AN APPROACH USING HIDDEN MARKOV MODEL FOR ESTIMATING AND REPLACING MISSING CATEGORICAL DATA

(Research-in-Progress)

Jianjun Cao
Nanjing Telecommunication Technology Institute, Nanjing, China
jianjuncao@yeah.net

Xingchun Diao
Nanjing Telecommunication Technology Institute, Nanjing, China
diaoxch640222@163.com

Ning Zhang
Nanjing Telecommunication Technology Institute, Nanjing, China
cjj_8@163.com

Ting Wang
Nanjing Telecommunication Technology Institute, Nanjing, China
wangting_sohu@sohu.com

Abstract: In order to process missing data, we propose a statistical relational learning approach for estimating and replacing missing categorical data. First, for a given data set, all categorical attributes are classified as a proper number of groups, and these groups are independent of each other. Second, principles for ordering attributes in one group are proposed and the attribute sequence of the group could be indexed by the principles. Third, a hidden Markov model for estimating missing categorical value is represented. According to complete record samples, probabilities of missing value belonging to each possible value are estimated by the model. The missing value can be replaced through referring to the probabilities. Finally, the implement process of the proposed approach is illustrated by an example.

Key Words: Data Quality, Data Cleaning, Missing Value, Incompleted Record, Hidden Markov Model, Statistical Relational Learning

1. INTRODUCTION

The data lower integrity is one of common forms of poor-quality data, and the treatment of this issue is an important task for improving data quality [3, 10]. For a relational data set, its hierarchical structure of the lower integrity is shown in Figure 1.

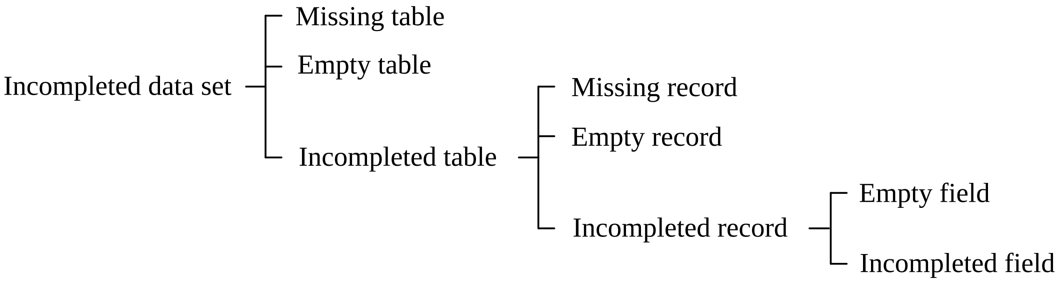

Figure 1 The hierarchical structure sketch map for the lower integrity of data set content

In figure 1, missing table means that some table is not to be created. Empty table means that there is no physical record in the table. Missing record indicates that the physical record is left out, when the table is not empty. Empty record indicates that the record fields are empty except those fields that are required ones or primary key. Empty field is the field whose value is not to be filled. Incompleted field means that the content of the field value is lower integrity. For example, Chinese name field is filled with just one Chinese character (a complete Chinese name is consisted of no less than two Chinese characters (one is surname and another is firstname)). Treatment approach for incompleted field is not included in this paper.

Just as shown in Figure 1, the lower integrity of data set content could own quite a few different forms, however, in general, field is the basic operation unit in relational data set. A record, which is the basic unit for completely describing data object, may include several fields. So, the lower integrity of relational data set mainly indicates incompleted records with missing fields.

We call the record with empty field as uncompleted record. Incompleted record may be empty record or record with just one empty field. Generally, incompleted records and empty fields are all called missing data, and attribute values of empty fields are called missing values.

In data mining, data warehouse and data quality management, missing data is an important aspect which affects data quality. Missing data are often included in real data. Missing values can lead to serious problems when the data is used for reporting, information sharing and decision support [11]. First, the data with missing value can provide incomplete information. For example, a survey question that is related to personal information will be more like to be left unanswered for those who are more sensitive to privacy. Second, many data modeling and analysis techniques cannot deal with missing values and have to cast out a whole record if one of the attribute values is missing [12, 16]. In some cases, this treatment approach is wasteful of data resources. Third, even though some data modeling and analysis tools can handle missing values, there are often restrictions in the domain of missing values. For example, classification systems typically do not allow missing values in the class feature [1, 14].

According to various application backgrounds, the requirements of missing data treatments are different. In many statistical techniques and some early machine learning systems, whole record with any empty field is disregarded [15, 16]. This kind of approaches are simple, but the main problem is that it will cause significant loss of the available information if the dataset contains many missing values. Another approach is to replace missing values with a default value or a global constant. This approach is widely adopted in the database community. For example, in relational database systems， it is a common practice to replace a missing value with a null value (labeled by “Null”) [7]. While this approach is useful in resolving some database problems such as the referential integrity issue, it is not very helpful when the data is used for analysis purposes.

In a representative information system, missing data treatment will become complex because of the limitation of specific domain background. From the perspective of data cleaning, missing data treatment includes data detection, classification, estimating and replacing missing values. These three steps are corresponding to three kinds of typical techniques in data cleaning. They are data detection, data analysis and data modification [2, 3].

When we detect incompleted record, complete records and incompleted records should be separated firstly. In this paper, incompleted records are classified as three classes, including incompleted and unmodifying records, incompleted and modifying records, incompleted and deleting records. Brief description of the classes is listed in Table 1.

In table 1, complete record indicates the record without empty field. Incompleted and unmodifying records includes incompleted records with complete information, that is, though the record contains empty fields, it still can describe the object wholly (For example, in unmarried personal record, fields about mate information are empty); also includes records that some available fields are not empty, while those empty fields don’t impact current application, or have no modifying value, or can’t be modified presently. Incompleted and modifying records indicate that incompleted records have some missing

NO.	Class name	Description
1	Complete records	Records without empty field.
2	Incompleted and unmodifying records	Records with empty field but can satisfy the specific

		application.
3	Incompleted and modifying records	Records with empty field and need to be modified.
4	Incompleted and deleting records	Empty records, or incompleted records has no modifying value.

Table 1 Record classes and brief description

values need to be estimated and replaced. Incompleted and deleting records include empty records, useless records which can’t be used because their important fields are empty, and records have no modifying value because they have too many empty fields. Above three classes of incompleted records can be defined by domain specialists.

In practice, we often use simple estimates to fill the empty fields in incompleted and modifying records. Usually, if the missing value is of a numeric type, the mean of the nonmissing values for the same field is used as the estimate; if missing value is categorical, the mode value (most frequent value) is used [12, 15, 6, 13]. This approach is convenient and provides a satisfying solution to missing data in many cases. A major problem is that the variability associated with missing data is biasedly represented when all missing values of an field are replaced with the same value. As a result, the statistical distribution of the data is altered and may impact data quality.

In order to resolve missing data issue better, some other advanced approaches have been proposed. These approaches were often applied to some specific data analysis task. For classifying analysis, a surrogate split method was proposed in reference [1], and a probabilistic weighting method was proposed in reference [14]. For numeric data analysis, methods based on regression techniques were discussed in reference [9, 13, 17].

Bayesian methods are important approaches for dealing with missing categorical data [8]. Chiu and Sedransk [5] proposed a Bayesian method for estimating and replacing missing data based on some prior knowledge about the distributions of the data. However, this method primarily was applied to univariate missing data and some special multivariate cases. Chen and Astebro [4] developed another Bayesian method for estimating and replacing missing categorical data, using the uniform prior distribution and a Dirichlet posterior distribution. Their method performed very well when the missing data is missing at random, but it remains to be tested for cases where data is missing not at random. Li [11] proposed a simple Bayesian approach for estimating and replacing missing categorical data. With this approach, the posterior probabilities of a missing attribute value belonging to a certain category are estimated. The approach is nonparametric and does not require prior knowledge about the distributions of the data. However, when the approach estimates missing values of any empty field, it must use all the other unmissing categorical values, and those fields which are irrelevant to the empty field are also included. For relational data, the hypothesis that the attributes are conditionally independent of each other under a given class value, is a basic precondition for computing estimate value. But the hypothesis lacks reasonable support.

In this article, based on the hidden Markov model, we propose a statistical relational learning approach for estimating and replacing missing categorical data. In missing value estimate process, this new approach explores and describes statistical dependency relationship among categorical attribute values by the hidden Markov model. Then, missing value could be replaced by quantified statistical dependency relationship.

The rest of this article is concisely organized as follows. In section 2 we describe the proposed approach. In section 3 we illustrate the approach by an example, and the article is completed with section 4 which contains our concluding remarks.

2. THE PROPOSED APPROACH

2.1 Grouping of Categorical Attributes

Employing categorical attributes in relational data is one of effective ways for standardized design and data quality management. In many tables of the relational data, the number of categorical attributes could be up to more than 50%. In our research, we properly extend the scope of categorical attributes. They not only include categorical attributes defined in data design phase, but also include attributes whose value domains include a certain number of values in fact (such as Birth year of a given person set). Generally, categorical attributes are used to describe key information of entity objects. So, the treatment of missing categorical data is very important to improve data quality.

For categorical attributes of a relational data set, there are two types of relationship between every two attributes. One type of relationship is that the tow attributes' values are independent of each other, another type of relationship is that one of the two attributes' values are dependent on another attribute's. So, all categorical attributes can be divided into a number of independent groups, that is, every two attributes belonging to different groups are independent of each other, while one attribute in a group is dependent on another one in the same group.

When a relational data set has *M* categorical attributes, $X_1, X_2, \cdots, X_M$, the grouping procedure could be shown in Figure 2.

```
Begin
create M groups, $\{X_1\},\{X_2\},\cdots,\{X_M\}$
for i=2:M
   {if $X_i$ is independent of $X_1,\cdots,X_{i-1}$
       continue
     else
     create a new group, $\{X_i\}\cup$ groups own atrribute which is dependent on $X_i$
   }
End
```

Figure 2 The grouping procedure of categorical attributes

The implementation of the grouping procedure in Figure 2 is often done by personals with special domain knowledge, and statistical dependency relationship also could be identified with the help of statistical analysis for attribute values.

2.2 Ordering of Attributes

The Markov model describes ordering stochastic process, so we should order attributes in one group. In practice, to identify the sort order of attributes with domain knowledge is a directed and effective way. We propose several principles for ordering attributes as follows.

Time sequence

Some entities described by attibutes own or imply time information. We can order attributes using time sequence. Such as Birth year and the Year when join in work, First qualification and Highest qualification, etc.

Space inclusion

Space inclusion relationship often exists among some entities described by attributes. Such as Province, City, Universiy, College and Department. According to these relationships, we can order attributes from big space to small space.

Concept hierarchy

Concept hierarcical relationship may be existed among some entities described by attibutes. Such as Information science, Computer Science, Compter software, Database theory and system, Data quality, and Data cleaning. Accordingly, concept hierarchy can help us to order attributes.

Business priority

In some business domain, values of an attribute may be restricted by values of another attribute. For example, in courrent education mechanism, Degree is restricted by Qualification, so, the Qualification should come before the Degree.

Other general knowledge

In some cases, we could get ordering information from other gereral knowledge. For instance, high Income is the advantage of owning home, then, the Income (low or high) should come before the HomeOwner (no or yes).

When need to rank a group of categorical attributes, we must reasonably use above principles. If more than one principle apply to some attributes, we should select the early introduced principle. Even so, dependency relationship among categorical attributes is often complex, so different attribute sequences could be gotten under certain circumstance. However, the sequence difference doesn't impact implement next step for estimating missing categorical values.

In addition, we should avoid that the first attribute in the sequence is empty unless all attributes are empty. So, for different incompleted records, more than one different attribute sequences may be needed.

2.3 The Hidden Markov Model

In this part, we introduce the hidden Markov model used to estimate missing values. Compared to classical Markov model, the hidden Markov model can describe statistical relationship among different types of states, which corresponds to different types of attribute values.

Considering the attribute sequence in one group, The hidden Markov model can describe the statistical dependency distribution of attributes' values.

$$P(X_{i+1} = v_{(i+1)k_{(i+1)}} \left| X_1 = v_{1k_1}, X_2 = v_{2k_2}, \cdots, X_i = v_{ik_i}\right.) = P\left(X_{i+1} = v_{(i+1)k_{(i+1)}} \middle| X_i = v_{ik_i} \right) \tag{1}$$

In formula (1), $i = 1,2,\cdots,M-1$, $k_i = 1,2,\cdots,L_i$, where M is the number of attributes. X_i is the ith categorical attribute, and $v_{i1}, v_{i2}, \cdots, v_{iL_i}$ are values of X_i , where L_i is the number of X_i's values.

2.4 Estimating and Replacing Missing Values

Given a sample data set with *N* records, all values of the group of attributes are all filled. Let N_{ik_i} be the number records whose $X_i = v_{ik_i}$, and let $N_{jk_j|ik_i}$ be the number of records whose $X_j = v_{jk_j}$, given $X_i = v_{ik_i}$. From the sample set, we can use formula (2) and formula (3) to estimate distribution probabilities of attribute values.

$$P\left(X_1 = v_{1k_1}\right) = \frac{N_{1k_1}}{N} \tag{2}$$

$$P\left(X_{i+1} = v_{(i+1)k_{(i+1)}} \middle| X_i = v_{ik_i} \right) = \frac{N_{(i+1)k_{(i+1)}|ik_i}}{N_{ik_i}} \tag{3}$$

If all attributes in the group are empty, we use formula (2) to estimate probabilities of values of X_1, and use formula (3) to estimate dependency probabilities of other attribute values. Otherwise, we use formula (3) to estimate dependency probabilities of empty attribute values.

Based on the estimated probabilities, two alternative methods for replacing the missing value are adopted [11]: The first replaces the missing value with the value having the maximum probability (MaxProp); the second uses a value that is selected with probability proportional to the estimated dependency distribution (PropProp). When replacing missing values, we should operate in turn of the attribute sequence.

3. AN ILLUSTRATION EXAMPLE

To illustrate the implement process of the proposed approach, consider the data set in Table 2 [11]. The data set contains 16 records with 4 attributes: Income, Age, Gender, and HomeOwner (whether or not the person owns a home). There are 8 empty fields in the data set.

NO.	Income	Age	Gender	HomeOwner
1	low	<30	female	no
2	low	<30	male	no
3	low	30-55	female	yes
4	low	30-55	female	no
5	low	>55	female	no
6	high	<30	male	yes
7	high	30-55	female	yes
8	high	30-55	male	yes
9	high	30-55	male	yes
10	high	30-55	male	no
11	high	>50	male	yes
12		30-55	female	yes
13		30-55	female	yes
14		<30	female	
15			male	no
16			male	no

Table 2 The illustrative data set

At first, the 4 attributes couldn't be divided into more than one independent groups on the grouping procedure of categorical attributes (see Figure 2). Then, we give an attribute sequence on ordering principles $\langle \text{Gender, Age, Income, HomeOwner} \rangle$.

Next, We select the top 11 records as a sample set and estimate dependency probabilities of missing values using formula (3).

$$P(<30|\text{male}) = \frac{2}{6} = 0.3333,\ P(30-55|\text{male}) = \frac{3}{6} = 0.5000,\ P(>55|\text{male}) = \frac{1}{6} = 0.1667$$

$$P(\text{low}|<30) = \frac{2}{3} = 0.6667,\ P(\text{high}|<30) = \frac{1}{3} = 0.3333$$

$$P(\text{low}|30\text{-}55) = \frac{2}{6} = 0.3333,\ P(\text{high}|<30\text{-}55) = \frac{1}{3} = 0.6667$$

$$P(\text{low}|>55) = \frac{1}{2} = 0.5000,\ P(\text{high}|>55) = \frac{1}{2} = 0.5000$$

$$P(\text{no}|\text{low}) = \frac{4}{5} = 0.8000,\ P(\text{yes}|\text{low}) = \frac{1}{5} = 0.2000$$

$$P(\text{no}|\text{high}) = \frac{1}{6} = 0.1667,\ P(\text{yes}|\text{high}) = \frac{5}{6} = 0.8333$$

At last, we replace missing values based on MaxProp and PropProp respectively. These results are listed in Table 3. Obviously, the results using PropProp are just one possible scenario.

NO.	Income		Age		HomeOwner	
	MaxPost	PropProp	MaxPost	PropProp	MaxPost	PropProp
12	high	high	-	-	-	-
13	high	high	-	-	-	-
14	low	low	-	-	no	no
15	high	low	30-55	<30	-	-
16	high	high	30-55	30-50	-	-

Table 3 Replaced Missing Values

4. CONCLUSIONS

In this article, a statistical relational learning approach for estimating and replacing missing categorical data is proposed. Due to categorical data are very important in relational data, the proposed approach used to deal with missing categorical data has high practical value. The hidden Markov model that we use to describe the dependency relationship is an effective model. When estimate a missing value, it just need no more than two attribute values, so, the estimating process is straightforward and fast. Further more, all categorical attributes are divided into several independent groups and they are treated respectively. This strategy could avoid processing too many attributes at beginning. In practice, the ordering of attributes is not very strict. In many situations, we only need to consider the dependency relationship between two attributes.

In the future, we need more experimental study to validate the approach comprehensively.

ACKNOWLEDGEMENTS

This research was supported in part by the China Postdoctoral Science Foundation Special Funded Project under Grant No. 201003797, the China Postdoctoral Science Foundation Funded Project under Grant No.20090461425, the Jiangsu Planned Projects for Postdoctoral Research Funds under Grant No.0901014B, and the PLA University of Science and Technology Pre-Research Funded Project under Grant No.20110604. We would like to thank Li X. B. for his research on Bayesian approach for missing data treatment.

REFERENCES

[1] Breiman, L., Friedman, J. H. and Olshen, R. A., *Classification and Regression Trees*. Wadsworth International Group, Belmont, 1984, pp.203-215.

[2] Cao, J. J., Diao, X. C., Wang, T. and Wang F. X., Research on Domain-independent Data Cleaning: A Survey. *Computer Science*, 37(5), 2010, pp.26-29.

[3] Cao, J. J., Diao, X. Ch., Wu J. M., Yuan, Zh. and Peng, C., Classification Detection Method for Uncompleted Records Based on Bit Operation. *Systems Engineering and Electronics*, 32(11), pp.2488-2492.

[4] Chen, G. and Astebro, T., How to Deal with Missing Categorical Data: Test of a Simple Bayesian Method. *Organ. Res.*, Methods, 6(3), 2003, pp.309-327.

[5] Chiu, H. Y. and Sedransk, J., A Bayesian Procedure for Imputing Missing Values in Sample Surveys. *J. Amer. Statist. Assoc.*, 81(3905), 1986, pp.5667-5676.

[6] Clark, P., and Niblett, T., The CN2 Induction Algorithm. *Mach. Learn*, 3(4), 1989, pp.261-283.

[7] Codd, E. F., Extending the Database Relational Model to Capture More Meaning. *ACM Trans. Database Syst.*, 4(4), 1979, pp.397-434.

[8] Congdon, P., Bayesian Models for Categorical Data. *John Wiley & Sons*, New York, 2005.

[9] Fan, W., Lu, H., Madnick, S. E., and Cheung, D. DIRECT: A System for Mining Data Value Conversion Rules from Disparate Data Sources. *Decis. Support Syst.*, 34(1), 2002, pp.19-39.

[10] Li, D. Y. and Du, Yi., *Artificial Intelligence with Uncertainty*. CRC Press, USA, 2007.

[11] Li X. B., A Bayesian Approach for Estimating and Replacing Missing Categorical Data. *ACM Journal of Data and Information Quality*, 1(1), 2009, pp.1-11.

[12] Michie, D., Spiegelhalter D. J. and Taylor, C. C., *Machine Learning, Neural, and Statistical Classification*. Prentice Hall, New York, 1994.

[13] Pyle, D., *Data Preparation for Data Mining*. Morgan Kaufmann, San Mateo, CA, 1999.

[14] Quinlan, J. R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann of Elsevier, San Mateo, 1993.

[15] Quinlan, J. R., Unknown Attribute Values in Induction. In Proceedings of the 6th International Workshop on Machine Learning. Morgan Kaufmann, San Mateo, CA, 1989, pp.164-168.

[16] SAS Institute, *SAS Procedure Guide*. SAS Institute Inc., Cary, NC. INC., 1990.

[17] Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann of Elsevier, San Francisco, CA, 2005.