Nearest Neighbor Search Algorithm

Esraa Hadi Obead Alwan
College of science for Gail / Babylon university
Isr.phd@gmail.com

Abstract

A fundamental activity common to image processing, pattern recognition, and clustering algorithm involves searching set of n , k-dimensional data for one which is
nearest to a given target data with respect to distance function .
Our goal is to find search algorithms with are full - search equivalent -which is resulting match as a good as we could obtain if we were to search the set exhausting.

1- Aim of the work .

We propose a framework made up of three components, namely
l. A technique for obtaining a good initial match.
2. An inexpensive method for determining whether the current match is a full- search equivalent match.
3. An effective technique for improving the current match.
Our approach is to consider a good solution for component in order to find an algorithm, which balances the overall complexity of the search.

Key word: Clustering algorithm , search algorithm, K-neavest neighbor classification algorithm .

الخلاصة

الفعاليات الاساسية لمعالجة الصور، تميز الانماط وخوارزميات المجاميع تتضمن عملية البحث لمجموعة من A و لـــ K من الابعاد للبحث عن نقطة هي الاخرى في مجاميع البيانات من حيث دالة البعد.

الهدف من البحث هو ايجاد خوارزمية بحث ذات نتائج جيدة مثل التي نحصل عليها في حالة البحث الشامل ويهدف البحث الى:-

1- تقنية جديدة لايجاد التطابق البدائي.

2- طريقة مكثفة لتحديد هل ان التطابق الحالي هو تطابق كلي مماثل .

3- اقتراح تقنية لتحسين البحث الحالي.

الكلمات المفتاحية: خوارزمية التصنيف، خوارزمية البحث، الخوارزميات العنقوية لاقرب k جار.

2- Introduction

A fundamental activity in many algorithms involves searching a set of high dimensional data items for one, which is nearest to a given item. This is often referred to as Nearest Neighbor (NN) search algorithm problem. For example we can find the NN search problem in algorithms for K-nearest neighbor classification [Coer *et al*., 1967;Blue *et al*., 1994] , vector quantisation [Robert , 1984 ;Allen *et al*., 1992], and K-means clustering [Yosehp *et al*., 1980;Xiafeng *et al*., 1995;Daernyo *et al*., 1997] where the bulk of the computational cost in each iteration is in mapping each element of the training set to its nearest cluster center . Algorithms which require NN search are in a variety of problem domains, such as image compression [Robert ,1984;Nasser *et al*., 1988;Akront, 1994] , conceptual clustering [Ketterlin *et al*., 1995] ,machine learning [Scott *et al*., 1993], pattern matching .

In NN search, we need to search a set of prototype vectors for one which most closely to vector to be searched .The set of prototype vectors is often called the codebook, and each vector in the codebook is codeword.

A straightforward approach to NN search is to search the codebook exhaustively by comparing the target vector with each codeword using some distinct distortion measure and choosing the one, which is nearest, the target vector.

The excessive cost of this full search approach, however, makes it ,impractical to use the large codebooks when the application requires high-dimensional vectors, especially if the codebooks are to be searched several times. For example, the cost exhaustively searching vectors of dimensions in a codebook with n codewords is of order kn. for each target vector.
A number of techniques trade accuracy for spread by approximating the NN match, and terminating the search as soon as an acceptable approximation has been fond .
Among the techniques considered are those which use K-d tree, triangle elimination criteria, predictive search, partial distance function evaluations.
Different search strategies are compared in terms of the expected number of full vector comparisons which to be done to find the NN match. That is if the dimension of vectors is K, the complexities are expressed in terms of the equivalent number of order K operations. Doing so allows us to compare how the search methods perform over vectors of large dimensions.
A number of good strategies can be used to find initial approximation, which is close to the NN match. K-d tree which are arbitrarylsplitting planes are very effeetive in terms of the nearest of the approximate decision stage requires O (K) Opel" 21110113- This results in an amount of work, which is required O (log n) distortion measure evaluations.

Another good technique exploits the spatial correlation of image blocks, often called the image coherence properties to predict the NN match.

Having found an initial approximation, we have to determine whether or no that approximation is already the NN match. This means that we need a set of critical with which we could eliminate from further consideration those codewords, which are no better than the current approximation. If there are no other candidate codewords left after applying the elimination criteria, then we know that We have found the NN match and the search is finished, otherwise we, select from among the remaining codewords another approximation which is most likely to be better than the current approximation.

For a set of elimination criteria to be effective, it must be able to eliminate a large number of codewords at a time if the distortion is a metric , a set of criteria based on the triangle inequality can be very effective.

If there are other candidate codewords left to be considered, we continue with search by selecting from among the remaining codewords another approximation to the NN match. We then apply the elimination criteria again, and repeat the process until no other code word is better than the current approximation.

Note, however, that the elimination criteria do not guarantee that all the remaining codewords are betterthan the current approximation .lt is necessary to have partial ordering of the codebook to determine which of the remaining codewords should be tried next. This partial coding must be able to give the remaining codewords, which is most likely to be a better approximation.

3-Background

An efficient searching algorithm over a set of data involves at least four Steps. First, the algorithm pre-processes the data set to be obtaining either away of listing the data items in a certain order, or a structure, which can be used to search efficiently for a data item.
In binary search, for example, either the list of numbers is sorted in ascending order, or the numbers are organized in a balanced binary tree.

Second the algorithm needs away of starting the search. Given a target item binary search begins by taking the number in the middle of the sorted list. If the numbers are organized in a

binary tree, the search takes the value at the root of the tree. We can regard this value as the initial match to the number we are searching for.

Third, the algorithm needs away of eliminating sum of data from further consideration without having to go through each item. Binary search does this by comparing the given target item to the initial match .lf the target item is less than the initial match, then all the number greater than the initial match are eliminated from the list. Similarly, if the target item is less than the number at the root binary tree, the search conditions to the left subtree, ignoring all numbers in the right subtree.

Fourth, after obtaining reduced set of numbers to be searched, the algorithm proceeds to find the next candidate match.

At each stage, the algorithm remembers the best match so far, and checks whether or not it has finished. Clearly, if no data item is left in the reduced set , then current match must be the best match to the given target item . ln binary search, the search terminates when a leaf node is reached. In some cases, we might want to terminate search earlier by calculating the difference between the target item and the current match, and terminating the search as soon as the difference goes below a certain threshold.

We also note that the procedure for choosing the next match needs not to be the same as the way we choose the initial match. For example, if the set of target item to be matched is correlated sequence of numbers, then we can use the matching values of preceding target items to predict the matching for the current target item, and take that predicted value as the initial match. ln that case, the search does not have to start at the value in the middle of the sorted list of data item, or at the root of the binary tree.

The steps we just described can be regarded as components of a general framework. These components:

1- Starting procedure, which finds an initial guess to the best match.

2- An eliminating procedure to reduce the size of the candidate set.

3- A procedure for selecting the next match from the reduced set of candidate codewords.

Several works are tried to add some improvement to this general framework.

The approximation and elimination search algorithm (AESA) proposed in [Maria *et al* ., 1994]recognize the importance. of having a good match at each stage of elimination process, but the initial guess is still _chosen arbitrarily. Poggi [Poggi , 1993], on the other hand m uses the NN match of the preceding target vector for the initial match, and a set of sorted lists of means to determines the next match after the elimination criteria have been applied. However, Poggi does not consider using other strategies for each sub-problem.

4-Projection Technique

Projection function can be used to search the codebook in manner, which can be regarded as generalization of binary search. The search can start by choosing the initial match to be codeword whose projection value is closest to the target vector.

The geometric properties of the projection function can be applied on the projection values to determine which vectors can be ignored for being further away than the current match from the target vector.

4-I Means Constrained Search (MCS)

A straightforward way of projection the vectors to scalar values is to map each vector to the mean value of its components. This projection is usually applied in image processing application where the mean represents the average brightness of an image block [Chay *et al*., 1996].

If the K is the dimension of vectors and .l\/lx, My are the means of the component of the current target vector X and a codeword Y, respectively then lf h is the mean squared error (MSE) between the target vector ,X , and the initial match ,Ycurr , the any codeword Y ,which satisfies

(Mx-My)^2 > h

can be closer than Ycurr to X .We can thus limit the search to those code whose means are within the interval (Mx - $\sqrt{h}, Mx + \sqrt{h}$) [12].

5- Proposed work (FrameWork)

We believe that efiicient full-search equivalent search algorithm can be achieved by decomposing the problem into smaller sub-problems, and Ending a good solution to each sub-problem in such away that the solutions complement each other towards an overall reduction in the computational and storage complexity of the search.

We suggest a general framework for NN search, which emphasizes the three
Components described so for.

The algorithm in figure (1) illustrates an implementation of proposed framework. The components are represented in the figure by the functions
PickGuess() , ReduceSet() , and PickNeXt(). At each stage, the algorithm keeps a record of the best match so far and its distance to the target vector.

Although some researches propose using a separate techniques on the reduced set after the elimination criteria has been applied, we choose to use the same elimination procedure, reduce set, at each stage of the algorithm.

Any search technique, which can be applied the reduced set must also be applicable to the original set of candidate codewords, and visa versa.

The performance of the algorithm will not be much worse than as exhaustive search of the reduced set. In order to avoid the adverse effect of the pathological here all codewords are equivalent to the target vector, we require the elimination criteria to discard codewords which can be determined to be of the same distance to the target vector as the current match. Clearly, a faster rate of reduction in the size of the remaining set at each stage of the algorithm means that fewer iteration are needed to get the sets size down to zero,

The terminating condition " Until $|S|$ ==O)" in the algorithm assumes that we want to fall-search equivalent NN match- that is, the algorithm should result in a matching codeword which is of the same distance to the target vector as the best possible match obtained by exhaustively searching the codebook. As such, we require that the search terminate only after all codewords have been ruled out as better than the current match has.

Let X be the target vector, A the codebook, Ycurr the best match so far, and hcurr = D (X, Y curr), the following functions correspond to the three components of our approach to the nearest neighbor search problem:

PickGuss (S) // return a codeword from the set S

ReduceSet (X, Y, H, S) // applies the elimination criteria and returns the resulting // reduced set of codewords

PickNext (X, Y, H, S) // returns codeword from the sets

The following function returns the closest matching codeword:

```
findMatc(X,S)
  S←——S;
  Y ←— PickGuess(S);
  h=D(X,Y)
  Ycurr ←—— Y;
  Hcurr ←—— h;
  Repeat
  {
  S ←—— S/Y;
  S ←—— ReduceSet(X,Y,h,S)
  If (|S̄|)0)
  {
    Ycurr ←—— Y;
    Hcurr ←—— h;
  }
} Until (|S̄|==0);
Return Ycurr}
```

Figure 1 Pseudo code of the algorithm which illustrates the three sub-problem involved in the NN Algorithm

6- Performance

Our proposed framework allows us to consider various solutions to each component of the NN search problem. However, the components are not independent of each other. A solution for a particular component can only be evaluated in terms of how well it helps improve the performance of the solutions in other components .ln order to evaluate the performance of the solutions to each component, we need to look at the performance of the of the NN search algorithm as whole.

For example, suppose a solution for PiekGuess () produces, on average, only moderately good initial matches for a moderate amount of computation on the other hand, another solution for PickGuess () produces outstanding initial matches, but it also requires a large amount of computation.

Choosing which solution to use depends on how much the performance of the solutions for ReduceSet () and PickNeXt () are effected by the improvement in the initial matches. lf there is only marginal increase in performance in these components, then the overall cost of the search may in fact increase rather than decline.

In this work, We describe the performance of solution in terms of the resulting number of (D (K) distance calculations. We would like to avoid having to evaluate the efficiency of an algorithm by timing the execution of the software implementation. A number of factors, such as program style, disk access, error trapping and diagnostic fragments of code affect the performance of the software implementation.

7-Future Work

There are a number of techniques, Which can be used to compare vectors without having performed an 0 (K) operation all the time. For example, partial distance search technique [15] can be used to reduce the arithmetic cost of the comparison. The technique works by computing the distance function progressively until the value either goes outside the bounds, or the value has been computed on all distance. The technique proposed in various ways to reduce the arithmetic cost of NN search algorithms [Wen *et al*, 1997;Wen-Jyi *et al*., 1997;Wen –Jyi*et al*.,1997; Ngwa *et al*., 1991].

Reference

Akront, N., Prost, R. and Goutte. R."Image Processing by vector quatization" l994.: a revievv focused on codebook generation. Image and Vision computing l2(9):627- 63 7, November,

Allen Gersho and Robert M.Gray."Vector Quatization and signal compression? Kluwer Academic Publishers. Boston/Dorderecht /London ,l992.

Blue, IL. Candel , G.T. C1rother, P.I. Chellappa, R. and Wilson, C.L." Evaluation of Pattern classiiier for Fingerprint & GCR applications". Pattern Recognition,27(4):485-50l , l 994.

Change-Da Bei and Robert M Gary."An improvement of the minimum distortion encoding algorithm for vector quantization "lEEE Transaction on Communication ,Com-33(1O):1 132-133, (10)ctober,1985. ~

Chay, S.M. and Lo, K.T. "East Clustering Process for vector quatization codebook design". Electronic Letters,32(4):3l I-312, febrevvary,l996.

Coer, T.M. and Hart, P.E., "Nearest Neighbor pattern classification ".lEE Transaction on Information Theory , IT-3(1):21-27,january ,l967.

Cruan, L. and Kane , "Equal-Average hyperplane partitioning method for vector quatization of image data". Pattern Recognition ,13;693-699,1992.

Daernyo Lee, Seongioon Back , and Koemngy O, Sung. "Modified K- means algorithm for vector quantizer Design ". IEEE Signal Processing Letters,4(l):2-4, January, l 997.

Ketterlin , A., Crancaski, P. and Korezak, I.I. "Conceptual clustering in structured database ".A particular approach .In Proceedings of the First International Conference of knowledge Discovery and Data Mining ,pages ISO-ISI, August l995.

Maria Luisa Mico [116] and Jose Oncim " A new version ofthe nearest neighbor approximation and elimination search algorithm (AESA) with linear processing time and memory requirement ", Pattern Recognition,15:9-17,]anuary 1994.

Nasser M Masrabed and Robert A, King." lmage coding using vector quatization ": A revieW.lEEE Transaction on communication ".36(8):957- 971, August,1988.

NgWa-Ndifor, J. and Tellis." Predictive Partial search algorithm for vector quatiztion". Electonic letters, 27(19):1722-1723,1991.

Poggi , G. [131] "Fast algorithm for Full search VQ encoding". Electronic letters 29(12): 1141-1142, Jun 1993.

Robert M.Gray "Vector Qautization". IEEE ASSP Magazine ages 4- 29,April,l984.

Scott Cost and Steven Salberg. "A wheighted nearest algorithm for learning with symbolic features '. Machine Leaning, l0:57-7S,l993.

Wen-Iyi Harang , Biing-Yan Chen , and Sen-shiang. "A fast vector quantization encoding method using Wavelet Transform ". Pattern Recognition letters ISZ73-76,l997.

Wen-Iyi Harang , Sen-shiang, and Biing-Yan Chen "Past codevvord search algorithm using Wavelet transform and partial distance search technique". Electronic letters,33(5):365-366,l997.

Wen-Tyi Hwangi , Yeong-chorg. Ln, Yi-Chig Zeng _ "East block- matching algorithm for video coding". Electronic letters,33(lO):833-935, May,l997.

Xiafeng Han, Peirre Kelsen, Vijay a Ramachandram , and Robert Trajans. "Computing minimal spanning subgraphs in linear time ".S1AM journal on computing,24(6):1332-1358, December, 1995.

Yosehp Linde, Anders Buzo and Robert M Gray. "An algorithm for vector quantization design "IEEE Transaction on communications ,Com- 28(10:28(1):84-95, January 1980.