International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue14
November 2017

An Efficient k-Nearest Neighbors Approach Based on Various-Widths clustering

[1]T . Jithendar , [2] Ch.Surendra, [3]Mesa Kalpana

[1-2] Assistant Professor, Department of Computer Science and Engineering, Guru nanak Institutions technical campus, Hyderbad, T.S,India.

[3] Assistant Professor, Department of Computer Science and Engineering ,Aarushi Group of Institution college of Engineering warangal, TS, India.

Email : kalpanamesa112@gmail.com , Email-: cstjithendar@gmail.com
Email : billu.suri@gmail.com

Abstract: *The approximate k-NN search algorithms are well-known for their high concert in high dimensional data. Thelocality-sensitive hashing (LSH) method, that uses a number of hash functions, is one of the most fascinating hash-based approaches. The k-nearest neighbour approaches based Various-Widths Clustering (kNNVWC) has been widely used as a prevailing non-parametric technique in many scientific and engineering applications. However, this approach incurs a huge pre-processing and the querying cost. Hence, this issue has become an active explore field. The proposed system presents a novel k-NN based Partitioning Around Medoids (KNNPAM) clustering algorithm to powerfully find k-NNs for a query object from a given data set to minimize the extend beyond among clusters; and grouping the centers of the clusters into a tree-like index to effectively trim more clusters. Experimental results demonstrate that KNNPAM perform well in finding k-NNs for query objects compared to a number of k-NN search algorithms, mainly for a banking domain and real world data set with high dimensions, various distributions and large size. The problem of quickly finding the "exact" k-NN for a query object in a large and high dimensional data set using metric reserve functions that satisfy the triangle inequality property.KD-tree: To organization the data set in a balanced binary-tree, where the data set is recursively split into two parts along one axis .*

KEYWORDS: **Clustering, Non-parametric, Medoids, Experimental, Partitioning**

1. INTRODUCTION

The k-nearest neighbor approach (kNN) has been extensively used as a powerful nonparametric technique in many scientific and engineering applications. However, this approach incurs a large computational cost. Hence, this issue has become an active research field. In this work, a novel kNN approach based on various widths clustering, named kNNVWC, to efficiently find kNNs for a query object from a given data set, is presented. kNNVWC does clustering using various widths, where a data set is clustered with a global width first and each produced cluster that meets the predefined criteria is recursively clustered with its own local width that suits its distribution. This reduces the clustering time, in addition to balancing the number of produced clusters and their respective sizes. Maximum efficiency is achieved by using triangle inequality to prune unlikely clusters.

Experimental results demonstrate that kNNVWC performs well in finding kNNs for query objects compared to a number of kNN search algorithms, especially for a data set with high dimensions, various distributions and large size. he*k*-Nearest Neighbors algorithm (or *k*-NN for short) is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the *k* closest training examples in the feature space. *k*-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/*d*, where *d* is the distance to the neighbor. Fixed width clustering creates a set of clusters of fixed radius (width) w. Here the width w is a parameter to be specified by the user. First, a data vector is taken and used as the centroid (center) of the first cluster with radius w. Then for each subsequent data vector the Euclidean distance between the centroid of the current clusters and this data vector is computed.

If the distance to the closest cluster center from the data vector is less than the radius w, the data vector is added into that cluster and the centroid of that cluster is adjusted to the mean of the data vectors it contains. If the distance to the closest cluster center is more than the radius w, then a new cluster is formed with that data vector as the centroid. This operation produces a set of disjoint, fixed width (radius of w) clusters in the feature space.

There is a huge number of clustering algorithms and also numerous possibilities for evaluating a clustering against a gold standard. The choice of a suitable clustering algorithm and of a suitable measure for the evaluation depends on the clustering objects and the clustering task. The clustering objects within this thesis are verbs, and the clustering task is a semantic classification of the verbs. Further cluster parameters are to be explored within the cluster analysis of the verbs.

II. LITERATURE SURVEY

P. Cunningham et al[1] addresses Perhaps the most straightforward classifier in the arsenal or machine learning techniques is the Nearest Neighbour Classifier – classification is achieved by identifying the nearest neighbours to a query example and using those neighbours to determine the class of the query. This approach to classification is of particular importance today because issues of poor run-time performance are not such a problem these days with the computational power that is available. This paper presents an overview of techniques for Nearest Neighbour classification focusing on; mechanisms for assessing similarity (distance), computational issues in identifying nearest neighbours and mechanisms for reducing the dimension of the data.

A.Vergara et al[2] have presented Sensor drift remains to be the most challenging problem in chemical sensing. To address this problem we have collected an extensive dataset for six different volatile organic compounds over a period of three years under tightly controlled operating conditions using an array of 16 metal-oxide gas sensors. The recordings were made using the same sensor array and a robust gas delivery system.. We introduced a machine learning approach, namely an ensemble of classifiers, to solve a gas discrimination problem over extended periods of time with high accuracy rates. Experiments clearly indicate the presence of drift in the sensors during the period of three years and that it degrades the performance of the classifiers. Our proposed ensemble method based on support vector machines uses a weighted combination of classifiers trained at different points of time. As our experimental results illustrate, the ensemble of classifiers is able to cope well with sensor drift and performs better than the baseline competing methods.

C. Silpa-Anan et al[3] have proposed In this paper, we look at improving the KD-tree for a specific usage: indexing a large number of SIFT and other types of image descriptors. We have extended priority search, to priority search among multiple trees. By creating multiple KD-trees from the same data set and simultaneously searching among these trees, we have improved the KD-tree's search performance significantly. We have also exploited the structure in SIFT descriptors (or structure in any data set) to reduce the time spent in backtracking. By using Principal Component Analysis to align the principal axes of the data with the coordinate axes, we have further increased the KD-tree's search performance.

M. Muja et al[4] discussed For many computer vision and machine learning problems, large training sets are key for good performance. However, the most computationally expensive part of many computer vision and machine learning algorithms consists of finding nearest neighbor matches to high dimensional vectors that represent the training data. We propose new algorithms for approximate nearest neighbour matching and evaluate and compare them with previous algorithms. For matching high dimensional features, we find two algorithms to be the most efficient: the randomized k-d forest and a new algorithm proposed in this paper, the priority search k-means tree. We also propose a new algorithm for matching binary features by searching multiple hierarchical clustering trees and show it outperforms methods typically used in the literature. We show that the optimal nearest neighbor algorithm and its parameters depend on the data set characteristics and describe an automated configuration procedure for finding the best algorithm to search a particular data set. In order to scale to very large data sets that would otherwise not fit in the memory of a single machine, we propose a distributed nearest neighbour matching framework that can be used with any of the algorithms described in the paper.

N. Kumar et al[5] addresses Many computer vision algorithms require searching a set of images for similar patches, which is a very expensive operation. In this work, we compare and evaluate a number of nearest neighbors algorithms for speeding up this task. Since image patches follow very different distributions from the uniform and Gaussian distributions that are typically used to evaluate nearest neighbors methods, we determine the method with the best performance via extensive experimentation on real images. Furthermore, we take advantage of the inherent structure and properties of images to achieve highly efficient implementations of these algorithms. Our results indicate that vantage point trees, which are not well known in the vision community, generally offer the best performance.

M. J. Prerau [6] discussed Most current intrusion detection methods cannot process large amounts of audit data for real-time operation. In this paper, anomaly network intrusion detection method based on Principal Component Analysis (PCA) for data reduction and Fuzzy Adaptive Resonance Theory (Fuzzy ART) for classifier is presented. Moreover, PCA is applied to reduce the high dimensional data vectors and distance between a vector and its projection onto the subspace reduced is used for anomaly detection. Using a set of benchmark data from KDD (Knowledge Discovery and Data Mining) Competition designed by DARPA for demonstrate to detection intrusions. Experimental results show the proposed model can classify the network connections with satisfying performance.

A. Almalawi et al[7],have presented Supervisory Control and Data Acquisition (SCADA) systems monitor and control infrastructures and industrial processes such as smart grid power and water distribution systems. Recently, such systems have been attacked, and traditional security solutions have failed to provide an appropriate level of protection. Therefore, it is important to develop security solutions tailored to SCADA systems. However, it is impractical to evaluate such solutions on actual live systems. This paper proposes a SCADA security testbed based on virtualization technology, and introduces a server which is used as a surrogate for water distribution systems. In addition, this paper presents a case study of two malicious attacks to demonstrate how the testbed can easily monitor and control any automatised processes, and also to show how malicious attacks can disrupt supervised processes.

S. Suthaharanet al8] addresses Security of wireless sensor networks (WSN) is an important research area in computer and communications sciences. Anomaly detection is a key challenge in ensuring the security of WSN. Several anomaly detection

algorithms have been proposed and validated recently using labeled datasets that are not publicly available. Our group proposed an ellipsoid-based anomaly detection algorithm but demonstrated its performance using synthetic datasets and real Intel Berkeley Research Laboratory and Grand St. Bernard datasets which are not labeled with anomalies. This approach requires manual assignment of the anomalies' positions based on visual estimates for performance evaluation. In this paper, we have implemented a single-hop and multi-hop sensor-data collection network. In both scenarios we generated real labeled data for anomaly detection and identified different types of anomalies. These labeled sensor data and types of anomalies are useful for research, such as machine learning, and this information will be disseminated to the research community.

M. Bawa, et al[9] have discussed We consider the problem of indexing high-dimensional data for answering (approximate) similarity-search queries. Similarity indexes prove to be important in a wide variety of settings: Web search engines desire fast, parallel, main-memory-based indexes for similarity search on text data; database systems desire disk-based similarity indexes for high-dimensional data, including text and images; peer-to-peer systems desire distributed similarity indexes with low communication cost. We propose an indexing scheme called LSH Forest which is applicable in all the above contexts. Our index uses the well-known technique of locality-sensitive hashing (LSH), but improves upon previous designs by (a) eliminating the different data-dependent parameters for which LSH must be constantly hand-tuned, and (b) improving on LSH's performance guarantees for skewed data distributions while retaining the same storage and query overhead. We show how to construct this index in main memory, on disk, in parallel systems, and in peer-to-peer systems. We evaluate the design with experiments on multiple text corpora and demonstrate both the self-tuning nature and the superior performance of LSH Forest.

T. Liu et al [10] addresses This paper is about non-approximate acceleration of high-dimensional nonparametric operations such as k nearest neighbor classifiers. We attempt to exploit the fact that even if we want exact answers to nonparametric queries, we usually do not need to explicitly find the data points close to the query, but merely need to answer questions about the properties of that set of data points. This offers a small amount of computational leeway, and we investigate how much that leeway can be exploited. This is applicable to many algorithms in nonparametric statistics, memory-based learning and kernel-based learning. But for clarity, this paper concentrates on pure k-NN classification. We introduce new ball-tree algorithms that on real-world data sets give accelerations from 2-fold to 100-fold compared against highly optimized traditional ball-tree-based k-NN . These results include data sets with up to 106 dimensions and 105 records, and demonstrate non-trivial speed-ups while giving exact answers.

M. Hall et al[11] have presented The novel notion of one-time outlier query is introduced in order to detect anomalies in the current window at arbitrary points-in-time. Three algorithms are presented. The first algorithm exactly answers to outlier queries, but has larger space requirements than the other two. The second algorithm is derived from the exact one, reduces memory requirements and returns an approximate answer based on estimations with a statistical guarantee. The third algorithm is a specialization of the approximate algorithm working with strictly fixed memory requirements. Accuracy properties and memory consumption of the algorithms have been theoretically assessed. Moreover experimental results have confirmed the effectiveness of the proposed approach and the good quality of the solutions.

C. T. Jr et al[12] addresses In this paper we present the Slim-tree, a dynamic tree for organizing metric datasetsin pages of fixed size. The Slim-tree usesthe “fat-factor” which provides a simple way to quantify the degree of overlap between the nodes in a metric tree. It is well-known that the degree of overlap directly affects the query performance of index structures. There are many suggestions to reduce overlap in multidimensional index structures, but the Slim-tree is the first metric structure explicitly designed to reduce the degree of overlap. Moreover, we present new algorithms for inserting objects and splitting nodes. The new insertion algorithm leads to a tree with high storage utilization and improved query performance, whereas the new split algorithm runs considerably faster than previous ones, generally without sacrificing search performance.

III.CONCULSION

An Efficient k-Nearest Neighbors Approach Based on Various-Widths clustering to handle a difficult data set for classification. The Number of feature can easily compared and performance of widths clustering will provide a effective results. K- Nearest Neighbour survey domain is proposed in this literature survey. k-NN based Partitioning Around Medoids (KNNPAM) clustering algorithm to efficiently find k-NNs for a query object from a given data set to minimize the overlap among clusters; and grouping the centers of the clusters into a tree-like index to effectively prune more clusters

REFERENCES

[1]. P. Cunningham and S. J. Delany, “k-nearest neighbourclassi- fiers,” Multiple Classifier Systems, pp. 1–17, 2007.

[2] .A.Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta, “Chemical gas sensor drift compensation using

classifier ensembles," Sens. Actuators B: Chemical, vol. 166, pp. 320–329, 2012.

[3]. C.Silpa-Anan and R. Hartley, "Optimisedkd-trees for fast image descriptor matching," in Proc. IEEE Conf. Comput. Vis.

Pattern Recog., 2008, pp. 1–8.

[4]. M. Muja and D. Lowe, "Scalable nearest neighbour algorithms for high dimensional data," IEEE Trans. Pattern Anal. Mach.

Intell., vol. 36, no. 11, pp. 2227–2240, Nov. 1, 2014.

[5]. N. Kumar, L. Zhang, and S. Nayar, "What is a good nearest neighbours algorithm for finding similar patches in images?" in

Proc.10th Eur. Conf. Comput. Vis., 2008, pp. 364–378.

[6]. M. J. Prerau and E. Eskin, "Unsupervised anomaly detection using an optimized k-nearest neighbors algorithm,"

Undergraduate thesis, Columbia Univ., New York, NY, USA, Dec. 2000.

[7]. A. Almalawi, Z. Tari, I. Khalil, and A. Fahad, "SCADAVT-a framework for SCADA security testbed based on virtualization technology," in Proc. IEEE 38th Conf. Local Comput. Netw., 2013,pp. 639–646.

[8]. S. Suthaharan, M. Alzahrani, S. Rajasegarar, C. Leckie, and M. Palaniswami, "Labelled data collection for anomaly detection in wireless sensor networks," in Proc. 6th Int. Conf. Intell. Sensors,SensorNetw. Inf. Process., Dec. 2010, pp. 269 –274.

[9]. M. Bawa, T. Condie, and P. Ganesan, "LSH forest: self-tuning indexes for similarity search," in Proc. 14th Int. Conf. World

Wide Web., 2005, pp. 651–660.

[10].T. Liu, A. W. Moore, and A. Gray, "New algorithms for efficient high-dimensional nonparametric classification," J. Mach.

Learning Res., vol. 7, pp. 1135–1158, 2006 .

[11]. F. Angiulli and F. Fassetti, "DOLPHIN: An efficient algorithm for mining distance-based outliers in very large datasets,"

ACMTrans.Knowl.Discovery Data, vol. 3, no. 1, p. 4, 2009.

[12]. C. T. Jr., A. Traina, B. Seeger, and C. Faloutsos, Slim-Trees: High Performance Metric Trees Minimizing Overlap Between Nodes. New York, NY, USA: Springer, 2000.