

A Novel Two-Phase Method for the Classification of Incomplete Data

Xiuyun Qu, Bo Yuan, Wenhua Liu

Graduate School at Shenzhen

Tsinghua University

Shenzhen 518055, P. R. China

xiuyunqu@gmail.com, {yuanb, liuwh}@sz.tsinghua.edu.cn

Abstract—The issue of incomplete data exists across the entire field of data mining. In this paper, a novel two-phase method is developed to deal with the challenge of incomplete data on classification problems. In phase I, the dataset is divided into disjoint subsets based on the attributes with missing values. In phase II, each subset is used to train appropriate classification algorithms respectively in parallel. Experimental results show that the proposed scheme works favorably compared to other techniques on both synthesized and real data sets.

Keywords—classification; incomplete data; missing values; imputation; feature deletion

I. INTRODUCTION

In real world, incomplete data in which part of the values are missing for certain features are an inevitable and common issue, particularly in the fields of medical, biological and social sciences. For instance, partial responses in surveys are common in social science, leading to incomplete datasets with arbitrary patterns of missing data. The missing data problem may be caused by various reasons such as nonresponse in social science studies [1] and faulty measurement in data acquisition. Since most machine learning and data mining techniques are designed with complete data in mind, incomplete data may severely affect the quality of learned patterns and the performance of algorithms. As a result, how to properly handle incomplete data is an important and challenging problem in the practice of machine learning and data mining.

In our study, we focused on the classification of incomplete data. There are different approaches to handling incomplete data as far as classification is concerned, from simply removing samples or features with missing values to completing the original data set by filling in specific values (usually referred to as data imputation [2]). However, the deletion of samples or features may result in the loss of useful information especially when a large portion of samples or features have missing values. In the meantime, imputation methods are often based on various assumptions with regard to the distribution of missing data, which may not always be feasible. In this paper, a novel scheme is developed for conducting classification on incomplete data without applying deletion or imputation.

The rest part of this paper is organized as follows. An overview of the previous research on incomplete data is given in Section II. Section III introduces the details of the

proposed method for handling incomplete data. Experimental results are presented in Section IV. This paper is concluded in Section V with a list of directions for future work.

II. AN OVERVIEW OF INCOMPLETE DATA

In incomplete datasets, certain values are missing for one or more features. When choosing the right techniques for dealing with this issue, it is necessary to have a good understanding of different reasons that lead to incomplete data.

A. Types of Missing Data

Little and Rubin [3] define a list of missing mechanisms, which are widely accepted by the community.

1) *Missing completely at random (MCAR)*. If subjects who have missing data are a random subset of the complete sample of subjects, this type of missing data is called MCAR. The reason for missing is completely random and, in other words, the probability that an observation is missing is not related to any other features. Typical examples of MCAR are when a tube containing a blood sample of a study subject is broken by accident (such that the blood parameters can not be measured) or when a questionnaire of a study subject is accidentally lost [4]. This situation is rare in real world and is usually discussed in statistical theory.

2) *Missing at random (MAR)*. The probability that an observation is missing commonly depends on some information about that subject and this type of missing data is called MAR. For example, suppose we want to evaluate the predictive value of a particular diagnostic test and the test results are known for all diseased subjects but unknown for a random sample of nondiseased subjects. In this case, the missing data belong to MAR, which are conditional on an observed patient characteristic (here the presence or absence of the disease) [4]. This mechanism is common in practice and is generally considered as the default type of missing data.

3) *Not missing at random (NMAR)*. If the probability that an observation is missing depends on information that is not observed, this type of missing data is called NMAR. For example, high incomers may be more reluctant to provide their income information [4]. This situation is relatively complicated and there is no universal solution.

In this paper, the missing mechanism considered in our method is MAR, which is the most studied missing type in machine learning and data mining.

B. A Review of Work on Incomplete Data

In the past decades, significant efforts have been devoted to this area from the point of view of statistical theory, machine learning and so on. Various methods for handling incomplete data have been introduced and these methods can be summarized as follows: samples or features deletion, missing values imputation and learning with missing data.

1) *Samples or features deletion.* Samples or features with missing values are simply removed from the dataset. This method is easy to implement and usually performs well when the missing rate is low. However, it is obvious that it may ignore some potentially valuable information and create bias in the dataset.

2) *Imputation of missing values.* Most studies on incomplete data focus on imputation. Imputation techniques are based on the idea that any subject in a study sample can be replaced by a new randomly chosen subject from the same source population. The imputation of missing data of a feature is to generate values drawn from an estimate of the distribution of this variable [4]. Common imputation schemes include completing missing data with specific values such as the unconditional mean or the conditional mean (if one has an estimate for the distribution of missing features given the observed features) [4] [5].

Dempster et al. [6] propose a maximum likelihood estimation method for incomplete data via the EM algorithm. Meng and Rubin [7] develop the Generalized EM algorithm replacing the complicated M-step of EM with several computationally simpler CM-steps. Taking into account the imprecision caused by the fact that the distributions of the variables with missing values are estimated, Rubin [8] introduces Multiple Imputation (MI) by creating several imputed data sets in which different imputations are based on a random draw from different estimated underlying distributions. In recent years, there are some new imputation methods based on machine learning techniques such as k-NN [9] and kernel-based methods [10].

3) *Learning with missing data.* Some classifiers can be customized in order to handle incomplete data directly, such as Artificial Neural Network (ANN) [11], C4.5 decision trees [12], Bayesian Networks (BN) [13], Rough sets [14] and Logistic regression algorithm [2].

Generally speaking, different approaches suit different datasets, which should be selected according to the property of the dataset at hand as well as the requirement on algorithm complexity and efficiency.

III. METHODOLOGY

Incomplete data problems usually occur in areas such as social sciences, bank or shop surveys and medical research. Suppose a set of diagnostic data of patients and normal people among different hospitals in different areas has been collected. The dataset is likely to be incomplete due to several reasons. For example, patients and normal people might be required to undertake different examinations (usually patients take more examinations) or the medical conditions were different among hospitals so that people in different places took different examinations. In this situation,

the missing rate of the data may be relatively high and the strategy based on samples or features deletion is not suitable because it may discard some potentially valuable information. The strategy of imputation may not be suitable either because the missing values may distribute differently from the existing ones (the diagnosis data of patients and normal people have different distributions) and imputation can bring in bias. In order to handle incomplete datasets where neither deletion nor imputation is appropriate, a novel two-phase method is proposed as follows.

In phase I, the original dataset is separated into disjoint subsets according to the information gain of the features with missing values, as in (1). In phase II, the data in each subset are used in the training of some standard classification algorithms respectively in parallel.

$$Gain(S, F) = E(S) - \sum_{v \in \{F_N, F_A\}} \frac{|S_v|}{|S|} E(S_v) \quad (1)$$

$$E(S) = - \sum_{i=1}^n p_i \log_2(p_i) \quad (2)$$

In (1), S is the dataset. F is one of the features with missing values while F_N and F_A refer to the two subsets of F depending on whether the value of F is missing (N) or available (A). In (2), $E(S)$ is the information entropy of S where n is the number of the categories and p_i is the probability that the samples in dataset S belong to the i^{th} class.

• Training

Phase I:

- (1) Given a training dataset D , calculate the information gain of each feature with missing data.
- (2) Select the feature F with the highest information gain. Split D into two subsets D_N and D_A based on the values of F (missing or available). In other words, D_A consists of samples where the values of F are available and D_N contains the rest samples. Since the values of F are all missing in D_N , F is removed and the dimension of D_N is reduced by one. A binary tree T is built in which D is the root node and the child nodes are D_A and D_N .
- (3) For datasets D_A and D_N , repeat step 1 and step 2 until there is no missing values in each child node.

Phase II:

The original dataset D has been split into several subsets D_i . A separate classifier C_i is trained on each subset respectively with appropriate classification algorithms such as Libsvm [15].

Table I shows an example of dataset with missing values. The first column is the ID of samples and each row represents a sample with features A , B and C . The last column is the class label and NaN means that the value is missing. Fig. 1 is a schematic drawing of the tree T after Phase I on this training set. After Phase I, the dimensions of most subsets are lower than that of the original dataset D , which may help improve the performance of the classifiers used in Phase II.

When the number of features with missing values is high, a minimum missing rate can be applied to avoid generating too many subsets (the number of samples in each subset

needs to be maintained at a reasonable level for classification). Note that, some subsets generated in this situation may contain incomplete data and the classifiers used on these subsets should be able to handle missing data, while other subsets without incomplete data can still be handled by standard classifiers.

TABLE I. A TRAINING SET WITH MISSING VALUES

Sample	Feature A	Feature B	Feature C	Label
1	0	1	NaN	1
2	0	2	NaN	0
3	NaN	3	NaN	1
4	NaN	1	1	0
5	NaN	2	1	1
6	1	2	1	1
7	2	4	0	1
8	2	1	2	1
9	3	5	3	0
10	1	3	5	0

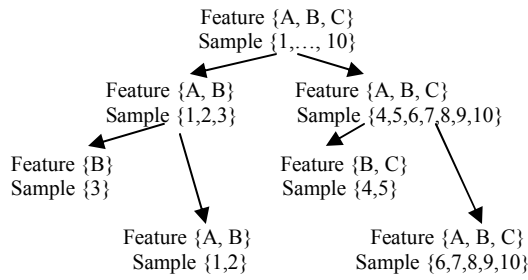


Figure 1. The Tree Representation of the Training Set in Table I.

- Testing

Phase I:

For each unknown sample X to be classified, find the corresponding subset D_i in the tree T generated above.

Phase II:

Apply C_i (trained on D_i) to assign a class label to X .

In summary, the proposed two-phase strategy converts the original incomplete dataset into a set of complete datasets so that standard classifiers that can only handle complete data can be also applied directly.

IV. EXPERIMENTS

This section presents the empirical studies on the proposed two-phase method compared to some existing methods described in Section II. The experiments were carried on a synthesized dataset as well as a few datasets in the real world. We adopted the overall classification accuracy as the performance measure. The classifiers used in all experiments were implemented in Weka 3.6.0 [16].

A. A synthesized dataset

A simple 2D dataset as shown in Fig. 2 was used to illustrate the performance of the two-phase method. There were 1400 samples ('o' vs. 'x') in the dataset and 400 of them marked along the x-axis had their Y values missing. The results are shown in Table II (10-fold cross validation).

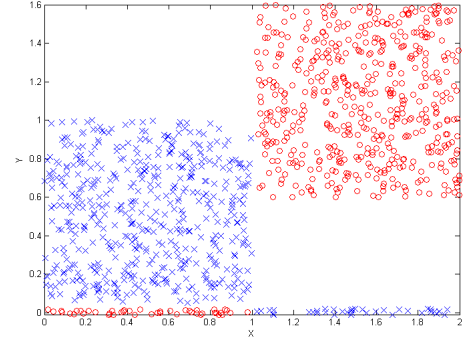


Figure 2. The synthesized dataset

TABLE II. THE ACCURACY ON THE SYNTHESIZED DATASET

Method	Incomplete Data			C4.5 after imputed with 2	C4.5 after mean imputation
	C4.5	BN	Tr-C4.5 ^a		
Accuracy	71.57%	71.29%	100%	99.64%	71.21%

a: Tr-C4.5 means the two-phase method combined with C4.5

It is clear that, on this simple dataset, the two-phase method (with C4.5) separated the dataset completely while, without imputation, both C4.5 and BN performed poorly. In the meantime, when being imputed with appropriate values, C4.5 also achieved a high accuracy. Fig. 3 shows the distribution of the data after being imputed with a fixed value (2.0). The classification boundaries by C4.5 are also shown in Fig. 3. This example shows the effectiveness of the two-phase method and there is no need to decide the imputation value in advance or by trial-and-error.

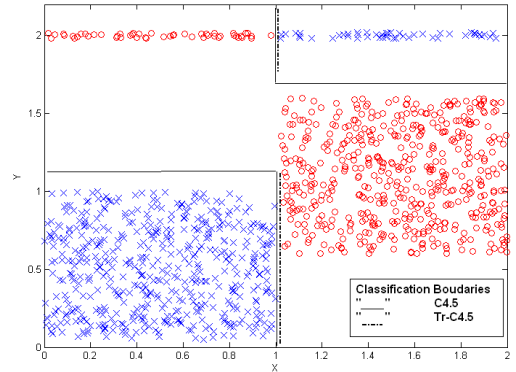


Figure 3. The synthesized dataset after being imputed with a fix value.

B. Real world datasets

The first dataset was Abalone from UCI datasets [17]. It is a multi-class dataset and, for the sake of clarity, a new two-class dataset was created by selecting a subset of samples and merging different classes. The resulting dataset was a relatively balanced one containing 689 class 1 samples and 709 class 2 samples. There was a feature 'sex' in the dataset with three possible values: 'M', 'F' and 'I' and all 'sex' values marked with 'I' were regarded as missing data. As a result, the incomplete dataset had 578 samples with

missing values. Again, the 10-fold cross validation was used and the experimental results are shown in Table III.

TABLE III. THE ACCURACY ON THE UCI DATASET

Accuracy	Libsvm	C4.5	BN
Incomplete Data		82.33%	75.46%
Imputed with 3 ^a	87.27%	83.12%	75.46%
Tr-Method	88.20%	84.12%	79.18%
Feature Deletion	86.12%	81.83%	75.68%

a: In the dataset, 'M' and 'F' were represented by 1 and 2 respectively.

Since the feature 'sex' only takes two possible nominal values, the imputation method based on the EM algorithm is not suitable. In this experiment, Libsvm, C4.5 and BN were employed as the classifiers on different datasets generated based on various strategies such as imputation and feature deletion. Note that Libsvm cannot be directly applied on incomplete dataset. It is clear that our method always achieved the highest accuracies when combined with different classifiers.

The second dataset was based on the credit card dataset from the PAKDD 2009 Data Mining Competition [18]. The dataset was provided by a major Brazilian retail chain containing customer information and credit levels (2-class problem). Firstly, the dataset was preprocessed based on our previous work [19]. Secondly, a subset of the original data was selected based on shops (ID_Shop). Thirdly, the values of the feature 'Months_In_Job' were discarded for some selected shops to create an incomplete dataset. This was to simulate the real world scenario where certain data from specific shops were missing. Finally, there were 10384 samples and 3171 of them had missing values. The 10-fold cross validation was used in this experiment and the results are shown in Table IV. Similar to the last case, the accuracies of the classifiers were the highest when combined with our method.

TABLE IV. THE ACCURACY ON THE PAKDD DATASET

Accuracy	Libsvm	C4.5	BN
Incomplete Data		81.90%	81.98%
EM Imputation	75.28%	81.03%	80.89%
Tr-Method	77.04%	82.82%	82.93%
Feature Deletion	74.32%	73.07%	74.30%

V. CONCLUSIONS

The major contribution of this paper is a conceptually simple and practically effective method for the classification of incomplete data without imputation or deletion. Experimental results showed that the proposed two-phase method performed reasonably well on a synthesized and two

real world datasets, compared to imputation and deletion. Certainly, there is still a lot of room for improvement. For example, it is important to better understand on what type of missing patterns this two-phase method can be expected to work favorably compared to other strategies. Also, it is possible to combine this new method with existing imputation methods in Phase I to estimate some of missing values that are suitable to be imputed, which may improve the overall performance of the classifier.

REFERENCES

- [1] G. B. Durrant, "Imputation methods for handling item-nonresponse in the social sciences: a methodological review," National Center for Research Methods: Methods Review Papers Series, 2005.
- [2] D. Williams, X. Liao, Y. Xue, and L. Carin, "On classification with incomplete data," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp. 427-436, 2007.
- [3] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 2nd ed., New York: John Wiley and Sons, 2002.
- [4] A. Donders, G. van der Heijden, T. Stijnen, and K. Moons, "Review: a gentle introduction to imputation of missing values," Journal of Clinical Epidemiology, vol. 59, pp. 1087-1091, 2006.
- [5] T. D. Pigott, "A review of methods for missing data," Educational Research and Evaluation, vol. 7, pp. 353-383, 2001.
- [6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, vol. 39, pp. 1-38, 1977.
- [7] X. Meng, D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," Biometrika, vol.80, pp. 267-278, 1993.
- [8] D. B. Rubin, Multiple Imputation for Non-Response in Surveys. New York: John Wiley and Sons, 1987.
- [9] P. Jonsson and C. Wohlin, "An evaluation of k-nearest neighbour imputation using likert data," Proc. 10th IEEE International Software Metrics Symposium, Los Alamitos, California, USA, pp. 108-118, 2004.
- [10] S. Zhang, Y. Qin, X. Zhu, J. Zhang, and C. Zhang, "Kernel-based multi-imputation for missing data," Proc. 4th international conference on Active Media Technology, pp. 106-111, 2006.
- [11] C. M. Ennett, M. Frize, and C. R. Walker, "Influence of missing values on artificial neural network performance," Medinfo, vol. 84, pp. 449-453, 2001.
- [12] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Kaufmann Publishers, 1993.
- [13] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," Machine Learning, vol. 29, pp. 131-163, 1997.
- [14] Z. Pawlak, Rough Set: Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic Publishers, 1991.
- [15] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001.
- [16] G. Holmes, A. Donkin, and I. H. Witten, "WEKA: a machine learning workbench," Proc. 2nd Australian and New Zealand Conference on Intelligent Information Brisbane, Qld., Australia, pp. 357-361, 1994.
- [17] A. Asuncion and D. J. Newman, "UCI Machine Learning Repository," University of California, 2007.
- [18] <http://sede.neurotech.com.br:443/PAKDD2009/>
- [19] Deep Think Team, "Report for the PAKDD 2009 Data Mining Competition," <http://sede.neurotech.com.br/PAKDD2009/files/47.PDF>, 2009.