Front. Comput. Sci., 2019, 13(1): 215–217
https://doi.org/10.1007/s11704-018-7066-4

LETTER

Statistical relational learning based automatic data cleaning

Weibang LI (✉)[1], Ling LI[2], Zhanhuai LI[3], Mengtian CUI[1]

1 School of Computer Science and Technology, Southwest Minzu University, Chengdu 610041, China
2 Archives of Southwest Minzu University, Chengdu 610041, China
3 College of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

1 Introduction

Data in real world is usually dirty, i.e., it may contain inconsistent, noisy, incomplete or duplicated values. Generally speaking, the identified dimensions of data quality management can be summarized as the following five types: accuracy, consistency, completeness, timeliness and reliability [1]. Dirty data can lead to incorrect conclusions and false decisions on both public and private scales (see Wikipedia). Thus it is urgent and crucial for organizations to improve the quality of data or to clean the data efficiently.

There are some approaches to improve the accuracy or efficiency of data cleaning by leveraging statistical and learning methods such as Bayesian belief network [2] and machine learning [3]. Although these existing data cleaning approaches are effective in their own scenarios, the drawbacks of these approaches are obvious: 1) State of the art approaches (e.g., [4]) depend on the involvement of human experts or external information to detect and repair data errors. 2) Many other approaches (e.g., [5]) depend on the availability of off-the-shelf patterns/rules when repairing the errors. 3) Some approaches (e.g., [2]) depend on the availability of a clean data table to learn data quality patterns/rules or learn data quality patterns/rules directly from the noisy data. Though several approaches can learn data quality patterns/rules from data, the accuracy of data cleaning is not high due to the lack of inference.

Received February 23, 2017; accepted September 18, 2018

E-mail: wbli2003@163.com

In this paper, we focus on the problem of cleaning dirty data without either existing data quality patterns/rules or involvement of human experts. We proposed an unsupervised data cleaning method based on statistical relational learning. Firstly, we learn a model of data in the form of Bayesian network [2], which reflects the dependency relationships between different attributes of the database table. Then, we translate the dependency relationships between attributes into first-order logic formulas, and convert first-order logic formulas into Markov logic networks by assigning a weight for each formula. Secondly, we transform the Markov logic networks into DeepDive inference rules and execute these rules on DeepDive platform. The results of inference are used to estimate the most likely repairs of dirty in data.

The main contributions of this paper are summarized as follows:

1) We present an unsupervised data cleaning framework based on statistical relational learning. Our approach involves converting the dependency relationships between attributes into Markov logic networks and inference on DeepDive platform.
2) We propose an algorithm to generate first-order logic formulas based on the dependency relationships between attributes, and present an approach to calculate the weights of first-order logic formulas based on the mutual information involved in the formulas.
3) The final contribution of this paper is an experimental study. We conduct several experiments to evaluate the accuracy and applicability of our approach. The experiments are performed on real-world datasets. We show

that our approach has higher accuracy in terms of different situations and is universal for different kinds of datasets.

The technical details, proofs and evaluations can be found in the support information.

2 Unsupervised data cleaning

Owing to the absence of explicit data quality patterns/rules, we propose to learn the dependencies between the attributes of data table and obtain a Bayesian network of the attributes. We transform the Bayesian network into Markov logic network and calculate the weight of each formula based on the mutual information of the attributes involved in the formula. We transform the Markov logic networks into DeepDive inference rules and execute these rules on DeepDive platform, then estimate the most likely repairs of dirty in data based on the inference results of DeepDive.

Figure 1 shows the framework of our data cleaning approach. We generate preliminary data quality rules from raw data set or the sample of raw data set in accordance with the volume of data. In the component of rules management, we transform the preliminary data quality rules into Markov logic network and construct DeepDive inference rules based on Markov logic network. In the component of statistical processing, we estimate the most likely data repairs by leveraging inference on DeepDive. The estimated data repairs can be used to clean the original dirty data.

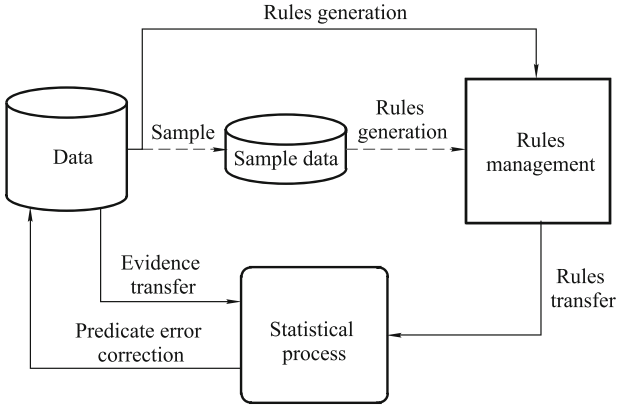

Fig. 1 Framework of the unsupervised data cleaning approach

In this paper we use GeNIe Modeler to learn the numerical parameters of Bayesian network. Ultimately, we get a complete Bayesian network and we generate data cleaning rules based achieved Bayesian network.

We generate the data cleaning rules based on first-order logic at first.

For a relation R with attributes set Attrs(R)=$\{A_1, A_2, \ldots, A_m\}$, we define the atomic sentences $attr\text{-}A_1(id, v_{A_1}), \ldots, attr\text{-}A_m(id, v_{A_m})$, where $attr\text{-}A_i(id, v_{A_i})$ means that the attribute value of the ith attribute A_i in Attrs(R) of the tuple with key=id in R is v_{A_i}. We define the relation constants as follows:

Equality: $equal\text{-}A(id_1, id_2)$ signifies that the attribute value of A in R of the tuple with key=id_1 equals that of the tuple with key=id_2.

Matching: $match\text{-}A(id_1, id_2)$ means that the attribute values of A in R of the tuple with key=id_1 and the tuple with key=id_2 are matched.

we specify the first-order logic formulas of a learned Bayesian network from data. Assuming that there is a directed edge between attribute A_1 and A_2 and pointing from A_1 to A_2, we formalize the dependency relationship between A_1 and A_2 in the form of first-order logic as follows:

$attr\text{-}A_1(id_1, v) \wedge attr\text{-}A_1(id_2, v) \Rightarrow equal\text{-}A_2(id_1, id_2)$

Here v is the value of attribute A_1 in the tuples of id_1 and id_2.

In another case, if there are more than one attribute pointing to the same attribute together, for instance, attribute A_1, $A_2, \ldots, A_i$ point to attribute A_j, then the dependency relationship between $A_1, A_2, \ldots, A_i$ and A_j can be formalized in the form of first-order logic as follows:

$attr\text{-}A_1(id_1, v_{A_1}) \wedge attr\text{-}A_1(id_2, v_{A_1}) \wedge attr\text{-}A_2(id_1, v_{A_2}) \wedge attr\text{-}A_2(id_2, v_{A_2}) \wedge \cdots \wedge attr\text{-}A_i(id_1, v_{A_i}) \wedge attr\text{-}A_i(id_2, v_{A_i}) \Rightarrow equal\text{-}A_j(id_1, id_2)$

Here $v_{A_1}, v_{A_2}, \ldots, v_{A_i}$ are the values of the attributes $A_1, A_2, \ldots, A_i$ in the tuples of id_1 and id_2.

We assign infinite weights to hard rules in the experiments, whereas set the weights of soft rules to positive real numbers. The weight of a soft rule is positively related to the degree of dependence between the attributes involved in the soft rule. We learn the weights of soft rules based on the information theory in this paper. The degree of dependence can be measured by the mutual information between the attributes. The mutual information of two attributes X, Y can be expressed in terms of divergence between marginal and conditional probability distributions of the values of X and Y as follows [6]:

$$I(X;Y) = \sum_{x \in V_X} \sum_{y \in V_Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}. \quad (1)$$

Here V_X is the set of X's values and V_Y is the set of Y's values, $P(x)$ and $P(y)$ are the marginal probability distributions of X and Y, and $P(x, y)$ is the joint probability distribution of X and Y.

Similarly, the conditional mutual information of three attributes X, Y and Z can be expressed as follows:

$$I(X;Y|Z) = \sum_{x\in V_X}\sum_{y\in V_Y}\sum_{z\in V_Z} P(x,y,z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)}. \quad (2)$$

Here $P(x,y,z)$ is the joint probability distribution of X, Y and Z, $P(x,y|z)$ is the conditional probability of X and Y given condition Z, and $P(x|z)$ is the conditional probability of X given condition Z, etc.

For each soft rule F with more than two attributes involved, assuming that there are n+1 attributes X, Y_1, Y_2, …, Y_n and X depends on Y_1, Y_2, …, Y_n. The mutual information of X given Y_1, Y_2, …, Y_n is $I(X;Y_1,Y_2,\ldots,Y_n)$. We define the weight of F based on the e exponential of $I(X;Y_1,Y_2,\ldots,Y_n)$ as follows:

$$w_F = e^{I(X;Y_1,Y_2,\ldots,Y_n)} - 1. \quad (3)$$

Now that $I(X;Y) \geqslant 0$ and $I(X;Y_1,Y_2,\ldots,Y_n) \geqslant 0$, thus $e^{I(X;Y)} - 1 \geqslant 0$ and $e^{I(X;Y_1,Y_2,\ldots,Y_n)} - 1 \geqslant 0$. The introduction of e exponential function in the weights calculating enforces the weights of soft rules with higher degree of dependence and enlarges the gap of weights with different dependence degrees meanwhile.

Markov logic network constructs a probabilistic knowledge base system that combines first-order logic formulas with probabilistic graphical model (undirected Markov networks). We perform probabilistic inference on top of the Markov networks.

In this paper we perform inference on DeepDive framework. An example of DeepDive rule is as follows:

$$q() : -S(x,y), weight = w, \quad (4)$$

where $S(x,y)$ is the body predicate of $q()$, and x, y are variables in $S(x,y)$. $S(x,y)$ can be grounded by replacing the variables with constants in the possible world of variables.

3 Experimental results

We evaluate the cleaning quality, scalability, and sensitivity to error types of our algorithm on two real-world datasets. To measure these three kinds of errors, we use the notions of *Precision*, *Recall* and *F-measure*. Our experiments were conducted on two real-life datasets as follows: HPT dataset is a real-world dataset published by the U.S. Centers for Medicare & Medicaid Services. CAR dataset is a real-life dataset that contains information on sales, prices and characteristics of the car models sold in Europe during 1970–1999.

We select a Bayesian data cleaning method [7] and a data repairing method [8] based on statistical technique as the baselines and marked as BAYC, BYWP respectively, while mark our data cleaning method as ADCS. We assess the scalability of our approach by varying sizes of datasets from 10K to 50K tuples when running the experiments. We leverage DeepDive as the inference engine of Markov logic rules.

The experimental results demonstrate the following: (i) ADCS outperforms BAYC and BYWP significantly in *Precision*, *Recall* and *F-measure* respectively on both datasets of different ratio noi% of noise ranging from 2% to 10%; (ii) ADCS outperforms BAYC and BYWP distinctly in *Precision*, *Recall* and *F-measure* on different datasets sizes ranging from 10K to 50K; (iii) the employment of e exponential function in calculating the weights of formulas is conducive to the improvement of accuracy; and (iv) ADCS_T surpasses ADCS_R remarkably in accuracy on different datasets sizes.

Acknowledgements The work was supported by the Fundamental Research Funds for the Central Universities, Southwest Minzu University (2018NQN32), the National High Technology Research and Development Program 863 of China (2015AA015307), the National Natural Science Foundation of China (Grant Nos. 61672432, 61702161), the Key Research and Development and Promotion Program of Henan Province of China (182102210213), and the Foundation of Henan Educational Committee (18A520003).

References

1. Carlo B, Monica S. Data Quality: Concepts, Methodologies and Techniques. Berlin: Springer Publishing Company, 2006
2. Doshi P, Greenwald L, Clarke J R. Using Bayesian networks for cleansing trauma data. In: Proceedings of the 6th International Florida Artificial Intelligence Research Society Conference. 2003, 72–76
3. Yakout M, Elmagarmid A K, Neville J, Ouzzani M, Ilyas I F. Guided data repair. Proceedings of the VLDB Endowment, 2011, 4(5): 279–289
4. Wang J, Kraska T, Franklin M J, Feng J. Crowder: crowdsourcing entity resolution. Proceedings of the VLDB Endowment, 2012, 5(11): 1483–1494
5. Fan W, Geerts F, Jia X, Kementsietsidis A. Conditional functional dependencies for capturing data inconsistencies. Journal of ACM Transactions on Database Systems, 2008, 33(2): 1–48
6. Smyth P, Goodman R M. Rule induction using information theory. In: Proceedings of the International Conference on Knowledge Discovery in Databases. 1991, 159–176
7. Hu Y, De S, Chen Y, Kambhampati S. Bayesian data cleaning for Web data. 2012, arXiv preprint arXiv:1204.3677
8. De S, Hu Y, Meduri V, Chen Y, Kambhampati S. Bayeswipe: a scalable probabilistic framework for improving data quality. Journal of Data and Information Quality, 2016, 8(1): 1–30