Stable Entity Resolution for High-Dimensional Data

Yi Liu[1)] Student, Xingchun Diao[1*)] Professor, Jianjun Cao[2)] Professor, Nianfeng Weng[2)] Engineer and Lei Zhang[1)] Student
[1)](PLA University of Science and Technology)
[2)](Nanjing Telecommunication Technology Institute)
diaoxch640222@163.com

Abstract: With the development of big data, the requirements for entity resolution in high dimensional data are becoming more and more pressing. It is a feasible way to handle it by feature selection, and we can choose relevant features to obtain similarity vector of two records and put it into classifiers to identify whether they are duplicate. We propose a new approach called stable entity resolution based on multiobjective ant colony optimization to resolve it. It combines three filter feature selection methods to provide stable feature information, and adopts fisher score and maximal information coefficient to generate heuristic information for multiobjective ant colony optimization. It employs two objectives which are classification accuracy and stability of feature selection to be optimized to get better classification performance. Two classic benchmark datasets are taken to validate our method which is compared with other two approaches, and the results show its superiority.

Keywords: entity resolution, multiobjective ant colony optimization,feature selection, feature selection stability, high dimensional data

1 INTRODUCTION

Entity resolution (ER) is the one of the most important stages of data cleaning and its task is to identity the different descriptions which refer to the same world entity. ER is also called record linkage, records deduplicate, data linkage, coreference resolution, and name disambiguation etc. (Wang et al. 2016; Yannik et al. 2016; Seyed et al. 2017)

The methods to resolve ER can be categorized into Feature Based Similarity (FBS) methods, Relationship Based Data Cleaning (ReIDC), Semantic Based Methods (SBM), and Crowdsourcing Based Methods (CBM). FBS intends to find out two records are whether matches (duplicate) or non-matches (unduplicated) through comparing records' similarity vector obtained by features' similarity, and it is used extensively. ReIDC estimates the matched pairs based on the strength valued of relationship connections between records (Rabia et al. 2013). SBM combines the semantic information between records to identity the matched pair of records (Evandro et al. 2016). CBM distributes the possible matched records to the human workers of Crowdsourcing platform to find out the true matched records (Chai et al. 2016).

The proportion of high dimensional data is growing faster and faster with the development of big data and the spread applications of machine learning, such as gene, text, picture, twitter and position information etc. In general, high dimensional data stands for the data which has more dimensions than instances (Masulli and Rovetta. 2015). Developing the ER methods for high dimensional data is becoming a new and important research direction. Feature selection is a key stage of data preprocessing, and its aim is to select a minimal subset of features to maximize the classification performance. Adopting feature selection can reduce the time for getting original data, compress data storage, obtain the classification model faster, and improve model's interpretation and classification performance. It is a feasible way to select appropriate features by employing feature selection and identify matched records based on FBS methods.

The stability of feature selection is the robustness of results with respect to small changes in the dataset composition. Improving the stability of feature selection can find out relevant features, increase confidence of experts to the results, and further reduce the complexity of getting original data and time costs.

This paper proposes an algorithm called Stable Entity Resolution Based on Multiobjective Ant Colony Optimization (SERMOACO) for entity resolution in high dimensional data. It combines three Filter ranking feature selection methods' results as the stability guidance information, and the values of Fisher score and Maximal Information Coefficient (MIC) are taken as the heuristic information of multiobjective ant colony optimization (MOACO). It takes classification accuracy and feature selection stability as two optimization objectives to balance its classification performance and stability, and experiments show the effectiveness of our methods.

2 THEORIES AND RELATED WORKS

2.1 ER's Process

ER can be regarded as a process of binary classification whose results have two types: matched (duplicate) and non-matched (non-duplicate) which are represented as class 1 and class 2 (Cao et al. 2016). In detail, the process can be descripted as follows: two selected records' similarity vector is calculated based on their corresponding feature's similarity value which is obtained by different functions according to its type. Then the vector is input into a binary classifier to classify them as class 1 or class 2. Its procedure can be depicted as Fig. 1.

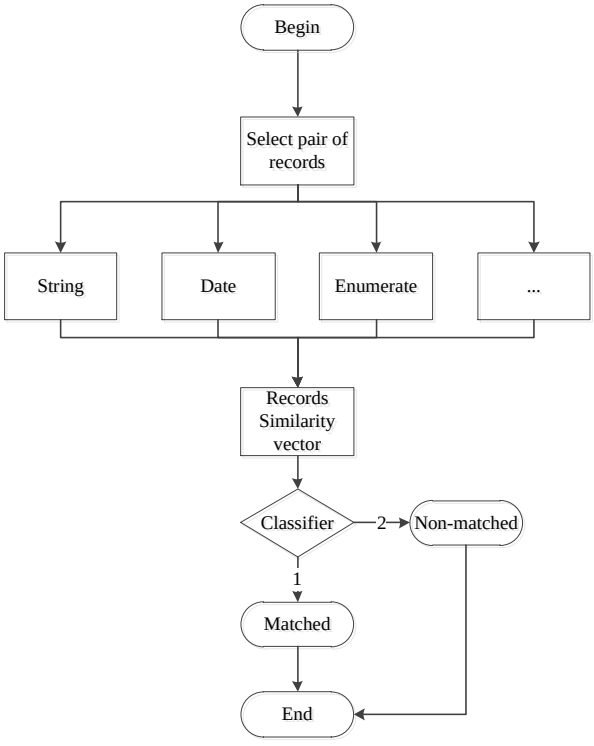

Figure 1: ER's process

2.2 Multiobjective Optimization Theory

Multiobjective optimization problems (MOPs) can be described as eq. (1) (Eckart et al. 2003)

$$\min \boldsymbol{F}(\boldsymbol{x}) = (f_1(\boldsymbol{x}),\ f_2(\boldsymbol{x}), ...,\ f_m(\boldsymbol{x}))^T,\ \ \boldsymbol{x} \in \Omega \quad (1)$$

where decision vector $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ belongs to nonempty decision space Ω, objective vector $\boldsymbol{F}: \Omega \to \Lambda$ is composed of m ($m \geq 2$) objectives, and Λ is objective space. The solutions of MOPs are called Pareto solutions (PS) which don't dominate each other by Pareto dominance relation. The objective vectors corresponding to all PS constitute Pareto front (PF) in decision space. The performance of a multiobjective evolutionary algorithm (MOEA) is measured by two objectives: convergence, i.e. the PS of the MOEA should be as close as possible to the PF, and diversity, i.e. the PS should be as diverse as possible in the objective space.

2.3 The Development of ER in High Dimensional Data

With the development of big data and machine learning, the volume of high dimensional data is growing faster and faster, and the methods based on FBS for entity resolution in high dimensional data is becoming a new hot point.

Detecting bioinformatic duplicates is important in bioinformatics, especially when two or more databases which have the same entities. Chen et al. (2015) made some experiments in gene datasets by using a traditional entity resolution, and the results show that traditional entity resolution methods have a weak performance in gene database, so we need to develop some new ways to resolve it.

Online social networks have become very popular in our daily lives, such as Wechat, Facebook, Twitter and so on. A person may use the same or different information to create accounts on multiple social applications. In order to identity them, Olga et al. (2016) proposed a machine learning method, and it can not only match users across two social networks, but also search for a user by similar name and de-anonymize a user's identity.

Forensics examiners need to find duplicate files during an investigation, and current forensic tools often use hashing method to realize it. But they may fail due to differences in file format. In order to overcome their disadvantages, Clay (2016) introduced a tool called sdtext which identifies similar files based on their textual contents. It uses inverse document frequency to create dictionary for files, and adopts bit vectors to represent the presence or absence of a dictionary term in file, and takes a cosine similarity measure to compare two vectors to identify whether they are identical or not.

Besides those scenes, some new entity resolution methods are proposed according to the concrete problems such as optical music recognition (Christophe et al. 2016), duplication detection in bug reports (Lin et al. 2016), and web entity deduplication (Vasilis et al. 2016) etc.

Though many approaches have been developed for entity resolution in high dimensional data, the features used for comparison are selected manually or not chosen. There are two problems in those ways. First, a user may have no idea which feature should be selected when they adopt those methods in new datasets. Second, it is not always effective by using all features especially they have redundant correlations. Feature selection can resolve both problems effectively, and also stability is another important aspect which we should carefully consider for improving the performance of entity resolution and the interpretability of model.

3 SERMOACO

The components of SERMOACO are show in Fig. 2.

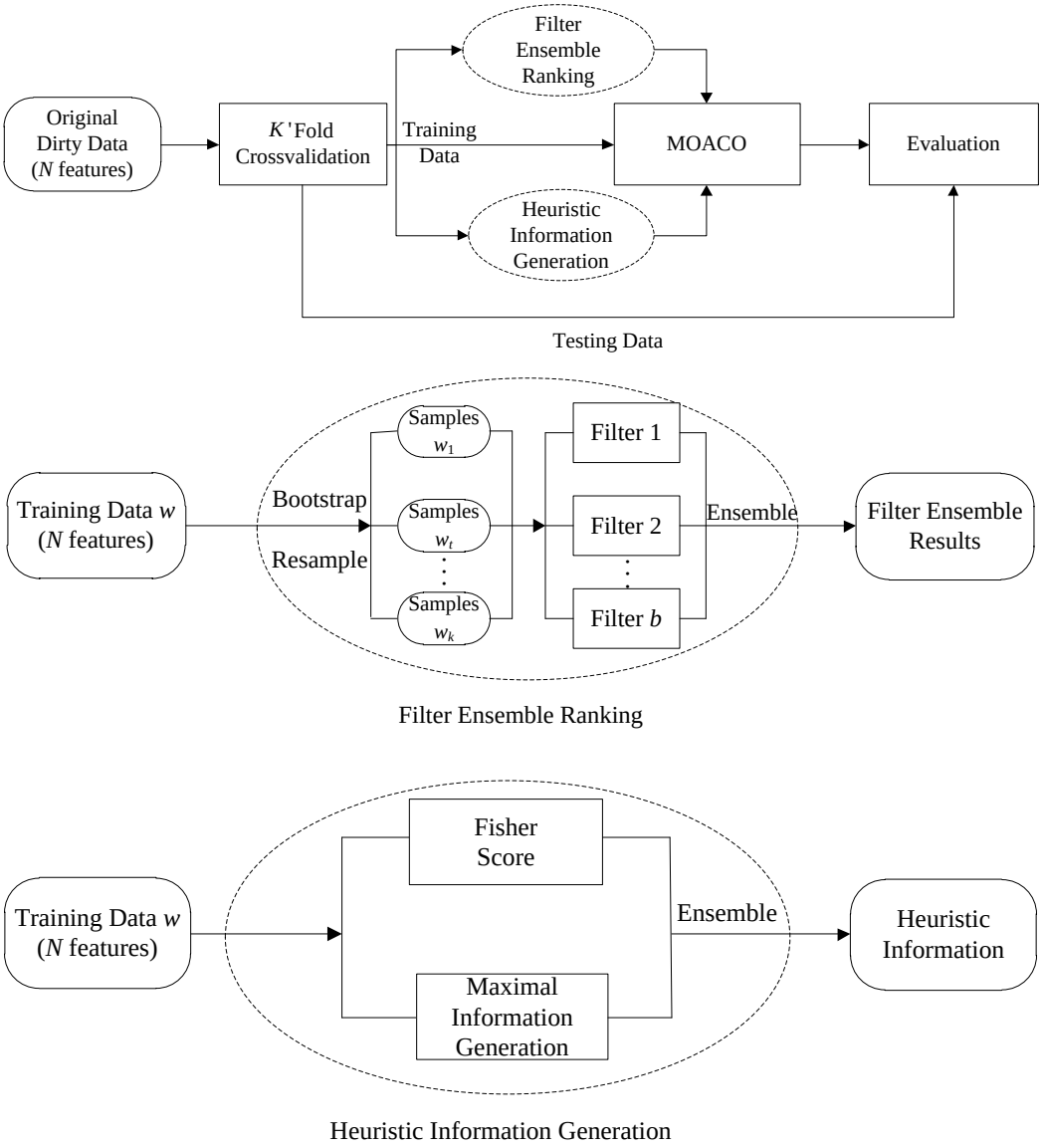

Figure 2: SERMOACO

SERMOACO is composed of three components, i.e. Filter Ensemble Ranking, Heuristic Information Generation, and MOACO. Filter Ensemble Ranking adopts some Filter feature selection methods to rank features in bootstrap samples and combine them to provide stability information for MOACO; Heuristic Information Generation which provides the heuristic information for MOACO is made up of features' Fisher score and their MIC values, which balances the features' discriminant ability and stability; Two objectives, classification accuracy and stability, are optimized by MOACO. Next, we will introduce them in detail.

3.1 MOACO

Feature selection is a classic combinatorial optimization problem and also a subset problem. MOACO is superior to other evolutionary algorithms at resolving combinatorial optimization problem (I.D.I.D. and Fernando 2015). Cao et al. (2008) proposed a graph-based ant system which converts problem to graph and puts pheromone on its edges, and experiment results show its effectiveness and superiority.

But it is used to resolve single objective optimization problem, we generalize it to settle multiobjective optimization problems and adopts it as an important element of SERMOACO to select feature subset.

The pseudo code of proposed MOACO is showed in algorithm 1.

Algorithm 1. MOACO pseudo code

01. **BEGIN**
02. **WHILE** (Not meet the maximal number of iteration)
03. Initialize Pareto archive, pheromone matrices, and heuristic information
04. **FOR** each ant
05. Select features based on probability distribution which is computed by pheromone values and heuristic information
06. **END FOR**
07. Evaluate solutions according to two objectives and update Pareto archive according to their Pareto dominance relations
08. Update pheromone matrices by solutions in Pareto archive
09. **END WHILE**
10. Output Pareto Solutions
11. **END**

In step 5, the probability distribution is given by (2)

$$p_{ij}^{u} = \frac{[\tau_{ij}]^{\alpha} \cdot [\eta_{ij}]^{\beta}}{\sum_{l \in N_i^k} [\tau_{ij}]^{\alpha} \cdot [\eta_{ij}]^{\beta}} \quad \text{if } j \in N_i^u \quad (2)$$

where p_{ij}^{u} denotes the probability of ant u selects feature *j* after selecting feature *i*, τ_{ij} denotes the pheromone value of edge (*i*, *j*) at current iteration, η_{ij} is the static heuristic information of the "goodness" of edge (*i*, *j*), and N_i^u is the feasible edges for ant *u* after it selects edge *i*.

Existing research has demonstrated that it can get more high quality Pareto solutions by employing more than one pheromone matrix. So the proposed MOACO adopts two pheromone matrices, one for each objective. And the values from two pheromone matrices need to be aggregated into a single pheromone value. We use weighted product method which is given by (3) to realize it.

$$\tau_{ij} = (\tau_{ij}^{1})^{(1-\lambda)} \cdot (\tau_{ij}^{2})^{\lambda} \quad (3)$$

Where λ is a weight and has $0 \le \lambda \le 1$.

We will discuss the Heuristic Information Generation which provides heuristic information for MOACO at section 3.3.

In step 8, we use (4) to update two pheromone matrices.

$$\tau_{ij}(t) = \begin{cases} (1-\rho)\tau_{ij}(t-1) + \Delta'(tabu^{t}) & e_{ij} \in \Psi(tabu^{t}) \\ (1-\rho)\tau_{ij}(t-1) & \text{otherwise} \end{cases} \quad (4)$$

where ρ is pheromone evaporation rate, $\Delta'(tabu^t)$ is the incremental values of pheromone. And $\Delta'(tabu^t)$ is given by (5)

$$\Delta'(tabu^t) = (\sum_{h=1}^{m} f_h(tabu^t)) / (\mathrm{Q} * m) \quad (5)$$

where *m* is the number of objectives， $f_h(tabu^t)$ denotes the h^{th} objective value of feature subset $tabu^t$, Q is a constant value.

3.2 Filter Ensemble Ranking

Ensemble method can improve feature selection stability effectively, and it is used widely (Adil et al. 2014; Ghadah and Wang 2015; Iman et al. 2016).

Filter feature selection methods contain two types, i.e. univariate and multivariate approaches: in the first case, each single feature is evaluated from others independently; In the second one, the inter-dependencies among features are taken into account. Univariate methods include Information Gain (IG), Symmetrical Uncertainty, Gain Ratio, Chi Squared (χ^2), and One Rule etc. Multivariate methods involve ReliefF, ReliefW, Support Vector Machine One, and Support Vector Machine Recursive Feature Elimination etc. Univariate methods have a better stability but a worse classification performance, and multivariate methods are the opposite.

We develop a new approach which combines the results of univariate and multivariate methods to provide a more quality feature ranking which contains both advantages.

Some previous researches have showed that IG, χ^2 and ReliefF have a good comprehensive performance (Barbara et al. 2017), so we adopt those three approaches to rank features in bootstrap samples. Besides, different aggregation strategies have no obvious distinction between each other, so we employ median way to combine the three feature rankings, i.e. assigning the median rank value to the corresponding feature across all the original lists.

At last, we choose the features by their descending values to make up of feature subset as the final output.

3.3 Heuristic Information Generation

In MOACO, heuristic information denotes prior experience defined by user according to the problem. Appropriate heuristic information can improve algorithm's ability effectively. Based on the characteristics of feature selection, we take Fisher score and MIC of features to compute heuristic values. The combination of both two ways can enhance the ability of MOACO as a higher Fisher score denotes a feature has better discriminant ability, and a feature has a stronger relationship with class labels if its MIC value is bigger.

3.3.1 Fisher Score

Fisher score is a measure which evaluates feature's discriminant ability, and it means that the same feature's values between different instances in the same class must be smaller, otherwise bigger. In a binary classification problem, the h^{th} feature's Fisher score is given by (6)

$$Fscore(h) = \frac{|\bar{\mu}_{1h} - \bar{\mu}_{2h}|}{\sqrt{\sigma_{1h}^2 + \sigma_{2h}^2}} \quad (6)$$

where $\bar{\mu}_{1h}, \bar{\mu}_{2h}$ denotes the mean value of the h^{th} feature in class 1 and class2, respectively. And σ_{1h}^2, σ_{2h}^2 denotes the h^{th} feature's variance in class 1 and class 2, respectively.

3.3.2 MIC

Some researches often use some measures to evaluate the relationship between features and class labels, such as Pearson's correlation coefficient, and Spearman's rank correlation coefficient etc. But they all have some disadvantages hardly to be resolved. Pearson's correlation coefficient can only measure the linear relationship, and it can't be used to evaluate nonlinear relationships and non-functional relations. Spearman's rank correlation coefficient could measure nonlinear relation, but it has a poor accuracy (Hauke and Kossowski. 2011).

The idea of MIC is that we can draw a grid on a scatterplot which is composed of two variables in a certain way if there is a relationship between the two variables, and make the most of their points spread in some grid cells (Zhang et al. 2014). MIC can be used in any relationship between two variables no matter whether it is linear or not.

Given a finite set *D* which is composed of two variables, we can partition the *x* values of *D* into *x* bins and the *y* values of *D* into y bins, and this pair of partitions is called an *x* by *y* grid *G*. Given a grid *G*, let $D|_G$ be the distribution induced by the points in *D* on *G*. And we can get different distributions under different *G* when *D* is certain.

Given a finite set $D \subset \square^2$ and positive integer *x* and *y*, we can get eq. (7)

$$I^*(D, x, y) = \max I(D|_G) \quad (7)$$

where $I(D|_G)$ is the mutual information of points' distribution $D|_G$ in grid *D*.

And we normalize *I** by (8)

$$M(D)_{x,y} = \frac{I^*(D, x, y)}{\log \min\{x, y\}} \quad (8)$$

At last, we can get the MIC value of the two variables by (9)

$$\mathrm{MIC}(D) = \max_{xy<B(n)}\{M(D)_{x,y}\} \quad (9)$$

MIC takes values in [0, 1], and the two variables have a stronger relationship when the value is higher. Besides, it is also symmetrical, i.e. MIC(*X*, *Y*)=MIC(*Y*, *X*).

In this paper, we adopt MIC to measure the correlation between feature and class labels, and take it as the heuristic information of MOACO. The procedure of measuring features by MIC is describes as algorithm 2.

Algorithm 2. MIC measurement pseudo code

```
01. BEGIN
02. FOR each feature
03. Use a vector to record current feature's values under all instances, and employ another vector to record the corresponding class labels
04. Get the correlation value of the two variables by MIC
05. END FOR
06. END
```

3.4 The Optimization Objectives

In SERMOACO, there are two objectives optimized by MOACO, i.e. classification accuracy and feature selection stability.

Supposing the number of instances is *Num* in a binary classification problem, the number of instances which are classified as class 1 correctly is *P_num*, the number of instances which are classified as class 2 correctly is *N_num*, then classification accuracy *P* is given by (10)

$$P = \frac{P\_num + N\_num}{Num} \quad (10)$$

There are many indicators measuring the stability of feature selection such as Tanimoto distance, Dunne Stability Index, Weighted Consistency and Extensions of Kuncheva Similarity Measure (EoKSM) etc. As EoKSM has a better performance and can be employed between two feature subsets which have different numbers, we take EoKSM to measure the similarity between the two feature subsets of Filter Ensemble Ranking and MOACO. Given two feature subsets $\boldsymbol{s}$ and $\boldsymbol{s}'$, their EoKSM value is given by (11)

$$EoKSM(\boldsymbol{s}, \boldsymbol{s}') = \frac{|\boldsymbol{s} \cap \boldsymbol{s}'| - \frac{|\boldsymbol{s}| \cdot |\boldsymbol{s}'|}{c}}{\max[-\max(0, |\boldsymbol{s}| + |\boldsymbol{s}'| - c) + \frac{|\boldsymbol{s}| \cdot |\boldsymbol{s}'|}{c}; \min(|\boldsymbol{s}|, |\boldsymbol{s}'|) - \frac{|\boldsymbol{s}| \cdot |\boldsymbol{s}'|}{c}]} \quad (11)$$

where EoKSM takes values in [-1, 1]; a value of 1 means that two sets are identical and the stability of feature selection algorithm is the best.

As SFSMOACO holds K' cross validation method to evaluate the results, we must take the mean of two objectives' values as the final output, and they are obtained by (12) and (13)

$$\bar{P} = \frac{1}{K'} \sum_{l=1}^{K'} P_l \quad (12)$$

$$D = \frac{2}{K'(K'-1)} \sum_{c_1=1}^{K'-1} \sum_{c_2=c_1+1}^{K'} EoKSM(\boldsymbol{s}_{c_1}, \boldsymbol{s}_{c_2}) \quad (13)$$

3.5 Pseudo Code

In detail, the proposed SERMOACO is showed in algorithm 3.

Algorithm 3. SERMOACO pseudo code

01. **BEGIN**

02. **FOR** *K'* cross validation

03. Use Bootstrap to generate *k* groups of samples in training data

04. Rank the features by the three Filter feature selection methods in *k* groups of samples

05. Aggregate the features' ranking results

06. Obtain each feature's Fisher score in training data

07. Get each feature's MIC value

08. Aggregate each feature's Fisher score and MIC value as the heuristic information of MOACO

09. Adopt MOACO to select feature subsets

10. **END FOR**

11. **END**

4 EXPERIMENTS AND DISCUSSIONS

4.1 Data and Pretreatment

In this section, we describe the experiments conducted to validate the effectiveness of our proposed algorithm. We use two clean high dimensional data from website[1] to generate corresponding dirty duplicate datasets. The two clean datasets are madelon which has 500 features and 2600 instances and leukemia which has 7070 features and 72 instances. We choose two instances from same class as duplicate pair and two instances from different classes as distinct pair to generate experiment datasets. The characteristics of synthetic datasets are show in table 1.

Table 1: Characteristics of synthetic experiment datasets

Datasets	Matched	Unmatched	Features
Madelon_ER	200	800	500
Leukemia_ER	100	400	7060

We use other two feature selection methods to make a comparison with our method, i.e. Duplication Detection based on Ant Colony Optimization (DDACO) (Cao et al. 2010), and Minimal Redundancy Maximal Relevance (MRMR) (Peng et al. 2005). We adopt the two approaches to select feature subsets and then generate corresponding feature similarity vectors to identity pair of records referring to the same entity to compare their performance with our method. We employ fivefold cross validation and Naive Bayes classifier, and the percentage of selected features is ranging from 1% to 5%.

4.2 Performance Analysis

The experiment results are shown in fig. 3, fig. 4, fig .5 and fig. 6.

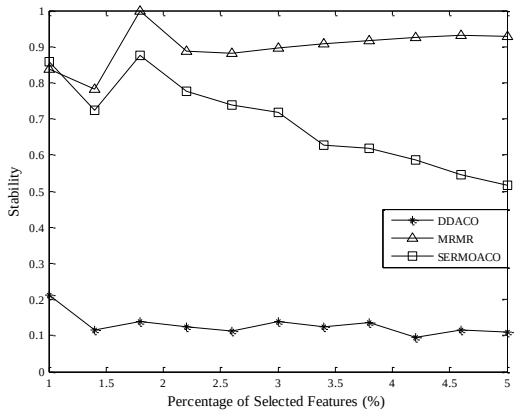

Fig. 3 Stability on Madelon_ER

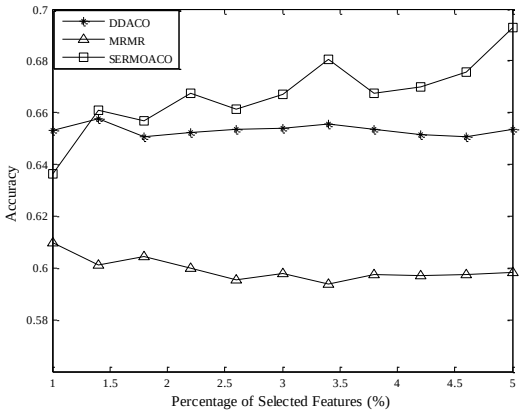

Fig. 4. Accuracy on Madelon_ER

[1] http://featureselection.asu.edu/datasets.php

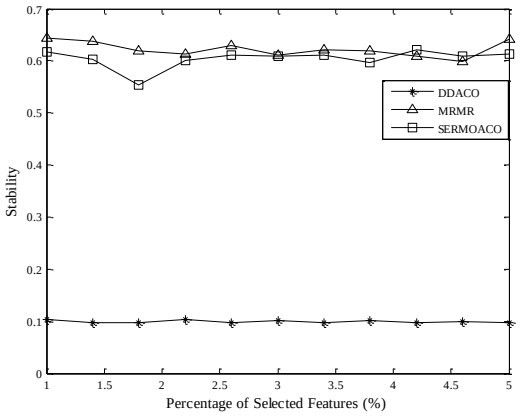

Fig. 5. Stability on Leukemia_ER

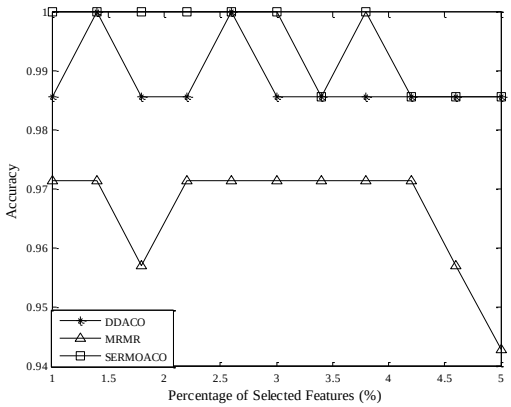

Fig. 6. Accuracy on Leukemia_ER

From fig. 3, we can find that the stability of SERMOACO is worse than MRMR but better than DDACO. It is because that MRMR is a filter feature selection method which is more stable for the small changes of training data. But fig. 4 shows that MRMR a lower accuracy than SERMOACO. And also we can conclude that SERMOACO is much better than DDACO both in stability and accuracy. DDACO and SERMOACO are both use ant colony optimization to select features. Swarm intelligence optimization algorithms are inherent random approaches which may get two distinct feature subsets if they run two times in the same data. But SERMOACO has a significantly better stability than DDACO. There are two reasons which lead to that. We adopt three stable filter feature selection to obtain the features' rankings of training data, and combine their results as guidance information to improve the stability of SERMOACO. Besides, the heuristic information which is gotten by fisher score and MIC can also improve the stability and classification performance of SERMOACO.

Fig. 5 and fig. 6 provide more information about the performance of SERMOACO. It is clear that the stability of SERMOACO is almost the same as MRMR, which means that our method has a better stability in a higher dimensional space. DDACO is still the worst in the view of stability. And fig. 6 shows that SERMOACO has the best classification ability though the three algorithms have some identical values in some situations.

5 CONCLUSIONS

In this paper, we propose a new approach called SERMOACO which is adopted for resolving entity resolution in high dimensional data. We can make some concludes through experiments.

(1) The proposed method can improve the classification performance for entity resolution in high dimensional data.

(2) Our approach has a better feature selection stability than DDACO and almost the same as MRMR in some situations.

Though the developed way is adopted for high dimensional data, it is also can be used in traditional entity resolution methods which were proposed for some problems which have a lot of features to further improve their performance.

An existing problem is that we haven't deployed it on real entity resolution system for validating its capability, and it is our future research.

ACKNOWLEDGEMENT

This work was supported by the Natural Science Foundation of China under Grant 61371196.

REFERENCES

Wang Q., Cui M., Liang H. 2016. "Semantic Aware Blocking for Entity Resolution,". *IEEE Transactions on Knowledge and Data Engineering* (28:1),pp. 166-180.

Yannik S., Jiang X., Volker K., Sebastian B. 2016. "Classification based Record linkage with Pseudonymized Data for Epidemiological Cancer Registries,". *IEEE Transactions on Multimedia,* (18:10),pp. 1929-1941.

Seyed M. R. B.,Boualem B., Srikumar V., Seung H. R., Hamid R. M., Wang W. 2017."A Systematic Review and Comparative Analysis of Cross Document Coreference Resolution Methods and Tools,". *Computing* (99:4),pp. 313-349.

Rabia N. T., Dmitri V. K., Sharad M. 2013. "Adaptive Connection Strength Models for Relationship based Entity Resolution,". *Data and Information Quality* (4:2),pp. 8:1-8:22.

Evandro F., Renata V., Aline V. 2016. "Improving Coreference Resolution with Semantic Knowledge," in *Proceedings of the PROPOR,* Springer, pp. 213-224.

Chai C., Li G., Li J., Deng D., Feng J. 2016. "Cost Effective Crowdsourced Entity Resolution: A Partial Order Approach," in *Proceeding of the SIGMOD,* pp. 969-984.

Masulli F, Rovetta S. 2015. "Clustering high-dimensional data,". *ACM Transactions on Knowledge Discovery from Data*, (3:1), pp. 1-58.

Cao J.J., Liu Y., Diao X.C., Zhang B., Peng C. 2016. "A New Design of Ensemble Classifiers for High-dimension Entity Resolution," in *Proceedings of the ICIQ 2016*.

Eckart Z., Lothar T. , Marco L., Carlos M. F., Viviane G. F. 2003. "Performance Assessment of Multiobjective Optimizers: An Analysis and Review," *IEEE Transactions on Evolutionary Computation* (7:2),pp. 117-132.

Chen Q. Y., Justin Z, Karin V. 2015"Evaluation of a Machine Learning Duplicate Detection Method for Bioinformatics Databases," in *Proceedings of the DTMBIO,* pp. 4-12.

Olga P., Michael F., Lior R., Yuval E. 2016. "Matching Entities Across Online Social Networks," *Neurocomputing* (210),pp. 91-106.

Clay S. 2016 "Test based Document Similarity Matching Using Sdtext," in *Proceeding of the 49th Hawaii International Conference on System Sciences,* pp. 5607-5616.

Christophe R., Tim C., Mark I. 2016. "Duplicate Detection in Facsimile Scans of Early Printed Music," in *Proceedings of the Analysis of Large and Complex Data Studies in Classification, Data Analysis, and Knowledge Organization,* Springer.

Lin M., Yang C., Lee C.Y., Chen C.C.. 2016 "Enhancements for Duplication Detection in Bug Reports with Manifold Correlation Features,". *Journal of System and Software* (121),pp. 223-233.

Vasilis E., Kostas S., Vassilis C. 2016 "Benchmarking Blocking Algorithms for Web Entities," *IEEE Transactions on Big Data,* (10:1109),pp 1-1.

I.D.I.D A., Fernando T.G.I. 2015. "Performance Analysis of the Multiobjective Ant Colony Optimization Algorithms for the Traveling Salesman Problem," *Swarm and Evolutionary Computation*(23),pp. 11-26.

Cao J.J., Zhang P.L., Wang Y.X., Ren G.Q., Fu J.P. 2008. "A Graph based Ant System for Subset Problems," *Journal of System Simulation* (20:22),pp. 6146-6150.

Adil F., Zahir T., Ibrahim K., Abdulmohsen A., Albert Y.Z. 2014. "An Optimal and Stable Feature Selection Approach for Traffic Classification Based on Multi-Criterion Fusion," *Future Generation Computer Systems* (36),pp.156-169.

Ghadah A., Wang W. 2015. "Weighted Heuristic Ensemble of Filters," in *Proceedings of SAI Intelligent Systems Conference,* London, UK, pp. 609-615.

Iman K., Sunil K. G., Dinh Q. P., Svetha V. 2016. “Stabilizing l1-norm Prediction Models by Supervised Feature Grouping,” *Journal of Biomedical Informatics* (59),pp. 149-168.

Barbara P., Nicoletta D., Marta A. 2017. “Exploiting the Ensemble Paradigm for Stable Feature Selection: A Case Study on High-Dimensional Genomic Data,” *Information Fusion* (35),pp. 132-147.

Hauke, J., Kossowski, T. 2011. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2), 87.

Zhang Y., Jia S.L., Huang H.Y., Qiu J.Q., Zhou C.J. 2014. “A Novel Algorithm for the Precise Calculation of the Maximal Information Coefficient,” *Scientific Reports* (4:4),pp. 6662.

Cao J.J., Diao X.C., Du Y., Wang F.X., Zhang X.Y. 2010. “Classification Detection of Approximately Duplicate Records based on Feature Selection Using Ant Colony Optimization,” *Acta Armamentrii* (31:9),pp. 1222-1227.

Peng H.C., Long F.H., Ding C. 2005. “Feature Selection based on Mutual Information Criteria of Max Dependency, Max Relevance and Min Redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* (27:8),pp. 1226-1238.