



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

IJIRSET

International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 3, March 2023

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.423

9940 572 462

6381 907 438

ijirset@gmail.com

www.ijirset.com

A Survey Article on Use of KNN for Variety of Applications

Jayesh Nalawade¹, Tejas Shete², Sangram Pokharkar³, Rashi Kaskar⁴, B. M. Borhade^{*}

U.G. Student, Department of Computer Engineering, SITS, Narhe, Pune, India¹⁻⁴

Associate Professor, Department of Computer Engineering, SITS, Narhe, Pune, India^{*}

ABSTRACT: Finding the requested object within a big and unclear dataset is a difficult process that requires both time and computational complexity. Several study methodologies were suggested to resolve these issues. Of them, using the K-Nearest Neighbor (kNN) algorithm is the most straightforward, effective, and efficient method. This method has uses in a number of areas, including pattern recognition, text classification, moving object detection, etc. Several scholars under varied circumstances have suggested numerous kNN algorithms. We divided these strategies into two categories in this paper: kNN approaches that are (1) structure-based and (2) non-structure-based. The purpose of this work is to examine each kNN technique's main premise, benefits, drawbacks, and target data. The structure based kNN techniques such as Ball Tree, k-d Tree, Principal Axis Tree, Orthogonal Structure Tree, Nearest Feature Line (NFL), Center Line and Non-structured kNN techniques such as Weighted kNN, Condensed NN, Model based k-NN, Ranked NN (RNN), Pseudo/Generalized NN, Clustered kNN (CkNN), Mutual kNN (MkNN), Constrained kNN etc., are analyzed in this paper. It has been noted that memory constraints affect structure-based kNN approaches, whereas computation complexity affects non-structure-based kNN techniques. As a result, whereas non-structure kNN approaches may be used to vast volumes of data, structure-based kNN techniques can be applied to tiny volumes of data.

KEYWORDS: KNN, Applications, Query Processing.

I. INTRODUCTION

The smallest datasets are often subjected to the query processing approach, which uses any in-memory algorithms like B+trees, R-trees, etc. to provide the desired results. Yet, it is incredibly difficult for typical query processing techniques to extract the necessary data within the allotted time when the datasets are vast, high dimensional, or unclear. Under these circumstances, the closest neighbor approaches are essential. In the areas of pattern recognition, text classification, object identification, and other areas, the closest neighbor (NN) approach is relatively straightforward, extremely successful, and efficient. It has several benefits, including simplicity, robustness to noisy training data, reduced query time and memory needs, etc., but it also has drawbacks, including computational complexity, memory scalability issues, and high algorithm execution costs.

The closest neighbor (NN) rule uses its nearest neighbor, whose class is already known, to determine the category of an unknown data point. This rule is often used in applications for object identification, event recognition, ranking models text classification, pattern recognition, and text categorization. For stationary locations, a variety of strategies have been put forth for processing closest neighbor queries quickly. The entire set of data is divided into training data and sample data points, with the k-nearest neighbor being within the first group. The distance between each training point and the sample point is measured, and the point with the lowest distance is referred to as the closest neighbor. Although this method is fairly simple to use, the outcome might occasionally be affected by the value of k.

The non-structured k-NN method has been enhanced to handle the growing dimensionality of the data space. In order to identify a class of sample data points, T. M. Cover et al. hypothesized that the closest neighbor may be determined based on the value of k, which determines the number of nearest neighbors. Eventually, weights were used to enhance it. According to how far away from the sample data point they are, the training points are given weights. The two fundamental problems with the aforementioned approaches are their computational complexity and memory needs. To overcome memory constraints, the size of the data set is decreased by repeating patterns that do not offer further information, and the data points that have no bearing on the outcome are removed from the training data set.

In addition to time and memory constraints, the value of k is primarily taken into account to identify the category of the unknown sample. Many approaches, including ranking, fake neighbor information, clustering, etc., are utilized to speed up traditional kNN. In the regions of moving objects, where query indexing rather than object indexing is used, non-

structured k-NN approaches are further enhanced. As the objects in a database are constantly moving, query indexing is used instead of object indexing.

II. KNN TECHNIQUES

Structure Based KNN Technique:

Tree structures are used by the structure-based k-NN approach to represent the training datasets. A Voronoi cell approach designed exclusively for closest neighbor searches was put forth by Berchtold. Index structures used by range queries are based on branch-and-bound techniques. An R-tree indexes the points in Roussopoulos's famous technique, which uses depth-first traversal of the tree and an R-tree to discover the k nearest neighbors. The items in the tree's nodes are sorted and pruned throughout the traverse depending on a variety of criteria. This method was streamlined by Cheung and Fu without sacrificing performance. Branch-and-bound algorithms are altered using a variety of techniques, particularly when used with highdimensional data, to better fit closest neighbor situations.

When the k nearest neighbors are obtained, a variety of incremental methods for similarity ranking have been put forth that can quickly compute the (k +1)-st nearest neighbor. They employ an R-global tree's priority queue for the items that need to be visited. Hjaltason et al.'s incremental 's closest neighbor approach, which employs a priority queue of the R+-tree's items to be visited, is more detailed. They demonstrate the effectiveness of this best-first traversal for the provided R-tree. Henrich suggested an approach that uses two priority queues and is quite similar to this one. Multi-step closest neighbor query processing algorithms are often utilized for high-dimensional data. Ting Liu proposed the Ball Tree idea. It is a binary tree in which the internal nodes are utilized to quickly search through the leaves, which also carry information. The Figure demonstrates this. 1. It is faster than kNN and uses a top-down method.

The training data are split into right node and left node for the k-dimensional trees. based on query history. The tree's left or right side is looked at. Records in the terminal node are then reviewed to determine which data node is the closest to the query record. The NFL idea, which separates the training data into planes, was suggested by Stan Z. Li et al. By leveraging the feature lines that run through each pair of samples from the same class, it is utilized to improve the representational capability of a sample set with a small size. The NFL categorization approach is shown with an example. The NFL distance is given rank 1 after the assessed distances are sorted in increasing order.

Local Nearest Neighbor, put out by W. Zheng et al., is an enhancement to NFL in that it assesses the feature line and feature point in each class, but only for points whose corresponding prototypes are neighbors of query point[23]. The "Tunable Measure," which Yongli Zhou et al. introduce [24], is used to evaluate distances for NFL rather than feature lines. It initially calculates distance using configurable metrics before implementing NFL steps.

Non Structure Based KNN:

For resolving closest neighbor problems for moving objects in one dimensional space, Kollios et al. put up a sophisticated method. The future trajectory of a moving point $x(t) = x_0 + v_x t$ is changed into a point (x_0, v_x) in a so-called dual space by their approach using a duality transformation. In the "1.5-dimensional" example, when objects are traveling in the plane but are only allowed to move along a few line segments, the solution is generalized to include all possible scenarios (e.g., corresponding to a road network). The processing of Continuous RNN (CRNN) requests when $k=1$ was, nevertheless, the subject of research by Tian Xia and Donghui Zhang. They use a technique called 60-degree pruning for their approach.

The monitoring region of a CRNN query is defined as six pie-regions and six cir-regions. A continuous nearest neighbor query is used to monitor the candidate in each sub-space, called continuous filter. Frequent Updated R-Tree (FUR) and hash tables are used to store the cir-region, and optimization techniques like lazy-update and partial-insert are proposed to avoid unnecessary NN searches and reduce the updates on the FUR-tree. Albers et al. [29] investigated Voronoi diagrams of continuously moving points, and proposed a solution for finding the k nearest neighbors for a moving query point. Song et al. and Tao et al. considered the nearest neighbor problem for a query point moving on a line segment for static and for moving data points. The time period is divided into n equal-length intervals, and the algorithm tries to reuse the information contained in the result sets of the previous samples. Bailey and proposed weighted kNN (WkNN) algorithms use weights to evaluate distances and assign weights to each calculated value. CNN and Reduced Nearest Neighbor (RNN) algorithms store the patterns one by one and eliminate duplicate ones. Model Based kNN selects similarity measures and creates a similarity matrix. Subash C et al. improved the kNN by introducing the concept of ranks. Modified kNN uses validity of all data samples in the training data set to classify the class of the sample data point.



Yong zeng et al., proposed a new concept to classify sample data points. The method introduces the pseudo neighbor, which is not the actual nearest neighbor. Euclidean distance is evaluated and pseudo neighbor with greater weight is found and classified for unknown sample. Clustering is used to calculate nearest neighbor. Each training dataset is clustered based on the $_k$ value and all cluster centers form a new training set. Weighting is assigned to each cluster according to number of training samples in clusters.

III. REVERSE KNN TECHNIQUE

Reverse k-nearest neighbor (RkNN) queries return the data objects that have the query object in the set of their nearest neighbors. The goal of aRkNN query is to identify the influence of a query object on the whole data set. However, the relationship between k-NN and RkNN is not symmetric and the number of the reverse K-nearest neighbors is not known in advance. RkNN queries are more expensive than k-NN queries, requiring the running time of $O(n^2)$ and $O(n)$.

Stanoi et al. has proved that the RNN problem can be reduced to the NN problem. Korn and Muthukrishnan use two R-trees for the querying, insertion, and deletion of points. Yang and Lin introduce an Rdn-tree, which makes it possible to answer both RNN queries and NN queries using a single tree. Singh et al. proposed an algorithm where RkNN candidates are found by performing a regular kNN query, but it does not always find all RkNN points. Tao et al. and Maheshwari et al. propose main memory data structures for answering RNN queries in two dimensions. Tobias Emrich et al. proposed a novel concept of Constrained Reverse k-Nearest Neighbor (CRkNN) search on mobile objects (clients) performed at a central server. The CRkNN query computes the set $RkNN(q)$ of objects, for a given query object q , if and only if the result set exceeds a specific threshold, say $_m$, else the query reports an empty result. This approach minimizes the amount of communication between clients and central server by using the approximation of the positions to identify true hits and true drops. An example of CRkNN for monochromatic and bichromatic cases is shown in [Figure.3]. The mono-chromatic CR1NN query returns points 1 and 4 if, for some threshold value $m \in \{1, 2\}$ or nothing if $m \leq 3$. In the bi-chromatic case, two object sets Dred (red objects) and Dblue (blue objects) are considered.

The algorithm automatically generates mask image without user interaction that contains only text regions to be inpainted.

IV. OTHER NEAREST NEIGHBOR TECHNIQUE

Dimitris Papadias et al. and YunjunGao et al. proposed two types of MNN queries: MNNP and MNNT queries, which utilize batch processing and reusing technology to reduce the I/O cost and CPU time. They also proposed a technique on finding the mutual nearest neighbor, which returns the mutual neighbors of q (for a specified k_1 and k_2) at any time. However, this algorithm doesn't deal with the bichromatic datasets and can be further pruned to increase the performance.

S · N o	Technique	Concept	Merits	Demerits	Applications
1	BallTreeknearest neighbor(BTKNN)[19,20]	To improve the speed	1. Compatible with high-dimensional objects. 2. Represented data are tuned well to structure. 3. Simple to implement. 4. Especially used for geometric learning.	Implementation cost is high. When distance is increased, performance is decreased.	Robotic, vision, speech, graphics



2	k-d treenearest neighbor(kdNN)[21]	To dividethe trainingdata setsintotwohalves	1.Perfect balancedtrees are formed2.It is fast andsimple	1.Computationalcomplexity isrequired 2.Exhaustive search 3.Chanceofmisleadingthe points as it blindlysplits the points into twohalves.	Multidimensi onal datapoints.
3	Nearestfeature LineNeighbor(NFL)[22]	To havemultipletemp late perclass forclassification	1.Accurateclassificatio n2.Effective algorithmforsmalldatas ets. 3.Ignored informationin nearest neighbor areused	1.Chance of failure ifthe model inNFL isfar away from querypoint2.Computational Complexity 3. Hard to illustrate thefeature point in straightline.	Face Recognitionpr oblems
4	Local NearestNeighbor[23]	To focus onnearestneighbo rprototypeof querypoint	1.Overcomes thelimitationsofNFL	1.Increase in number ofcomputations	FaceRecogniti on
5	Tunable NearestNeighbor(TNN)[24]	Calculatethedista ncefirst andthenimplemen tstheseofNFL	1.Effectiveforsmalldata sets	1.Large number ofcomputations	Biasproblems
6	CenterbasedNearestNeighbor(CNN)[25]	To connectsamplepoi nt withknownlabel e dpointsona center line	1.Highlyefficientforsm alldatasets	1.Largenumberofcomputatio ns	PatternRecog nition
7	Orthogonal SearchTreeNearestNeighbor	Tospeedupthepro cess ,Orthogonalsearch treesareused	1.LessComputationitim e 2.Effectivefor large datasets	1.Querytimeismore	PatternRecog nition
8	Principal AxisTree Nearest Neighbor (PAT)[27]	Uses PATconstructionand PATsearch	1.Goodperformance2.F astSearch	1.Computational Timeismore	PatternRecog nition
9	kNearestNeighbor(k NN)[9]	To findthenearestnei ghbor based on 'k' value	1.Trainingisveryfast2.S imple and easy tolearn 3.Robust to noisyytrainingdata 4 .Effective if training data islarge5.Itissymmetric.	1.Biasedbyvalueofk2.Comp utationComplexity 3.Memorylimitation4.Being asupervisedlearning lazyalgorithm i.e. runsslowly 5.Easilyfooledby	Largesampled ata



10	Weightedknearestneighbor (WkNN) [33]	To assignweightston eighborsbased ondistancecalculated	1. Overcomeslimitatio nsofkNNby assigning equalweighttokneighbo rs implicitly. 2. Uses all trainingsamplesnot justk. 3. Makes the algorithmglobalone	1.Computationalcomplexity increasesincalculatingweigh ts2.Slowin execution	Largesampled ata
11	Condensed nearest neighbor (CNN)[34]	To eliminate data setswhich showsimilaritywit houtadding extra information	1. Reducesizeoftrai ningdata 2. Improvequerytimeand memory requirements 3. Reduce the recognition rate	1.CNN is orderdependent; it isunlikelytopickuppoin tson boundary. 2. Computational Complexity	Datasetwhere memory requirementis amainconcern
12	Reduced Nearest Neighbor (RNN)[35]	To removepatternsw hichdonot affect the training data setresults	1. Reduced size oftrainingdata and eliminatetemplates 2. Improved querytimeand memory requirements 3. Reduced recognition rate	1.ComputationalComplexity 2.Cost is high3.TimeConsuming	Largedataset
13	Model based k Nearest Neighbor(MkNN)[36]	To construct a modelfrom dataand classifynewdataus ing thismodel	1. More classification accuracy 2.Value of k is selectedautomatically 3.Highlyefficient due to reducednumber ofdatapoints	1.Donotconsidermarginaldat aoutsidetheregion	Dynamic web miningforlarg erepository
14	Rank nearest neighbor(kRNN)[37]	To assignrankstotrain ing datafor eachcategory	1.Performsbetterwhent here are too muchvariations betweenfeatures 2.Robust asbased onrank	1.MultivariatekRNNdepend sondistributionofthedata	Class distribution ofGaussian nature
15	Modifiedknearestnei ghor(MkNN)[38]	To classifynearestnei ghorbased on weights andvalidityofdata point	1.Partially overcomeslow accuracy ofWkNN 2.Stableandrobust	1.ComputationalComplexity	Methods facingoutlets
16	Pseudo/Generalized Nearest Neighbor(GNN)[39]	To utilize information of (n-1)neighborsalso	1.Uses(n-1) classeswhich consider thewholtrainingdatase t	1.Does not holdgoodfor small data2.Computationalcomple xity	Largedataset
17	Clusteredk nearestneighbor[40]	To select thenearestneighbo rfrom theclusters	1.Overcome defectsof unevenistribut ions of trainingsamples 2.Robust innature	1.Selection of thresholdparameter is difficultbefore runningalgorithm 2.Biased byvalueofkforclustering	TextClassifica tion



18	Reverseknearestneighbor[41-46]	Objects that have the query object as their nearest Neighbour, have to be found.	1. Approximate results can be obtained veryfast. 2. Wellsuitedfor2-Dimensionalsets 3. Wellsuitedforfinite,storeddatasets 4. Provides decision support	1. requires $O(n^2)$ time 2. do not supportarbitraryvaluesof k 3. cannotdealefficientlywithdatabaseupdates, 4. areapplicableonlyto2D	Spatialdataset
19	Continuous RkNN[47]	To monitor the regions upon updates using FUR tree	1. Overcomes thedifficulties of using thekNNandRkNNqueries onmovingobjects. 2. Best suited for monochromatic cases	1. Not suited forbichromaticcases 2. Notsuitedforlargepopulation ofcontinuously movingobjects. 3. MemoryLimitation	Movingobject dataset
20	Constrained RkNN[48]	To find the RkNN on moving objects based on constrains	1. Communicationloadis minimized. 2. CRkNNcan beapplied to both monochromatic and bichromaticcases.	1.Approximate resultcan be obtained forbichromaticcases.	Moving object datasetespeciallyinGPS
21	AggregatekNN[49]	To use aggregate function for finding the nearest neighbor	1.Provides memory-residentqueriesandcost models thataccurately predict theirperformanceinterms ofnodeaccesses 2. Approximate resultcan beobtainedfor	1. Cost for evaluatingthe disk-resident querymodelis high. 2. Lazyalgorithm	Spatialdataset
22	Mutual NearestNeighbor[50]	To find the Mutual Nearest Neighbor using TB- tree.	1.Usesbatchprocessing and reuse technology forreducing I/O cost andCPUtime. 2.HCMNNisusedto reduce the searchingtime of all the dataagain and again.	1. Appliedonlytothemonochromaticdatasets. 2. Computational complexity	Movingobject dataset

V. CONCLUSION

This paper compares the merits and demerits of structure and non-structure k-Nearest Neighbor techniques. Structure based algorithms such as Ball Tree, k-d Tree, Principal Axis Tree (PAT), and Orthogonal Structure Tree (OST), Nearest Feature Line (NFL), and Center Line (CL) are easy to construct and cost effective. Non structure based kNN techniques like simple k-NN technique have the drawback of memory limitation and biased by 'k' value. Constrained RkNN, Continuous RkNN and Mutual Nearest Neighbor are widely used in the moving object datasets, Aggregate kNN and Reverse kNN techniques are used in spatial datasets, and various kNN algorithms are proposed to reduce the time complexity and computational complexity. Only few algorithms are there to reduce the dimensionality and is yet to be analyzed by the researchers.

REFERENCES

[1] V.Vaidehi, S.Vasuhi, — Person Authentication using Face Recognition || , Proceedings of the world Congress on Engineering and Computer Science, 2008.
 [2] Shizen, Y. Wu, —An Algorithm for Remote Sensing Image Classification based on Artificial Immune b-cell Network || , Springer Berlin, Vol 40.

- [3] G. Toker, O. Kirmemis, —Text Categorization using k Nearest Neighbor Classification || , Survey Paper, Middle East Technical University.
- [4] Y. Liao, V. R. Vemuri, —Using Text Categorization Technique for Intrusion detection || , Survey Paper, University of California.
- [5] E. M. Elnahrawy, —Log Based Chat Room Monitoring Using Text Categorization: A Comparative Study || , University of Maryland.
- [6] X. Geng et. al, —Query Dependent Ranking Using k Nearest Neighbor || , SIGIR, 2008.
- [7] F. Bajramoviec. al —A Comparison of Nearest Neighbor Search Algorithms for Generic Object Recognition || , ACIVS 2006, LNCS 4179, pp 1186-1197.
- [8] Y. Yang and T. Ault, —Improving Text Categorization Methods for event tracking || , Carnegie Mellon University.
- [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] Berchtold, S., Ertl, B., Keim, D.A., Kriegel, H.P., Seidl, T. —Fast nearest neighbor search in high-dimensional space || in Proceedings of the International Conference on Data Engineering, pp. 209 - 218 (1998)
- [11] Roussopoulos, N., Kelley, S., Vincent, F.: Nearest neighbor queries. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 71–79 (1995)
- [12] Guttman, A.: R-trees: A dynamic index structure for spatial searching. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 47–57 (1984)
- [13] Cheung, K.L., Fu, A.W.-C.: Enhanced nearest neighbor search on the R-tree. ACM SIGMOD Record 27(3), 16–21 (1998)
- [14] Katayama, N., Satoh, S.: The SR-tree: An index structure for high-dimensional nearest neighbor queries. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp.369–380(1997)
- [15] Henrich, A.: A distance scan algorithm for spatial access structures. In: Proceedings of the Second ACM Workshop on Geographic Information Systems, pp. 136– 143 (1994)
- [16] Hjaltason, G.R., Samet, H.: Distance browsing in spatial databases. ACM Trans. Database Sys. 24(2), 265–318 (1999)
- [17] Beckmann, N., Kriegel, H.P., Schneider, R., Seeger, B.: The R*- tree: An efficient and robust access method for points and rectangles. In: Proceedings of the ACM SIGMOD International Conference On Management of Data, pp. 322–331 (1990)
- [18] Seidl, T., Kriegel, H.P.: Optimal multi-step k-nearest neighbor search. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 154–165 (1998)
- [19] T. Liu, A. W. Moore, A. Gray, —New Algorithms for Efficient High Dimensional Non-Parametric Classification || , Journal of Machine Learning Research, 2006, pp 1135-1158.
- [20] S. N. Omohundro, —Five Ball Tree Construction Algorithms || , 1989, Technical Report.
- [21] . F Sproull, —Refinements to Nearest Neighbor Searching || , Technical Report, International Computer Science, ACM (18) 9, pp 507-517. International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010
- [22] S. Z Li, K. L. Chan, —Performance Evaluation of The NFL Method in Image Classification and Retrieval || , IEEE Trans On Pattern Analysis and Machine Intelligence, Vol 22, 2000.
- [23] W. Zheng, L. Zhao, C. Zou, —Locally Nearest Neighbor Classifier for Pattern Classification || , Pattern Recognition, 2004, pp 1307-1309.
- [24] Y. Zhou, C. Zhang, —Tunable Nearest Neighbor Classifier || , DAGM 2004, LNCS 3175, pp 204-211.
- [25] Q. B. Gao, Z. Z. Wang, —Center Based Nearest Neighbor Class || , Pattern Recognition, 2007, pp 346-349.
- [26] Y. C. Liaw, M. L. Leou, —Fast Exact k Nearest Neighbor Search using Orthogonal Search Tree || , Pattern Recognition 43 No. 6, pp 2351-2358.
- [27] J.Mcname, —Fast Nearest Neighbor Algorithm based on Principal Axis Search Tree || , IEEE Trans on Pattern Analysis and Machine Intelligence, Vol 23, pp 964-976.



INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118



ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 **9940 572 462**  **6381 907 438**  **ijirset@gmail.com**



www.ijirset.com

Scan to save the contact details