IOP Conference Series: Materials Science and Engineering

Research and Implementation of Machine Learning Classifier Based on KNN

Research and Implementation of Machine Learning Classifier Based on KNN

Lishan Wang

Fujian Normal University, Fuzhou 350108, China

Abstract. Machine learning classifier is an important part of pattern recognition system; it is also an important research field of machine learning. The main research object of this paper is K data mining (KNN, K Nearest Neighbor) classification method, using KNN to classify the data, and compare the classification results. The research work of this paper mainly discusses the implementation of KNN-based machine learning classifier, mainly focusing on the theoretical analysis of K-data mining, algorithm implementation, and implementing KNN-based machine learning classifier.

1. Introduction

Machine learning classifier definition: The input data contains thousands of records, each record has many attributes, and one special attribute is called class (such as high, medium and low credit). The purpose of the machine learning classifier is to analyze the input data, and build a model, and use this model to classify future data. Data classification technology in credit card approval, target market positioning, medical diagnosis, fault detection, effectiveness analysis, graphics processing And in the field of insurance fraud analysis, you can see that machine learning classifiers are widely used.

The data used for classification is a set of samples of a known category, each sample containing a set of identical attributes. According to the role in the classification, attributes can be divided into conditional attributes and target attributes. Thus, a sample can be expressed in the form of (X_1, X_2,...$X_{m,}$ Y), where Xi is a conditional attribute and Y is a target attribute. The purpose of classification is to discover the dependencies between X_1, X_2, X_m… and Y, which are also called classification models or machine learning classifiers. It can be considered that the machine learning classifier is a function whose input is a sample of an unknown category and the output is the category of the sample.

2. K-data mining concept

KNN stands for k nearest neighbor classifications, identifying new records by a combination of K's most recent historical records. KNN is a well-known statistical method that has been studied intensively in pattern recognition over the past 40 years. KNN has been applied to text categorization in early research strategies and is one of the highly operational methods of the benchmark Reuters body. Other methods, such as LLSF, decision trees, and neural networks.

The idea of KNN is as follows: First, calculate the distance between the new sample and the training sample, find the nearest K neighbors; then, according to the category to which the neighbor belongs, determine the category of the new sample, if they all belong to the same category, then The new sample also falls into this category; otherwise, each post-selection category is scored and the new sample category is determined according to certain rules.

Take the K neighbors of the unknown sample X, and look at which category the K neighbors belong to, and classify X into which category. That is, among the K samples of X, K neighbors of X are found. The KNN grows from the test sample X, continuously expanding the area until it contains K training samples, and classifies the test sample X as the most frequently occurring category among the most recent K training samples. For example, in the case of K=6 in Fig. 1, the test sample X is classified into a black category according to the decision rule.

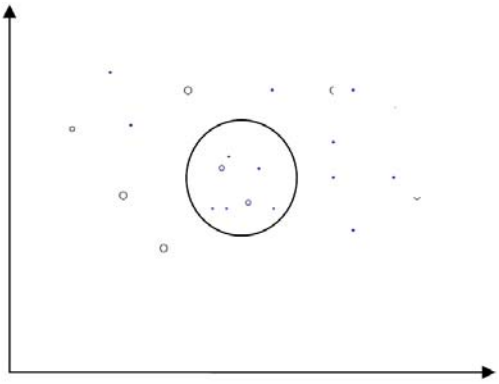


Figure 1. K Nearest Neighbor

The neighborhood classification is a lazy learning method based on the eyeball, that is, it stores all the training samples and knows that the new samples need to be classified to establish the classification. This is in stark contrast to decision numbers and backpropagation algorithms, which need to construct a general model before accepting new samples to be classified. Lazy learning is faster in training than in eager learning, but slower in classification because all calculations are postponed until then.

3. Mathematical model of KNN algorithm

The reason for prediction using the nearest neighbor method is based on the assumption that objects of neighbors have similar prediction values. The basic idea of the nearest neighbor algorithm is to find k points nearest to the unknown sample in the multidimensional space R^n, and judge the class of the unknown sample according to the categories of the k points. These k points are the k-nearest neighbors of the unknown samples. The algorithm assumes that all instances correspond to points in n-dimensional space. The nearest neighbor of an instance is defined according to the standard Euclidean distance. Let the eigenvector of x be:

$$<a_1(x), a_2(x), \ldots, a_n(x)>$$

Where $a_r(x)$ represents the rth attribute value of instance x. The distance between the two instances x_i and x_j is defined as d (x_i, x_j), where:

$$d(x_i, x_j)=\sqrt{\sum_{r=1}^{n}(ar(xi)-ar(xj))2}$$

In nearest neighbor learning, the discrete object classification function is f: R^n->V where V is a finite set $\{v_1, v_2, \ldots v_s\}$, ie different sets of categories. The selection of the nearest neighbor k value is based on the number and degree of dispersion in each type of sample, and different k values can be selected for different applications.

If the number of sample points around the unknown sample s_i is small, the area covered by the k points will be large, and vice versa. Therefore, the nearest neighbor algorithm is susceptible to noise data, especially the effects of isolated points in the sample space. The root cause lies in the basic KNN algorithm, in which the positions of the k nearest neighbor samples of the sample to be predicted are equal. In a natural society, usually an object is affected by its neighbors, and the closer the object is, the more influence it has.

4. KNN research method

The algorithm has no learning process, and predicts the category of the new sample by the samples with known categories at the time of classification, so it belongs to the instance-based reasoning method. If K is equal to 1, the category of the sample to be divided is the category of the nearest neighbor, called the NN algorithm.

As long as there are enough training samples, the NN algorithm can achieve a good classification effect. When the number of training samples approaches -∞, the classification error of the NN algorithm is twice the optimal Bayesian error; in addition, when K approaches ∞, the classification error of the KNN algorithm converges to the optimal Bayesian Error. The following describes the KNN algorithm:

```
Input: training data set D = {($X_i$, $Y_i$), $1 \le i \le N$}, where $X_i$ is the conditional attribute of the ith sample, $Y_i$ is the category, new sample X, distance function d.
Output: Category Y of X.
For i=1 to N do
Calculate the distance d ($X_i$, X) between X and $X_i$;
End for
Sort the distance and get d (X, $X_{i1}$) $\le$d(X, $X_{i2}$) $\le$… $\le$d(X, $X_{iN}$);
Select the first K samples: S= {($X_{i1}$, $Y_{i1}$)…($X_{iK}$, $Y_{iK}$)};
Count the number of occurrences of each category in S and determine the category Y of X.
```

5. Program interface design

In the C# integrated development environment, use the form designer, control toolbox, and properties window to create an application interface.

The requirements for each control property setting are as follows:

The form contains 4 groupBox controls, 6 TextBox controls, 2 ListBox controls, 3 Button controls, 8 label controls, 5 radioButton controls, and 1 checkBox control. The groupBox control, the TextBox control, the label control, and the ListBox control are named by default, and the value values of the remaining controls are as shown in Table 1.

Table 1. Name attribute value of each control

Control	Name attribute value
determine	ok
next	next
calculation	solve
numerical	numeric
type value	category
normalization	normalization
total	summation
Euclid	euclidean
weights	weighted

6. Database linkage

This article uses the data provided by the UCI machine learning library to test the program. The letter dataset is used, with 20,000 rows of data, 16 attributes, and 26 classification labels.

Create a database in the SQL server, named "datamin_problem", and then import the dataset downloaded from UCI, letter (text form), into the SQL server by importing the database. After importing, name the table "problem". The design view of the table is shown in Table 2.

Table 2. Training set data design table

Field name (attribute)	Type
col000	varchar
col001	varchar
col002	varchar
col003	varchar
…	…
col015	varchar
col016	varchar
col017	varchar

Since there are many UCI data, 1000 data is selected as the training set in this program. And select the other 100 data in the letter data for testing.

7. Program operation and debugging

Press the F5 key to run the program and enter the following values in each input box:

(a). Enter col000 in the "property name" input box;

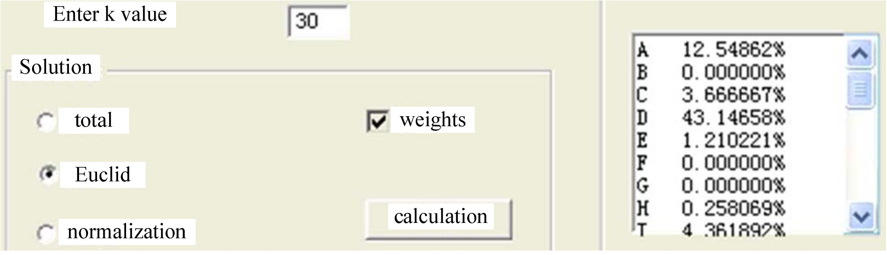
(b). In the "Classification Properties" group box, enter col017 in the "Name" box; enter A, B, C, ..., Z in the "Value" box; then click the "OK" button;

(c). In the "Attribute Data" group box, select the "Value" radio button; "Name" and "New Record" enter the value of the new record; each time you enter an attribute name and corresponding data, click the "Next" button; Enter a test set.

(d). In the Enter K Value text box, enter 30;

(e). Select "Euclidean" in "Workaround"; click the "Calculate" button.

The results of the operation are shown in Figure 2:



Enter k value: 30

Solution

○ total

● Euclid

○ normalization

☑ weights

```
A    12.54862%
B    0.000000%
C    3.666667%
D    43.14658%
E    1.210221%
F    0.000000%
G    0.000000%
H    0.258069%
T    4.361892%
```

Figure 2. Program running result graph

This result is the probability that the program judges that the data is a classification label such as A, B, C..., and the highest probability is that the program judges that the data belongs to that category. This finds 100 data from the letter data set to continue testing.

From the results of the program operation, we can see that the data obtained for the 100 data we entered during the test is compared with the data set of the letter. The final result is different in 28 data tests and data sets, so the correct rate reaches 82%. Therefore, the design requirements are basically met, and the KNN machine learning classifier is basically realized. This program can be used to classify such data.

8. Conclusion

This paper implements the KNN machine learning classifier, and the results of the data test show that the basic goal is achieved and the classification effect is achieved. The KNN classification algorithm is subjective because a distance scale must be defined. Since the understanding of the distance is not profound, the result of the classification depends entirely on the distance used. Thus, with a set of data, two different classification algorithms will produce two A completely different classification result usually requires experts to evaluate whether the results are valid. Since the recognition of results is often empirical, this limits the use of various distances.

References

[1] Ji S, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35 (1): 221.

[2] Silla Jr C N, Freitas A A. A survey of hierarchical classification across different application domains [J].Data Mining and Knowledge Discovery, 2011, 22 (2): 31-72.

[3] Han J, Kamber M. Data Mining: Concepts and Techniques[J]. Data Mining Concepts Models Methods &Algorithms Second Edition, 2011, 5 (4): 1-18.

[4] Liu Y, Bi J W, Fan Z P. A method for multi-class sentiment classification based on an improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm [J]. Information Sciences, 2017, 394 (9): 38-52.

[5] Gong M, Liang Y, Shi J, et al. Fuzzy C-means clustering with local information and kernel metric for image segmentation.[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2013, 22 (2): 573.

[6] Liu Z G, Pan Q, Dezert J, et al. Credal c-means clustering method based on belief functions [J].Knowledge-Based Systems, 2015, 74 (1): 119-132.

[7] Fernandez-Gago C, Agudo I, Lopez J. Building trust from context similarity measures [J]. Computer Standards & Interfaces, 2014, 36 (4): 792-800.