Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, 10-13 July, 2011

THE CONDENSED FUZZY K-NEAREST NEIGHBOR RULE BASED ON SAMPLE FUZZY ENTROPY

JUN-HAI ZHAI, NA LI, MENG-YAO ZHAI

Key Lab. of Machine Learning and Computational Intelligence, College of Mathematics and Computer Science, Hebei University, Baoding 071002, China
E-MAIL: mczjh@hbu.cn

Abstract:

The fuzzy k-nearest neighbor (F-KNN) algorithm was originally developed by Keller in 1985, which generalized the k-nearest neighbor (KNN) algorithm and could overcome the drawback of KNN in which all of instances were considered equally important. However, the F-KNN algorithm still suffers from the problem of large memory requirement same as the KNN. In order to deal with the problem, this paper proposes the condensed fuzzy k-nearest neighbor rule (CFKNN) which selects the important instances based on sample fuzzy entropy. The experimental results show that our proposed method is feasible and effective.

Keywords:

Nearest neighbor; Condensed nearest neighbor; Fuzzy nearest neighbor; Instance selection; Sample fuzzy entropy

1. Introduction

The nearest neighbor rule (NN) was originally proposed by Cover and Hart[l]，and is widely used in many fields such as pattern recognition[2], data mining[3], and machine learning[4-7]. The reasons for the use of this rule are its conceptual simplicity and easy to understand. However, the nearest neighbor rule suffers from the following three problems:

(1) To classify a test instance *x*, it is required to store all instances in training set, i.e. the computational space complexity is $O(n)$, where *n* is the number of instances in training set;

(2) The distances between *x* and all instances in training set are needed to compute, i.e. the computational time complexity is also $O(n)$;

(3) Each of the training samples is given equal importance in classifying an unseen sample, regardless of their different contribution to classification.

In order to deal with the problems mentioned above, many researchers did a great deal of research works. Hart proposed the condensed nearest neighbor rule (CNN) to deal with the problem (1) and (2) [8]. Keller proposed F-KNN algorithm to deal with the problem (3) [9]. However, F-KNN algorithm still suffers from the problem (1) and (2). In this paper, we propose the method of the condensed fuzzy nearest neighbor rule (CFKNN) based on sample fuzzy entropy. The experimental results show that the proposed method is feasible and effective.

The paper is organized as follows. Preliminaries on our proposed method are given in section 2. In section 3 the CFNN is presented and experimental results and analysis are provided in Section 4, Section 5 concludes the paper.

2. Preliminaries

In this section, we will introduce several basic concepts and algorithms related to our method, mainly including the concept of decision table, fuzzy entropy, the algorithm used for determining the fuzzy membership degree of instances in training set, and the FNN algorithm.

Definition 1 A decision table (*DT* in short) is a 2-tuple $DT = (U, A \cup C)$ where $U = \{x_1, x_2, \cdots, x_N\}$ is a non-empty finite set of objects (instances) called training set and *A* is a set of real-valued conditional attributes. C is the decision attribute, without loss of generality we suppose that the instances in *U* are classified into *p* categories $C_1, C_2, \cdots, C_p$.

Definition 2 Given a decision $DT = (U, A \cup C)$, $\forall x_i \in U, \forall C_j (1 \le i \le n; 1 \le j \le p)$, let $\mu(x_i, C_j)$ be the fuzzy membership degree of instance x_i belong to class C_j, the fuzzy entropy of instance x_i is defined as follows

$$Entr(x_i) = -\sum_{j=1}^{p} \mu(x_i, C_j) \log_2 \mu(x_i, C_j) \qquad (1)$$

978-1-4577-0308-9/11/$26.00 ©2011 IEEE

where the $\mu(x_i, C_j)$ denotes the fuzzy membership degree of instances belong to class C_j.

In this paper, we use the following algorithm to determine the fuzzy membership degree of instances in training set U [10].

Algorithm 1

Input: A *DT* with real-valued conditional attributes.

Output: the fuzzy membership degree of instances.

STEP 1: For each class C_j, calculating the center c_j of class C_j;

STEP 2: For each instance $x_i \in U(1 \le i \le n)$, calculating the distance d_{ij} between x_i and $C_j(1 \le j \le p)$;

STEP 3: For each instance $x_i \in U$, calculating the fuzzy membership degree $\mu(x_i, C_j)$ as follows,

$$\mu(x_i, C_j) = \frac{(d_{ij}^2)^{-1}}{\sum_{j=1}^{p} (d_{ij}^2)^{-1}} \tag{2}$$

For the convenience, we list the F-KNN algorithm as follows.

Algorithm 2

Input: A *DT* with real-valued conditional attributes, and a test instance *x*.

Output: Fuzzy K-nearest neighbor rule.

STEP 1: Initialize K=K_0;

STEP 2: For each test *x*, found K-nearest neighbors of *x* in DT;

STEP 2.1 Initialize i =1;

STEP 2.2 Do

STEP 2.3 Compute distance from *x* to x_i;

STEP 2.4 IF ($i \le K_0$) THEN

Include x_i in the set of K_0-nearest neighbors;

ELSE

IF (x_i closer to *x* than any previous nearest neighbor) THEN

Delete the farthest of the k-nearest neighbors;

Include x_i in the set of K_0-nearest neighbors;

STEP 3: For each test *x*, compute $\mu(x, C_j)$ using formula (2) and (3);

$$\mu(x, C_j) = \frac{\sum_{i=1}^{K_0} \mu(x_i, C_j) \times \left(\frac{1}{\|x - x_i\|^{\frac{2}{m-1}}} \right)}{\sum_{i=1}^{K_0} \frac{1}{\|x - x_i\|^{\frac{2}{m-1}}}} \tag{3}$$

3. The Condensed Fuzzy K-Nearest Neighbor Rule Based on Sample Fuzzy Entropy

In this section, we will present our method of the CFKNN, In KNN and F-KNN, the problem of high computational complexity is encountered unavoidably due to storing all instances in training set. In fact, for a given fuzzy information system, different instance in training set has different important degree, and has different contribution to classification. Some instances may be more important than the others. In our method we select the set of the important samples based on the fuzzy entropy of the sample in training set. The bigger the fuzzy entropy is, the more important the sample, because the sample with bigger fuzzy entropy can provide more information for classification and they are closer to the boundaries of class. So the set of the important samples usually contains the same information with original dataset in classification. If the important sample set is used as training set to classify an unseen test sample, the efficiency of classification can be increased, and computational complexity degree can be decreased.

In the following, we provide the CFKNN algorithm for our method proposed above.

Algorithm 3

Input: a *DT*, parameter K and α (suppose $|DT|=n$, and the samples in *T* are classified into *p* classes)

Output: $S \subset DT$

STEP 1: For each instance $x \in DT$, d determine the fuzzy membership degree of x with (2);

STEP 2: Randomly select one instance belonging to each class from training set, and put the selected samples into *S*.

STEP 3: Repeat the following process for each *x* remained in *DT*.

STEP 3.1: Find K nearest neighbors in *S*;

STEP 3.2: Determine the class membership degree of *x* with (3) $(\mu(x, C_1), \mu(x, C_2), \cdots, \mu(x, C_p))$;

STEP 3.3: Compute the fuzzy entropy of instance *x*

with (1) $Entr(x_i) = -\sum_{j=1}^{p} \mu(x_i, C_j) \log_2 \mu(x_i, C_j)$;

STEP 3.4: If $Entr(x) > \alpha$, then $S = S \cup \{x\}$; Else discard x;

STEP 3: Return S.

4. Experimental results

The effectiveness of our proposed method is demonstrated through numerical experiments in the environment of Matlab 7.0 on a Pentium 4 PC. In our experiments we totally select 9 datasets including 7 UCI datasets [8] and 2 real world datasets []. The 7 UCI datasets are Iris Dataset, Breast Cancer Dataset-WDBC, Breast Cancer Dataset-WPBC, Glass Dataset, Image Segmentation Dataset, Parkinsons Dataset, Pima Dataset. The 2 real world datasets are CT Image Dataset and RenRu Dataset. The CT Dataset is obtained by collecting 212 medical CT images from Baoding local hospital. All instances with 35 numerical attributes are classified into 2 classes (i.e., normal class and abnormal class). The RenRu Dataset is created by the key laboratory of machine learning and computational intelligence of Hebei Province, China. The RenRu Dataset is obtained by collecting 148 Chinese characters REN and RU with different typeface, font and size, in which there are 92 Chinese characters REN and 56 Chinese characters RU. For each Chinese character, it is described by 26 numerical features. The basic information of the 10 datasets is listed in table 1.

In the experiment, we set K=5, and randomly select 70% data in each dataset as training set, other 30% data as testing set, For each dataset, we run 10-fold cross-validation ten times, the experimental results are the average of the 10 outputs and listed in table 2. The experimental results demonstrate that the proposed method is effective and efficient.

In the experiment, we also explore the relation between the value of α and the testing accuracy in CFKNN. We change the parameter α from 0.5 to 1.0, and each time adds 0.05. With α set to different values we record the classification accuracies of the CFKNN, the changing curves are shown in figure 1. From the curves, we can see that the value of α does affect the classification result. For Iris dataset, it is appropriate for α take values in the interval [0.5, 0.75]. For WDBC and WDBC dataset, the appropriate interval is [0.5, 0.7] and [0.5, 0.85] respectively. For the other datasets, it is appropriate for α take values in the interval [0.5, 0.95].

In addition, in the experiment, we study the relationship between the value of α and the number of instances selected by CFKNN. The curves describing the relationship between α and the number of instances selected are obtained and shown in figure 2. From the curves, we have observed that the value of α does affect the number of instances selected by CFKNN, with the increase of the value of α, the number of instances selected by CFKNN is decreased continually, when $\alpha > 0.5$, most of curves become more and more smoothly except WPBC. So the number of instances selected by CFKNN will have little change with the increase of the value of α when $\alpha > 0.5$. Considering the testing accuracy, it is reasonable that α takes different value between 0.5 and 0.95 for different dataset.

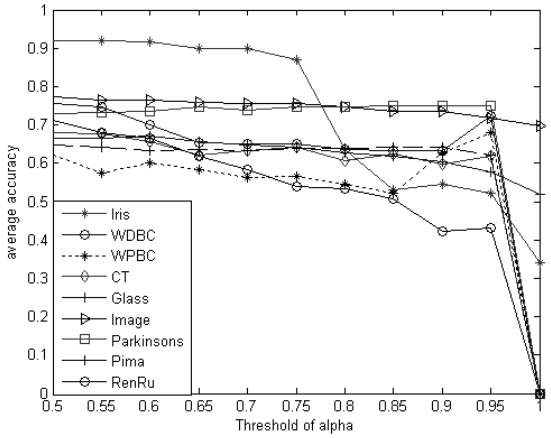

Figure 1. The average accuracy of CFNN on 9 datasets with different thresholds

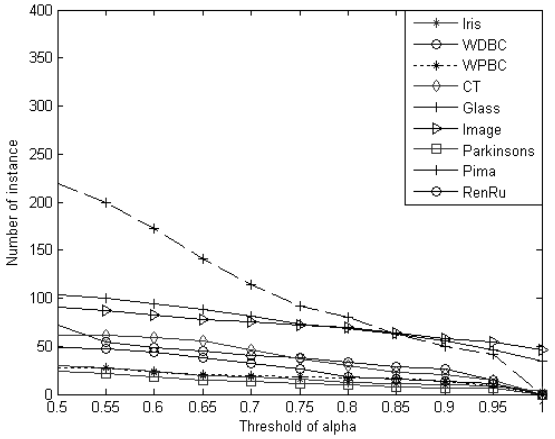

Figure 2. The number of instance selected by CFNN on 9 datasets with different thresholds

5. Conclusions

In this paper, in order to overcome the drawback of large computational complexity requirement of F-KNN, based on the fuzzy entropy of instance, we propose the condensed fuzzy k-nearest neighbor rule (CFKNN). The experimental results show that our proposed method is feasible and effective.

Acknowledgments

This research is supported by the national natural science foundation of China (60903088, 60903089), by the natural science foundation of Hebei Province (F2010000323, F2011201063), by the Key Scientific Research Foundation of Education Department of Hebei Province (ZD2010139), by the Scientific Research Foundation of Education Department of Hebei Province (2009312, 2009410), and by the Undergraduate Science and Technology Innovation Projects of Hebei University (2011043).

References

[1]. T. Cover, P. Hart. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 1967, 13(1):21-27.

[2]. B. Dasarathy. Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. Computer Society Press, 1991.

[3]. X. Wu, V. Kumar, J. R. Quinlan et al. Top 10 algorithms in data mining. Knowledge Information System, 2008, 14(1):1-37.

[4]. T. M. Mitchell. Machine learning. McGraw-Hill Companies, Inc. 2003.

[5]. K. Small, D. Roth. Margin-based active learning for structured predictions. International Journal of Machine Learning and Cybernetics, 2010, 1(1-4):3-25.

[6]. L. Wang. An improved multiple fuzzy NNC system based on mutual information and fuzzy integral. International Journal of Machine Learning and Cybernetics, 2011, 2(1):25-36.

[7]. Z. Liu, Q. Wu, Y. Zhang et al. Adaptive least squares support vector machines filter for hand tremor canceling in microsurgery. International Journal of Machine Learning and Cybernetics, 2011, 2(1):37-47.

[8]. P. Hart. The condensed nearest neighbor rule. IEEE Transaction on Information Theory, 1968, 14(5):515-516.

[9]. J. M. Keller, M. R. Gray, J. A. Givens. A fuzzy k-nearest neighbor algorithm. IEEE trans. on SMC, 1985, 15(4):580-585.

[10]. J. H. Zhai. Fuzzy decision tree based on fuzzy-rough technique [J]. Soft Computing, 2010, DOI: 10.1007/s00500-010-0584-0.

[11]. C. L. Blake, C. J. Merz. UCI Repository of machine learning databases. 1996, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[12]. X. Z. Wang, J. H. Zhai, S. X. Lu, Induction of multiple fuzzy decision trees based on rough set technique, Information Sciences, 2008,178(16):3188-3202.

Table 1. The basic information of the 10 datasets used in our experiments

DB	Number of instances	Number of attributes	Number of classes
Iris	150	4	3
WDBC	555	30	2
WPBC	191	33	2
Glass	160	9	6
Image	194	19	7
Parkinsons	195	22	2
Pima	768	8	2
CT Image	212	35	2
RenRu	148	26	2

Table 2. Experimental results with *K*=5

Dataset	□	The number of selected instances	The average accuracy	CPU Time(s)
Iris	0.65	50	0.96	0.0261
WDBC	0.55	174	0.92	0.2184
WPBC	0.80	90	0.68	0.0183
Glass	0.90	113	0.68	0.0580
Image	0.90	103	0.80	0.0561
Parkinsons	0.95	68	0.75	0.0182
Pima	0.75	377	0.69	0.4828
CT Image	0.90	83	0.84	0.0491
RenRu	0.95	59	0.81	0.0256