JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

Website: http://www.jmrr.org Vol.2, Iss.3 Pages:11-17

A SURVEY ON K NEAREST NEIGHBORS ALGORITHMS IN DATA MINING

[1]S Devi, [2]A Emima
[1]BSc Computer Science, [2]Assistant Professor
Holy Cross College, Trichy, Tamilnadu

Article Received: April 2021 Published: October 2021

Abstract

The k-nearest neighbor Algorithms is a non-parametric technique utilized for classification and regression. The k-nearest neighbors (KNN) Classification technique has been prepared to be utilized on-line and in Real Time to recognize customers / guests click stream information, coordinating it to a specific client gathering and suggest a customized perusing alternative that addresses the issue of the particular client at a specific time. The regression algorithm to diminish the size of the preparation set of KNN regression. Right now, initially evacuate the exception occurrences that's way the presentation of regression, and afterwards or the left cases and their nearest neighbors.

Keywords: KNN classification, structure less NN techniques, structured based NN technique, KNN regression, and regression algorithm

INTRODUCTION

Data mining is the procedure of naturally removing educated data from colossal measures of information. It has gotten progressively significant as genuine information colossally expanding [1]. The enormous measure of information that is put away in databases contains important concealed information which causes the client to improve the exhibition of dynamic procedure [2]. Data Mining or Knowledge revelation is expected to utilize information [11]. These Nearest neighbor classifiers depend on learning by similarity. The preparation tests are portrayed by n dimensional numeric properties. Each example speaks to a point in an-dimensional space. A case is arranged by a larger part vote of its neighbors, with the case being appointed to the class generally regular among its K neighbors estimated by a separation work. In the event that K=1, at that point the case is essentially allotted to the class of its closest neighbor [3][4]. A straightforward execution of KNN regression is to compute the normal of the numerical objective of k closest neighbors. Another methodology utilizes are verse separation weighted normal of the k closest neighbors. KNN relapse utilizes a similar separation works as KNN characterization [5]. At last it is examined about how to characterize the classification and regression strategies.

In this paper we present an experimental examination with sample data set on k-nearest calculation for high exactness grouping with the underlying centroids determination in a further developed way. The exactness of unique K-means calculation vigorously relies upon centroids at the starting and it has high computational intricacy.

LITERATURE SURVEY

M. Akhil and et.al., to describe the data mining is the process of automatically extracting knowledgeable information from huge amounts of data [1]. Biak's and et.al., to define the nearest neighbor classifiers are based on learned by analogy [2]. Tolvi used a genetic algorithm that is an evolutionary based to detect the outlier in linear regression models. In this method, the corrected BIC criterion is selected as the fitness function. Antonelli et al. also proposed an instance selection algorithm in the framework of a multi objective evolutionary learning of fuzzy rule-based systems [3]. Yun sheng song and et.al., to describe the Nearest Neighbors is a sort of instance based learning, or lazy learning, Where the capacity is just approximated locally and all calculation is conceded until classification [4]. Dr. Rajesh Verma and et.al., to describes the data Mining or Knowledge discovery is needed to make sense and use of data [5]. A.Arnaiz-Gonzalez and et.al., to describes Instance selection algorithms for regression are divided into two categories: evolutionary-based and Nearest neighbours [6]. Dr. Saed Sayed, to called the different name of KNN algorithm [7]. Mrutynjaya Panda and et.al., to describes the large amount of data that are stored in databases contains valuable hidden knowledge [8]. Hong Kong to call the all of the training samples is stored in an n-dimensional pattern space. When given an unknown sample, a k nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample [9]. AT&T Bell Laboratories to describe the KNN regression uses the same distance functions as KNN classification [10].

K NEAREST NEIGHBORS CLASSIFIER

K nearest neighbor (KNN) is a simple algorithm, which stores all cases and classifies new cases based on similarity measure. KNN algorithm also called as, Case based reasoning, K nearest neighbor, Example based reasoning, Memory based reasoning, Lazy learning, Instance based learning[6]

KNN is a non parametric classification method which is broadly classified into two types: Structure less NN techniques and Structure based NN techniques.

In structure less NN techniques whole data is classified into training and test sample data. From training point to sample point distance is evaluated, and the point with lowest distance is called nearest neighbor. Structure based NN techniques are based on structures of data like orthogonal structure tree (OST), ball tree, k-d tree, axis tree, nearest future line and central line [7]. Nearest neighbor classification is used mainly when all the attributes are continues. Simple K nearest neighbor algorithm is shown below

Step 1: find the K training instances which are closest to unknown instance

Step 2: pick the most commonly occurring classification for these k instances

There are various ways of measuring the similarity between two instances with n attribute values. Every measure has the three requirements. Let dist(A,B) be the distance between two points A, B then

1. $dist(A,B) \geq 0$ and $dist(A,B)=0$ if $A=B$

2. $dist(A,B)=dist(B,A)$

3. $dist(A,C) \leq dist(A,B)+dist(B,C)$

Property3 is called as “Triangle inequality”. It states that the shortest distance between any two points is a straight line. Most common distance measures used is Euclidean distance. For continuous variables Z score standardization and min max normalization is used. KNN is used in many applications such as: Classification and interpretation, Problem solving, Function learning and teaching and training.

Drawbacks of KNN are Low efficiency, Dependency on the selection of good values fork. Further research is required to improve the accuracy of KNN with good values of K[1].

Distance Metric

The entirety of the preparation tests are put away in a n-dimensional example space. At the point when given an obscure example, ak closest neighbor classifiers can s the example space for the k preparing tests that are nearest to the obscure example. "Closeness" is characterized as far as Euclidean separation, where the Euclidean separation, where the Euclidean separation between two focuses. The entirety of the preparation tests are put away in an-dimensional

example space. At the point when given an obscure example, a k closest neighbor classifiers cans the example space for the k preparing tests that are nearest to the obscure example. "Closeness" is characterized as far as Euclidean separation, where the Euclidean separation, where the Euclidean separation between two focuses,

$X= (x_1,x_2,\ldots\ldots,x_n)$ and $Y=(y1,y2,\ldots.,yn)$ is

$D(X, Y)= \sqrt{\sum(x_i-y_i)^2}$

X and Y are the two compared objects and n is their number of attributes. However, for distance calculations involving symbolic(nominal) data, other methods are necessary; for symbolic data a similarity metric is used to measure distance, the most basic of which is the overlap metric. The overlap metric simply tests for equality between two values, so that different values get distance 1 whereas equal values get distance 0 [3].

$\mathbf{d}_{overlap}(x , y) = 0$ when $x=y$ *And*
$\mathbf{d}_{overlap}(x, y)=1$, when $x\neq y$

K NEAREST NEIGHBORS REGRESSION

KNN regression is basic calculation and now and then called a lethargic calculation in view of its moderate learning. In regression, it takes k neighbors and returns their normal to new esteem. To ascertain the normal, Euclidean separation is for the most part utilized.

Instance Selection algorithms for regression

Instance selection algorithms for regression are divided into two categories: Evolutionary-based and Nearest neighbor-based[8][12].

A) Evolutionary based

This concept utilized a hereditary calculation that is a developmental based to distinguish the anomaly in direct relapse models. Right now, amended BIC basis is chosen as the wellness work. Every individual is completely depicted by a double vector (z1,z2,⋯,zn), where zi=0 shows the case xi isn't chosen as an anomaly else it is chosen, and i=1,2,3,⋯,n. The consequences of this analysis on little datasets have demonstrated that it was ready to identify the exception, yet additionally kept away from the potential issue that one anomaly keeps another from being identified [9]. This is additionally proposed a case choice calculation in the structure of a multi objective developmental learning of fluffy guideline based frameworks. This calculation ran for enormous scope datasets and showed signs of improvement execution. Inspite of the fact that the calculations based developmental have better information decrease rate and higher forecast precision, their computational expenses are around 3 to 4 sets of extent higher than the ones dependent on the neighbor for medium size informational collections.

Besides, the distinction in computational expense increases as the size of informational collections is developing [10]. So these techniques are difficult to apply to the issues in their ability.

B) Nearest neighbors based

Nearest Neighbors is a sort of instance based learning, or lazy learning, where the capacity is just approximated locally and all calculation is conceded until classification. Both for classification and regression, a helpful strategy can be to dole out loads to the commitments of the neighbors, so that the closer neighbors contributes more to the normal than the more far off ones. For instance, a typically weighting plan comprises in giving each neighbor a load of 1/d, where d is the separation to the neighbor. The neighbors are taken from a lot of items for which the class (for KNN order) or the article property estimation (for KNN replace) is known. This can be thought of as the preparation set for the calculation, however no unequivocal preparing step is required. An idiosyncrasy of the KNN calculation is that it is touchy to the neighborhood structure of the information [13].

Example of K-Nearest Neighbors Algorithm

We have data from the questionnaires and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples,

Table 1: Classification Table

X1=Acid Durability (seconds)	X2=Strength	Y=classification
7	7	Bad
7	5	Bad
3	5	Good
1	5	Good

Now the factory produces a news paper tissue that pass laboratory test with x1=3and x2=7, without another expensive survey, can we guess what the classification of this new tissue is?

1. Determine parameter K=number of nearest neighbors Suppose use K=3

2. Calculate the distance between the query-instance and all the training samples Coordinate of query instances is (3,7), instead of calculating the distance we compute square distance which is faster to Calculate(without square root)

Table 2: Square Distance to Query Instance

X1=Acid Durability (seconds)	X2=strength	Square distance to query instance(3,7)
7	7	$(7-3)^2+(7-7)^2=16$
7	5	$(7-3)^2+(5-7)^2=20$
3	5	$(3-3)^2+(5-7)^2=4$
1	5	$(1-3)^2+(5-7)^2=8$

3. Sort the distance and determine nearest neighbors based on the k-th minimum distance

Table 3: Rank Minimum Distance

X1=Acid Durability (seconds)	X2=strength	Square distance to query instance (3,7)	Rank minimum Distance	Is it included in 3-nearest neighbor
7	7	$(7-3)^2+(7-7)^2=16$	3	Yes
7	4	$(7-3)^2+(4-7)^2=20$	4	No
3	4	$(3-3)^2+(4-7)^2=4$	1	Yes
1	4	$(1-3)^2+(4-7)^2=8$	2	Yes

4. Gather the category Y of the nearest neighbors, notice in these cond row last column that the category of nearest neighbor(Y) is not include because the rank of this data is more than 3(=k)

Table 4: Category of Nearest Neighbor

X1=Acid Durability (seconds)	X2=strength	Square distance to query instance (3,7)	Rank minimum Distance	Is it included in 3-nearest neighbor	Y=category of nearest neighbor
7	7	$(7-3)^2+(7-7)^2=16$	3	Yes	Bad
7	4	$(7-3)^2+(4-7)^2=20$	4	No	Bad
3	4	$(3-3)^2+(4-7)^2=4$	1	Yes	Good
1	4	$(1-3)^2+(4-7)^2=8$	2	Yes	Good

5. Use simple majority of the category of nearest neighbors as the prediction value of the query instance. We have 2 good and 1 bad, since 2>1 then we conclude that a new paper tissue that pass laboratory test with x1=3and x2=7 is included in Good category [12].

RESULT AND OBSERVATION

Result: We have data from the questionnaires and objective testing with two attributes (acid

durability and strength) to classify whether a special paper tissue is good or not. And we can find this problem using K-Nearest Neighbors Algorithms Classification method. Because of this method is easy to compare the Regression method. And the final result is tissue paper is good.

Observation: KNN is a supervised gaining knowledge of set of rules used for each regression and classification. In operation may be in comparison to the subsequent analogy. To make a prediction, the KNN set of rules would not calculate a predictive version from a education dataset like in logistic or linear regression.

CONCLUSION

In this paper we have presented a KNN classifier and regression method. Classification methods are typically strong in modeling interactions. Most of the instance selection algorithms are mainly concerned with KNN classification, and less focused on KNN regression. Data mining offers promising ways to uncover hidden patterns can potentially be used future behavior. In this paper, we present the basic KNN classifiers techniques, KNN regression and the example of the K-Nearest Neighbors Algorithms.

REFERENCES

1. M. Akhil Jabbar B.L. Deekshatulu, Priti Chandra, "Data mining is the process of automatically extracting knowledgeable information from huge amounts of data", Vol. 340, pp.85-94, December2013.
2. Mrutynjaya Panda, Ajith Abraham, "The large amount of data that are stored in databases contains valuable hidden knowledge", August.10.2014.
3. Baik,S, Bla,J.,"Nearest neighbor classifiers are based on learning by analogy", Vol.3046, pp. 206-212, 2014.
4. S.Neelamegam, Dr.E.Ramraj, "The training samples are described by n dimensional numeric attributes", Vol.3, issue.5 , 2013.
5. AT&T Bell Laboratories, Holmdel, NJ, "KNN regression uses the same distance functions as KNN classification", September11, 1996.
6. Dr.Saed Sayed, "Different name of KNN algorithm", http://chem-eng.
7. NitiBhatia,Vandhana, "survey on nearest neighbors techniques" IJCSIS, Volume 80, 2010.
8. A. Arnaiz-Gonzalez, M. Kordos, "Instance selection algorithm for regression", C. Garcia-Osorio, (2016).
9. J.Tolvi, Generic algorithms for outlier detection and variable selection methods in regression tasks", Vol.8, PP.527-533, 2004.
10. M.Antonelli, P.Ducange, F.Marcelloni, "the difference in computational cost becomes larger as the scale of data sets is growing", pp.276-290, 2012.
11. Dr.RajeshVerma, RajKumar", Data Mining or Knowledge discovery is needed to make sense and use of data", Vol. 1, Issue.2, pp.2319-1058, August 2012.
12. http://youtu.be/IMxWxyhKEf1.
13. Yunsheng Song, Jiye Liang Jing Lu, Xingwang Zhao, "Evolutionary based and nearest neighbors based", pp1-8, April 12 2017.

Cite this article:

S Devi, A Emima, "A Survey on K Nearest Neighbors Algorithms in Data Mining", Journal of Multidimensional Research and Review (JMRR), Vol.2, Iss.3, pp.11-17, 2021