Incremental Gradient Descent Imputation Method For Missing Data In Learning Classifier Systems

Daqian Gu
National Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
gudaqian@ai.nju.edu.cn

Yang Gao
National Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
gaoy@nju.edu.cn

ABSTRACT

Learning with incomplete or missing data has been a major challenge in learning classifier system. One method for covering missing data is imputing missing values based on the statistic of known values. Another is marking them matching arbitrary case. A new approach using incremental gradient descent imputation model is proposed in this paper, which use the relationship among variables to estimate the missing value. And, some experiments are conducted in order to compare the performance of new approach and other classical covering methods.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms

Keywords

LCS, InGrImputation method, missing data

1. BACKGROUND

Learning Classifier Systems (LCS)[**?**] is a kind of self-adaptive, online learning systems, which was proposed by Holland in 1970s. In LCS, the classifiers are evaluated through reinforcement learning. And the population is evolved using genetic algorithm. All the fields LCS applied in are potentially plagued by the problem of missing, or incomplete, data. In LCS research, missing data has been one of the focuses.

Schafer et al. described that missing data mechanisms are commonly described as falling into one of three categories[**?**]. The first type of missing data is referred to as missing completely at random (MCAR). When data are MCAR, missing cases are no different than non-missing cases, in terms of the analysis being performed. The second type of missing data is referred to as missing at random (MAR). In this case, missing data depends on known values and thus is described fully by variables observed in the data set. The last type of missing data is referred to not missing at random (NMAR). Since the missing data depends on events or items which the researcher has not measured, this is the worst situation.

Unfortunately, missing data in LCS usually belongs to MAR or NMAR type. In LCS, there are already three traditional methods to cover missing data[**?**]: the first is called Wild-to-Wild, in which any classifiers in the population that match the specific variables of an input case are added to match set; the second is to impute mean value calculated from a mode of this variable; the third is to impute a value randomly selected within the range of the variable of the missing data. Holmes shows that, the three types of missing value covering mechanism exhibit similar efficiency in an XCS-based learning classifier system, with respect to learning rate and classification accuracy. All the above methods drop the relationship between the variable with missing data and other known variables. This paper advances a new incremental gradient descent imputation model (InGrImputation Model) based on the relationship to impute the missing data which LCS operates on.

In this paper, we design an incremental gradient descent imputation model to replace the missing data. The method is abbreviated to InGrImputation model. In Section 2, we describe this InGrImputation model in LCS. In Section 3, some experiments were conducted in order to compare the performance between InGrImputation method and other covering methods for missing data.

2. INGRIMPUTATION MODEL

As we know, LCS adapt themselves to environments while evolving their classifiers which represent the relationship between input variables and classification. In the learning process, relationship among the input variables has not been utilized directly; can we impute missing data based on the relationship? To answer these questions, we propose the InGrImputation Model to impute missing data based on this relationship to handle LCS missing data.

InGrImputation Model creates an universal model for the variable with missing data based on the relationship between the variable and other known variables. It may be inaccurate without any prior knowledge at the beginning. In each episode of training or testing in LCS, InGrImputation Model checks whether there is any missing data in the input, if the

GECCO'05, June25–29, 2005, Washington, DC, USA.

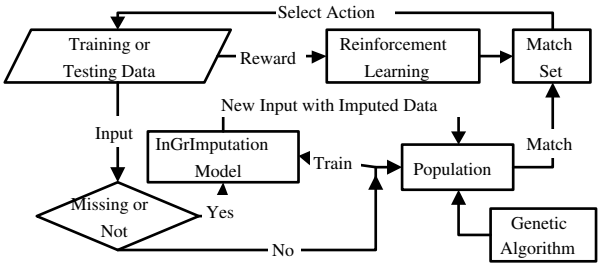

Figure 1: Architecture of LCS with InGrImputation Model(.eps format).

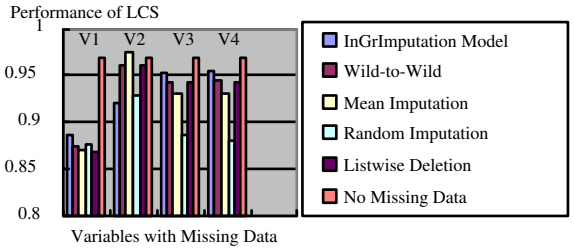

Figure 2: Performance of LCS with Missing Data on 'Iris' data(.eps format).

answer is 'no', InGrImputation Model adapts itself by gradient descent method according to the current input; if the answer is 'yes', InGrImputation Model simply creates a new value for the missing data based on other input variables and the current model.

The gradient descent algorithm for InGrImputation Model is shown below. Suppose F_{t-1} is InGrImputation Model at the end of the (t-1)-th training episode of LCS process, that means LCS can impute an estimation value I'_n of the missing data I_n based on the current InGrImputation Model F_{t-1} and other known variables $(a_{n1}, a_{n2}, \ldots, a_{np})$ for the input (n) at this time. This is the way of using InGrImputation Model during LCS training and testing processes.

$$I'_n = F_{t-1(w_1,w_2,\ldots,w_q)}\left(a_{n1}, a_{n2}, \ldots, a_{np}\right). \quad (1)$$

In the n-th episode, LCS receives a new input (m) without missing data, which is used to train InGrImputation Model. Firstly, LCS computes an error (δ_t) between I'_m and I_m.

$$\delta_t = \left(I_m - I'_m\right)^2 \quad (2)$$

$$\delta_t = \left(I_m - F_{t-1(w_1,w_2,\ldots,w_q)}\left(a_{m1}, a_{m2}, \ldots, a_{mp}\right)\right)^2 \quad (3)$$

In order to adjust the coefficients $(W = (w_1, w_2, \ldots, w_q))$ to improve the model, InGrImputation Model computes the partial derivative for each w_i at this episode. The vector V_t shows the reverse direction where W should be adjusted to at this time.

$$V_t = (v_{t1}, v_{t2}, \ldots, v_{tq}) = \left(\frac{\partial\delta_t}{\partial w_1}, \frac{\partial\delta_t}{\partial w_2}, \ldots, \frac{\partial\delta_t}{\partial w_q}\right). \quad (4)$$

The last step is adjusting W of the model, and replacing the older model F_{t-1} by the new model F_t with new W.

$$W' \quad = \quad W - \gamma * V_(t) \quad (5)$$

InGrImputation Model adapts itself along with the learning of LCS, after enough episodes of training it shows an accurate relationship between the variable with missing data and other known variables, and imputes an accurate estimation of missing data, which improves the performance of LCS.

3. EXPERIMENTAL PROCEDURE

The experiments here use an XCS- style LCS[**?**], and the data from from UCI repository. For each variable with missing data LCS uses InGrImputation model, Wild-to-Wild method, mean imputation method, random imputation method and listwise deletion method. Fig.2 shows the results of these experiments.

Similar experiments were conducted on other UCI data, and got the same results: InGrImputation Model shows better performance than any other methods for most cases, except on those variables (as V2 in Fig.2) with less importance or some conflicts to the classification of LCS. It is explicable that variables with less importance or some conflicts contribute poor or even negative effect to the classification, so imputation the missing data with more accuracy by InGrImputation Model helps less to the result.

4. CONCLUSIONS

This paper proposes InGrImputation Model which imputes missing data with higher accuracy and improves the performance of LCS. We hope that further research on more perfect methods like InGrImputation Model will contribute to handling missing data in LCS research.

5. ACKNOWLEDGEMENT

The paper is supported by the Natural Science Foundation of China (No.60475026), the National Outstanding Youth Foundation of China (No.60325207), the National Grand Fundamental Research 973 Program of China (No.2002CB312002) and the Natural Science Foundation of Jiangsu Province, China(No.BK2003409).

6. REFERENCES

[1] Holland J: A Mathematical Framework for Studying Learning in Classifier Systems. Evolution games and learning. Physica, 1986, 22D(1-3) :page: 307-317.

[2] Wilson, S.W: State of XCS Classifier System Research. Technical Report 99.1.1, Prediction Dynamics, Concord, MA, 18 March 1999.

[3] Schafer. J. L. and Graham, J. W.: Missing data: Our view of the state of the art. Psychological Methods, 7(2), 147-177, 2002.

[4] Holmes JH and Bilker WB: The effect of missing data on learning classifier system classification and prediction performance. Advances in Learning Classifier Systems. Springer Verlag, 2003, Vol. 2661: 46-60.

[5] John H. Holmes, Jennifer A. Sager, and Warren B. Bilker: A Comparison of Three Methods for Covering Missing Data in XCS. IWLCS-2004 (GECCO 2004), Seattle, Washington, June 2004.