A new design of ensemble classifiers for high-dimension entity resolution

CAO JIANJUN*, Nanjing Telecommunication Technology Institute
LIU YI, PLA University of Science and Technology
DIAO XINGCHUN, Nanjing Telecommunication Technology Institute
ZHANG BIN, Nanjing Telecommunication Technology Institute
PENG CONG, Nanjing Telecommunication Technology Institute

Abstract: For high-dimension Entity Resolution (ER), a new design of ensemble classifiers based on feature selection is proposed, which regards ER as a binary classification problem. Binary classifier's classification performance and similarity measurement are defined, Support Vector Machine is adopted as the base classifier. Classification accuracy, output dissimilarity of classifiers and feature subset's number are used as optimization objects. Feature similarity vector of two candidate records is calculated as input data. Based on ER's characteristics, the multiobjective problem is translated into a single objective optimization and graph-based ant colony optimization to solve it. The proposed method is validated by high-dimension datasets.

• Technologies for IQ improvement and assurance→Data scrubbing and cleansing

Additional Key Words and Phrases: entity resolution, ant colony optimization, feature selection, support vector machine.

1. INTRODUCTION

Entity resolution (ER) is a very regular and important issue in data cleaning. ER's task is to find out the ambiguous denotations which refer to a unique entity [Tan Mingchao 2015]. ER's classical methods include FBS (Feature Based Similarity) and ReIDC (Relationship Based Data Cleaning) [Mugan et al. 2014], in which FBS is the earliest and essential way. It uses classifiers to compare the similarity values of each pair of records' attributes, and identify whether they are matched based on records' similarity measurement.

In recent years, a number of researchers have paid attention to FBS methods. For ER on incomplete information datasets, Qi Gu *et al*. (2014) introduced the concept of confidence as an effective complement to similarity to measure the trustworthy of the data records' similarity. They proposed an adaptive rule-based approach to calculate the similarity and Confidence Based Approach to focus on pairwise ER. Besides, they made confidence to propagate on the entity relational graph until six points are reached to dig deeper relationship between entities, similarity and confidence which achieved rock-solid performance. Doaa Medhat *et al*. (2015) developed a modified Cross-Language Levenshtein Distance algorithm for solving matching names across different writing scripts. They also used phonetic matching technique mixed with their algorithm to improve the overall f-measure and speed up the matching process after extracting phonetic and pattern similarity of two records. Liu Dong *et al*. (2015) investigated username's co-reference resolution in internet. They developed a method to link user identities across multiple websites relying only on usernames. They classified username features into surface features and comparison features after formulating the problem, and calculated the identification score to find out whether two usernames refer to the same owner. Zhou Xing *et al*. (2015) developed a multiple classifier system using resampling and ensemble selection. Instead of developing an empirical optimal resampling ratio, the proposed system varied the ratio in a range to generate multiple resampled data, and used the resampled data to train multiple classifiers and ensemble selection to select the best classifiers subset, which is also the best resampling ratio combination.

However, in big data era, high-dimension characteristic has become one of the most important issues, and it also brings some special problems in ER compared to former situations. As traditional FBS methods use some of the attributes, they may not be suitable and able to improve the system's classification accuracy when data's dimensions are high. Though some researchers have developed feature selection methods to select record's attributes, they used no more than 20 attributes [Kastner et al. 2013] which lost a lot of useful information in high-dimension data. In order to overcome those problems, we propose an ensemble classifiers, where we use Support Vector Machine (SVM) as base classifier for ER based on FBS after defining classification performance and similarity measurement of classifiers, then graph-based ant colony optimization is adopted to design classifiers' model whose outputs are combined through majority voting method. Some benchmark datasets are adopted to validate our method.

2. CLASSIFIER'S CLASSFYING PERFORMANCE AND SIMILARITY MEASUREMENT

2.1 Classifier's classifying performance measurement

The binary classifier's classification performance measurement for ER can be defined as follows:

Classification accuracy P

$$P = \frac{\text{Number of match samples}}{\text{Number of duplicate samples}} \times 100\% \quad (1)$$

The distribution matrix of output results of binary classifier is defined as:

$$p = [p_{ii'}], i, i' = 1, 2 \quad (2)$$

Where $p_{ii'}$ represents probability of class i labeled as class i' wrongly, it has two labels named 1 and 2.

In equation (2), $p_{ii'}$ is defined as:

$$p_{ii'} = \frac{\text{Number of samples of class } i \text{ classified } i'}{\text{Number of samples of class } i} \times 100\% \quad (3)$$

If $p_{ii}\,(i = 1, 2)$ represents the classification accuracy of samples of class i, then it can be calculated by (4)

$$p_{ii} = 1 - p_{ii'} \quad (4)$$

So classification accuracy P can be computed by (5)

$$P = P_i p_{ii} + P_{i'} p_{i'i'} \quad (5)$$

Where P_i is the prior probability of class i. Given testing sample sets, P_i can be figured out by (6)

$$P_i = \frac{N_i}{N_i + N_{i'}} \quad (6)$$

Where N_i is the number of sample of class i, and $N_{i'}$ is the number of sample of class i'.

2.2 Similarity measurement

Compared to any local single classifier, ensemble classifier owns a higher accuracy of classification in most cases [Whalen & Pandev 2013]. In this paper, we use similarity

measurement (similar to diversity [Yu & Ni 2014])between classifiers to design base classifier.

Given a sample set, if the training and testing samples are the same, and the same type of classifier Λ (SVM in this paper) has the same function (i.e. mapping the same states' samples to the same class space) and parameters (i.e. σ and C are the same in SVM), then for a feature subset "*subset*", the corresponding samples' feature similarity vectors will be determined. The "*subset*" can be mapped into a certain classifier Λ_{subset} and obtain an output distribution matrix p after using training samples and testing samples(feature similarity vectors) to train and test classifiers (determine α_l^* and w_0^* in SVM). We have

$$\Lambda(subset) = (\Lambda_{subset}, p) \tag{7}$$

The similarity of classifier Λ_{subset} can be measured by similarity degree of "*subset*" and p called input similarity and output similarity respectively.

Definition 1 (Input similarity of classifiers). The similarity of classifiers' input feature subsets is called input similarity of classifiers, which can be calculated by Tanimoto distance [Li et al. 2010]. Supposing $subset_1$ and $subset_2$ are two nonempty binary classifiers' input feature subsets, we have

$$S_t(subset_1, subset_2) = 1 - \frac{|subset_1| + |subset_2| - 2|subset_1 \cap subset_2|}{|subset_1| + |subset_2| - |subset_1 \cap subset_2|} \tag{8}$$

In which $S_t \in [0,1]$ .When $S_t = 0$, it means that there is no identical element between two subsets; When $S_t = 1$, it means that the two subsets are the same, and the binary classifiers trained by the corresponding training samples are also the same. I.e., if S_t is closer to 1, the two subsets will have stronger similarity degree, and two binary classifiers' input similarity degree is stronger.

Definition 2 (Output similarity of classifiers). The similarity degree of classifiers' output distribution matrix is called output similarity of classifiers, which can be measured by normalized Pearson's correlation coefficient. Supposing two binary classifiers' output distribution matrices are $p' = [p'_{ii'}], p'' = [p''_{ii'}]$ respectively, where $i = 1,2$, $i' = 1,2$, we have

$$S_c(p', p'') = \frac{1}{2}\left(1 + \frac{\sum_{i=1}^{2}\sum_{i'=1}^{2}(p'_{ii'} - \overline{p'})(p''_{ii'} - \overline{p''})}{\sqrt{\sum_{i=1}^{2}\sum_{i'=1}^{2}(p'_{ii'} - \overline{p'})^2 \sum_{i=1}^{2}\sum_{i'=1}^{2}(p''_{ii'} - \overline{p''})^2}}\right) \tag{9}$$

Where $\overline{p'}, \overline{p''}$ is the mean value of p', p'' respectively, which can be written

$$\overline{p'} = \frac{1}{4}\sum_{i=1}^{2}\sum_{i'=1}^{2} p'_{ii'} \tag{10}$$

Where $S_c \in [0,1]$. When $S_c = 1$, it means the two matrices are fully positive correlation, and the corresponding classifiers' output distribution matrices are the same; When $S_c = 0$, it means the two matrices are fully negative correlation, and the corresponding classifiers' output distribution matrices have the worst similarity.

LEMMA 1. If $\Lambda(subset_1) = (\Lambda_{subset_1}, p_1)$, $\Lambda(subset_2) = (\Lambda_{subset_2}, p_2)$, and $S_c(p_1, p_2) < 1$, then $S_t(subset_1, subset_2) < 1$.

PROOF. Supposing $subset_1 = subset_2$, based on equation (12) and assumptions, we have $p_1 = p_2$, i.e., if $S_t(subset_1, subset_2) = 1$, then $S_c(p_1, p_2) = 1$. So the assumption is wrong and lemma is true.

From lemma 2.3, the condition of output dissimilarity of classifiers is stronger than the input dissimilarity, so we can use output similarity of classifiers to measure the classifiers' similarity.

3. DESIGN CLASSIFIERS BASED ON FEATURE SELECTION

Kankanala *et al*. (2014) proposed AdaBoost algorithm to design ensemble classifiers, which selected training samples' feature subsets to train and designed classifiers to obtain different classifiers fit for different samples.

The method we propose for designing binary classifiers' model based on feature selection for ER can be described as follows : Given ensemble classifiers composing of L binary classifiers, and P_l denotes classification accuracy of classifier l, and q_l denotes input feature subsets' number of classifier l, then the object fuction for classifier l to select input feature subsets and construction way can be determined by follows :

$$\max P_l \tag{11}$$

$$\max\{1 - \max_{j=1}^{l-1}\{S_c(p_j, p_l)\}\} \tag{12}$$

$$\min q_l \tag{13}$$

It means that we hope the classifier l has the highest classification accuracy, and the biggest dissimilarity from other classifiers and the least number of feature subsets. That is a multiobjective combinatorial optimization problem, and the priority of those three objects is descending.

There is no global optimal solution in multiobjective problem with respect to all objects generally. There are a set of solutions that are superior to the rest of the solutions in the search space when all objects are considered, but they are inferior to other solutions in the space in one or more objects. These solutions are known as Pareto-optimal solutions or nondominated solutions [Zitzler et al. 2010].

Ant colony optimization is a classical meta-heuristic algorithm, and it is developed fast and widely used in recent years especially in resolving complicate multiobjective combinatorial optimization problems. Cao Jianjun *et al.* (2008) developed a graph-based ant system for resolving subset problems. They defined the structure graph and equivalent routes, and proposed a new updating pheromone policy based on strengthening the pheromone on equivalent routes which balanced the convergence speed and searching ability. In this paper, we use it to design binary classifiers' model. Based on problem's characteristic, the algorithm's heuristic information η_k is defined as Fisher standard discriminant rate of k^{th} similarity feature:

$$\eta_k = \frac{\left|\bar{\mu}_{1k} - \bar{\mu}_{2k}\right|}{\sqrt{\sigma_{1k}^2 + \sigma_{2k}^2}} \tag{14}$$

Where $\bar{\mu}_{1k}, \bar{\mu}_{2k}$ denotes k^{th} similarity feature's mean value while σ_{1k}^2, σ_{2k}^2 denotes k^{th} similarity feature's variance value in match and non-match class respectively. Equation (14) shows that the similarity feature value owning a higher Fisher discriminant rate has a prior selected chance, which means these samples are easy to distinguish.

Based on the previous analysis, we can design model to construct classifier l as follows:
i) A multiclass classifier has better computation efficiency and classification accuracy when its feature subset's number q is between 5 to 10 [Fung et al. 2011]. Though each classifier is binary classifier in our method which can obtain high accuracy through less feature subsets, our data has a high-dimension characteristic. For the objective function (13), we can vary q from 10 to 20.
ii) For a fixed q, the objective functions (11) and (12) can be translated into single objective function equation (15) by weights' aggregation:

$$\max \alpha_1 P_l + \alpha_2 (1 - \max_{j=1}^{l-1} \{S_c(p_j, p_l)\}) \quad (15)$$

Where $\alpha_1 > 0, \alpha_2 > 0, \alpha_1 + \alpha_2 = 1$. When $P_l \to 1, l = 1,2,\cdots,L$ the equation (12) result is approaching 0(i.e. $(1 - \max_{j=1}^{l-1} \{S_c(p_j, p_l)\}) \to 0$), so there is a conflict between (11) and (12). In order to avoid obtaining high dissimilarity between classifiers leading to decreasing the accuracy in return, we need to set a high α_1 in model because our goal is to get a higher classification accuracy, so we set α_1=0.7, α_2=0.3 in our experiments. Each ant constructs a path and a solution of (15), and algorithm's parameters are updated by comparing the values of objective function (15).
iii) Supposing P_b and q_b denotes the global optimal objective solutions of classifier l after designing former classifiers, and $P_{k'}$ and $q_{k'}$ denotes current optimal objective solutions after using all q values to search, then the solutions' evaluation method can be described as follows: If $P_{k'} > P_b$ or $P_{k'} = P_b$, $q_k = q_b$, then using current optimal solution to replace global optimal solution; Otherwise, the current optimal solution is inferior to global optimal solution, do nothing.
For classifier l, the final global optimal solutions contain optimal feature subsets and the training results. The ensemble binary classifiers' results are determined through majority voting.

4. EXPERIMENT AND DISCUSSIONS

4.1 Data and pre-processing

In order to test our method on high-dimension datasets, we use two datasets warpAR10P and warpPIE10P provided from [Hassanzadeh et al. 2009]. warpAR10P has 130 samples whose dimensions are 2400, which has been assigned 10 labels and each label has 13 samples. warpPIE10P has 210 samples whose dimensions are 2420, which has been classified 10 labels and each label has 21 samples.
For a given dataset, we choose two different records in the same label as a pair of similarity records, and two records in different labels as a pair of dissimilarity records. Feature similarity vector of two records is calculated through small number divided by large number in corresponding dimension which constrains value under 1.
For a test and each dataset, we choose 140 pairs of similarity records and 200 pairs of dissimilarity records as the first data and another 100 pairs of similarity records and 120 pairs of dissimilarity records as the second data. For the first data, the 340 feature similarity vectors are calculated as training data, and 220 feature similarity vectors as testing data for the second data.

4.2 Results and discussion

As discussed before, we choose SVM as the base classifier, and set its parameters as default set as follows: Kernel function is 'rbf', δ=0.4 and C=100. The number of SVM is set to be 5.

The parameters of ant colony optimization are set as follows: initial pheromone $\tau_{ij}(0)$=100, the important factor of pheromone and heuristic information α=1, β=1, pheromone decay coefficient ρ=0.2, evaporation rate Q=0.02, ant's number is 20, the generation's number 40. The feature subset's searching bounds vary from 10 to 20. We compare our method with a single SVM, and use classification accuracy to evaluate results. Every time we choose the same number as discussed but different samples from each datasets. Algorithm 1 denotes our algorithm, Algorithm 2 denotes compared algorithm, and some simulation results are showed in table 1 and 2.

Table I. warpAR10P Results

No.	Algorithm 1		Algorithm 2	
	Feature Subsets	Classification Accuracy	Feature Subsets	Classification Accuracy
1	12/17/12/12/13	0.927	12	0.877
2	16/16/15/16/15	0.926	15	0.821
3	19/15/17/15/17	0.991	19	0.878
4	14/11/13/16/18	0.951	16	0.896
5	12/13/18/10/18	0.975	16	0.901
6	12/19/10/16/16	0.966	13	0.893
7	15/19/11/13/14	0.995	10	0.880
8	12/18/18/17/18	0.951	12	0.834

Table 2. warpPIE10P Results

No.	Algorithm 1		Algorithm 2	
	Feature Subsets	Classification Accuracy	Feature Subsets	Classification Accuracy
1	15/19/14/13/15	0.994	17	0.913
2	16/16/15/16/15	0.926	16	0.821
3	12/16/17/14/19	0.989	12	0.875
4	12/11/12/12/12	0.928	12	0.886
5	15/19/19/18/12	0.964	15	0.846
6	12/18/12/10/17	0.968	14	0.915
7	19/19/18/19/16	0.951	19	0.906
8	18/11/11/15/14	0.988	17	0.916

Table 1 and 2 show that our proposed method improves classification accuracy obviously. Though feature subset's searching bounds are constrained less than 20, none of 5 SVMs selected more than 19 features, which demonstrates that only few features are enough to classify the records despite thousands of dimensions. On the other hand, the feature subsets' number is different from each other in 5 SVMs, which indicates that the model utilizes high-dimension data's information from different combination of features. Though the two tables did not list the contents of each feature subsets because of limited space, the features in different subsets are not the same despite their same number, which demonstrates our model's validity from another point of view.

5. CONCLUSIONS

To efficiently resolve high-dimension ER's problem, we regarded it as a binary classification problem. We defined binary classifier's classification performance measurement and similarity measurement index, and constructed an ensemble classifiers model which used SVM as the base classifier. In order to design each classifier's model, we used ant colony optimization to resolve it. Two high-dimension datasets were adopted to evaluate our proposed method, the conclusions are

summarized as follows: We obtain an expectable results that the classification accuracy is improved compared with a single SVM, which could guide our future research in high-dimension ER; The designed ensemble classifier model selects different feature subsets which utilize attributes' information efficiently compared with a single feature subset; Our model regards ER's feature selection as a multiobjective problem whose objects are highest classification accuracy, and highest output dissimilarity of classifiers and least feature subset's number, which balances classification's efficiency and accuracy.

In the future, we intend to have a research on the further issue: how many feature subsets and classifiers is the best choice based on different datasets.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China under Grant 61371196, the China Postdoctoral Science Special Foundation under No. 201003797, the China Postdoctoral Science Foundation under No. 2015M582832 and No. 20090461425, the Jiangsu Post-doctor Research Fund of China under No. 1402138C.

REFERENCES

Tan, M. C. 2015. The key technologies for entity resolution. (Doctoral dissertation, PLA University of Science and Technology, NanJing).

MUGAN, J., CHARI, R., HITT, L., MCDERMID, E., SOWELL, M., QU, Y., and COFFMAN, T., 2014. Entity resolution using inferred relationships and behavior. In *2014 IEEE International Conference on Big Data,* 555-560.

GU, Q., ZHANG, Y., CAO, J., and XU, G., 2015. A confidence-based entity resolution approach with incomplete information. In *International Conference on Data Science and Advanced Analytics*, 97-103.

MEDHAT, D., HASSAN, A., and SALAMA, C., 2015. A hybrid cross-language name matching technique using novel modified Levenshtein Distance. In *Tenth International Conference on Computer Engineering & Systems*.

Liu, D., Wu, Q., Han, W. and Zhou, B. 2015. User identification across multiple websites based on username features. *Chinese Journal of Computers, 38*, 10, 2028-2040.

ZHOU, X., DIAO, X., and CAO, J., 2015. A High Accurate Multiple Classifier System for Entity Resolution Using Resampling and Ensemble Selection. *Mathematical Problems in Engineering 2015*, 2, 1-6.

KASTNER, S., CHOI, S.-P., and JUNG, H., 2013. Author Name Disambiguation in Technology Trend Analysis Using SVM and Random Forests and Novel Topic Based Features. In Green Computing and Communications (GreenCom), 2013 *IEEE and Internet of Things* (*iThings/CPSCom*), *IEEE International Conference on and IEEE Cyber*, *Physical and Social Computing IEEE*, 2141-2144.

WHALEN, S. and PANDEY, G., 2013. A Comparative Analysis of Ensemble Classifiers: Case Studies in Genomics. In *2013 IEEE 13th International Conference on Data Mining*, 807-816.

NI, J., 2014. An Improved Ensemble Learning Method for Classifying High-Dimensional and Imbalanced Biomedicine Data. *IEEE/ACM Transactions on Computational Biology & Bioinformatics 11*, 4, 657-666.

KAI-QI, L.I., DIAO, X.C., CAO, J.J., and FENG, L.I., 2010. High precision method for text feature selection based on improved ant colony optimization algorithm. *Jiefangjun Ligong Daxue Xuebao/journal of Pla University of Science & Technology 11*, 6, 634-639.

KANKANALA, P., DAS, S., and PAHWA, A., 2014. ADABOOST(+): An Ensemble Learning Approach for Estimating Weather-Related Outages in Distribution Systems. *IEEE Transactions on Power Systems 29*, 1, 359-367.

ZITZLER, E., THIELE, L., LAUMANNS, M., FONSECA, C.M., and FONSECA, V.G.D., 2003. Performance assessment of multiobjective optimizers: an analysis and review. *IEEE Transactions on Evolutionary Computation 7*, 2, 117-132.

CAO, J.J., ZHANG, P.L., WANG, Y.X., REN, G.Q., and JIAN-PING, F.U., 2008. Graph-based Ant System for Subset Problems. *Journal of System Simulation 20*, 22, 6146-6150.

FUNG, P.C.G., MORSTATTER, F., and LIU, H., 2011. *Feature Selection Strategy in Text Classification*. Springer Berlin Heidelberg.

HASSANZADEH, O., CHIANG, F., LEE, H.C., MILLER, R., and J., E., 2009. Framework for evaluating clustering algorithms in duplicate detection. *Proceedings of the Vldb Endowment 2*, 1, 1282-1293.