International Journal of Database Theory and Application
Vol.8, No.2 (2015), pp. 267-274
http://dx.doi.org/10.14257/ijdta.2015.8.2.25

Dynamic Cost-Sensitive Fussy Clustering for Uncertain Data Based on the Genetic Algorithm

Yuwen Huang

Department of Computer and Information Engineering, Heze University, Heze 274015, Shandong, China
Key Laboratory of computer Information Processing, Heze University, Heze 274015, Shandong, China
hzxy_hyw@163.com

Abstract

The existing fussy clustering algorithms for uncertain data don't consider the dynamic cost and the treatment effect is lower, so this paper proposes the dynamic cost-sensitive fussy clustering approach for uncertain data based on the genetic algorithm (GADCSFA). Firstly, this paper gives the definition of dynamic cost and adjacent interval, and the uncertain attributes are disposed as the interval number. Secondly, we give the method of fuzzy c-means clustering based on the interval data, and the interval numbers of fussy clustering solution and cost space are coded by its centre and radius. At last, the dynamic fussy clustering approach for uncertain data based on the genetic algorithm is structured, which uses the genetic algorithm to search the optimal clustering centre and cost by the hybridization, the mutation and selection. The experiments show that, compared to the other fussy clustering algorithm for uncertain data, GADCSFA has higher classification accuracy and performance, and the total expenditure is lower.

Keywords: *Dynamic cost; Fussy clustering; uncertain data; Genetic algorithm*

1. Introduction

In recent years, with the development of computer technology, there is more and more uncertainty in data integration, the data extraction, the scientific data management, the multimedia applications, and knowledge application, so many agencies and organizations have amassed the huge uncertainty data. In traditional data mining, the existence and accuracy of the data are true, and the traditional mining techniques are unable to manage effectively the uncertain data, so the research for uncertain data is a hot spot. At present, the clustering and classification methods for uncertain data include mainly the support vector machine (SVM), the Bayesian theory, the decision tree etc. The paper [1] developed a new dynamic communication distance estimation method using uncertain interval data stream clustering. The paper [2] investigated different data migration types and proposed a technique to generate artificial non-stationary data which follows different migration types. The paper [3] proposed pruning methods to speed up UK-means for clustering data with uncertainty. The paper [4] gave the approximate UK-means to identify heuristically objects of boundary cases and re-assign them to better clusters. Bi and Zhang proposed the total Support vector classification (TSVC) for uncertain data [5]. Bhattacharyya put forward to structure the binary classifiers by the second-order cone programming and the chebyshev inequality [6]. Yang proposed the USVC iteration method based on the linear model [7]. Demichelis proposed the model of extended Bayesian classification based on Bayesian hierarchical model (BHM) [8]. The current uncertain data mining researches are in pursuit of high accuracy and low error rate, and assume that all uncertain costs of classification and prediction are the same, so the existing mining methods can't complete the task if they face with the different type cost in mining uncertain data. Cost-sensitive data mining considers different types cost, and it can produce the minimum cost by the optimal decision behavior. The cost-sensitive uncertain data mining provides the bridge between uncertainty theory and the social practice, and can promote the development of the data mining. Palacios proposed a genetic fuzzy classifier, which is able to extract fuzzy rules from interval or fuzzy valued data, is extended to this type of classification [9]. Fan proposed a cost-sensitive learning algorithm to train hierarchical tree classifiers for large-scale image classification application [10]. The paper [11] is to examine whether classification cost is affected both by the cost-sensitive approach and by skewed distribution of class. The paper [12] proposed two greedy wrapper forward cost-sensitive selective naive Bayes approaches. The existing uncertain data clustering don't take into account the dynamic cost factors and deal with data with low efficency, so this paper proposes a dynamic cost based on genetic algorithm for the clustering of uncertain data.

2. Dynamic Cost

Static cost mechanism uses the fixed value misclassification and test cost in the same data sets, but each cost is different in real life. For example, the cost for blood analysis is different in one hospital compared to the rest of the hospital in one area. The construction process of dynamic cost is as follows.

Step 1: Determine the application domain of data sets. For example, we will be blood test.

Step 2: Find the different approaches for tests. For example, choose different hospital to do the blood test.

Step 3: Combine with expert's experience and background knowledge, and determine the different cost in different application domain. For example, the costs of blood test are different in different hospital.

Step 4: Sort the different cost, and get an ascending sequence. For example, All test costs are sort as $\left[cost_1, cost_2, cost_3, \ldots, cost_m\right]$.

Step 5: $cost_a$ is the lower limit, and $cost_b$ is the upper limit. All $cost_x \in \left[cost_a, cost_b\right]$ are feasible.

In feasible cost space, in order to find the optimal cost and improve the correct classification rate of positive and negative samples, so this paper uses the feasible misclassification and test costs to structure the cost space, and extends G-mean for searching the optimal cost. The cost function is defined as follows.

$$f(C) = G\left(RE(\mathrm{c}), PR(\mathrm{c})\right) = \sqrt{RE(\mathrm{c}) \times PR(\mathrm{c})} \ .$$

$$RE = \frac{TP}{TP + FN} \ .$$

$$PR = \frac{TP}{TP + FP} \ .$$

c is a point of cost space, and G is G-mean function. Re is the response rate, and Pr is the precision rate.

3. Definition of Adjacent Interval

The uncertain data influences significantly on the clustering result. In recent years, the nearest neighbor rule has applied widely to estimate uncertain data in the field of pattern recognition. This paper selects the nearest samples as the estimated value of the uncertain data, and the nearest samples of the uncertain data are chosen by the similarity. $X = \{x_1, x_2, \ldots, x_n\}$ is uncertain data set, $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$, and the similarity of sample x_i and x_j is as follows.

$$S(x_i, x_j) = \sum_{x=1}^{m}\left(1\Big/(x_{ix} - x_{jx}) \times I_x\right).$$

I_x is the significance of the xth attribute. In order to calculate the uncertain attribute $\overline{x_{jb}}$, search K similar samples by calculating the similarity, and find the minimum x_{jb}^- and the maximum x_{jb}^+ of the corresponding attributes. In other word, the uncertain attribute $\overline{x_{jb}}$ is described as the interval $\left[x_{jb}^-, x_{jb}^+\right]$, and the certain attribute is described by the interval $\left[x_{jb}^-, x_{jb}^+\right]$, $x_{jw}^- = x_{jw}^- = \overline{x_{jw}}$.

4. Fuzzy C-Means Clustering Based on the Interval Data

Fuzzy C-means clustering algorithm is widely used, and Euclidean distance is selected usually as distance measure. For the interval data clustering, this paper generalizes Euclidean distance to the interval data.

$\mathrm{I}_i = \left[a_i^-, a_i^+\right] \in \mathrm{W}$, Euclidean distance of I_1 and I_2 is as follows.

$$d_1(I_1, I_2) = \sqrt{(a_1^- - a_2^1)^2 + (a_1^+ - a_2^+)^2}$$

Dataset $X = \{x_1, x_2, \ldots, x_n\}$, $x_k = (x_{k1}, x_{k2}, \ldots, x_{kp})$, $x_{kj} = \left[x_{kj}^-, x_{kj}^+\right]$, $k = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$. $\mathrm{V} = \{v_1, v_2, \ldots, v_c\}$ is the clustering prototype, $v_i = (v_{i1}, v_{i2}, \ldots, \mathrm{v}_{\mathrm{ip}})$, $v_{ij} = \left[v_{ij}^-, v_{ij}^+\right]$, $i = 1, 2, \ldots, c$, $j = 1, 2, \ldots, p$. $U = (u_{ki})_{n \times c}$ is fuzzy division matrix, and u_{ki} is the membership that sample k belongs to classification i.

The objective function of fuzzy clustering based on the separation distance is as follows.

$$F(u,v) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ki})^m \sum_{j=1}^{p}\left[(x_{kj}^- - v_{ij}^-)^2 + (x_{kj}^+ - v_{ij}^+)^2\right].$$

If $I_i = \left[a_i^-, a_i^+\right]$, $m_i = \dfrac{a_i^- + a_i^+}{2}$ is its middle, and $w_i = \dfrac{a_i^+ - a_i^-}{2}$ is its width. The distance of I_1 and I_2 based on the middle and width is as follows.

$$F(u,v) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ki})^m \sum_{j=1}^{p}\left[(M_{kj} - m_{ij})^2 + q \times (W_{kj} - v_{ij})^2\right],$$

$M_{kj} = \dfrac{x_{kj}^- + x_{kj}^+}{2}$, $W_{kj} = \dfrac{x_{kj}^+ - x_{kj}^-}{2}$, $m_{ij} = \dfrac{v_{ij}^- + v_{ij}^+}{2}$, $w_{ij} = \dfrac{v_{ij}^+ - v_{ij}^-}{2}$, $k = 1, 2, \ldots, n$, $i = 1, 2, \ldots, c$,

$j = 1, 2, \ldots, \mathrm{p}$. The iterative clustering prototype of fuzzy partition matrix is as follows.

$$u_{ki} = \left[\sum_{l=1}^{c}\left(\frac{\sum_{j=1}^{p}\left[(M_{kj} - m_{ij})^2 + q(W_{kj} - w_{ij})^2\right]}{\sum_{j=1}^{p}\left[(M_{kj} - m_{ij})^2 + q(W_{kj} - w_{lj})^2\right]}\right)\right]$$

$$v_{ij}^{-} = \frac{\sum_{k=1}^{n} (u_{ki})^{m} M_{kj}}{\sum_{k=1}^{n} (u_{ki})^{m}}$$

$$v_{ij}^{+} = \frac{\sum_{k=1}^{n} (u_{ki})^{m} W_{kj}}{\sum_{k=1}^{n} (u_{ki})^{m}}$$

5. Dynamic Cost-sensitive Fuzzy Clustering for Uncertain Data Based on the Genetic Algorithm

5.1. Encoding Scheme

The cost matrix $Cost_{m \times m} = \begin{bmatrix} c_{11}, c_{12}, \ldots, c_{1m} \\ c_{21}, c_{22}, \ldots, c_{2m} \\ \ldots \quad \ldots \quad \ldots \\ c_{m1}, c_{m2}, \ldots, c_{mm} \end{bmatrix}$, $c_{ij} = test\_cost_{ij} + mis\_cost_{ij}$, $test\_cost_{ij}$ is test cost, and mis_cost_{ij} is misclassification cost. $test\_cost_{1j} = test\_cost_{2j} = \ldots = test\_cost_{mj}, 1 \le j \le m$. The solution space of the fuzzy clustering for uncertain data is $V = \{V_1, V_2, \ldots V_c\}$, $Vi = (v_{i1}, v_{i2}, v_{i3}, \ldots, v_{id})$. If the upper and lower bounds of the interval are encoded, the condition that the upper bound is larger than the lower bound may not been met. In order to encode the interval number, encode respectively the middle and radius. This paper uses a real coding, and the chromosome is consisted by the real parameter variables as follows. Choose the real to encode each chromosome, and each chromosome is $4d \times c$ dimension.

$$c\left\{\overbrace{\begin{matrix} m_{11}, r_{11}, cm_{11}, cr_{11}, \ldots, m_{1d}, r_{1d}, cm_{1d}, cr_{1d}, \\ m_{21}, r_{21}, cm_{21}, cr_{21}, \ldots, m_{2d}, r_{2d}, cm_{2d}, cr_{2d}, \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ m_{c1}, r_{c1}, cm_{c1}, cr_{c1}, \ldots, m_{cd}, r_{cd}, cm_{cd}, cr_{cd}, \end{matrix}}^{4d}\right.$$

If the chromosome is decode, the lower bound $v_{ij}^{-} = v_{ijl} = m_{ij} - r_{ij}$, and the upper bound $v_{ij}^{+} = v_{ijh} = m_{ij} + r_{ij}$. The lower bound $c_{ijl} = cm_{ij} - cr_{ij}$, the upper bound $c_{ijh} = cm_{ij} + cr_{ij}$.

5. 2. Fitness Function

In order to evaluate the fitness of individuals, the fitness function is defined. Choose the next generation with high fitness, and the cost function is transformed as follow.

$$f(x) = G(RE(x), PR(x)) = \sqrt{RE(x) \times PR(x)}$$

After the cost is applied to a data set, $RE(x)$ and $PR(x)$ are the response and precision rate that are gotten to use the classifier by the cross validation. This paper considers the change of misclassification and test before and after clustering.

There are many costs in the clustering process, and the misclassification and test costs are common. The misclassification cost is the costs when a real category is misclassified as the wrong label, and test cost is the consumption for obtaining the test attribute value. The cost change is as follows.

$$\cos t^{chang}\left(\overline{A_i}\right) = \cos t^{change}_{misclass}\left(\mathrm{A}, \overline{A_i}\right) - C_{test}\left(\overline{A_i}\right).$$

$$\cos t^{change}_{misclass}\left(\mathrm{A}, \overline{A_i}\right) = mis\_\cos t\left(\mathrm{A}\right) - mis\_\cos t\left(\mathrm{A} \cup \overline{A_i}\right)$$

A is $\cos t^{chang}\left(\overline{A_i}\right)$ is cost change after attribute $\overline{A_i}$ is test, $\cos t^{change}_{misclass}$ is the misclassification cost change after clustering. $C_{test}\left(\overline{A_i}\right)$ is cost for testing the attribute $\overline{A_i}$. Combine the objective function of fuzzy clustering, and the fitness function is defined as follows.

$$fitness = \alpha \times F\left(u,v\right) + \beta \times f\left(x\right) + \gamma \times \cos t^{chang}\left(\overline{A_i}\right)$$

$$= \alpha \times \sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ki}\right)^{m}\sum_{j=1}^{p}\left[\left(M_{kj} - m_{ij}\right)^{2} + \theta \times \left(W_{kj} - v_{ij}\right)^{2}\right] + \beta \times \sqrt{RE\left(x\right) \times PR\left(x\right)} + \gamma \times \cos t^{change}_{misclass}\left(\mathring{\mathrm{A}}, \overline{A_i}\right) - C_{test}\left(\overline{A_i}\right)$$

α , β , γ are the regulatory factor.

5.3. Description of Dynamic Cost-sensitive Fuzzy Clustering for Uncertain Data Based on the Genetic Algorithm

The description of the dynamic cost-sensitive fuzzy clustering for uncertain data based on the genetic algorithm is as follows.

Input: Uncertain data set $X = \{X_1, X_2, ..., X_m\}$, $X_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$, the clustering number c , the population size m , the crossover probability Pc, the mutation probability Pm, the maximum iterations MAX=500, the cost matrix $Cost_{m \times m}$,$c_{ij} \in [c_{ijl}, c_{ijh}]$, the parameter α , β , θ , γ .

Output: The clustering result of $X = \{X_1, X_2, ..., X_m\}$.

Step 1: Data Initialization. Transform $X = \{x_1, x_2, ..., x_n\}$ into the interval number. If x_{iw} is a certain value, $x_{iw} = \left[x^-_{jw}, x^+_{jw}\right]$, $x^-_{jw} = x^+_{jw} = x_{jw}$. If x_{iw} is uncertain data, find the nearest attributes value, $x_{iw} = \left[x^-_{jw}, x^+_{jw}\right]$. $c_{ij} \in Cost_{m \times m}$, $c_{ij} = [c_{ijl}, c_{ijh}]$.

Step 2: Population initialization. Encode the solution and cost space, and initialize the population and the fuzzy matrix $\overline{U}$.

Step 3: Use fuzzy C-means clustering algorithm based on the interval data to build classifier M , and get the response and precision rate by 10-fold cross validation. According to fitness function, calculate the fitness function of each individual. Save the chromosome with the maximum fitness value $G_{\max}$ and classifier $M_{\max}$, and calculate the clustering center $\overline{V}^{(i)}$ for each chromosome in population.

Step 4: Selecting operation. According to the fitness of the chromosomes, select N individuals for crossover by roulette selection.

Step 5: Crossover operation. Cross the chromosomes by the uniform distribution function, and N individuals are made pairs respectively to produce the next generation.

Step 6: Mutation operation. Produce the random numbers p and q by the uniform distribution. The individual $U^{(1)} = \left(U_1^{(1)}, U_2^{(1)}, \ldots, U_n^{(1)}\right)$, $U_p^{(1)} = wU_p^{(1)} + (1 - w)U_q^{(1)}$, $w \in \left[0, 1\right]$.

Step 7: If the iterations are the maximum, and the algorithm ends. Output the optimal individual $G_{\max}$. Otherwise, turn to step 3.

Step 8: Output the clustering result of $X = \{X_1, X_2, \ldots, X_m\}$ by the classifier $M_{\max}$

6. Simulation Experiment

At present, there aren't standard uncertain data sets, and the certain data sets from standard UCI are turned into the uncertain data in the experiments. If the attributes with probability $x\%$ are uncertain, the ratio of certain attributes is $(1 - x)\%$. In order to analyze the uncertainty, the uncertain probability is increased form 10% to 50% by each 10% increase in the experiments. Use Java as the programming language, and choose Linux 3.1.2+GCC4.1.2+RAM 2.0G+Weka3.6.4 as the experiment platform. Each experiment is performed 50 times in each data set, and the results are the average.

Choose Cover type, Ecolin, Nursery and Yeast as test data, and the uncertain ratio is set as 10%. Initialize the population size m = 200, the crossover probability Pc = 0.8, the mutation probability Pm = 0.1, the maximum iterations MAX = 500. CK-means [13], P-DBSCAN [14] and EM [15] are common algorithms used for uncertain data clustering, and the clustering accuracy is $\frac{TP + TN}{TP + FP + TN + FN} \times 100\%$. The common algorithms compare with the dynamic cost-sensitive fuzzy clustering algorithm based on genetic algorithm (GADCSFA), and the clustering accuracy results are as follow.

Table 1. Clustering Accuracy for Uncertain Data

Clustering	Clustering accuracy			
algorithm	Covertype	Ecolin	Nursery	Yeast
CK-means	67.4%	70.7%	65.7%	74.4%
P-DBSCAN	72.3%	76.3%	79.3%	77.6%
EM	76.5%	85.6%	83.7%	80.4%
GADCSFA	85.1%	91.2%	89.8%	86.2%

In all experiments, choose the equal fixed value for test and misclassification cost in CK-means, P–DBSCAN and EM, and GADCSFA uses the interval number near the fixed. As can seen form table 1, GADCSFA this paper proposes has higher clustering accuracy for uncertain data than the other algorithms, so the dynamic cost is feasible.

In order to verify the dynamic efficiency of GADCSFA, take away the dynamic cost-sensitive character to get the fuzzy clustering for uncertain data based on the genetic algorithm (GAFA), and turn the dynamics cost into the static cost-sensitive fuzzy clustering based on genetic algorithm (GACSFA). Choose the area under the ROC curve (AUC) as the measure index for the classifiers. In UCI data sets, use Cover type, Diabetes, Ecolin, Nursery, Sponge and Yeast as test dates, and the uncertain ratio of each data sets is 10%. AUC of three classifiers are as follow.

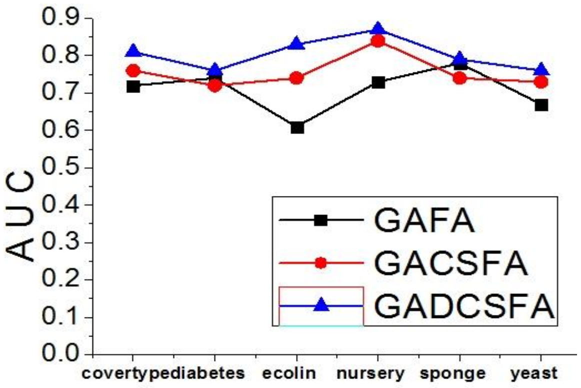

Figure 1. AUC of GAFA, GACSFA and GADCSFA

The simulation results show that AUC of GADCSFA is higher than GAFA and GACSFA form Figure 1, so its performance is more excellent than the others. Therefore, the dynamic cost is superior to static cost.

In order to verify the cost consumption, we set the fixed test cost and misclassification in CK-means, P–DBSCAN, and the cost of GADCSFA is the interval number. The total cost is constituted by test cost and misclassification, and GADCSFA uses the genetic algorithm to search the optimal cost. The total costs of CK-means, P-DBSCAN, and GADCSFA are as follows.

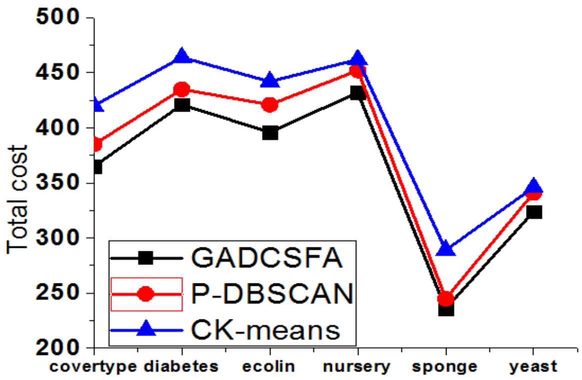

Figure 2. Total Cost of CK-means, P-DBSCAN, GADCSFA

The total cost of GADCSFA is lower than CK-means, P-DBSCAN form the Figure 2, and GADCSFA can reduce effectively the total costs.

6. Conclusion

The fussy clustering for uncertain data is an important research in pattern recognition, and it is prevalent in many real-world applications. This paper puts forward the dynamic cost-sensitive fuzzy clustering algorithm based on genetic algorithm for uncertain data, which not only considers the cost, but also can search the optimal solution by the genetic algorithm. The algorithm can dispose the continuous and discrete attribute for uncertain data by the interval number. Experimental result shows, the fussy clustering we proposed has higher clustering accuracy and performance than the other algorithms for uncertain data, which can save especially the total cost, and is suitable for uncertain data.

Acknowledgements

This work was supported by the science and technology projects of the Shandong province universities (No. J13LN53, No. J14LN65). This work has been partially supported by grants from the research fund of haze university (No. XY12KJ03, No. XY14KJ06).

References

[1] Q. H. Luo, X. Z. Yan, J. B. Li and Y. Peng, "DDEUDSC: A Dynamic Distance Estimation using Uncertain Data Stream Clustering in mobile wireless sensor networks", Measurement, vol, 55, **(2014)** September, pp: 423-433.

[2] A. J. Graaff and A. P. Engelbrecht, "Clustering data in an uncertain environment using an artificial immune system", Pattern Recognition Letters, vol. 32, Issue 2, **(2011)** January 15, pp: 42-351.

[3] W. Ngai, B. Kao, R. Cheng, M. Chau, S. D. Lee, D. W. Cheung and K. Y. Yip, "Metric and trigonometric pruning for clustering of uncertain data in 2D geometric space", Information Systems, vol. 36, Issue 2, **(2011)** April, pp. 476-497.

[4] L. Xu, Q. H. Hu, E. Hung and C. C. Szeto, "A heuristic approach to effective and efficient clustering on uncertain objects", Knowledge-Based Systems, vol. 66, **(2014)** August, pp. 112-125.

[5] P. Agrawal and J. Widom, "Generalized Uncertain Databases: First Steps", Proceedings of the 4th International VLDB Workshop on Management of Uncertain Data (MUD 2010) in conjunction with VLDB, Singapore, **(2010)**, pp. 99-111.

[6] C. Bhattaeharyya, K. S. Pannagadatta and A. J. Smola, "A second order cone programming formulation for classifying missing data", Proceedings of the 2004 Conference on Advances in Neural Information Processing Systems, **(2004)**, pp. 153-160.

[7] J. Q. Yang and S. Gunn, "Exploiting uncertain data in support vector classification", Knowledge-Based Intelligent Information and Engineering Systems, **(2007)**, pp. 148-155.

[8] F. Demichelis, P. Magni and P. Piergiorgi, M. Rubin and R. Bellazzi, "A hierarchical naïve bayes model for handling sample heterogeneity in classification problems: an application to tissue micro-arrays", *BMC Bioinformatics*, **(2006)**, vol. 7, p. 514.

[9] M. Ana, L. Sánchez and I. Couso, "Linguistic cost-sensitive learning of genetic fuzzy classifiers for imprecise data", International Journal of Approximate Reasoning, vol.52, Issue 6, **(2011)** September, pp. 841-862.

[10] J. P. Fan, J. Zhang, K. Z. Mei and J. Y. Peng and L. Gao, "Cost-sensitive learning of hierarchical tree classifiers for large-scale image classification and novel category detection", Pattern Recognition, In Press, Corrected Proof, Available online, **(2014)** October 31.

[11] J. Kim, K. Choi, G. Kim and Y. Suh, "Classification cost: An empirical comparison among traditional classifier", Cost-Sensitive Classifier, and Meta Cost, Expert Systems with Applications, vol. 39, Issue 4, **(2012)** March, pp. 4013-4019.

[12] I. Alfonso, B. Concha and L. Pedro, "Cost-sensitive selective naive Bayes classifiers for predicting the increase of the h-index for scientific journals", Neurocomputing, vol. 135, no. 5, **(2014)** July, pp. 42-52.

[13] S. D. Leed, B. Kao and R. Cheng, "Reducing uk-means to K-means", The 1st Workshop on Data Mining of Uncertain Data, Omaha, **(2007)** October 28, pp. 483-488.

[14] C. Michael, C. Reynold and K. Ben and N. Jackey, "Uncertain data mining: an example in clustering location data", Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006), Berlin, Springer Verlag, **(2006)**, pp. 199-204.

[15] H. Hamdan and G. Govaert, "Mixture Model Clustering of Uncertain Data", Proceedings of the IEEE International Conference on Fuzzy Systems, **(2005)**, pp. 879-884.

Author

Yuwen Huang, he was born in 1978 at Shanxian, and received the Master of Engineering in Computer Science from the "Guangxi Normal University" in 2009. He is now a lecturer at the Department of Computer and Information Engineering, Heze University. His research interests include the data-mining, intelligence Calculation.