CUDAP: A Novel Clustering Algorithm for Uncertain Data Based on Approximate Backbone

Ping Jin
Information and Engineering School, West Anhui University, Luan, China
School of Computer Science, University of Science and Technology of China, Hefei, China
jinping@wxc.edu.cn

Shichao Qu
School of Software, Dalian University of Technology, Dalian, China
School of Computer Science, University of Science and Technology of China, Hefei, China
Shichaoqu@dlut.edu.cn

Yu Zong*
Information and Engineering School, West Anhui University, Luan, China
zongyu@gmail.com

Xin Li
School of Computer Science, University of Science and Technology of China, Hefei, China
leexin@mail.ustc.edu.cn

***Abstract*—Clustering for uncertain data is an interesting research topic in data mining. Researchers prefer to define uncertain data clustering problem by using combinatorial optimization model. Heuristic clustering algorithm is an efficient way to deal with this kind of clustering problem, but initialization sensitivity is one of inevitable drawbacks. In this paper, we propose a novel clustering algorithm named CUDAP (Clustering algorithm for Uncertain Data based on Approximate backbone). In CUDAP, we (1) make *M* times random sampling on the original uncertain data set D^m to generate *M* sampled data sets $DS=\{Ds_1,Ds_2,\ldots,Ds_M\}$; (2) capture the *M* local optimal clustering results $P=\{C_1,C_2,\ldots,C_M\}$ from *DS* by running UK-Medoids algorithm on each sample data set Ds_i, $i=1,\ldots M$; (3) design a greedy search algorithm to find out the approximate backbone(*APB*) from *P*; (4) run UK-Medoids again on the original uncertain data set D^m guided by new initialization which was generated from *APB*. Experimental results on synthetic and real world data sets demonstrate the superiority of the proposed approach in terms of clustering quality measures.**

***Index Terms*—NP-hard Problem; Uncertain Data Clustering Problem; Heuristic Clustering Algorithm; Approximate Backbone**

I. INTRODUCTION

In recent years, data analysis and knowledge discovery in uncertain data become more and more important in many applications, such as, sensor network, biomedical measurement, financial market analysis and weather predictions, etc. In real applications, many reasons such as the error in physical measurements, the randomness in data transmission and data staling, lead to data uncertainty [1]. Generally, data uncertainty includes three levels: table, tuple, and attribute level. For each uncertainty level, we need different models to represent it. In this paper, we focus on the uncertain data with attribute level uncertainty. Traditional data usually have a finite value in each attribute and get a precision position in space. On the contrary, uncertain data is not an authentic point in space and located in a finite region represented by uncertain data object. The intrinsic characteristics of uncertain data make it difficulty to the data management, and also raise challenges to data mining and knowledge discovery. Generally, clustering for uncertain data described by combinatorial optimization model as following:

Given a set of uncertain data set $D^m=\{d_1^m,d_2^m,...,d_N^m\}$ and the number of cluster centers, K, where d_i^m $(i=1,...,N)$ is defined as Probability Density Function $f_i: R^m \to R_i^m$, ($f_i(x)\geq 0$, $\forall x\in R_i^m$, $\int_{x\in R_i^m} f_i(x)dx=1$). Uncertain clustering algorithm attempts to seek K cluster centers $\{c_1,c_2,...,c_K\}$ ($K\leq N$), such that the quality measure function $\Phi^m(\{c_1,c_2,...,c_K\})=\sum_{k=1}^{K}\sum_{d_i^m\in c_k} dist(d_i^m,c_k)$ is minimized, where $dist(\cdot)$ is the expectation distance function defined as $dist(x,y)=\int_{x\in R^m{}_i}\int_{y\in R_j{}^m} d(x,y)f_i(x)f_j(y)dxdy$.

Drineas et al. have proved that this clustering problem is NP-hard [2]. For large scale uncertain dataset, it is hard to get optimal clustering results for this kind of clustering problem in polynomial time. Researchers introduce local search methods and devise a lot of heuristic clustering algorithms [3-5]. Generally, the essence of heuristic clustering algorithm is to find a sub-optimal solution by a heuristic searching process in a local space. The existing uncertain data clustering algorithms can be divided into two kinds, density-based uncertain data clustering and

doi:10.4304/jsw.9.3.732-737

heuristic uncertain data clustering. A density-based clustering algorithm, FDBSCAN, extends the traditional clustering algorithm DBSCAN to uncertain data clustering [6]. For heuristic clustering algorithm in uncertain data mining, the expectation distance function was defined at first to capture the uncertain characteristics embedded in data attribute, and then the traditional heuristic clustering algorithms are invoked. UK-means is an important heuristic clustering algorithm for uncertain data, it has many advantages, such as simplification, easy implementation. Due to K-means is the basic idea of UK-means, so initialization sensitivity and outlier sensitivity problem are the drawbacks of UK-Means[7]. In order to deal with the outlier sensitivity problem of UK-means, UK-Medoids introduce K-Medoids into uncertain clustering [8]. CK-means improves the quality of UK-means clustering result by using a new expectation distance function.

Recently, researchers described an interesting phenomenon in heuristic algorithms [9-12]. In TSP, Max-SAT, and GBP problem, nearly 80% of the sub-optimal solutions are distributed near the optimal solution, i.e. the "big valley" phenomenon. "Big valley" phenomenon means that most sub-optimal solutions have high similarity with the optimal solution in NP-hard problem. In NP-hard problem, backbone analysis is a popular method, it is used to design heuristic algorithm with high effectiveness. Backbone is defined as the common parts of all optimal solutions of the NP-hard problem. As the optimal solutions are hard to get for NP-hard problem, it is difficult to obtain the backbone of the solution. Due to the "big valley" phenomenon, researchers tend to approximate backbone structure with sub-optimal solutions. Approximate backbone is defined as the common parts of several sub-optimal solutions in NP-hard problem. Researches can use the approximate backbone to design more effective heuristic algorithm, and get clustering solution with higher quality. From the above research, it is intuitive that we can get approximate backbone from sub-optimal solutions to help us to design better algorithm.

In this paper, we make use of the approximate backbone to uncertain data clustering problem and propose a novel clustering algorithm named CUDAP (Clustering algorithm for Uncertain Data based on Approximate backbone) . In the framework of CUDAP, we first make *M* times sampling in original uncertain dataset, and then we call UK-Medoids with randomly initialization in each sampled dataset to obtain *M* local optimal clustering results; The approximate backbone was derived from these local optimal clustering results; Eventually, we run UK-Medoids again in original uncertain dataset with a new initialization which was generated from approximate backbone. Experimental results on synthetic datasets and UCI uncertain datasets show that CUDAP has the ability to get better clustering results than other compared clustering algorithm. The Proposed clustering algorithm could be used in applications that require interaction with the physical world, such as location-based services and sensor monitoring data mining.

The rest of the paper is structured as follows. Section II we introduce the definition of approximate backbone for uncertain data. Section III proposes the details of CUDAP algorithm. Section IV discusses evaluation set up and metrics as well as analysis of experiments performed on synthetic as well as real world data sets. Section V summarizes the paper and discusses open questions.

II. DEFINITION OF APPROXIMATE BACKBONE

In this part, we first introduce the definition global optimal solutions and sub-optimal solutions of heuristic clustering for uncertain clustering problem, and then we propose the definition of backbone and approximate backbone.

Given an uncertain dataset $D^n = \{d_1^n, d_2^n, ..., d_N^n\}$, and the number of cluster centers, K, Uncertain clustering algorithm attempts to seek K cluster centers $\{c_1, c_2, ..., c_K\}$ ($K \le N$), such that the quality measure function $\Phi^n(\{c_1, c_2, ..., c_K\})$ is minimized. As discussed above, this clustering problem is actually a typical combinatorial optimization problem. The search space *S* of this clustering problem consists of all the possible combinations of the data objects. For optimization, we need to traverse *S* to find out a set of cluster centers $\{c_1^*, c_2^*, ..., c_K^*\}$ such that the Φ^n value is minimized. This cluster center set is defined as the optimal clustering result. Obviously, it is almost impractical to thoroughly traverse the *S* of a very large data set due to the NP-hard nature. Recently, to deal with time cost many researchers have proposed heuristic clustering that only search a subset $S' \subset S$ to discover the approximation of the optimal solution. The cluster centers $\{c_1, c_2, ..., c_K\}$ corresponding to the smallest Φ^n value in *S'* is considered as the sub-optimal clustering result. In uncertain clustering problem, UK-Medoids and UK-means are two traditional heuristic clustering algorithms.

Based on optimal clustering results and sub-optimal clustering results, we propose the definition of backbone and approximate backbone respectively.

DEFINITION 1. For an uncertain data clustering problem, all the optimal solutions are $P^* = \{C_1^*, C_2^*, ..., C_M^*\}$, and each solution is represented by $C_m^* = \{c_1^*, c_2^*, ..., c_K^*\}$, where c_k^*, $k = 1, ..., K$ is an uncertain object. The corresponding backbone cluster $B_m C^*$, $m = 1, ..., M$ has two properties: (1) $|B_m C^*| \ge 2$; (2) All the uncertain objects belong to the same cluster. The backbone of P^* is defined as the collection of $B_m C^*$, $m = 1, ..., M$, e,g. $Bone = \bigcup_{m=1}^{M} B_m C^*$.

Generally, it is difficult to obtain the optimal solutions in polynomial time for a NP-hard problem. According to the "big valley" phenomena [13], the researchers use the sub-optimal solutions to approximate the optimal solutions.

DEFINITION 2. For an uncertain data clustering problem, several sub-optimal solutions are

$P=\{C_1,C_2,...,C_M\}$, and each sub-optimal solution is represented by $C_m=\{c_1,c_2,...,c_K\}$, c_k, $k=1,...,K$ is an uncertain object. The corresponding approximate backbone cluster APB_mC, $m=1,...,M$ has two properties: (1) $|APB_mC|\geq 2$; (2) All the uncertain objects in the approximate backbone clusters belong to the same cluster. The approximate backbone of P is defined as the collection of APB_mC , e,g. $APB=\bigcup_{m=1}^{M} APB_mC$.

III. Clustering algorithm for Uncertain Data based on Approximate Backbone

A. The Framework of CUDAP

In this section, we introduce the framework of Clustering algorithm for uncertain data based on approximate backbone, as shown in algorithm 1. CUDAP includes four main parts: (1) we firstly generate *M* sampled data sets from the original uncertain data set, and then, run UK-Medoids algorithm on these sampled data sets to generate *M* sub-optimal clustering results. (2) we find the approximate backbone from *M* sub-optimal clustering results to capture the approximate backbone by running FAB_GS algorithm; (3) we generate a new initialization from approximate by running Find_Init algorithm; (4) a better sub-optimal clustering result is derived by running UK-Medoids again with new initialization on the original uncertain data set.

Algorithm 1. CUDAP
Input: uncertain dataset D^m , clusters number K , the number of sub-optimal solutions M and sampling rate γ Output: The clustering result
(1) $m\leftarrow 0$, $P\leftarrow\varnothing$; (2) repeat (2.1) Make sampling on uncertain dataset D^m to get the sampled dataset Ds_p ; (2.2) Run UK-Medoids on Ds_m with random initialization to get the sub-optimal clustering result C_m ; (2.3) $P=P\cup C_m$, $m\leftarrow m+1$; Until $m\geq M$; (3) Call FAB_GS to get the approximate backbone APB from *P*; (4) Call Find_Init to generate a new initialization C_{org} from approximate backbone APB ; (5) run UK-Medoids again on D^m with new initialization C_{org} to capture better clustering results; (6) Return clustering results.

B. FAB_GS

It is an important step of CUDAP for finding approximate backbone from *M* sub-optimal clustering results. We assume that each cluster $C_m=\{c_1,c_2,...,c_K\}$, $m=1,...,M$ stores the data object number that belong to it. Based on this assumption, we use the set intersection method to find the co-occurrence data objects and the approximate backbone is generated. In algorithm 2, we propose a greedy set intersection method FAB_GS (Find Approximate Backbone using Greedy Search). In FAB_GS framework, we first randomly select a sub-optimal result $C_m=\{c_1,c_2,...,c_K\}$, $m=1,...,M$ from *P*; and then, intersect C_m with the rest $M-1$ sub-optimal clustering results to generate the corresponding approximate backbone cluster APB_mC ; Iterate this two steps until all sub-optimal clustering results are intersected.

Algorithm 2. FAB_GS
Input: *P* Output: Approximate Backbone APB
(1) $APB\leftarrow\varnothing$; (2) Randomly select a sub-optimal clustering result $C_m=\{c_1,c_2,...,c_K\}$ from P ; (3) Intersect C_m with the rest $P-1$ sub-optimal clustering results to generate the corresponding approximate backbone cluster APB_mC ; (4) $APB=APB\bigcup APB_mC$; (5) Continue to select another sub-optimal clustering C_m , and rerun step (3)-(4) until all sub-optimal clustering result are intersected; (6) return APB ;

C. Find_Init

After we capture the approximate backbone APB from *M* sub-optimal clustering results, Find_Init (Find Initialized cluster centers form approximate backbone) algorithm is run to find a new initialization for guiding the K-Medoids clustering algorithm on original uncertain data set. Algorithm 3 gives the main steps of Find_Init: we first assign the data objects in D^m to each approximate backbone cluster APB_mC, $m=1,...,M$, and then merge these approximate backbone clusters to generate *K* clusters, eventually, we regard the center of these *K* clusters as the new initialization.

Algorithm 3. Find_Init
Input: Approximate backbone APB , uncertain dataset D^m Output: new initialization $C_{org}=\{c_1,c_2,...,c_K\}$
(1) $C_{org}\leftarrow\varnothing$; (2) According to APB , assign the data objects in D^m to each approximate backbone cluster APB_mC, $m=1,...,M$; (3) merge APB_mC, $m=1,...,M$ to generate *K* clusters by using Single-linkage algorithm; (4) $C_{org}=C_{org}\bigcup c_k$,where c_k is the center of the k cluster; (5) return C_{org}

IV. Experimental Evaluation

In this paper, all the experiments are conducted on computer with Intel 1.6GHz Core2 CPU, RAM 2GB and

with Windows XP environment. The compared clustering algorithms are implemented by Microsoft Visual Studio 6.0, C++ programming language.

A. Uncertain Datasets and Metrics

Synthetic uncertain datasets: the method for generating synthetic dataset was first proposed by Wang [10]. In synthetic uncertain datasets, all the uncertain objects are distributed in a 100×100 two dimensional space. Each object has a finite rectangular region MBR (Minimum Bounding Rectangle) with random size. We first randomly select K points as the centre of K uncertain objects, and then generate K uncertain objects. The distance between every two objects' centre at least $100/\sqrt{2K}$. Then randomly assign the remaining $N-K$ uncertain objects to the K clusters. For each object, the distance between its centre and cluster centre is at least $100/\sqrt{K}$. Thus we can make sure that each object is closer to its cluster centre than other clusters centre. The PDF of each uncertain object is represented as follows: in each dimension, we set $I^h = 10$ as the distribution interval, and each dimension of the MBR of each uncertain object is in this interval, and we use a distribution parameter β $(0 < \beta < 1)$ to regulate the size of MBR. And in MBR of each object, we randomly generate N points. As this way, the synthetic uncertain data set with random distribution are created.

TABLE I.

DEFAULT VALUE OF PARAMETERS

Parameter	Default value
n	100
S	100
K	10
β	0.5

The synthetic uncertain dataset is affected by 4 parameters. n, the number of uncertain objects; S, the number of points in each uncertain object; K, the number of clusters; β, distribution parameter, which affects the size of MBR of each uncertain objects. The default value of these parameters are shown in TABLE I. We create a series of synthetic datasets by changing the value of each parameter

TABLEII.

STANDARD UCI DATASETS

Dataset	Points	Attributes	Clusters
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6

Standard UCI uncertain datasets: In this paper, we use 3 standard UCI benchmark data sets (Glass, Iris, and Wine) to generate UCI uncertain datasets. TABLE II describes the properties of these datasets. Each dimension of the three datasets are numerical. For each UCI dataset, we generate uncertain object through distribution from the points in them. Distribution depends on the PDF of each object: in each dimension of UCI datasets, we set a distribution interval $I^h = 0.1 * \max\_length$, where $\max\_length$ the max length of h dimension. Distribution parameter $\beta(0 < \beta < 1)$ is used to regulate the MBR of each uncertain object. In the MBR of each uncertain object, randomly generate n points, and each point has the same distribution. Thus, we can get the standard UCI uncertain datasets with random distribution.

There are two parameters which could affects the generation of standard UCI uncertain datasets: S, the number of points in each uncertain object; β, distribution parameter which decide the size of distribution. We first set default value of S and β (as shown in TABLE II) and then generate a series of uncertain datasets by using different parameters values.

TABEL III.

DEFAULT PARAMETER VALUE OF STANDARD UCI UNCERTAIN DATASETS

Parameter	Default value
S	100
β	0.5

In this paper, we use F-Score to measure the precision and recall rate of our methods and the compared clustering algorithms.

DEFINITION 3. Given an uncertain data set D^m, $C^* = \{C_1^*, C_2^*, ..., C_k^*\}$ is the benchmark clustering result, and $C = \{C_1, C_2, ..., C_k\}$ is the clustering result derived by clustering algorithm. The precision rate of $C_j (j = 0,1,...,K)$ and $C_i^* (i = 0,1,...,K)$ is defined as:

$$\Pr ecision(i, j) = \frac{|C_j \cap C_i^*|}{|C_j|}$$

DEFINITION 4. Given an uncertain data set D^m, $C^* = \{C_1^*, C_2^*, ..., C_k^*\}$ is the benchmark clustering result, and $C = \{C_1, C_2, ..., C_k\}$ is the clustering result derived by clustering algorithm. The recall rate of $C_j (j = 0,1,...,K)$ and $C_i^* (i = 0,1,...,K)$ is defined as:

$$\mathrm{Re} call(i, j) = \frac{|C_j \cap C_i^*|}{|C_i^*|}$$

DEFINITION 5. F-Score is defined as a linear combination of precision rate and recall rate: $F = \frac{2PR}{P+R}$, where $P = \frac{1}{k}\sum_{i=1}^{k} \max(\Pr ecison(i, j))$,

$$R = \frac{1}{k}\sum_{i=1}^{k} \max(\mathrm{Re} call(i, j)),\ i = 1,2,...,K$$

According to DEFINTION 5, the value of F-Score is between 0 and 1. The bigger value denotes the better quality of clustering results.

B. The Setting of Parameter γ

In this section, we conduct experiments on the influence of sampling parameter γ. Figure 1 shows the experimental results on three UCI uncertain data sets. In

this Figure, X axis represents the γ value and Y axis denotes the corresponding F-Score value. According to the changing line in Figure 1, the F-Score value curve has the same changing trend. Three lines are rising when $\gamma \le 0.05$, and then they are changed to flat. This phenomena show that the influence of sampling rate has reduced when $\gamma > 0.05$.

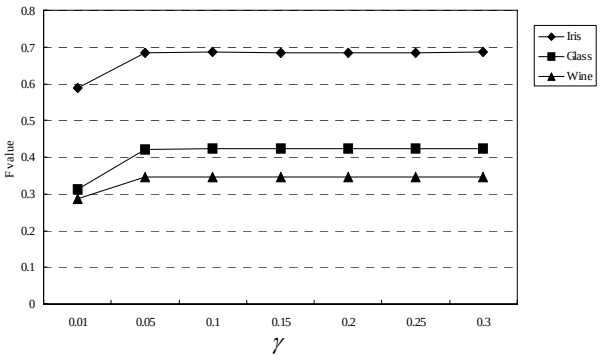

Figure 1. The experimental results on three UCI uncertain data sets by changing γ value

Figure 2 shows the experimental results of 5 compared clustering algorithms on 3 UCI uncertain data sets with $S = 120$ and $\beta = 0.75$. From Figure 2, we can find that the clustering results derived by CUDAP on 3 UCI uncertain data sets are better than those of other compared clustering algorithms. This experiment shows the ability of CUDAP on deal with the real world uncertain data clustering problem.

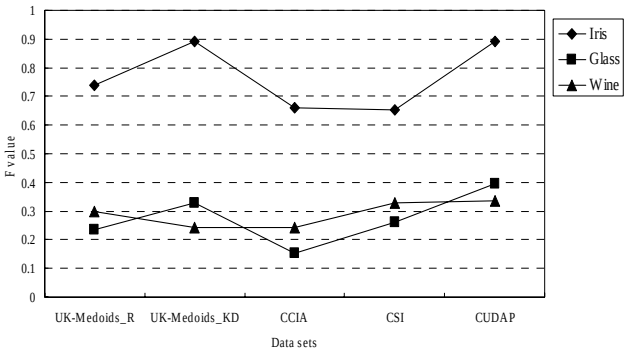

Figure 2. The comparison of 5 algorithm on 3 UCI uncertain data sets

C. *Expremental Results and Analysis*

In order to value the efficiency of our proposed method, we compare CUDAP with 4 clustering algorithms: UK-Medoids_R(UK-Medoids with random initialization), UK-Medoids_KD(UK-Medoids with initialization derived by using kd-tree [14]), CCIA [15], CSI [12].

TABLE IV shows the experimental results on synthetic data set ($n = 100$, $s = 100$, $K = 20$, $\beta = 0.75$). From TABLE IV, we can find that the quality of clustering results derived by CUDAP is over the compared clustering algorithm. UK-Medoids_R and UK-Medoids_KD has similar quality. This phenomenon shows the fact that random initialization and kd-tree based initialization method cannot deal with the initialization sensitivity problem of clustering algorithm for uncertain data. CUDAP and CSI use the common information (captured by approximate backbone) to guided the search steps of heuristic clustering, so the quality of clustering results have significant improvements.

Because of the sampling method used in our method, the time cost of CUDAP is less than that of other 4 compared clustering algorithms.

TABLE IV.
THE COMPARISON OF FIVE ALGORITHMS ON SYNTHETIC DATA SET

	FMeasure	Time(s)
UK-Medoids_R	0.63522	10.781
UK-Medoids_KD	0.665738	8.25
CCIA	0.556172	6.704
CSI	0.830377	6.453
CUDAP	0.851897	6.2734

CONCLUSION

In this paper, we focus on the modification of initialization sensitivity problem of heuristic clustering algorithms for uncertain data. We make use of the common part of several sub-optimal clustering results to design a new way to improve the quality of clustering results. Approximate backbone is used to capture the common information and a new algorithm named CUDAP is proposed. Experimental results show that CUDAP has ability to find better clustering results.

ACKNOWLEDGMENT

The work described in this paper was supported by grants from Natural Science Foundation of China (Grant No. 60775037), the Key Program of National Natural Science Foundation of China (Grant No. 60933013), the Nature Science Research of Anhui.(Grant No. 1208085MF 95), the Nature Science Foundation of Anhui Education Department(Grant No. KJ2012A273 and KJ2012A274) and the EU FP7 ICT project M-Eco: Medical Ecosystem Personalized Event-based Surveillance (No.247829). The authors would like to thank the reviewers for their valuable comments.

REFERENCES

[1] Y. Tao. X. Xiao, and R. Cheng, et al. Range search on multidimensional uncertain data[J], ACM Transactions on Database Systems, 2007, 32(3): 15-62.

[2] P. Drineas, R. Frieze, S. Vempala, et al. Clustering large graphs via singular value decomposition. Machine Learning, 2004: 56(1-3): 9-33.

[3] Y. Zong, P.Jin, D.G. Xu, R. Pan. A Clustering Algorithm based on Local Accumulative Knowledge. Journal of Computers,2013,8(2):365-370.
[4] C.Y Ren. Heuristic Algorithm for Min-max Vehicle Routing Problems. Journal of Computers,2012, 7(4):923-928.
[5] J.Y. Xie, S. Jiang, W.X. Xie, X.B. Gao. An Efficient Global K-means Clustering Algorithm.Journal of Computers,2011,6(2):271-279.
[6] H. P. Kriegel, M. Pfeifle. Density based clustering of uncertain data[C]. Proceedings of Knowledge Discovery and Data Mining, 2005, 672-677.
[7] M. Chau, R. Cheng, B. Kao, and J. Ng. Uncertain data mining: An example in clustering location data. In Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2006), volume 3918 of Lecture Notes in Computer Science, pages 199–204, Singapore, 9–12 Apr. 2006. Springer
[8] F. Gullo, G. Ponti and A. Tagarelli, et al. Clustering uncertain data via k-medoids[C]. Proceedings of the 2th International Conference on Scalable Uncertainty Management, 2008: 229-242.
[9] W. X. Zhang, Configuration landscape analysis and backbone guided local search: Part I: satisfiability and maximum satisfiability[J]. Artificial Intelligence, 2004, 158(1):1-26.
[10] S. Kirkpatrick, G. Toulouse, Configuration space analysis of travelling salesman problem[J], Journal de Physique, 1985, 46: 1277-1292.
[11] J. He, Q. Tie, H. Yan, et al. Backbone analysis and applications in heuristic algorithm design[J]. Journal of Acta Automatica Sinica, 2011, 37(3): 257-269.
[12] Y. Zong, G.D. Xu, and P Jin, et al. HC_AB: a new heuristic clustering algorithm based on approximate backbone[J]. Information Processing Letters, 2011, 111: 857-863.
[13] K.Wang. Ngai, Ben. Kao, and Chun, Kit, Chui, et al. Efficient clustering of uncertain data[C]. Proceedings of Sixth Information Conference on Data Mining, 2006: 436-445.
[14] S. J. Redmond, C. Heneghan. A method for initializing the K-Mediods clustering algorithm using kd-tree[J]. Journal of Pattern Recognition Letters, 2007, 28(8): 965-973.
[15] M. Chau, R. Cheng, and B. Kao, et al. Uncertain data mining: an example in clustering location data[C]. Proceedings of Knowledge Discovery and Data Mining, 2006, 199-204.

Ping Jin received B.E in Computer Application from Agriculture University of Anhui China in 1999 and received her M.E in Computer Science and Technology from Hefei University of Technology. She is also a senior visiting scholar of university of Science and Technology of China. Her researches interesting include data mining, recommendation and computer application.

Shichao Qu received B.E. and M.E. in Software Engineering from Dalian University of Technology China in 2011 and 2013 respectively. His researches interesting include data mining, recommendation and computer application.

Yu Zong is a Joint PhD of Dalian University of Technology China and Victory University Australian, and he obtained a PhD degree in Computer Science from Dalian University of Technology China in 2010. Prior to this, he received M.E degree in Software Engineering from Dalian University of Technology, and he also received B.E in Computer Application from Agriculture University of Anhui China in 2005 and 1999, respectively. His research interests include data mining, social network mining, recommendation, intelligent algorithm, as well as the computer applications.

Xin Li received B.E. in Computer Science from University of Science and Technology of China in 2010 and now a PhD candidate student in USTC. His research interests include data mining and machine learning.