A Method Based on the Edited FKNN by the Threshold Value

Chao Zhang, Jianmei Cheng, Liangzhong Yi
Sichuan Police College/the Department of Road Traffic Management, Luzhou, China
Email: galoiszhang@163.com

***Abstract*—The edited technique is of great importance in pattern recognition. The classical edited fuzzy technique use fuzzy k nearest neighbors(FKNN) to take out some useless samples which was classified erroneously in the editing process. In this paper, a proposed edited fuzzy k nearest neighbors based on threshold is developed, which not only consider the maximum membership value but also consider that whether the maximum value is bigger than the given threshold value. We use refer samples to classifier the test samples by FKNN, in which we not only select samples classified correctly but also consider the maximum membership value. That is, threshold value is used to take out some samples that was unreliable. Several comparisons are made between the proposed edited FKNN and the classical edited FKNN, which shown that the proposed method is better then the classical method.**

***Index Terms*—FKNN, edited FKNN, the threshold value**

I. Introduction

Pattern classification is a method capable of discriminating patterns, it is an approach to supervised learning in pattern recognition[13]. In pattern classification, instance selection plays a great role in improving recognition rate and reducing the size of samples set. Instance selection is to reduce the original training set to smaller one. Through the process of instance selection, some unreliable samples are removed, and it also reduce the computation complexity and storage space.

At present, there are many schloars researching the theory of instance selection. Several methods such as edited KNN [4], edited technique for genetic algorithm[7], unlabel data[6], evolutionary prototype selection[5], comparison result[8] [9], edited FKNN [13], NNR through proximity graphs[14], NNR using a two-layer perception[16], were proposed to reduce the number of instances in the training set. Some methods extract only unreliable samples while others try to remove as many instances as possible without significant degradation of the reduced dataset for learning[8].

In this paper, we mainly study and develop edited FKNN algorithm[10]. Edited FKNN is to improve recognition performance by using an editing procedure to reduce the number of preclassified samples. It first divided the training set into two subset: refer set and test set. Then each sample in the reference set is classified using FKNN. Finally, all the samples misclassified are then deleted from the reference set. In the procedure of edited FKNN, its editing criterion is whether the estimate class label, i.e., the class label of maximum membership value, is equal to class label or not.

In this paper, we adopt edited FKNN by the threshold value to filter the data sets. In Edited FKNN, it just only consider what category does the maximum membership value come into. In this paper, besides what category does the maximum membership value come into, we also consider if the maximum membership value is large enough. If the maximum membership value is large enough, it means that the sample almost belong to certain class, otherwise the sample may be on the boundary of multiple classes. We use refer samples to classifier the test samples by FKNN, in which we not only select samples classified correctly but also consider the maximum membership value is large enough or not.

II. Preliminaries

In the following we review briefly two preliminary theory: FKNN and Edited FKNN.

A. FKNN

FKNN was first proposed by Keller[10] to solve a problem that each of the labeled samples is given equal importance in deciding the class memberships of the patterns to be classified. The advantage in FKNN is that the degree of membership in a set can be specified, rather than just the binary [2], [11], [12]. This algorithm firstly finds K nearest neighbors to each testing sample according to dissimilarity measure, give a initial membership of labeled samples, and then makes a decision according to the labeled neighbors, usually by assigning the label of the most voted class among these K neighbors.

Let K be the number of nearest neighbors, and $D=\{(y_1,\theta),(y_2,\theta),\cdots,(y_t,\theta)\}=\{C_1,C_2,\cdots,C_c\}$ be the labeled training set, x be a testing sample, where t is the number of all labeled samples. The algorithm is summarized as follows[2]:

Step1. Input the value K, the labeled sample set D, and Initialize the membership matrix

Manuscript received October 20, 2012; revised November 1, 2012; accepted November 1, 2012.

doi:10.4304/jcp.8.7.1821-1825

$$U=(u_j(y_i))_{t\times c}=\begin{pmatrix} u_1(y_1) & u_2(y_1) & \cdots & u_c(y_1) \\ u_1(y_2) & u_2(y_2) & \cdots & u_c(y_2) \\ \vdots & \vdots & \ddots & \vdots \\ u_1(y_t) & u_2(y_t) & \cdots & u_c(y_t) \end{pmatrix}$$

, where $u_j(y_i)=u_{ij}$ denotes the degree of sample vector y_i in j-th classes.

Step2. Calculate the distance between x and all labeled samples y_i, i.e., $d(x,y_i)=\left|x-y_i\right|$.

Step3. Sort the distances $\{d(x,y_i)\,|\,i=1,2,\cdots,t\}$, determine K nearest neighbors that are closest to x, list K nearest neighbors $D_K=\{y_1',y_2',\cdots,y_K'\}$ according to the sort.

Step4. Gather the labels of the K nearest neighbors $D_K=\{y_1',y_2',\cdots,y_K'\}$, there are $K_j, j=1,2,\cdots,c$ neighbors which respectively belong to c different subset C_j. The degree of the membership of x in the j-th class is $u_j(x)=\frac{\sum_{i=1}^{K}u_{ij}w_i}{\sum_{i=1}^{K}w_i}$, where w_i represents the importance of y_i', if $j_{\max}$ satisfies $u_{j_{\max}}(x)=\max\{u_j(x)\,|\,j=1,2,\cdots,c\}$, i.e., $j_{\max}=\arg\max\{u_j(x)\}$, then x is assigned to $C_{j_{\max}}$.

B. Edited FKNN

Edited K Nearest Neighbor (EKNN) algorithm was created in 1972, whose main idea is to remove the given instance if its class does not agree with majority class of its neighbors. Let $D=\{(X_1,\theta_1),(X_2,\theta_2),\cdots,(X_n,\theta_n)\}=\{C_1,C_2,\cdots,C_c\}$ be the training set, and each sample X_i has a given class label θ_i. In EKNN, θ_i of each X_i is first estimated using the KNN, and the data set D is edited by deleting (X_i,θ_i) whenever θ_i does not coincide with its estimate . Then the k-NNR is used again to estimate θ of X by using the edited data [13].

We divided D into two subsets $D=D_1\bigcup D_2$: reference set $D_1=\{Y_1,Y_2,\cdots,Y_t\}$ and testing set $D_2=\{Z_1,Z_2,\cdots,Z_m\}$, in which, $\{Y_i\in D_1, i=1,2,\cdots,t\}$ and $\{Z_i\in D_2, i=1,2,\cdots,m\}$, $n=t+m$. The reference set D_1 is used to estimate the class label of the testing set D_2 by FKNN. The samples must be deleted from the testing set whenever θ does not coincide with its estimate. Finally, FKNN is used again to estimate θ of the unlabeled x by using the edited reference set. The EFKNN is as follows:

Step1. Input the value K, and the labeled sample set D. Then divided D into two subsets $D=D_1\bigcup D_2$: reference set $D_1=\{Y_1,Y_2,\cdots,Y_t\}$ and testing set $D_2=\{Z_1,Z_2,\cdots,Z_m\}$.

Step2. Initialize the membership matrix $U=(u_{ij})_{tc}$ as (1), where $u_{ij}=u_j(Y_i)$ denotes the degree of sample vector Y_i in j-th class.

Step3. For the given testing sample $Z\in D_2$, calculate the distance between n-dimensional testing sample Z and the all labeled samples Y_i, i.e., for any $i\in\{1,2,\cdots,t\}$, $d(Z,Y_i)=|Z-Y_i|$.

Step4. According to Step3, Step4, Step5 in the section II.(A).

Step5. Delete Z from testing set D_2 if the estimated category $j_{\max}$ is different from the class label θ of Z, otherwise reserve Z in testing set D_2.

Step6. Do Step3, Step4, Step5 in this section until all the samples in D_2 are estimated. Then the updated testing set is the edited set that we are looking for.

From the above algorithm, we see that through the edited technique testing set can be updated continuously.

III. The Edited FKNN Classifier Based on the Threshold

In this section, the novel parts of Edited FKNN based on the threshold is described. It is the extension of EFKNN. In EFKNN, membership value was used to assign a value to every sample in each class, which indicates the degree to which the element belongs to a fuzzy set. Next, we estimate the category of the sample according to the maximum membership value. However, in that process, some samples may be in the boundary of multiple class, and its maximum membership value may be small. From the classification view of point, these samples are usually unreliable or useless which should be deleted from the testing set. If we still follow EFKNN technique, i.e., follow the rule of the maximum membership, some boundary samples will be reserved in the testing set. That isn't what we wanted and isn't the desired result.

According to the algorithm in section 2.2, delete Z from testing set D_2 if the estimated category $j_{\max}$ is different from the class label θ of Z and $u_{j_{\max}}(x)=\max\{u_j(x)\,|\,j=1,2,\cdots,c\}\geq\alpha$, otherwise reserve Z in testing set D_2. Do Step3, Step4, Step5 in this section until all the samples in D_2 are estimated. Then the updated testing set is the edited set that we are looking for. The algorithm is summarized as follows:

Step1. Input the value of nearest neighbors K, and the labeled sample set $D=\{(X_1,\theta_1),(X_2,\theta_2),\cdots,(X_n,\theta_n)\}=\{C_1,C_2,\cdots,C_c\}$. Then divided D into two subsets $D=D_1\bigcup D_2$: reference set $D_1=\{Y_1,Y_2,\cdots,Y_t\}$ and testing set $D_2=\{Z_1,Z_2,\cdots,Z_m\}$.

Step2. Initialize the membership matrix $U=(u_{ij})_{tc}$ as (1), where $u_{ij}=u_j(Y_i)$ denotes the degree of sample vector Y_i in j-th class.

Step3. For the given testing sample $Z\in D_2$, calculate the distance between n-dimensional testing sample Z and the all labeled samples Y_i , i.e., for any $i\in\{1,2,\cdots,t\}$, $d(Z,Y_i)=|Z-Y_i|$.

Step4. Do Step3, Step4, Step5 according to algorithm in the part B of section II.

Step5. Delete Z from testing set D_2 if the estimated category $j_{\max}$ is different from the class label θ of Z, otherwise reserve Z in testing set D_2.

Step6. Do Step3, Step4, Step5 in this section until all the samples in D_2 are estimated. Then the updated testing set is the edited set that we are looking for.

We use reference set to classify the testing set by FKNN, in which we not only select samples classified correctly but also consider the maximum membership value. Threshold value α is used to control the size of the maximum membership value. Through the use of threshold value some boundary samples or noise samples can be taken out of reference set. On the whole, the updated reference set is more "clean", and each labeled sample in reference set can accurately represent its class. For the edited technique, the use of threshold value must be more reasonable and suitable. In the next section, experiment will demonstrated its superiority.

IV. Experiment

The classical EFKNN and the proposed EFKNN respectively are used to classify the same testing set in the experiment. We desire to show the proposed method by the comparison experiment.

A. Date set

In this section, we will evaluate the performance of EFKNN classifier based on threshold value using two different synthetic data. A two(three)-dimensional synthetic data set DA(DB) for a three(four)-class problem is generated as follows. Each class has 2000(4000) patterns which were independent and identically distributed (i.i.d), drawn from a normal distribution having mean as $(0,0),(3,0),(1.5,3)$ $((0,0,0),(3,0,0),(0,3,0),(0,0,3))$, and the same covariance identity matrix as $I_{2\times2}$ ($I_{3\times3}$) (see details in Table.1).

TABLE I.
EXPERIMENT DATA

	DA			DB			
Samples number	6000			16000			
Class number	3			4			
Attribute number	2			3			
Class	C1	C2	C3	C1	C2	C3	C4
Testing number	DATE(3000)			DATE(8000)			
	1000	1000	1000	2000	2000	2000	2000
Training number	DATR(3000)			DATR(8000)			
	1000	1000	1000	2000	2000	2000	2000

In the experiment, 3000 samples in DA are used to be the training set DATR, and the other 3000 samples are used to be the testing set DATE. Likewise, 8000 samples in DB are used to be the training set DBTR, and the other 8000 samples are used to be the testing set DBTE. In the training set DATR(DBTR), 1500(4000) samples are used to be the reference set DATRRE(DBTRRE), and the other 1500(4000) samples are used to be the testing set DATRTE(DBTRTE), i.e., $DATR=DATRRE\bigcup DATRTE$ ($DBTR=DBTRRE\bigcup DBTRTE$) (note that the testing set DATE(DBTE) is different from the testing set DATRTE(DBTRTE), DATRTE(DBTRTE) is used to edit samples in training procedure, while DATE(DBTE) is used to be the testing samples for getting recognition rate). (see details in Table. 1, Table. 2).

B. Ddiscussion

We respectively use edited testing samples and the modified edited testing samples to classify the same testing samples DATE(DBTE) so as to compare their recognition rate. The followingTable.3, Table.4, Figure.1 show two algorithms' recognition rate for the same data on adjusted values of neighbors k, where k from 1 to 10.

We first edit testing set DATRTE(DBTRTE) by EFKNN and EFKNN based on the threshold. In the experiment, the threshold α that we assign is 0.5 and 0.6. Then the edited testing set DATRTE(DBTRTE) was used to classify the testing set DATE(DBTE) by FKNN. Their respective retention rate and recognition results are shown in Table.3, Table.4.

From the Figure.1, we see that the recognition rate of EFKNN based on the threshold is higher than EFKNN, where K from 1 to 10. As the threshold value α increase, the requirement of edit is been advanced. If its class does not agree with majority class of its neighbors

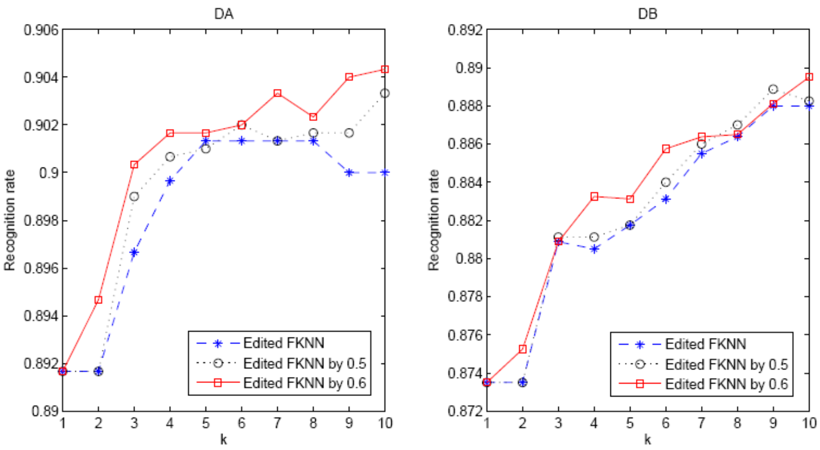

Figure 1. the comparsion of three different method on two different dataset.

and its size of membership value is not big enough, i.e., the degree of membership for the sample of the estimated class must big enough such as α , the sample will be deleted from the testing set. That is, those who remained absolutely belong to certain class. It is clearly that the improved method does eliminate a lot of boundary samples and unreliable samples. The experimental results show the proposed method outperforms the original methods with respect to each K .

V. Conclusion

In this paper, we have presented an algorithm based on the view that the threshold value is used to control the size of the maximum membership value in EFKNN, i.e., the maximum membership should be big enough. Through the use of threshold value some unreliable samples literally taken out from the training set. As seen from the experiment, the method based on the threshold value is really more reasonable and suitable.

Acknowledgment

The authors would like to thank the research fund of sichuan key laboratory of intelligent network information processing (SGXZD1002-10) and key laboratory of the radio signals intelligent processing (Xihua university) (XZD0818-09).

References

[1] Brighton, H., Mellish, C., "Advances in instance selection for instancebased learning algorithms," *Data Mining and Knowledge Discover*, vol. 6(2) , pp. 158-172, 2002.

[2] Jianmei Cheng, Li Yan, Zheng Pei, Chao Zhang, "A combined method to deal with uncertain data in fuzzy k-nearest neighbor classifier," *Proceedings of the 9th International FLINS Conference*, pp. 282-287, 2010.

[3] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*. vol.12, pp. 21-27, 1967.

[4] Wilson D. L., "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 2(3), pp. 408-421 , 1972.

[5] Garc´ıa, S., Ram´on Cano, J., Herrera, F., "A memetic algorithm for evolutionary prototype selection: A scaling up approach," *Pattern Recognition*, vol. 41(8), pp. 2693-2709, 2008.

[6] Guan, D., Yuan, W. Lee, Y.K., Lee, S.Y., "Nearest neighbor editing aided by unlabeled data," *Information Sciences*, vol. 179(13), pp. 2273 -2282, 2009.

[7] Ludmila I. Kuncheva, "Editing for the k-nearest neighbors rule by a genetic algorithm," *Pattern Recognition Letters*, vol. 16(8), pp. 809-814, 1995.

[8] Jankowski, N., Data regularization. In: L. Rutkowski and R. Tadeusiewicz (editors), *Neural Networks and Soft Computing*, pp. 209-214, 2000.

[9] N. Jankowski, M. Grochowski., "Comparison of instances selection algorithms: results and comments," *Artificial Intelligence and Soft Computing*, Springer, pp. 580-585, 2004.

[10] Keller, J.M., Gray, M.R., Givens, J.A., "Fuzzy k−nearest neighbor algorithm," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 15(4), pp. 580-584, 1985.

[11] L.A. Zadeh. "Fuzzy sets," *Information and Control*. Vol. 8 pp. 338-353, 1965.

[12] W. Hung, M. Yang, D. Chen. "Bootstrapping approach to feature-weight selection in fuzzy c-means algorithms with an application in color image segmentation," *Pattern Recognition Letters*, vol. 29, 1317-1325, 2008.

[13] Yang, M.S., Chen, C.H., "On the edited fuzzy k−nearest neighbor rule," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 28(3), pp. 461-466, 1998.

[14] S´anchez, J.S., Pla, F., Ferri, F.J., "Prototype selection for the nearest neighbour rule through proximity graphs," *Pattern Recognition Letters*, vol. 18(6), pp. 507-513, 1997.

[15] Ruiqin Chang, Zheng Pei, Chao Zhang, "A modified editing k-nearest neighbor rule," *Journal of Computers*, vol. 6(7), pp. 1493-1500, 2011.

[16] H. Yan, "Prototype optimization for nearest neighbor classifiers using a two-layer perception," *Pattern Recognition*, vol. 26, pp. 317-324, 1993.

Chao Zhang was born in Hubei Province, China, in 1986. He received the M.S. degree from Xihua University, Chengdu, in 2011, both in mathematics. He is currently an assistant with the Department of Road Traffic Management, Sichuan Police College. His research interests include pattern recognition, intelligent transportation systems, and information theories.

Jian-mei Cheng is currently an assistant with the Department of Road Traffic Management, Sichuan Police College. She received the M.S. degree in mathematics from Xihua University. Her research interests include pattern recognition, intelligent transportation systems, and information theories.

Liang-zhong Yi was born in Sichuan Province, China. He received the PH.D. degree from Southwest Jiaotong university, Chengdu. His research interests include applied mathematics, intelligent information processing, logical reasoning.

TABLE II.
TWO SUBSET:REFERENCE SET AND TESTING SET

class		C_1	C_2	C_3		C_1	C_2	C_3	C_4
Training number	DATR(3000)	1000	1000	1000	DATR(8000)	2000	2000	2000	2000
Training number	DATRRE(1500)	500	500	500	DBTRRE(4000)	1000	1000	1000	1000
	DATRTE(1500)	500	500	500	DBTRTE(4000)	1000	1000	1000	1000

TABLE IV.
THE REMAINING RATE AND RECOGNITION RATE IN ORIGINAL METHOD

the retention rate										
	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
DA	1287	1287	1300	1311	1320	1322	1323	1327	1333	1334
DB	3397	3397	3475	3507	3516	3530	3533	3551	3546	3551
the recognition rate										
DA	2675	2675	2690	2699	2704	2704	2704	2704	2700	2700
DB	6988	6988	7047	7044	7054	7065	7084	7091	7104	7104

TABLE III.
THE REMAINING RATE AND RECOGNITION RATE IN PROPOSED METHOD

the retention rate											
		k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
DA	α_1	1287	1287	1299	1306	1315	1318	1321	1325	1327	1327
	α_2	1287	1261	1278	1281	1283	1290	1292	1295	1293	1296
DB	α_1	3397	3397	3464	3486	3496	3502	3506	3516	3512	3511
	α_2	3397	3239	3356	3368	3372	3380	3373	3379	3379	3387
the recognition rate											
		k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
α_1	DA	2675	2675	2697	2702	2703	2706	2704	2705	2705	2710
	DB	6988	6988	7049	7049	7054	7072	7088	7096	7111	7106
α_2	DA	2675	2684	2701	2705	2705	2706	2710	2707	2712	2713
	DB	6988	7002	7047	7066	7065	7086	7091	7092	7105	7116