Survey of Nearest neighbor classifiers technique in data mining

Hiba Mohamed Suliemansalieam
Dept of Computer Science
Elneelaien university

Abstract:

Classification task in data mining is distinguish between objects of different classes and predictive modeling to predict the class label of unknown record. In this paper, we present the k-nearest neighbor classifier technique(KNN) which is the simplest Machine Learning algorithms method of classification task including basic concept of KNN method, KNN advantages ,Limitation of KNN and application of KNN . We show Several major studies and researches that had been done to enhance the KNN method. The goal of this survey is to provide a comprehensive review of k-nearest neighbor classifier technique(KNN) in Data mining.
keywords— classification technique, analogy, k-nearest neighbor classifier.

المستخلص:

مهمة التصنيف في التنقيب عن البيانات هي التمييز بين كائنات من فئات مختلفة والنمذجة التنبؤية للتنبؤ بتسمية فئة السجل (الكائن)غير المعروف. هذا البحث يقدم دراسه عن تقنية مصنف الجار الأقرب (KNN) وهي أبسط طرق خوارزميات التعلم الآلي لمهمة التصنيف في التنقيب عن البيانات. وقد تم تناول الحديث في هذه الورقه عن المفهوم الأساسي لتقنية KNN المزايا التي تميزهذه التقنيه , تطبيقات KNN في جميع مجالات الحياه , وقد تم تناول قصورهذه التقنيه في بعض المواضع مما يقلل من فعاليتها وتم عرض العديد من الدراسات والبحوث التي تم إجراؤها للحد من هذا القصور. الهدف الرئيسي في هذه الورقه هو الاستطلاع و تقديم مراجعة شاملة لتقنية المصنف الأقرب لـ (KNN) في تنقيب البيانات.

1.Introduction:

Data mining is [10] a process of extracting and discovering patterns in large data sets . It has many Tasks divided to Prediction Methods – Use some variables to predict unknown or future values of other variables like Classification and Regression tasks. The

other method of data mining is description Methods – Find human interpretable patterns that describe the data like Clustering and Association Rule tasks. Classification task is assigning objects to one of several predefined categories. It learning a target function that maps each attribute set to one of predefined class labels . This function also know classification model it useful for both descriptive modeling that can serve as an explanatory tool to distinguish between objects of different classes and predictive modeling to predict the class label of unknown record[13] a general approach for solving Classification problem is systematic approach for building classification model from input set. Numerous classification methods involve techniques like decision tree induction, Bayesian networks and k-nearest neighbor classifier. In the present paper, we have concentrated on the techniques necessary to do this. In particular, this work is concerned with k-nearest neighbor classifier.

2. Concept of k-nearest neighbor classifier:

The k-nearest neighbors (KNN) algorithm one [12] of the simplest Machine Learning algorithms based on Supervised Learning technique. It can be used to solve both classification and regression problems. The KNN algorithm is based on learning by similarity. The training samples are described by n dimensional numeric attributes .all of the training samples are stored in an n-dimensional pattern space. to predict the target label by finding the nearest neighbor class. The closest class will be identified using the distance measures like Euclidean distance .

where the Euclidean distance between two points. X=(x1,x2,……,xn) and Y=(y1,y2,….,yn)

$$(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)2}$$

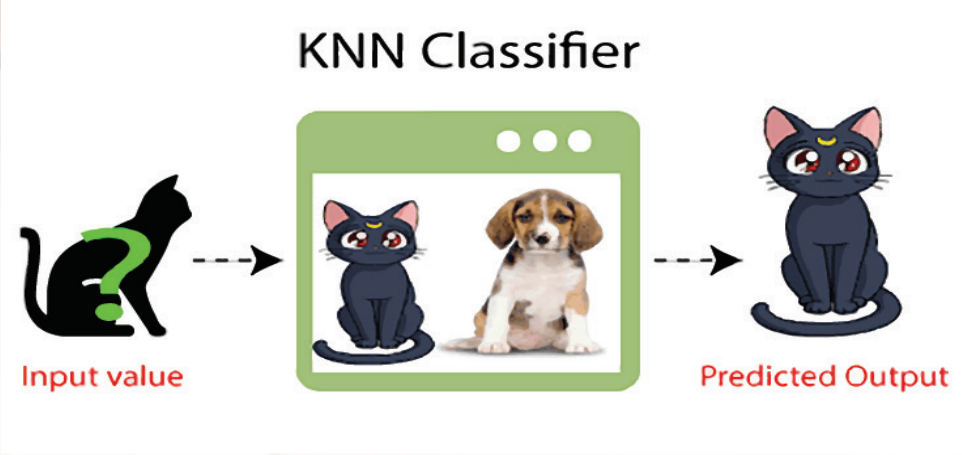

Fig.1 shaw k-nearest neighbor classifier work

3. The advantages of KNN algorithm, summarized as follows:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

This algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

Without points or numbering

4 .Application of of KNN algorithm in real life

K-nearest algorithm has various uses in day life. scientists and the beginner of machine learning use this algorithm for a simple task. Some of the uses of the k nearest neighbor algorithm are:

Finding diabetics ratio

We can use K Nearest Neighbor Algorithm to judge the ratio of diabetes , If we figure out the data of age, pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index and other required data ,we can easily plot the probability of diabetes at a certain age.

Concept Search:

Concept search is the industrial application of the K Nearest Neighbor Algorithm. It means searching for similar documents simultaneously. The data on the internet is increasing every single second .The main problem is extracting concepts from the large set of databases. K-nearest neighbor helps to find the concept from the simple approach.

Finding The Ratio of Breast Cancer In the medical sector.

The KNN algorithm is widely used. It is used to predict breast cancer. Here KNN algorithm is used as the classifier. The K nearest neighbor is the easiest algorithm to apply here. Based on the previous history of the locality, age and other conditions KNN is suitable for labeled data.

Recommendation System

All search engines use the algorithms of k-nearest neighbor. The 35% revenue of Amazon comes from the recommendation system. Decide the online store, YouTube, Netflix, and all search engines use the algorithms of k-nearest neighbor.

Without points or numbering:

5. Limitation of KNN Algorithm

It is called a lazy learner algorithm because it does not learn from the training set immediately instead, it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category dataset using the testing dataset can be computationally expensive. Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm. Accuracy depends on the quality of the data. With large

data, the prediction stage might be slow. It Require high memory – need to store all of the training data.

Without spaces:

Many researchers have studied methods for enhance and improve the e efficiency of K Nearest neighbor (KNN) algorithm for classification.

Charu Gupta,Kratika Goyal , Dharna Gureja , Enhance k nearest neighbor using dynamic selected, 2004 , their idea KNN algorithm has large memory requirements as well as high time complexity. Several techniques have been proposed to improve these shortcomings in literature. they have first proposed novel improved algorithm. It is a combination of dynamic selected, attribute weighted and distance weighted techniques. Experimental results have proved that their proposed algorithm performs better than conventional KNN algorithm.

Parvin et al ,Modified K-Nearest Neighbor,2008 , present a modified weighted kNN algorithm to enhance the performance of kNN. The algorithm preprocesses the training dataset using the testing dataset. The preprocessing first determines the validity of each data point by measuring its similarity to its k neighbors and then measures its distance weight to each data point in the testing dataset. The product of the validity and distance weight for each data point produces a weighted training dataset. This reduces a multi-dimensional dataset into one dimensional dataset, which improves the efficiency of kNN.

Karina Giberta,b, Miquel Sànchez-Marrèa,c, Víctor Codinaa (2010) Presented a study entitled (Choosing the Right Data Mining Technique) Classification of Methods and Intelligent Recommendation One of the most difficult tasks in the whole KDD process is to choose the right data mining technique, as the commercial software tools provide more and more possibilities together and the decision requires more and more expertise on the methodological point of view. Indeed, there are a lot of data mining techniques available for an environmental scientist

wishing to discover some model from her/his data. This diversity can cause some troubles to the scientist who often have not a clear idea of what are the available methods, and moreover, use to have doubts about the most suitable method to be applied to solve a concrete domain problem. Within the data mining literature there is not a common terminology. A classification of the data mining methods would greatly simplify the understanding of the whole space of available methods. Furthermore, most data mining products either do not provide intelligent assistance for addressing the data mining process or tend do so in the form of rudimentary "wizard-like" interfaces that make hard assumptions about the user's background knowledge. In this work, a classification of most common data mining methods is presented in a conceptual map which makes easier the selection process. Also an intelligent data mining assistant is presented. It is oriented to provide model/ algorithm selection support, suggesting the user the most suitable data mining techniques for a given problem.

Lingyun Wei ,Xiaoli Zhao, , An Enhanced Entropy-K-Nearest Neighbor Algorithm Based on Attribute Reduction, 2014 , presented a scalable kNN algorithm to reduce the computation time for a large training dataset by clustering the dataset to N clusters and distributing them to N machines where each machine is assigned an equal amount of data to process. The master server distributes the query for a testing data to predict its class so that each machine can perform the kNN algorithm execution in parallel and return the results to the master server for consolidation.

Selahaddin Batuhan Akben1 and Ahmet Alkan , An Improved KNN Algorithm Based on Kernel Methods and Attribute Reduction, 2015 , they proposed to reducing the effect of noisy data. In the first stage of the proposed method, they coefficient of density of each element in the training set was obtained by Parzen window method. And then, the membership of each test element was determined according to the total of density coefficients (weights) of neighbors belonging to the same class. As for the last stage, the

performance results of the frequently used KNN methods and the proposed method (Density-weighted KNN, Classical KNN and Distance-weighted KNN) were compared.

Dhrgam AL Kafaf, , B-kNN to Improve the Efficiency of kNN , 2017 , he present the B-kNN algorithm to improve the efficiency of kNN using a two-fold preprocess scheme built upon the notion of minimum and maximum points and boundary subsets. For a given training dataset ,B-kNN first identifies classes and for each class ,it further identifies the minimum and maximum points (MMP) of the class. A given testing object is evaluated to the MMP of each class. If the object belongs to the MMP, the object is predicted belonging to the class. If not, a boundary subset (BS) is defined for each class. Then, BSs are fed into kNN for determining the class of the object. As BSs are significantly smaller in size than their classes, the efficiency of kNN improves.

Yu et al , Indexing the Distance: An Efficient to KNN Processing, 2001 introduce a distance-based kNN algorithm to improve the efficiency of kNN by preprocessing the training dataset. The preprocessing involves partitioning the training dataset and identifying the centroid of each partition to be a reference point to The partition. Then ,they compute the distance of each data point in the partition to the reference point and index the distances in a B+ tree. For a testing data, the closest partition is found by computing the distance of the data to the centroids of partitions. Once the closet partition is identified , the B+ tree of the partition is used to search the nearest neighbor to the data in the partition.

Najat Ali (2019) Presented a study entitled (Evaluation of k-nearest neighbor classifier performance for heterogeneous data sets) she work on measuring the distances between the test sample and the training samples to determine the final classification output. The traditional k-NN classifier works naturally with numerical data. where data can be described as a mixture of numerical

and categorical features. For the sake of simplicity, her work considers only one type of categorical data, which is binary data. In her proposed, several similarity measures have been defined based on a combination between well-known distances for both numerical and binary data, and to investigate k-NN performances for classifying such heterogeneous data sets. The experiments used six heterogeneous datasets from different domains and two categories of measures. Experimental results showed that the proposed measures performed better for heterogeneous data than Euclidean distance, and that the challenges raised by the nature of heterogeneous data need personalised similarity measures adapted to the data characteristics.

6. Conclusion:

The k-nearest neighbors (KNN) algorithm is one of the important introductory supervised classifier algorithms. K-NN algorithm stores all the available data and classifies a new data point based on the similarity, class. The closest class will be identified using the distance measures like Euclidean distance. It is simple Machine Learning algorithms and it does not make any assumption on underlying data , these led Motivation to use it. on the other side it has many Limitation like it does not learn from the training set immediately, it stores all dataset spending time , that is why it is called a lazy algorithm which is useless in big data . Therefore, a lot of efforts are made to improve its performance .

In future work, This algorithm is one of the simplest machine learning algorithms and it has many uses in practical life as it enables us to create models with great benefits. Therefore, many studies must be conducted to improve its performance by addressing its shortcomings.

References:

(1) Dorina Kabakchieva, Predicting Student Performance by Using Data Mining Methods for Classification,2013
(2) Data aspirant ,Knn Classifier, Introduction to K-Nearest Neighbor, Algorithm, December 23, 2016.
(3) Dhrgam AL Kafaf, Dae-Kyoo Kim and Lunjin Lu, B-kNN to Improve the Efficiency of kNN, 2017.
(4) Hamid Saadatfar , A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data,2020.
(5) J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Elsevier, 2006.
(6) Mohammed J. Zaki , Data Mining and Analysis, Rensselaer Polytechnic Institute, Troy, New York,2014
(7) Muja,M. and Lowe, D.GScalable Nearest Neighbor Algorithms for high Dimensional Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014
(8) Najat Ali ,Evaluation of k-nearest neighbor classifier performance for heterogeneous data sets,2019.
(9) Pang-Ning Tan ,Introduction to Data Mining (2nd Edition) , Michigan State University, 2005
(10) Parvin, H., Alizadeh, H., and Minaei-Bidgoli, B. MKNN: Modified K-Nearest Neighbor. In Proceedings of the World Congress on Engineering and Computer Science, 2008.
(11) Thair Nu Phyu ,Survey of Classification Techniques in Data Mining, 2009.
(12) Pang-Ning Tan, Michael Steinbach, Anuj Karpatne ,Vipin Kumar, Introduction to Data Mining (Second Edition),2020
(13) Witten, I. H. and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 2nd edition 2005.
(14) Witten, I. & Frank, E. (, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005
(15) Wilson, D. R. & Martinez, T. Reduction Techniques for

Instance-Based Learning Algorithms. Machine Learning 2000.
(16) Yu, C., Ooi, B. C., Tan, K.-L., and Jagadish, H. Indexing the Distance: An Efficient Method to Knn Processing 2001.
(17) Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan, Top 10 algorithms in data mining, 8 October 2007.
(18) Zhou, Y., Li, Y., and Xia, S. An Improved KNN Text Classification Algorithm Based on Clustering. Journal of Computers, 2009.