ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Survey On The Applications Of K-Nearest Neighbour Algorithm And Its Variants.

S Preethi*, and PLK Priyadarsini.

School of Computing, SASTRA University, Tamil Nadu, India.

ABSTRACT

Big data Analytics are a developing field that analyses huge amount of structured, semi-structured and unstructured data that has a possibility to be mined for information. Many techniques for analytics have been evolving in this field. One of such developments is the use of K- Nearest Neighbor algorithm. This technique with different variations is being used to perform Big Data Analytics in several research fields such as medical data analysis, Social Networks and the like. In this paper, we present a survey of some of the works on using K-Nearest Neighbor algorithm and its variants in different fields of research. We also present some of the theoretical results on K-Nearest Neighbor classification and detection to improve its effectiveness.

Keywords: Bigdata, KNN Classification, Data Center, Medical Application

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Big data Analytics [1] is a developing field with a conceptual idea to analyze large amounts of structured, semi-structured and unstructured data that has a possibility to be mined for information to allow improved decision making, insight, and process optimization. It normally deals with data sets with sizes ranging from terabytes to many petabytes. Further Big data Analytics field explores the capacity of frequently used computer program implement to occupy, create method and handle statistics in an acceptable expire time.

Big data, classify has been moved from a few dozen terabytes to many petabytes. Hence there is a necessity of techniques to have a better insight of data sets that are various, compound, and of a massive scale. The challenges and opportunities in this field can be viewed as being three-dimensional namely with increasing volume (quantity of data), velocity (speed of data in and out), and variety (extent of data types and sources). Later it was identified that there are few more factors [1] that are important in Big Data Analytics, namely Veracity, viscosity, Value and Variability.

Many of the Big Data challenges are attempting to solve by using several techniques like Association rule learning, Classification tree analysis, Genetic algorithms, Machine learning, Regression analysis and Natural Language Processing and the like. In data mining, classification that can be used a supervised learning by the report of consequential statistics classes, where class is an aspect to classify in the analysis of complicated. The Nearest Neighbor is a very simply its hugely able to the algorithm for pattern recognition.

Nearest neighbor classification is utilized all the feature is continual. It performs classification purely on similarity basis. As the name of the algorithm indicates, for classification, of a novel tuple, it focuses on k nearest tuples. Nearest Neighbor (NN) techniques can be generally categorize into two types 1) structure less NN techniques 2) structure based NN techniques. The structure less NN technique the entire information is typed into training and test sampling data. Distance is found out from a training point to sample point and it is called an interval. Intervals are calculated into the point with the minimum intervals are named nearest neighbor. K nearest neighbors (KNN) falls into this category. Structure-based NN techniques are located on arrange of information like orthogonal structure tree (OST), ball tree, k-d tree, axis tree.

K-nearest neighbors(KNN) algorithm is an algorithm used very often for classification and regression. The KNN algorithm also named as case-based reasoning, example-based reasoning, instance based learning, memory based reasoning or lazy learning.

The method of classification in KNN begins with a data point. The data point is the construct of a certain number of attributes that define a data point. The data point is separated into two sets: the training set and test set. The training set is used as input to the algorithm while test set is used to calculate the efficiency of an algorithm. The separation of the data set can be done using various methods such as arbitrator method, random sampling, and cross-validation [2] . KNN classifies any new tuple by using training data tuples similar to it. In KNN algorithm the training tuples can be held as a set of data points *n* dimensions are set *n* attributes reported in the data set. To find out the *k* closest(nearest) data points to a tuple in the *n*-dimensional space, different distance metrics are used and some of them are Euclidean distance, Minkowski distance, Manhattan distance [2].

There are several variations of KNN algorithm and they are all widely used in different fields of research involving real-time environments like medical, manufacturing field retails application, network analysis, etc [3]. In this paper, we review some of the publications, which use KNN-classification algorithm and its variants for finding solutions to problems in different fields like medicine, networking.

In the next section, we review the publications with Medical applications. In section 3, we review a collection of publications covering several other applications. In section 4, we present some of the theoretical works on KNN algorithm. We conclude the review in section 5.

ISSN: 0975-8585

MEDICAL APPLICATIONS

Classification of heart diseasesposea challenging problem and research communities are doing various works. The first two papers in this section deal with the usage of KNN with other machine learning and evolutionary algorithms for analysis which help in diagnosis.

Disorders is one of the most among critical heart diseases of the heart valve. This is One of the several methods to identify this disorder is through doppler technique. But for obese people and people with calcified disease, diagnosis through this technique is not satisfactory unless the spectrogram of the Doppler shift signals to find out the degree of the disease. In medical systems, the effect of artificial intelligence plays very important role in detecting the problems. In [4], a decision support system to detect the heart valve diseases is proposed. Wavelet transform and short time Fourier transform are applied on the Doppler ultrasound to extract the features. An artificial immune based fuzzy k-NN algorithm was hybridized to for classification. The results were compared with other works using Artificial Neural Networks, SVM and so on. The results obtained are comparatively better than ANN and SVM.

In [5], the authors use KNN combined with a genetic algorithm for effective classification The main benefit of a genetic algorithm is used to enhance the level of accuracy of findings. Here the nearest neighbor technique is also implemented where classification is done without requiring any additional data instead it uses only its training samples. The genetic algorithm is an evolutionary algorithm for finding the global optimum solution for an optimization problem [6]. In the proposed model the authors claim that the accuracy level of the results was improved by reducing the redundancy using a combination of genetic algorithm with KNN. They prune some of the irrelevant attribute using a genetic algorithm. The set of final attributes is used to classify the patients with heart disease and without the disease. Tests were conducted on medical data from UCI repository [5].

In [7], the authors use the same combination of genetic algorithm and KNN helps to identify the modules present in the lung in the diagnosis of lung cancer. Comparing to other classification methodologies, this proved to be more accurate. The paper suggests a three-phase methodology - preprocessing and gray conversion, feature extraction using Gabor filter, then a combination of KNN classification and genetic algorithm.

Chemical burns would cause serious damage to the skin of the human body system and various research works are being carried out to identify the severity of burns. The severity is classified in three ways, named like superficial, partial thickness and full thickness. In [8], the authors use a KNN classifier to analyze the burn injuries after preprocessing. As a first step, various images were collected from hospitals and maintained as a medical image database. The histogram processed utilizing the difference equalization to improve. Then change the color images from RGB arrange to the L*a*b separate to space. The A and B element of the L*a*b color split represents the chrominance of the image. The sub-samples of this were used to extract appearance, such represent the Discrete co-sign transform (DCT), Skewness, Kurtosis. These appearances were utilized as input data are KNN classifier.

Autoimmune diseases are diagnosed through Indirect Immunofluorescence (IIF) tests using Human Epithelial (HEp-2) cells. Patients' serum contains Anti-Nuclear Antibodies (ANAs). Auto Immune diseases are identified by manually searching and classifying the fluorescents training patterns in the cells generated by ANAs. The validity of test usually depends on the experience and expertise of the physician. In [9], HEp-2cellsclassifier that works on cell-level is proposed. The classifier for similarity searches is based KNN algorithm on MESSIF framework. They used Haralick features, Local Binary Patterns, scale Invariant Feature Transform, Granulometry-base descriptor, Surface descriptor and Color Structure (CS) descriptor from the set of MPEG-7 as image descriptions.

In [10] a variant of KNN namely citation-KNN [11] is improved and used to identify a malignant mass image in the Breast ultrasound images. Multiple instance learning (MIL) is normally to resolve the training data issue with uncompleted instruction almost labels of training data. In the conventional supervised research issue with each training data is a stable length of point aspect with a label. Every MIL named are multiple instances Citation-KNN is an enhanced with KNN algorithm is acceptable for MIL access. KNN is a type

of lazy learning technique is delay rectification of training data through a query required to be a recapitulation. In the case of MIL, the Euclidean distance can't be used and Hausdorff distance also can't be used directly. Hence, in this work, to resolve this problem, an enhanced Citation-KNN algorithm named is a locally weighted Citation-KNN (LWCKNN) is suggested.

Microarray technology is being broadly used in the learning of gene explanation in cancer diagnosis. Microarray technology is realizing a design of normal and abnormal cells because it can calculate a thousand of genes. Parametric statistical methods sometimes will not be useful for microarray data as any of the assumptions will not work for this kind of data and as the genes are in thousands, type I error will be increased [12], genetic algorithm is combined with k-nearest neighbor (GA/KNN) to consider the data and they obtained the acceptable results. This method name as Adaptive Genetic Algorithm/k-nearest neighbor (AGA/KNN). This paper suggests that the naturalist considers to classify the suitable genes simply from the sub-gene set and analyze the test samples exactly.

In online unstructured textual information is increasing, especially in the biomedical domain, such as biomedical system implements, medical announces, patient dismissal summaries. To facilitate access to relevant information, documents are annotated with thesaurus or ontologies. In [13], the authors used KNN to retrieve relevant documents. They MeSH descriptors to decide the topics and the vector space model (VSM) [14] is used to find k most similar documents in the target document. one idea of using machine learning algorithm for ranking labels documents.

OTHER APPLICATION

KNN classification algorithm is used in several varieties of application areas other than medical research. In this section, we review some of the publications which cover applications in Networks, social networking, mobile computing, human behavioral models and the like

In [15] , a hybrid approach using fuzzy distance measure along with KNN classifier is proposed. They used the fuzzy distance between training and test data sets. The efficiency is access to established by contrast their performances of the classic and the society placed heuristic approach to the important real-world classification is issues to apply from the UCI machine-learning benchmark archive. The exploratory is a consequence of show the suggest that hybrid algorithms significantly explore more optimal weight vectors specially and allocate more exact of the classification result than the effective of well-known occasion locate on the intuitive and heuristic classification algorithms and classic access to real data sets.

In [16], the authors used k-nearest neighbor (KNN) algorithm along with a different similarity function for distance calculate between a test point and a training point. This different similarity function access to locate on limited research. A different similarity function is used for the classifying the test patterns. To improve this KNN algorithm is used in fifteen numerical datasets from the UCI machine learning data store. In every one of the 5-fold and 10-fold cross-validations are utilized. They compared the classifying to the efficiency of this approach with other well-known clustering algorithms and there was an improved accuracy.

Network attacks are expected to extend into the recent past and intrusion detection and more than fetching an uncertain section to the protected secure information systems. The difficulty in implementing the supervised network intrusion detection is acquired to sufficient the supervised network attacks data to classifying the type of the attack patterns. The obtaining data is a time-absorbing effort and depend on the field. In [17], a supervised network intrusion detection is approached to suggest the located on on Transductive Confidence Machines for K-Nearest Neighbors (TCM-KNN). This is a machine learning algorithm. They have conducted experiments on the KDD Cup 1999 data set. They approached to suggest that the most robust to success with the intrusion detection method. The demand and expand with the recognition of smartphones, SoLoMo (Social-Location-Mobile) method are likely to be aggressive and develop into a well-known mobile social networking stage. These SoLoMo systems continuously return k-nearest Neighbor based on their geo-locations for friendship. This recommended approach is easy, but abort to make continuous friendships. In Social networks, likes Facebook and Twitter can be utilized to support better friend recommendations. In [18], an another metric, co-space is an expense to proposed by considering to everyone user distances in the natural world and the virtual world (social distance). Distances between users in the

conventional social networks are calculated by using a hardly any of MapReduce jobs. The efficient to computing the social expense is acquired by the construction of a distributed index on top of the key-value store, and the distances located on the geo-locations utilized an R-tree. Then they used kNN based on co-space distances. They conducted experiments with Gowalla dataset and the results were effective and efficient.

A different application of KNN is shown in [19]. The fault diagnosis is a rolling element bearing can be utilized for the compound of weighted k nearest neighbor _WKNN_classifiers. This process utilizes a wavelet packet transform located on the lifting system to the preprocess of vibration signals before aspect the extraction. Then aspect to all extractions of the Time- and frequency-domain to perform the operation conditions of the manner. After extracting the sensitive aspect, multiple classifiers located on WKNN are combined to enhance the classification accuracy.

Automatic detection of objects in domains such as bioinformatics ,traffic supervision access control, and authentication system involves challenging tasks. The main problem may occur in identifying the objects in extreme diversity where the largest variety of appearance form dimension and color, positioning are available. In [20], the authors proposed a model which has three main phases such as detection of interest points, local descriptor and object model. In this paper, they used Gabor Jet Local Descriptors. The proposed method also identifies interior object parts with more classification methods.

In [21] the authors propose a simple but effective clustering algorithm called CLUB. This algorithm uses mutual k nearest neighbor to identify initial clusters. And the k nearest neighbors for identifying density backbones of these initial clusters. Then, each unlabeled point is assigned to the cluster, which is its nearest high-density-neighbour. CLUB was compared with three classical and three state-of-the-art methods. They tried this algorithm on two-dimensional and multi-dimensional datasets. They also demonstrated the algorithm's performance for face recognition using Olive it Face dataset.

Reverse Nearest Neighbor(RNN) Queries are a kind of Spatial queries, RNN queries are solved through region approach in which every single object in the space has a certain region associated with it and all the objects belonging to this region to identify the query objects to their nearest neighbor. This method works efficiently in a mobile environment, but for only a single object. There may be queried for several objects. In [22], the authors propose Group ReversekNN methodology using the concepts of computational Geometry. Experiments conducted by them proved the performance, efficiency, and accuracy of the proposed algorithms.

The manufacturing industry is a determinant of the key factor is Human performance. If the act of design is a valuable and complicated problem for reporting a human activity. It is useful for managers to predict the employee's capabilities and to the recruitment of new staff with required skills. The classification method is a very effective of the K-nearest neighbor (KNN) algorithm, there is no initial learning of data dispersion. In [23], enhance the KNN algorithm was for this problem. This method utilizes the neighboring distance calculation is located on entropy, the classification is resolving strategy and the quantitative illustration method of human performance. The performance is compared with five classic classification algorithms utilized the data sets from the University of California, Irvine(UCI) machine learning repository and the effectiveness of the algorithm is ascertained.

The text categorization problem was innovated through analysis the Multi-label learning. Many of the online resources have documents belonging to several domains. In training sets that are related to the multi-label learning is a set label. We need to estimate the label sets of the undetectable occasion by analyzing training instances with well-known label sets. In [11] a multi-label lazy learning access to the named ML-KNN is given by an utmost posterior (MAP) is essential to utilize the label set for the undetectable instance. Three different real-world issuing namely Yeast gene functional analysis, natural scene classification and automatic web page classification were utilized to test the algorithm's performance. The results show that ML-KNN attains superior execution its approach to other multi-label learning algorithms.

In [24], KNN is used along with three different methodologies for feature extraction namely Autocorrelation, Preprocessing and Framing for speech recognition. The authors claim that this method provides 80 % accuracy level.

CONCLUSION

In this paper, we tried to present many kinds of works based on K-Nearest Neighbor classification and its variants. Apart from the medical field, we showed how the KNN algorithm is effectively used in information and communication technologies, social networking, Face Recognition, query optimization, bug report assignment.

REFERENCES

[1] Text Book: Bill Franks, "Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics", John Wiley & sons, 2012

[2] Nitin Bhatia and Vandana (2010): Survey of Nearest Neighbor Techniques, International Journal of Computer Science and Information Security, Vol. 8, No. 2, 302-305

[3] Abdulkadir Sengur, & Ibrahim Turkoglu (2008). A hybrid method based on artificial immune system and fuzzy k-NN algorithm for diagnosis of heart valve diseases. Expert Systems with Applications,35, pp 1011–1020.

[4] M.Akhil jabbar, B.L Deekshatulu & Priti Chandra. Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. In International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) , Procedia Technology 10 (2013), pp 85 – 94

[5] S.N Sivanandam,S.N Deepa,"Introduction to genetic algorithms", *Springer(2008)*

[6] P . Bhuvaneswari , &Dr. A. Brintha Therese. Detection of Cancer in Lung With K-NN Classification. International Conference on Nanomaterials and Technologies (CNT 2014), Procedia Materials Science 10 (2015), pp 433 – 44.

[7] Dr. Malini suvarna & Mr.Venkategowda N(2015). Performance Measure andEfficiency of Chemical Skin Burn Classification Using KNN Method. In International Conference on Eco-friendly Computing and Communication Systems, ICECCS, Procedia Computer Science 70 (2015) pp 48 – 54

[8] RomanStoklasa , Tomas Majtner, & DavidSvoboda(2014).Efficient k-NN basedHEp-2cellsclassifier. Pattern Recognition,47 , pp 2409–2418.

[9] Jianrui Ding, H.D. Cheng, Min Xian, Yingtao Zhang, & Fei Xu(2015).Local-weighted Citation-kNN algorithm for breast ultrasound imageclassification. Optik 126, pp 5188–5193

[10] J. Wang, J.-D. Zucker, Solving the multiple-instance problem: a lazy learning approach, in: Proc. 17th Int'l Conf. Machine Learning, 2000, pp. 1119–1125. Khadim Drame, Fleur Mougin and Gayo Diallo(2016).

[11] Chien-Pang Lee , Wen-Shin Lin , Yuh-Min Chen and Bo-Jein Kuo(2011),Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method, Expert Systems with Applications 38,(2011),pp 4661–4667

[12] Large scale biomedical texts classification: a kNN and an ESA-based approaches. Journal of Biomedical Semantics (2016).

[13] Salton G, Wong A, Yang CS. A vector space model for automatic indexing, Commun ACM. 1975;18(no 11):613–20.

[14] Hamdi Tolga Kahraman(2016). A novel and powerful hybrid classifier method: Development and testing of heuristic k-nn algorithm with fuzzy distance metric. Data & Knowledge Engineering, 103 ,pp 44–59

[15] Gautam Bhattacharya , Koushik Ghosh , Ananda & S. Chowdhury(2012). An affinity-based new local distance function and similarity measure for kNN algorithm. Pattern Recognition Letters ,33, pp 356–363.

[16] Yang Li, & Li Guo(2006). An active learning based TCM-KNN algorithm forsupervised network intrusion detection. Computers &security 26 , pp 459 –467

[17] 11. Xianke Zhou, Sai Wu, Gang Chen, & Lidan Shou (2014). kNN processing with co-space distance in SOLOMO systems. Expert Systems with Applications 41, pp 6967–6982

[18] Szidonia Lefkovits & Laszlo Lefkovits(2015).Distance based k-NN Classification of Gabor Jet Local Descriptors. In International Conference Interdisciplinarity in Engineering INTER-Engg , Procedia Technology 19, pp 780-785.

[19] Mei Chen , Longjie Li, Bo Wang , Jianjun Cheng , Lina Pan , & Xiaoyun Chen(2016). Effectively clustering by finding density backbone based-on kNN. Pattern Recognition 60, pp 486–498

[20] Anasthasia Agnes Haryantoa , David Taniara , & Kiki Maulana Adhinugraha(2015). Group Reverse kNN Query optimisation. Journal of Computational Science, 11, pp 205–221

[21] Ni Li, Haipeng Kong , Yaofei Ma, Guanghong Gong1 & Wenqing Huai(2016). Human performance modeling for manufacturing based on an improved KNN algorithm. International Journal Advanced Manufacture Technology (2016) 84, pp 473–483

[22] Min-Ling Zhang, & Zhi-Hua Zhou. ML-KNN:Alazy learning approach to multi-label learning. Pattern Recognition, 40 (2007), pp 2038 – 2048.

[23] T.Lakshmi Priya, N.R.Raajan ,P.Preethi & S.Mathini(2012).Speech and Non-Speech identification and classification using KNN algorithm.Procedia Engineering, 38,pp 952-958.