Improving Analogy learning performance by reduce Computation time

Hiba Mohamed Suliemansalieam

Dept of Computer Science ,Elneelaien university

Abstract:

The analogy learning technique is the most public classification task in data mining. It depends to classify objects at storing the dataset and when it gets new data classified to a category that is much similar to the new data and predicts the class of an object by calculating the distance of the object to each sample in the training dataset. Working on a big dataset can be an expensive task, therefore it was been necessary to improve this method. In this paper, we present improvement Knn algorithm which it is learning based on Analogy. This improved algorithm built upon two strategies, first it identifies the ranges (minimum and maximum points MMP) of the classes, If the object belongs to the rang then the object is predicted belonging to the class. Another strategy is only used in one case when the object falls out of ranges, we calculate the distance between this object and the MMP of classes. We present experiment to evaluate the improvement method. The results show the developed method is better in performance than traditional Knn algorithm and the efficiency had been improved.

keywords—classification technique, Bayesian Network, analog learning, DAG structure.

المستخلص:

التعلم التناظري هو واحد من ابسط تقنيات التصنيف في التنقيب عن البيانات. وهى تعتمد في تصنيف الكائنات على تخزين مجموعة البيانات اولا وعندما تحصل على بيانات جديدة يتم تصنيفها إلى فئة تشبه إلى حد كبير البيانات الجديدة . وتتنبأ بفئة الكائن الجديد عن طريق حساب مسافة بين هذا الكائن وكل عينة في مجموعة البيانات التدريب . عند العمل على مجموعة بيانات كبيرة تكون مكلفة لأنها تأخذ الكثير من الوقت في التحسيب ، ولذلك

كان لابـد مـن تحسـين أداء هـذه الخوارزميـه , حيـث قامـت كثـير مـن الجهـود لتحسـين أداء هـذه الخوارزميـه أعتـمادا عـلى مفاهيـم مختلفـه .وفي هـذه الورقـه قـام الباحـث بتحسـين أداء خوارزميـة knn التـي تعتمـد عـلى Analogy learning وذلـك باتبـاع اسـتراتيجيتين هـما نطاقـات الحـد الأدنى والأقـصى لنقـاط الفئـات(MMP) ، وعندمـا تقـع نقـاط الكائـن الـذي نريـد تصنيفـه داخـل هـذا النطـاق فهـو ينتمـي لهـذه الفئـه . ومـن المتوقـع ان يقـع الكائـن خـارج جميـع نطاقـات الفئـات في هـذه الحالـه فقـط يتـم اسـتخدام الإسـتراتيجية الاخـرى وهـى حسـاب المسـافة بـين هـذا الكائـن و MMP للفئـات. وقـد تـم اجـراء تجربـه experiment لهـذه الخوارزميـه المحسـنه لقيـاس أدائهـا مقارنـه مـع خوارزيـة knn التقليديـه . وقـد تـم تقيـم الخوارزميـات اعتـمادا عـلى وقـت التنفيـذ أي الوقـت الـذي تسـتغرقه الخوارزميـه قـي التنفيـذ واظهـرت النتائـج ان الخوارزميـه المحسـنه هـى الافضـل في الاداء.

الكلـمات المفتاحيـه :التعلـم التناظـري، تنقيـب البيانـات ، تقنيـة تصنيـف الكائنـات ، شـبكات بايسـون، هيكليـة Dag.

1. Introduction

Classification problem is (Zaki, Wagner Meira,2020)one of Data Mining task assigning objects to one of several predefined categories. It learning a target function that maps each attribute set to one of predefined class labels. There are different techniques to learning classification function for example, Decision trees are trees that classify instances by sorting them based on feature values , A Bayesian Network (BN) is (Thair Nu Phyu,2009) a graphical model for probability relationships among a set of variables features ; Typically the task of learning a Bayesian network can be divided into two subtasks: initially, the learning of the DAG structure of the network, and then the determination of its parameters .

There another learning by analogy (k-nearest neighbor classifier algorithm), that we want to focus on it in our study. This algorithm describes training samples by n dimensional numeric attributes, each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample to predict, we need to calculate the similarity distance using similarity measures like Euclidean distance. Although learning by analogy is easy and simple technique, it is expensive and erodes efficiency in the large training dataset when we considering every element in

the training to calculate the distance.

To address this ,we suggest defined the ranges of classes that built upon the notion of minimum and maximum points first, then the testing element is predict if it is within the rang of class , it predicted belonging to the class and when the object fall out of ranges , we calculate the distance between this object and the MMP of classes .The least distance is the closest to the category. This will reduce the calculation distance time and improve the efficiency of algorithm.

the paper is organized as follows: Section 2 gives an overview of related works, Section 3 preprocessing of Training Dataset. Section 4 prediction, Section 5 presents evaluating of improving . Section 6 concludes the paper with a discussion on the future work.

2 Related Works

Many researchers have studied methods for improving the Knn efficiency.

Hamid Saadatfar (2020) Presented a study entitled (A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data) the K-nearest neighbors (KNN) machine learning algorithm is a well-known non-parametric classification method. However, like other traditional data mining methods, applying it on big data comes with computational challenges. Indeed, KNN determines the class of a new sample based on the class of its nearest neighbors; however, identifying the neighbors in a large amount of data imposes a large computational cost so that it is no longer applicable by a single computing machine. One of the proposed techniques to make classification methods applicable on large datasets is pruning. LC-KNN is an improved KNN method which first clusters the data into some smaller partitions using the K-means clustering method; and then applies the KNN for each new sample on the partition which its center is the nearest one. However, because the clusters have different shapes and densities, selection of the appropriate cluster is a challenge. In this paper, an approach has been proposed

to improve the pruning phase of the LC-KNN method by taking into account these factors. The proposed approach helps to choose a more appropriate cluster of data for looking for the neighbors, thus, increasing the classification accuracy. The performance of the proposed approach is evaluated on different real datasets. The experimental results show the effectiveness of the proposed approach and its higher classification accuracy and lower time cost in comparison to other recent relevant methods.

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 our proposed algorithm performs better than conventional KNN algorithm.

Parvin et al ,Modified K-Nearest Neighbor,2008 , presented 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.

Zhou et al, An Improved KNN Text Classification Algorithm Based on Clustering, 2009 , presented a modified kNN algorithm to preprocess training datasets using the k-mean clustering algorithm (MacQueen et al., 1967) to find the centroid of each known class. After identifying the centroid, the algorithm eliminates far-most points in the class to avoid the multi-peak distribution effect which involves multiple classes overlapping. After the elimination, the

k-mean clustering algorithm is used to identify sub classes and their centroids which form sa new training set.

Changyin Zhou (2006) Presented a study entitled (Improving nearest neighbor classification with cam weighted distance) propose a novel cam weighted distance to ameliorate the curse of dimensionality it different from the existing neighborhood-based methods which only analyze a small space emanating from the query sample, the proposed nearest neighbor classification using the com weighted distance CamNN) optimizes the distance measure based on the analysis of f inter-prototype relationship.

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

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.

Lin et al , An Enhanced Entropy-K-Nearest Neighbor Algorithm Based on Attribute Reduction, 2015 , combine the kNN algorithm with the k-mean clustering algorithm to improve

the accuracy and efficiency of kNN for intrusion detection by preprocessing the training dataset. The preprocessing involves finding the centroid of each class using the k-mean clustering and computing the distance between each point in the class to its neighbors and the class centroid. The same preprocessing applies to the testing dataset. Similar to Parvin et al.'s work, it can be considered as converting the n-dimensional dataset into one-dimensional dataset.

3.Preprocessing Training Dataset

We fallow two steps For Preprocessing Training Data

1. defining classes and sorting them in lists .

A given training data set contains data elements which are definedin terms of attributes .we separate classes and sorting them in lists .

Algorithm 1:shows the separation and sorting of classes.

```
1. procedure SEPARATE CLASSES(td:
   in Training Dataset, ct: out Class
   Type, cd: out Class Dataset)
2. for each instance in td, do
3. if instance label ∈ ct then
4. class dataset ←instance
5. else
6. ct← instance label
7. cd← instance
8. end if
9. end for
10. end procedure
```

2.determined the minimum and maximum value of each dimension (attribute) of the class. consider :

C1={(1,1,6),(2,5,7),(3,7,8),(4,9,3),(5,2,10)} C2={(11,19,18),(10, 26,17),(25,25,19),(29,17,18),(31,12,20)}

C1 min ={(1,1,3)} C1max={(5,9,10)}

C2min=(10,12,17)} C2max=(31,26,20)

Algrithm2 show the defining minimum and maximum value of

classes

Procedure a: FIND MAX POINT(cd: in Class Dataset)

1. For each instance in cd do
2. if instance > max Instance then
3. max Instance← instance
4. end if
5. end for
6. CMAX ←max Instance

Procedure b: FIND MIN POINT (cd: in Class Dataset)

1. For each instance in cd do
2. if instance < min Instance then
3. min Instance← instance
4. end if
5. end for
6. CMIN ←min Instance
7. End procedure

4. Prediction

The class of a testing element is predicted based on the arranged list that consist minimum and maximum points of classes. the testing element is evaluated if it is within the range of class.

consider :

C1={(1,1,6),(2,5,7),(3,7,8),(4,9,3),(5,2,10)} C2={(11,19,18),(10, 26,17),(25,25,19),(29,17,18),(31,12,20)}

C1 min ={(1,1,3)} C1max={(5,9,10)}

C2min=(10,12,17)} C2max=(31,26,20)

For example, consider a testing data T2={(2,8,4)}. PerC1min={(1,1,3)} andC1max={(5,9,10)}of C1, T2 is within the range of the MMP, and thus it is predicted belonging to C1. If a testing data fall out of range. suppose a testing element T1=(6,10,2) also we need to predicted it , first if fall out of ranges

C1 min ={(1,1,3)} C1max={(5,9,10)}

C2min=(10,12,17)} C2max=(31,26,20)

Then we used the second strategy , calculated the distance of T1 to MMP of the classes measured as : the distance of T1 to C1min(1,1,3) measured as 13.0384 ,to C1max(5,9,10) measured as 11.91638 , to C2min(10,12,17) measured as 23.08679 , to C2max (31,26,20) measured as13.0384 . The shortest distance is 11.91638 , Therefore, T1 is predicted belonging to C1.

Algorithm 3: shows the Prediction and evaluation.

```
1: procedure PREDICTION(RS: in arranged Subset, CMAX :in Class Maximum
Point, CMIN :in Class Minimum Point ,CME in  class Mediator point  ,pl: out
Predictions List, el: out Evaluation List))
2: initialize kNN (RS)
3: for each class do
4: for each instance in testing dataset do
5: if instance(<Mediator and instance≥ CMIN) or (instance< CMIN) then
6: pl ← class type
7: else
8: pl ← kNN. classify(instance)
9: end if
11: end for
12: el← evaluate (predictions list)
13: end procedure
```

We apply the improvement method to the Iris data and measure the execution time .The study was carried out comparatively by comparing the results of applying the traditional kNN algorithm to the original training dataset with the results.

5.1Results

After we implementing the improvement algorithm against traditional Knn . we take these result and measure its execution time, blow table had show us that our method made and improvement in the process of classification.

Table 1: Results implementation of kNN

Alg.	Time cost
Knn	1.254

Table 2: Results implementation of Improvement method

Alg.	Time cost
Improvement method	0.066

Table3: Comparison results with other approach

approaches	Techniques method	Time cost
Improvement method	minimum and maximum points (MMP) and boundary subsets (BS)	1.254
Knn	calculated distance	0.066

5.2 Discussion

One of the most important results that have been achieved reducing the implementation time, predication and recall . It will be feasible when used in large data. One of the challenges that are expected to face many improvement of this theory is the overlap between classes. When we followed the two suggested strategies in predicting test element in our experiment the overlap was avoided. we used in the case studies the value1for k in the traditional kNN algorithm and the same data set was used for two methods. We Compared the performance of two algorithms , The results had shown the improved algorithm has the best performance.

6. Conclusion

We improved the efficiency of knn algorithm, which is one of machine learning algorithm based on learning by analogy. We improve the efficiency of KNN algorithm by reduce the computation time for a large training dataset. The improvement base on preprocessing the training dataset built upon adjust the

boundary of classes. The case studies presented in this paper validate the improve method by demonstrating its improvement one efficiency over the kNN algorithm. The results show a significant enhancement in efficiency with little sacrifice of accuracy compared to the traditional kNN algorithm.

In the future work, the improvement that had been developed in this method through this paper to perform a binary classification of the data. We plan to multiple classification task to perform.

References:

(1) Anany Levitin, *Introduction* to the *design* & *analysis of algorithms(* (2nd Edition)) , 2007

(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) Hall,M.,Frank,E.,Holmes,G.,Pfahringer,B.,Reutemann, P., and Witten, I. H The WEKA Data Mining Software: an Update. ACM SIGKDD Explorations Newsletter, 2009.

(5) Karina Giberta , Miquel Sànchez-Marrèa, Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation , 2010.

(6) Kubat, Miroslav Cooperson Martin), A reduction technique for nearest-neighbor classification: Small groups of examples. Intell. Data Anal , 2001

(7) Muja,M. and Lowe, D.GScalable Nearest Neighbor Algorithms for high Dimensional Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014

(8) 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.

(9) S.-W., and Tsai, C.-F. CANN: An Intrusion Detection System Based on Combining Cluster Centers and Nearest Neighbors. Knowledge Based Systems, 2015.

(10) Thair Nu Phyu ,Survey of Classification Techniques in Data Mining, 2009.

(11) 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.

(12) Wilson, D. R. & Martinez, T. Reduction Techniques for Instance-Based Learning Algorithms. Machine Learning 2000.

(13) Yu, C., Ooi, B. C., Tan, K.-L., and Jagadish, H. Indexing the Distance: An Efficient Method to Knn Processing 2001.

(14) Zhou, Y., Li, Y., and Xia, S. An Improved KNN Text Classification Algorithm Based on Clustering. Journal of Computers, 2009.
(15) Zheng, Z. Constructing X-of-N Attributes for Decision Tree Learning, Machine Learning2000.
(16) ZakiWagne,Meira,Jr.,Data mining and machine learning fundamental concept,2020
(17) Zhou, Y., Li, Y., and Xia, S. An Improved KNN Text Classification Algorithm Based on Clustering. Journal of Computers, 2009.
(18) Z., Qiu, M., Chen, Y., and Liu, H. Coarse to Fine KNearest Neighbor Classifier. Pattern Recognition Letters, (2013).
(19) Zhao, Learning Structured Representation for Text Classification via Reinforcement Learning Tianyang 2004.