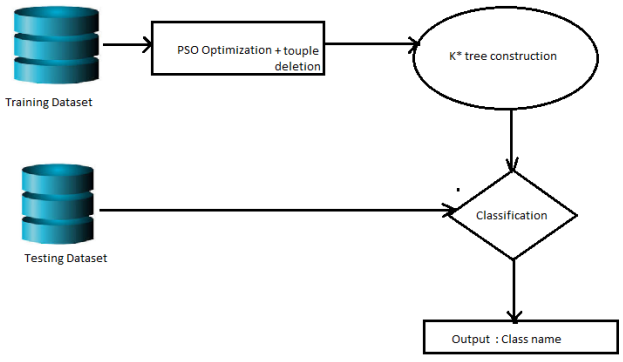

Fig 1: System Architecture

As shown in fig1, we implemented a new Knn classification approach to handle classification problem and find optimal solution. First we use PSO improvement to find optimum solution. Then we minimizes by using PSO based optimal solution i.e tuple deletion. Then K*tree construction take place. After than main classification take place.

Mathematical Model

Models describe about how the System functions. The following points are used as an input for a local server and consequently the output is displayed.

Let our system be S, represented as:

S={s, e, I, Pp, F, NDD, DD, H, Success, Failure}

The notations are given below: -

s = Starting state of the system

e = Ending state of the system

I = Set of input commands

I = {Training Dataset, Testing Dataset}

Pp = Pre-processing module

Pp = P1, P2, P3

Where,

P1 = PSO (Particle Swarm Optimization)

P2 = K*tree Method

P3 = kNN Algorithm

F = Set of used functions

F = {PSO(), DecisionTree(), Classification() }

Where,

PSO(): This method is used to provide optimal *K*-value.

DecisionTree(): This method is used to make decision tree using k*tree method.

Classification(): This method is used to classify the dataset and provide the resultant class for the dataset value.

NDD = Non-Deterministic Data

DD = Deterministic Data

H = Halt state

Success = If the class name is classified correctly.

Failure = If the class name is not classified correctly.

Algorithm:

Input: Training Dataset and Testing Dataset.

Output: Class names for Testing Dataset.

1) Start

2) Initialize Training dataset and testing dataset.

3 Pass the training dataset to PSO algorithm.

4) Find the optimal K value for the classification using PSO algorithm.

5) Make a decision tree using k*tree method for optimal k value.

6) Pass the optimal k value to the kNN algorithm.

7) Using kNN algorithm find nearest neighbors for that particular dataset value

8) Show result as a class name for that testing dataset.

9) End.

3.1 k-means Algorithm

(1) Arbitrarily choose k objects from D as the initial class centres.

(2) Repeat.

(3 (re) assign each object to the class to which the object is the most similar, based on the mean value of the objects in the class.

(4) Update the class means, i.e., calculate the mean value of the objects for each class.

(5) Until no change.

3.2 K*tree Algorithm

1) Start with all training instances associated with the root node

2) Use information gain to select which attribute to label each node with

Note: No root-to-leaf path should contain the same discrete attribute twice

3) Recursively constructs each sub-tree on the subset of training instances that would be classified down that path in the tree.

The border cases:

a) If all positive or all negative training instances remain, label that node “positive” or “negative” accordingly

b) If no attributes remain, label with a majority vote of training instances left at that node

c) If no instances remain, label with a majority vote of the parent’s training instances

IV. EXPERIMENTAL SETUP

Instead of classic dataset we use Abalone dataset for the testing and training. In this dataset there are some numeric values to define the class of that data. We will see some data records from the dataset to understand how data will look like in dataset.

0.615,0.47,0.175,1.2985,0.5135,0.343,0.32,14,Female

0.605,0.49,0.145,1.3,0.517,0.3285,0.31,14,Male

0.59,0.455,0.165,1.161,0.38,0.2455,0.28,12,Female

Attributes are Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight, Rings, Sex. Predicting the sex of abalone from physical measurements. The sex of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the sex. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.

V. RESULT AND DISCUSSIONS

By using PSO weak data discarded from dataset and only strong data will passed to k-means. Therefore K-means training time and testing time reduced and classification rate increases.

In this section we present the results of evaluating classification result of our implemented system based on PSO and KNN on dataset. finally check the results for Detection accuracy, false positive rate.

Detection Accuracy = (TP + TN) / (TP + TN + FP + FN)

False Positive Rate = FP / (FP + TN)

VI. CONCLUSION AND FUTURE WORK

In this paper, we have implemented new kNN classification algorithms, i.e. PSO, and the k*Tree methods, to select optimal-k-value for each test sample for efficient and effective kNN classification. The key idea of our implemented methods or system is to design a training stage with additional algorithms for reducing the running cost of test stage and improving the classification performance.

In future, we will focus on improving the performance of the implemented methods.

REFERENCES

[1] S. Zhang, “Shell-neighbor method and its application in missing data imputation,” Appl. Intell, vol. 35, no. 1, pp. 123–133, 2011.

[2] L. Shao, L. Liu, and X. Li, “Feature learning for image classification via multiobjective genetic programming,” IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 7, pp. 1359–1371, Jul. 2014.

[3] D. Tao, X. Tang, X. Li, and X. Wu, “Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 7, pp. 1088–1099, Jul. 2006.

[4] J. Yu, X. Gao, D. Tao, X. Li, and K. Zhang, “A unified learning framework for single image super-resolution,” IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 4, pp. 780–792, Apr. 2014.

[5] Q. Zhu, L. Shao, X. Li, and L. Wang, "Targeting accurate object extraction from an image: A comprehensive study of natural image matting," IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 2, pp. 185–207, Feb. 2015.

[6] K. Huang, D. Tao, Y. Yuan, X. Li, and T. Tan, "Biologically inspired features for scene classification in video surveillance," IEEE Trans. Syst., Man, Cybern., Part B, vol. 41, no. 1, pp. 307–313, Feb. 2011.

[7] S. Zhang, "Nearest neighbor selection for iteratively KNN imputation," J. Syst. Softw., vol. 85, no. 11, pp. 2541–2552, 2012.

[8] T. Wang, Z. Qin, S. Zhang, and C. Zhang, "Cost-sensitive classification with inadequate labeled data," Inf. Syst., vol. 37, no. 5, pp. 508–516, 2012.

[9] X. Zhu, X. Li, and S. Zhang, "Block-row sparse multiview multilabel learning for image classification," IEEE Trans. Cybern., vol. 46, no. 2, pp. 450–461, Feb. 2015.

[10] X. Wu et al., "Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, no. 1, pp. 1–37, 2008.

[11] K. Zhang, X. Gao, D. Tao, and X. Li, "Single image super-resolution with multiscale similarity learning," IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 10, pp. 1648–1659, Oct. 2013.

[12] D. Landgrebe, "A survey of decision tree classifier methodology," IEEE Trans. Syst., Man Cybern., vol. 21, no. 3, pp. 660–674, May 1991.

[13] H. Liu, X. Li, and S. Zhang, "Learning instance correlation functions for multi-label classification," IEEE Trans. Cybern., vol. 47, no. 2, pp. 499–510, Feb. 2017.

[14] S. Zhang, "Parimputation: From imputation and null-imputation to partially imputation," IEEE Intell. Inform. Bull., vol. 9, no. 1, pp. 32–38, Jan. 2008.

[15] G. Góra and A. Wojna, "RIONA: A classifier combining rule induction and k-NN method with automated selection of optimal neighbourhood," in Proc. ECML, 2002, pp. 111–123.

[16] B. Li, Y. W. Chen, and Y. Q. Chen, "The nearest neighbor algorithm of local probability centers," IEEE Trans. Syst., Man, B, vol. 38, no. 1, pp. 141–154, Feb. 2008.

[17] X. Zhu, H.-I. Suk, and D. Shen, "Multi-modality canonical feature selection for Alzheimer's disease diagnosis," in Proc. MICCAI, 2014, pp. 162–169.

[18] J. Wang, P. Neskovic, and L. N. Cooper, "Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence," Pattern Recognit., vol. 39, no. 3, pp. 417–423, 2006.

[19] U. Lall and A. Sharma, "A nearest neighbor bootstrap for resampling hydrologic time series," Water Resour. Res., vol. 32, no. 3, pp. 679–693, 1996.

[20] D. Cheng, S. Zhang, Z. Deng, Y. Zhu, and M. Zong, "KNN algorithm with data-driven k value," in Proc. ADMA, 2014, pp. 499–512.

[21] X. Zhu, S. Zhang, Z. Jin, Z. Zhang, and Z. Xu, "Missing value estimation for mixed-attribute data sets," IEEE Trans. Knowl. Data Eng., vol. 23, no. 1, pp. 110–121, Jan. 2011.

[22] H. A. Fayed and A. F. Atiya, "A novel template reduction approach for the K-nearest neighbor method," IEEE Trans. Neural Netw., vol. 20, no. 5, pp. 890–896, May 2009.

[23] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in Proc. ACM MM, 2013, pp. 143–152.

[24] H. Wang, "Nearest neighbors by neighborhood counting," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 6, pp. 942–953, Jun. 2006.

[25] Q. Liu and C. Liu, "A novel locally linear KNN method with applications to visual recognition," IEEE Trans. Neural Netw. Learn. Syst., to be published.

[26] X. Zhu, S. Zhang, J. Zhang, and C. Zhang, "Cost-sensitive imputing missing values with ordering," in Proc. AAAI, 2007, pp. 1922–1923.

[27] J. Hou, H. Gao, Q. Xia, and N. Qi, "Feature combination and the kNN framework in object classification," IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 6, pp. 1368–1378, Jun. 2016.

[28] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning k for kNN classification," ACM Trans. Intell. Syst. Technol., vol. 8, no. 3, pp. 1–19, 2017.

[29] D. Tao, J. Cheng, X. Gao, X. Li, and C. Deng, "Robust sparse coding for mobile image labeling on the cloud," IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 1, pp. 62–72, Jan. 2017.

[30] S. Zhang, M. Zong, K. Sun, Y. Liu, and D. Cheng, "Efficient kNN algorithm based on graph sparse reconstruction," in Proc. ADMA, 2014, pp. 356–369.

[31] X. Zhu, L. Zhang, and Z.Huang, "A sparse embedding and least variance encoding approach to hashing," IEEE Trans. Image Process., vol. 23, no. 9, pp. 3737–3750, Sep. 2014.

[32] B. Li, S. Yu, and Q. Lu. (2003). "An improved k-nearest neighbor algorithm for text categorization." [Online]. Available: https://arxiv.org/abs/cs/0306099

[33] F. Sahigara, D. Ballabio, R. Todeschini, and V. Consonni, "Assessing the validity of QSARS for ready biodegradability of chemicals: An applicability domain perspective," Current Comput.-Aided Drug Design, vol. 10, no. 2, pp. 137–147, 2013.

[34] X. Zhu, Z. Huang, Y. Yang, H. T. Shen, C. Xu, and J. Luo, "Selftaught dimensionality reduction on the high-dimensional small-sized data," Pattern Recognit., vol. 46, no. 1, pp. 215–229, 2013.