Evolutionary-based feature construction with substitution for data summarization using DARA

Florence Sia[a]*; Rayner Alfred*[b]

[a]Faculty of Computer System and Software Engineering, University Malaysia Pahang, Lebuhraya Tun Razak, 26300 Kuantan, Pahang, Malaysia

[b]School of Engineering and Information Technology, University Malaysia Sabah, Locked Bag 2073, 88999, Kota Kinabalu, Sabah, Malaysia

ABSTRACT

The representation of input data set is important for learning task. In data summarization, the representation of the multi-instances stored in non-target tables that have many-to-one relationship with record stored in target table influences the descriptive accuracy of the summarized data. If the summarized data is fed into a classifier as one of the input features, the predictive accuracy of the classifier will also be affected. This paper proposes an evolutionary-based feature construction approach namely Fixed-Length Feature Construction with Substitution (FLFCWS) to address the problem by means of optimizing the feature construction for relational data summarization. This approach allows initial features to be used more than once in constructing newly constructed features. This is performed in order to exploit all possible interactions among attributes which involves an application of genetic algorithm to find a relevant set of features. The constructed features will be used to generate relevant patterns that characterize non-target records associated to the target record as an input representation for data summarization process. Several feature scoring measures are used as fitness function to find the best set of constructed features. The experimental results show that there is an improvement of predictive accuracy for classifying data summarized based on FLFCWS approach which indirectly improves the descriptive accuracy of the summarized data. It shows that FLFCWS approach can generate promising set of constructed features to describe the characteristics of non-target records for data summarization.

KEYWORDS:

relational data mining; relational data summarization; feature construction; genetic algorithm

REFERENCES

1. D. Minnie and S. Srinivasan, “Application of knowledge discovery in database to blood cell counter data to improve quality control in clinical pathology,” Bio-Inspired Computing: Theories and Applications (BICTA), 2011 Sixth International Conference, pp. 338-342, 2011.

2. A. Sharafi, P. Wolf, and H. Krcmar, “Knowledge discovery in database on the example of engineering change management,” Industrial Conference on Data Mining – ICDM, pp. 9-16, 2010.

3. U. Fayyad, “Data mining and knowledge discovery: Making Sense Out of Data,” IEEE Intelligent System, 1996.

4. O. Maimon and L. Rokach, “Introduction to knowledge discovery and data mining,” The Data Mining and Knowledge Discovery HandBook, Springer, Heidelbergh, pp. 1-5, 2010.

5. K.J. Cios, W. Pedrycz, R.W. Swiniarski, and L.A. Kurgan, “The knowledge discovery process,” Data Mini