Directly Constructing Multiple Features for Classification with Missing Data using Genetic Programming with Interval Functions

Cao Truong Tran, Mengjie Zhang, Peter Andreae and Bing Xue
School of Engineering and Computer Science
Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand
{cao.truong.tran, mengjie.zhang, peter.andreae, bing.xue}@ecs.vuw.ac.nz

ABSTRACT

Missing values are a common issue in many industrial and real-world datasets. Genetic programming-based multiple feature construction (GPMFC) is a recent promising filter approach to constructing multiple features for classification using genetic programming (GP). GPMFC has been demonstrated to improve classification performance and reduce the complexity of many decision trees and rule-based classifiers, but it cannot work with missing data. To deal with missing data, this paper propose IGPMFC, an extension of GPMFC that use interval functions as the GP function set to directly construct multiple features for classification with missing data. Empirical results on five datasets and four classifiers show that IGPMFC can substantially improve the performance and reduce the complexity of the classifiers when faced with missing data.

Keywords

Missing Data; Classification; Feature Construction; Genetic Programming; Interval Functions

1. METHODS

GP-based multiple feature construction (GPMFC) [3] is a recent promising filter approach using GP for feature construction. GPMFC is able to evolve multiple high-level features from original features. The empirical results show that, in almost all cases, GPMFC can not only improve the classification performance, but can also reduce the complexity of many decision trees and rule-based classifiers.

Although GPMFC is an effective feature construction algorithm for complete data, it cannot deal with datasets containing missing values. To handle this issue, we designed IGPMFC, an extension of GPMFC that performs multiple feature construction using GP with interval functions. The key idea of IGPMFC is to use interval functions as the function set of GP. When a feature value is missing, it will be replaced by the interval associated with the feature. When a feature value is complete, it will still be treated as an interval, in which both the lower bound and upper bound are equal to the feature value. The purpose of using interval functions is that missing values are unknown; therefore replacing a missing value with an interval instead of a single value is likely to reflect better the uncertainty of the missingness.

1.1 Interval Function Set

Assume that the lower bound and the upper bound of each feature x are x_l and x_u, respectively. In this approach, we use four interval arithmetic operations as the function set of GP, taken from [2], defined as follows using two features x and y as an example:

$$x + y = \begin{cases} lower: & x_l + y_l \\ upper: & x_u + y_u \end{cases}$$

$$-x = \begin{cases} lower: & -x_u \\ upper: & -x_l \end{cases}$$

$$x * y = \begin{cases} lower: & min(x_l * y_l, x_l * y_u, x_u * y_l, x_u * y_u) \\ upper: & max(x_l * y_l, x_l * y_u, x_u * y_l, x_u * y_u) \end{cases}$$

$$1/x = \begin{cases} lower: & min(1/x_l, 1/x_u) \\ upper: & max(1/x_l, 1/x_u) \end{cases}$$

We note that the division operation is not well behaved when the interval spans zero, and requires the assumption that the lower bound and the upper bound of the denominator have the same sign. However, we allow the GP search to eliminate trees which violate this assumption.

1.2 Estimating the Real Output of an Individual

The output of an individual with interval functions is an interval. However, in order to put constructed features into classifiers, single values are required. Therefore, in the experiments, the real output of an individual is calculated as the middle point of the interval. Assume that $[out_l, out_u]$ is the output of an individual, the real output is defined as follows:

$$out = \frac{out_l + out_u}{2}$$

GECCO '16 July 20-24, 2016, Denver, CO, USA

ACM ISBN 978-1-4503-4323-7/16/07.
DOI: http://dx.doi.org/10.1145/2908961.2909002

2. EXPERIMENT DESIGN

The main objective of this study is to evaluate the impact of IGPMFC on classification with missing data. To achieve this, three experimental setups are designed to evaluate the impact of IGPMFC on classification with missing data: classification with missing data by using classifiers able to deal with missing data; classification with missing data by using imputation methods combined with GPMFC before using classifiers and classification with missing data by using IGPMFC to construct new features from missing data before using classifiers.

The experiments used five benchmark datasets selected from the UCI [1]: Balance Scale, Banknote, Breast Cancer, Iris Plant and Liver Disorders. With each dataset, perform 30 times: put randomly 10% missing values into relevant features. After that, ten-fold cross-validation was performed on the 30 incomplete datasets.

The experiment used two imputation methods which are mean imputation and MICE imputation [5]. The experiments used four decision trees that are able to classify missing data: C4.5, CART, REPTree and BFTree.

3. RESULTS AND ANALYSIS

Fig. 1 summarises the accuracy comparison of IGPMFC with Baseline, MeanGPMFC and MiceGPMFC. It is clear from Fig. 1 that in almost all cases, IGPMFC not only can achieve better classification accuracy compared to using original features, but also can achieve better classification accuracy than using GPMFC combined with mean imputation in most cases. Moreover, IGPMFC is comparable with GPMFC combined with using MICE Imputation that is expensive for classification task [4].

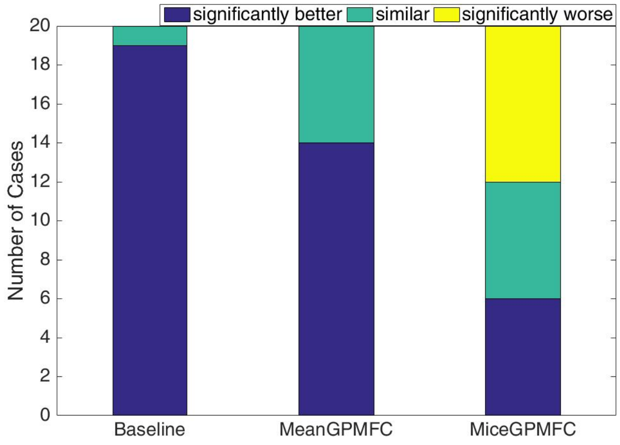

Figure 1: Accuracy comparison of IGPMFC with Baseline, MeanGPMFC and MiceGPMFC

Fig. 2 shows the average of ratio tree size of the other methods over IGPMFC. On average, Baseline generates about 4.5 times bigger trees than those using IGPMFC, and both MeanGPMFC and MiceGPMFC generates bigger trees than those using IGPMFC. In summary, in all cases, IGPMFC can dramatically reduce the complexity of the classifiers by using original features. Furthermore, IGPMFC can better reduce the complexity of the classifiers by using both simple and sophisticated imputations combined with GPMFC.

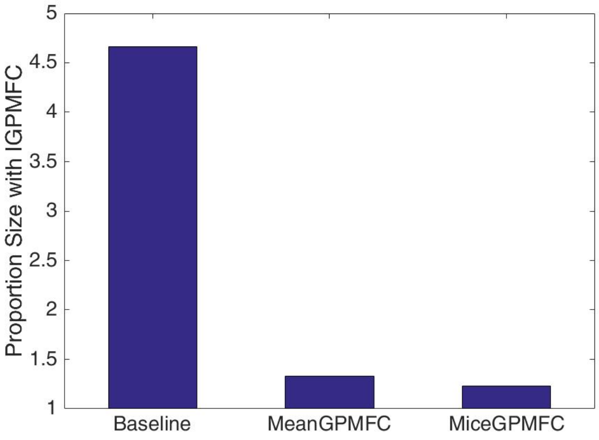

Figure 2: The average of ratio tree sizes of Baseline, MeanGPMFC and MiceGPMFC over IGPMFC

4. CONCLUSIONS

This paper proposed IGPMFC that is a multiple feature construction for classification with missing data using GP with interval functions. IGPMFC is an extension of GPMFC that is a recent promising feature construction method, but cannot deal with missing data. To tackle missing data, IGPMFC uses a set of interval functions as the function set of GP. When a feature value is missing, the value is replaced by the feature interval. Experimental results show that using IGPMFC achieves better classification accuracy compared to using original features or using a simple imputation method combined with GPMFC, and it is comparable with using an expensive imputation method combined with GPMFC. Moreover, IGPMFC can reduce the complexity of classifiers compared to the other methods.

5. REFERENCES

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[2] M. Keijzer. Improving symbolic regression with interval arithmetic and linear scaling. In *Genetic programming*, pages 70–82. 2003.

[3] K. Neshatian, M. Zhang, and P. Andreae. A filter approach to multiple feature construction for symbolic learning classifiers using genetic programming. *Evolutionary Computation, IEEE Transactions on*, 16:645–661, 2012.

[4] C. T. Tran, M. Zhang, and P. Andreae. A genetic programming-based imputation method for classification with missing data. In *Genetic Programming - 19th European Conference, EuroGP 2016, Porto, Portugal, March 30 - April 1, 2016, Proceedings*, pages 149–163, 2016.

[5] I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine*, 30:377–399, 2011.