International Journal of Security and Its Applications
Vol.9, No.2 (2015), pp.149-158
http://dx.doi.org/10.14257/ijsia.2015.9.2.14

Similarity Distance Noise Reduction of Entropy Based on Lifting KNN Classification Performance

Liu Jin-sheng, Guoxi Sun, Qinghua Zhang and He jun*

College of computer and electronic information, Guangdong University of Petrochemical Technology, Mao-Ming City, Guangdong Province, China 525000
450285590@qq.com; hejun_723@126.com

Abstract

To overcome the drawback of KNN algorithms based on distance measure which did not measure the contributions for each feature accurately. In this paper, a K-Nearest Neighbor (KNN) de-noise method based on likelihood distance entropy is proposed. The relations of feature parameters are used to measure their contributions for de-noise energy, then according to the contributions for each feature leading de-noise of the feature parameters. In order to compare the performance of these relative methods, the Letter corpora and Pima Indians Diabetes data-base are employ to carry out the experiments, the experiment results show that comparing with the other de-noise methods mentioned in this paper, this proposed method have a better ability for de-noise.

Keywords: *K-Nearest Neighbor; likelihood distance entropy; feature parameter contribution; de-noise*

1. Introduction

KNN (K-Nearest Neighbor) Algorithm, is a simple and effective non-parametric method, can find the nearest K neighbors according to its distance measure, and then on the base of decision attributes to determine the type by voting for a given unclassified samples [1]. KNN was widely used in the fields of pattern recognition, such as text classification, intrusion detection, fault analysis and image recognition and obtained some achievement.

In the early research, many researchers paid their attentions on the degree of adaptation for KNN neighborhood area conditions and the parameter K screening mainly. With the fusion mechanism of characteristic transform and pattern recognition has sprung up, and is becoming more and more mature, then lots of scholars began to be attracted in focusing on the relationship between characteristic transform and similarity distance measure. KNN regarded as the mainly research object, a method, determining range of the confidence interval for the similarity distance measure of two parameters according to analyze the differences between the two parameters probability distribution entropy, was reported in the literature [5]. Then based on achievement of [5], to modify the error of traditional method for calculating the distance, a method for obtaining the confidence interval of classes was proposed in [7]. To solve the trouble of the entities matching, a method for entities matching based on attribute information entropy of characteristic parameters was published in [6]. A concept for entropy relevance differences was introduced in [8], in the published, entropy characteristic transform index of characteristic parameters was defined as relevance degree of classes, to weigh the classified affects of characteristic parameters, thereby to establish a simple intrinsic association between the absolute distance metric and category characteristics.

From what has been discussed above, many research results mainly concentrated in the process of entropy feature transform processing redundancy feature, research on the

Corresponding author Tel: 86+06682923232; E_mail:hejun_723@126.com

ISSN: 1738-9976 IJSIA

construction fusion point between characteristic information and distance mechanism. But most of them ignore the characteristic parameters for certain categories of their own defects confused noise features. Generalization of the classification performance will be affected obviously by a lot of the correlation calculation of noise characteristics, or two pretreatment process of characteristic entropy transform index, For KNN algorithm, based on entropy characteristic transform index, a new sample similarity measurement mechanism is putted forward according to analyzing representation of category characteristic relevance difference between entropy characteristic transform indexes, UCI standard test data set Letter, Pima Indians Diabetes, and the practical application of intrusion detection data set KDD CUP '99 show that noise reduction effect based on optimizing similarity distance used entropy is obvious, classification performance is high.

2. Theoretical Background

2.1. Similarity Distance Noise Reduction Entropy

Definition 1: Assumed that the sample set Z, which has T classes defined as $C_1, C_2, \cdots, C_T$. Any given sample X_i which has $n-1$ condition attributes, each attribute condition have n different characteristic parameters such as $\{X_{i1}, X_{i2}, \cdots, X_{ij}, \cdots, X_{in}\}$, $\left|X_{ij}^k\right|$ used to descript the number of instances belonged to the class C_k.The entropy characteristic indexes for class C_k is defined as:

$$e_{ij}^k = -\frac{\left|x_{ij}^k\right|}{\left|x_{ij}\right|} \cdot \ln \frac{\left|x_{ij}^k\right|}{\left|x_{ij}\right|} \tag{1}$$

Eq (1) has the following properties:

(1) If $\left|X_{ij}^k\right| = X_{ij}$ and $e_{ij}^k = 0$, then X_{ij} belongs to Class C_k; If $\left|X_{ij}^k\right| = 0$ and $e_{ij}^k = 0$,then X_{ij} belong to other classes, the smaller of e_{ij}^k value, the higher filter degree of redundancy of X_{ij} to the characteristics of class C_k, which can effect the classification performance.

(2) If $\left|X_{ij}^k\right| \neq X_{ij}$ and $e_{ij}^k \neq 0$,then assumed X_{ij}^1 belongs to class C_1, denoted as ϕ_1, X_{ij}^2 belongs to class C_2,denoted as ϕ_2,…, X_{ij}^T belong to class C_T,denoted as ϕ_T, and then overall filter degree of redundancy of Eq (1) to class's T obtained a minimum value 0.25, the overall distinguish effect entropy characteristics index for T class given as:

$$e_{ij} = \sum_{e_{ij}^i \in \phi_k}^{T} e_{ij}^i = 0.25 + c \tag{2}$$

where c is a constant.

Proof: Let $p_{ij}^k = \frac{\left|x_{ij}^k\right|}{\left|x_{ij}\right|}, 0 \le p_{ij} \le 1, \quad k = 1 \ldots T$

$$\begin{aligned}\sum_{e_{ij}^i \in \phi_k}^{T} e_{ij}^i &= -\int_0^1 p_{ij}^k \ln p_{ij}^k d\left(p_{ij}^k\right) \\ &= -\frac{1}{2}\int_0^1 \ln p_{ij}^k d\left[\left(p_{ij}^k\right)^2\right] \\ &= -\frac{1}{2}\left(p_{ij}^k\right)^2 \ln p_{ij}^k - \frac{1}{4}\left(p_{ij}^k\right)^2\Big|_0^1 \\ &= 0.25 + c\end{aligned} \tag{3}$$

Definition 2: Set x_{ij} and x_{tj} as the characteristic parameter of given sample X_i and X_t, respectively. Under any condition attribute of j, them belonged to the class C_i and C_t, if $x_{ij} \neq x_{tj}$ and $i \neq t$,then the between class similarity of x_{ij}, x_{tj} can defined as:

$$d\left(x_{ij}, x_{tj}\right) = \begin{cases} \dfrac{\left(e_{ij}^i\right)^2}{e_{tj}^t} + \dfrac{\left(e_{tj}^t\right)^2}{e_{ij}^i}, & e_{ij}^i \neq 0,\ e_{tj}^t \neq 0,\ C_i \neq C_t \\ 0, & e_{ij}^i = 0 \text{或} e_{tj}^t = 0 \end{cases} \tag{4}$$

Eq (4), in the first case presents that there exists uncertainty for x_{ij} to X_i belonged to class C_i, to exclude the unclearly category information obtained from x_{ij} and x_{tj}, we employed the mutual category differences representation of entropy characteristic transform index method to the character of x_{ij} and x_{tj} ,respectively. Analyzing the verification processing of Eq (1), when the entropy characteristic transform index do not belonged to certain class, and for two entropy characteristic transform indexes under the same condition attributes, the ratio of $\left(e_{ij}^i\right)^2$ to e_{tj}^t , $\left(e_{tj}^t\right)^2$ to e_{ij}^i presents the magnitude of reducing the class noise, the higher amplitude values, the lower category noise level of entropy characteristic transform for decision attribute parameter values in the denominator. For more detailed, if x_{ij} to the C_i noise reduction degree e_{ij}^i is a reference, then $\frac{\left(e_{ij}^i\right)^2}{e_{tj}^t}$ presents the eliminating differences amount of the uncertainty degree of the characteristic class C_i and C_t appeared with x_{ij} compared for x_{tj}, if x_{tj} to C_t noise reduction degree e_{tj}^t is a reference, then $\frac{\left(e_{tj}^t\right)^2}{e_{ij}^i}$ presents the eliminating differences amount of the uncertainty degree of the characteristic class C_i and C_t appeared with x_{tj} compared with x_{ij}; If $\frac{\left(e_{ij}^i\right)^2}{e_{tj}^t} > \frac{\left(e_{tj}^t\right)^2}{e_{ij}^i}$,then $d(X_{ij}, X_{tj})$ is nearly to the clustering center of class C_t ,if $\frac{\left(e_{ij}^i\right)^2}{e_{tj}^t} < \frac{\left(e_{tj}^t\right)^2}{e_{ij}^i}$,then $d(X_{ij}, X_{tj})$ is near to the clustering center of class C_t. The second case x_{ij} to X_i belonged to class C_i do not exist uncertainty, or x_{tj} to X_j belonged to class C_t do not exist uncertainty . Therefore the Eq (5) presents under the same condition attribute j, the characteristic parameters x_{ij}, x_{tj} is the

category characteristics offset of C_t and C_t . So the category distance for X_i and X_t given as follows:

$$D(X_i, X_t) = \sum_{i=1,t=1}^{m} d(x_{ij}, x_{tj}) \tag{5}$$

2.2. Implementation

Step1: According to range of sample type index, retrieves and sums up the number of all the characteristic parameters appeared in the training data, and for any decision attribute parameters of given instance numbers, the single-category entropy characteristic transform index of the original characteristic parameters is calculated by Eq (1) .

Step2: For the detecting samples set $Y = \{y_{ij}\}$, defines the entropy characteristic transform factors of y_{ij} is $|y_{ij}|$ and $|y_{ij}^k|$, the entropy characteristic index of y_{ij} to T overall complex categories is calculated by Eq(6), which given as follow:

$$e_{ij}(Y) = -\sum_{k=1}^{T} \frac{|y_{ij}^k|}{|y_{ij}|} \cdot \ln \frac{|y_{ij}^k|}{|y_{ij}|} \tag{6}$$

The smaller value of $e_{ij}(Y)$, the greater effect for judge the y_{ij} to T overall complex categories.

Step3: According to Eq(4) and Eq(5), calculate the similarity between Y and the entropy characteristic index of training samples:

$$D(X_i, Y_i) = \sum_{\substack{i=1\cdots m \\ j=1\cdots n-1}} d(x_{ij}, y_{ij}) = \sum_{\substack{i=1\cdots m \\ j=1\cdots n-1 \\ k=1\cdots T}} \frac{|e_{ij}^k|^2}{e_{ij}(Y)} + \frac{[e_{ij}(Y)]^2}{|e_{ij}^k|} \tag{7}$$

Step4: Combining with the K-Nearest Neighbor rule, use the above steps to define the value of K and extracts K samples the most similar to Y.

Step5: Counting the N-decision attribute parameters in the Kth samples, it can define Y type which parameters appear most.

2.3. Accuracy Analysis

For any entropy characteristic transform index of the sample X_i and X_t in the Eq (4) , when $e_{tj}^i = 0$ or $e_{tj}^t = 0$, the effective degree provided by characteristic parameters in the sample classification processing is a determined value, then it only to test the measurement method of Eq (4) effect in the Eq (5). From the Euclidean distance view, the same characteristic parameters metrics, the Eq (5) has more advantage to make the distance between categories as large as possible.

Proof: Euclidean distance metric, $\because e_{ij}^k > 0, e_{tj}^r > 0, C_k \neq C_r$

$$\therefore \begin{cases} \dfrac{(e_{ij}^k)^2}{e_{tj}^r} + e_{tj}^r \geq 2e_{tj}^r, & e_{ij}^k > e_{tj}^r \\ \dfrac{(e_{tj}^r)^2}{e_{ij}^k} + e_{ij}^k \geq 2e_{ij}^k, & e_{ij}^k < e_{tj}^r \end{cases}, \tag{8}$$

Only if $e_{ij}^{i} = e_{tj}^{t}$, the equality was obtained.

$$\therefore \frac{\left(e_{ij}^{i}\right)^2}{e_{tj}^{t}} + e_{tj}^{t} + \frac{\left(e_{tj}^{t}\right)^2}{e_{ij}^{i}} + e_{ij}^{i} \geq 2e_{tj}^{t} + 2e_{ij}^{i} \text{ ,that } d(x_{ij}, x_{tj}) \geq e_{ij}^{i} + e_{tj}^{t}$$

$$\because e_{ij}^{i} + e_{tj}^{t} \geq \sqrt{\left(e_{ij}^{i} - e_{tj}^{t}\right)^2}, \therefore d\left(x_{ij}, x_{tj}\right) \geq \sqrt{\left(e_{ij}^{i} - e_{tj}^{t}\right)^2},$$

$$\therefore D\left(X_i, X_t\right) \geq \sqrt{\sum_{\substack{i=1 \\ t=1}}^{m} \left(e_{ij}^{k} - e_{tj}^{r}\right)^2}, k = 1, \cdots, T, r = 1, \cdots, T$$

For the same characteristic parameters metrics, if there are the number of the same characteristic value for X_i, X_t.

$$\because a \leq n-1$$

$$\therefore \sum_{j=1}^{a} e_{ij}^{i} \leq \sum_{j=1}^{n} \left(e_{ij}^{i} + e_{tj}^{t}\right) \leq d\left(x_{ij}, x_{tj}\right)$$

That means Eq (4) is more accurate to measure the distance between categories.

3. Experiment Setup and Analysis

3.1. Experiment Setup

The Pima Indians Diabetes(PID) and Letter provided by UCI Machine Learning The Pima Indians Diabetes dataset provided by UCI Machine Learning Repository(the number of categories is 2, number of condition attributes is 8, sample size is 768) used to the proposed experiment, any 10 samples of this database selected, X1 to X8 used as the training data set, X9 and X10 used as the test samples, A-H is condition attributes, I is decision attribute, Sample data are shown in Table I, Table I (a) is the original sample data, the entropy characteristic indexes given in table I (b) calculated by Eq(1) . X9 and X10 are calculated as overall complex categories entropy characteristic index. Table II is the process and result used five kinds algorithm to judge the category of X9 and X10.

Table I. Pima Indians Diabetes' Training and Testing Dataset

(a) Original sample

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
X_1	5	189	64	33	325	31	0.6	29	1
X_2	5	158	70	0	0	30	0.2	63	1
X_3	5	103	108	37	0	39	0.5	67	0
X_4	4	146	78	0	0	39	0.5	67	1
X_5	4	147	74	25	293	35	0.4	30	0
X_6	5	99	54	28	83	34	0.5	30	0
X_7	6	124	72	0	0	28	0.4	29	1
X_8	0	101	64	17	0	21	0.3	21	0
X_9	4	156	75	0	0	48	0.2	32	0
X_{10}	6	87	80	0	0	23	0.1	32	0

(b) Entropy characteristic index

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
X_1	0.35	0	0.35	0	0	0	0	0
X_2	0.22	0	0	0.35	0.31	0	0	0
X_3	0.22	0	0	0	0.31	0.35	0	0
X_4	0.35	0	0	0	0.37	0.35	0.35	0
X_5	0.35	0	0	0	0	0	0.35	0
X_6	0.22	0	0	0.35	0	0	0.35	0
X_7	0	0	0	0	0.37	0	0.35	0
X_8	0	0	0.35	00	0.31	0	0	0
X_9	0.7	0		0.7	0.68	0	0	0
X_{10}	0	0		0.7	0.68	0	0	0

3.2. Result Analysis

Table II is the process and result used five kinds algorithm to judge the category of X9 and X10.

Table II. Five Kinds Distance Metric KNN Contrast

Distance metric algorithm	Distance between X1-X8 and X9	X9 Neighbors samples and results(K=3,4,5)	Distance between X1-X8 and X10	X10 Neighbors samples and results(K=3,4,5)
Noise reduction of entropy	1.59,5.57,3.98,3.04 1.59,2.34,3.04,1.65	{X1,X5,X8,X6,X4} Definitioned as 0 class	0,2.304,0.718,1.459,0,0 3.045,1.647	{X1,X5,X6,X3,X4} Definitioned as 0 class
entropy relevance differences	1.42,2.75,3.73,2.6 1.57,1.63,2.09,2.83	{X1,X5,X6,X7,X4} Definitioned as 1 class	2,1.69,1.68,1.69,1.69 2,1.69,1.78	{X3,X8,X2,X4,X5} K=3,5,0 class K=4 undefined
category reliability	T(0,X9)=0.3175 T(1,X9)=0.2311	{X5,X2,X7,X4,X3,X8} Defined as 1 class	T(0,X10)=0.273 T(1,X10)=0.414	{X3,X8,X2,X4,X5} Definitioned as 0 class
the number statistics of the same characteristic parameters	8,3,1,3,1,8,2,1	{X3,X5,X8,X7,X2} K=3, 1 class K=4,undefined K=5,1 class	8,2,1,2,8,8,2,1	{X3,X8,X2,X4,X7} K=3 0 class K=4 undefined K=5 1 class
Euclidean distance	328.96,19.29,73.15 14.32,294.51,107.56 ,38.2864.78	{X4,X2,X7,X8,X3} Defined as 1 class	342.9,72.01,51.63,61.0230 0.42,92.79,38.11,27.9	{X8,X7,X3,X4,X2} K=3 1 class K=4 undefined K=5 0 class

From the table II, employed the proposed method, the X9 and X10 were classified correctly. However, only when meet the condition $k=3$ and $k=5$, the entropy relevance differences distance metric method can classify the X10 well. When $k=4$, the same number of samples neighbor point's only two classes. Considering the reliability calculation for the form of entropy for each same parameter, X9 were not judged correctly. While the traditional metrics for the number statistics of the same characteristic parameters method were employed, only in the case $k=3$ and $k=5$, X9 and X10 can be classified well. From the above example, for the same characteristic parameters category reliability and the number statistics of the same characteristic parameters, they ignore the role related to the categories of characteristic parameters under different conditions, it makes the samples distance is much smaller than the actual value, the larger the error, especially when the sample distribution is not suitable for conditions which related conditional probability of category in the neighborhood area, the risk of wrong and refusing classification are very big. Entropy relevance differences distance metric although pays attention to the category relevance of characteristic parameters under different conditions attribute, but for one characteristic value totally belonging to one class, it exaggerates the characteristic parameter small changes, even confuses the classification role of characteristic noise, it will make the sample distance much greater than the actual value. The disadvantages for the Euclidean distance are obviously, it ignores the differences of sample condition attribute; it cannot exclude the noise of correlation between parameters. This algorithm's main consideration, under certain condition attribute, the role for any two entropy characteristic transform index can cancel each other out, which can ensure contribution size for characteristic parameters to category features and

lower the effect on the vague characteristic parameters. Its accuracy is higher than the other four algorithms.

The performance of all the classification method mentioned in this proposed given in Table III.

Table III. The Performance of These Mentioned Methods

Dataset	Random range of K value	Category average accuracy rate(%)/ Algorithm time(s)				
		noise reduction of entropy	entropy relevance differences	category reliability	the number statistics of the same characteristic parameters	Euclidean distance
Letter	[18,35]	89.91/25	86.21/42	85.27/37	68.96/14	72.14/41
		91.58/28	90.27/45	86.69/38	80.25/15	81.36/56
	[25,48]	89.86/36	86.52/48	84.87/42	78.47/18	80.27/69
	[45,72]	87.63/42	85.24/53	83.91/45	75.69/20	86.26/78
		90.17/45	87.69/49	84.26/49	84.25/22	79.74/83
APID	[12,25]	89.69/14	86.91/25	85.67/21	72.13/7	76.98/31
	[32,52]	89.21/18	87.36/28	83.09/25	74.62/9	80.21/41
		90.25/29	88.74/41	85.71/36	69.89/13	75.32/55
	[45,63]	91.87/24	87.06/38	86.92/31	75.21/11	80.11/45
		88.57/27	86.12/43	84.12/37	73.24/10	79.34/46

It can be seen from the Table III, the first three algorithms overall classification performance higher than the traditional KNN algorithm after entropy characteristic index conversion. This proposed method achieved a high classification performance on APID small sample data sets, also in the classification accuracy of a large sample letter data set. This shows that the algorithm is less affected by the distribution of training samples; it can avoid the risk of misclassification of KNN in the conditional probability of category in the neighborhood area which is not the same. Meanwhile, classification accuracy remained relatively stable and the accuracy rate of maximum difference does not exceed 2% under the K values in the adaptation range greater than 50%. The other algorithms in the same K value range, classification accuracy steepness angle greater than this algorithm, the K value is more sensitive, especially for the number statistics of the same characteristic parameters KNN and Euclidean distance KNN algorithm, the accuracy of the maximum difference of nearly 15%. On the efficiency of classification, since this algorithm training set use entropy characteristic transform way of single category and the noise reduction is happened in the calculation process without secondary data preprocessing, therefore more efficient than other KNN algorithms of overall complex categories entropy characteristic transform.

In order to illustrate the performance of the mentioned methods, in this proposed we use the KDD CUP'99 to do experiments, the following performance figure obtained.

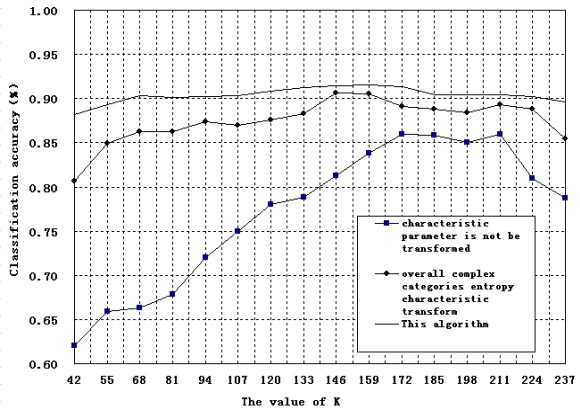

Figure 1. The Trend of the Classification Accuracy with K Increasing

KDD CUP′99 provided 10% data set includes four categories of network intrusion. In the experiments, 17854 records were randomly selected as the training set from the 10% of the data set, Normal is 4698,DOS is 9014,Probe is 42,U2R is 2057,R2L is 2043; 6854 records was selected as the test sample set, Normal is 1784, DOS is 3861,Probe is 79,U2R is 109 and R2L is 1021.The classification average of accuracy from the entropy relevance differences KNN and the category reliability KNN(overall complex categories entropy characteristic transform), the number statistics of the same characteristic parameters KNN and Euclidean distance KNN(characteristic parameter is not be transformed) calculated contrast with the classification average of entropy noise optimization KNN(single-category entropy characteristic transform). The results are shown in Figure 1. From the view of the overall accuracy, the entropy characteristic transform KNN algorithm is performed better than the traditional KNN algorithm. This algorithm is performed better than the entropy characteristic transform KNN. This shows that the algorithm has high stability and accuracy on the practical application of intrusion detection, in the process of distance calculation , it fully consider the difference of every characteristic parameter reflected the characteristic information, the de-noise results obviously and the accuracy is higher than the overall average two types of overall complex categories entropy characteristic transform algorithm, especially serious imbalance in the Probe class, the category noise mutual interference serious, the classification accuracy can be close to 90% .

4. Conclusion

Similarity distance measure is the key technology to improve the KNN algorithm classification generalization. According to analyzing the intrinsic properties of entropy characteristic transform index of theistic parameters, the similarity distance express is represented surrounding the category characteristic relativity differences between entropy characteristic transform indexes, greatly to minimize the category noise in the distance calculation process. According to theoretical analysis and UCI, KDD CUP'99's dataset classification experiment proved the algorithm is rationality, effectiveness and practicality.

Acknowledgements

This work was supported by Maoming industrial research project (2014009, 2012B01009) and Open fund of Guangdong key laboratories on petrochemical equipment fault diagnosis (714022), a major research project of Guangdong Provincial Department of Education (631054).

References

[1] J.-w. Han and M. Kamber, "Data Mining Concepts and Technology [M]", F. Ming and X. Meng, translated. Beijing: Mechanical Industry Press, (**2004**).
[2] P. Michalis, B. Francesco and G. Aristides, "K-Nearest neighbors in uncertain graphs [J]", Proc of the VlDB Endowment, vol. 3, no. 1, (**2010**), pp. 997-1008.
[3] A. K. Ghosh, P. Chaudhuri and C. A. Murthy, "Multiscale classification using nearest neighbor density estimates [J]", IEEE Transactions on Systems, man, and Cybernetics-part b:cybernetics, vol. 36, no. 5, (**2006**), pp. 1139-1148.
[4] Y. Wang, Z. O. Wang and S. Bai, "An Improved KNN Algorithm Applied to Text Categorization [J]", Journal of Chinese Information Processing, vol. 21, no. 3, (**2007**), pp. 76-81.
[5] S.-W. Ho and R. W. Yeung, "The Interplay between Entropy and Variational Distance [J]", IEEE Transactions on Information Theory, vol. 56, no. 12, (**2010**), pp. 5906-2929.
[6] B. H. Qiang, Z. F. Wu and J. Q. Yu, "Methodology for Entities Matching Based on Attribute Information Entropy [J]", Computer Engineering, vol. 31, no. 21, (**2005**), pp. 31-33.

[7] X. Q. Tong and Z. M. Zhou, "The KNN Improved Algorithm based on information entropy property values [J]", Computer Engineering and Applications, vol. 46, no. 3, (**2010**), pp. 115-117.
[8] J. Zhou and J. S. Liu, "KNN Algorithm Based on Feature Entropy Correlation Difference", Computer Engineering [J], vol. 37, no. 17, (**2011**), pp. 146-148.