Reduction methods for multi-label datasets based on Granular Computing

Marilyn Bello[1,2], Gonzalo Nápoles[2], Koen Vanhoof[2], and Rafael Bello[1]

[1] Computer Science Department, Universidad Central de Las Villas, Cuba
[2] Faculty of Business Economics, Hasselt University, Belgium
mbgarcia@uclv.cu

1 Introduction

Multi-label classification (MLC) refers to one specific type of classification in which an object can belong to several classes at the same time [10]. While this approach allows modeling a wide variety of real-world problems, it also increases the difficulty of correctly classifying the patterns.

A fundamental stage in the process of discovering knowledge consists in the *pre-processing* operations. A recent review [8] divides it into data preparation and data reduction methods. Data preparation steps convert raw data to an appropriate format by cleaning or transforming the data. Data reduction methods include popular techniques such as feature selection (reduction of the number of descriptive features, [13]), instance selection (reduction of the number of observations, [12]), feature extraction (creation of new features, [11]), instance generation (creation of new objects, [16]) and discretization (transformation of the continuous features to categorical features, [9]).

Prototype selection algorithms select a set of representative objects according to a well-defined criterion, while prototype generation algorithms are capable of generating a set of new objects in the application domain from the initial objects. In the MLC context, most data reduction methods have been focused on the selection and extraction of features [7], and discretization [5]. To the best of our knowledge, the only relevant work related to prototype selection in the MLC field is the kNNc method described in [4].

Hence, we propose three methods of instance selection [1], and three methods of instance generation (prototypes) [2]. These algorithms will be briefly described in the following sections. It is worth mentioning that, unlike the approach in [4], our methods are independent from the MLC algorithm.

2 Methods for instance selection

Our methods uses the lower and upper approximations as computed in the *Rough Set Theory* [14] to determine a proper granularity degree in the training set. The

first method builds a new training set as the union of the lower approximations attached with decision classes. The second method additionally includes objects that are in the boundary region, which have been relabeled by taking into account their membership degree to each decision class. The third method is similar to the second one, however, it does not consider the connection among decisions classes, so that labels are treated independently.

Aiming at exploring the global performance of our algorithms, we adopt the ML-kNN algorithm [18] while the Hamming Loss (HL) [15] is used as the evaluation measure. We leaned upon several multi-label training sets taken from the MULAN [17] and RUMDR [6] repositories.

Numerical simulations showed that our methods allow a reduction of up to 80% of the number of instances in some of the training sets without affecting the classification performance. Actually, in some cases we observed an increase the discriminatory power of the ML-kNN algorithm when operating on edited datasets.

3 Methods for prototype generation

In these methods, two classical granulation approaches are used: condition granulation and decision granulation. The former refers to the granularity of the universe according to the conditional attributes, while the latter is based on the decision classes. In the case of the first two methods, the granulation process is performed by using a similarity relation that builds similarity classes on the basis of the conditional attributes. This means that each similarity class represents an information granule that is used to build a prototype. The advantage of using similarity relations is that our methods can be used in the presence of both numerical and nominal attributes. The third method performs a granulation of the universe by using an equivalence relation, and taking into account the different labels existing in the universe of discourse. Hence, an equivalence class is built for each label, thus leading to a granular prototype.

After analyzing the reduction coefficient [3], it could be concluded that our methods achieve a significant reduction ($40-80\%$ reduction) in most case studies, while preserving the efficacy of the ML-kNN method. The set of generated prototypes can be used as a learning set for other learning algorithms, even those which are not intended for example-based learning.

4 Concluding Remarks

The efficacy and efficiency of the Machine Learning models depend on the quantity and quality of data. One alternative to deal with these issues is to edit the training set as a pre-processing step with the intention either to reduce the number of instances or improve data quality. In this research, we proposed methods for the reduction of datasets in MLC environments. The results showed that our proposal provides a suitable trade-off between algorithms performance and the number of training examples in the dataset.

References

1. Bello, M., Nápoles, G., Vanhoof, K., Bello, R.: Methods to edit multi-label training sets using rough sets theory. In: International Joint Conference on Rough Sets. pp. 369–380. Springer (2019)
2. Bello, M., Nápoles, G., Vanhoof, K., Bello, R.: Prototypes generation from multi-label datasets based on granular computing. In: 24th Iberoamerican Congress on Pattern Recognition (CIARP 2019). Springer (2019)
3. Bermejo, S., Cabestany, J.: A batch learning vector quantization algorithm for nearest neighbour classification. Neural Processing Letters **11**(3), 173–184 (2000)
4. Calvo-Zaragoza, J., Valero-Mas, J.J., Rico-Juan, J.R.: Improving knn multi-label classification in prototype selection scenarios using class proposals. Pattern Recognition **48**(5), 1608–1622 (2015)
5. Cano, A., Luna, J.M., Gibaja, E.L., Ventura, S.: Laim discretization for multi-label data. Information Sciences **330**, 370–384 (2016)
6. Charte, F., Charte, D., Rivera, A., del Jesus, M.J., Herrera, F.: R ultimate multi-label dataset repository. In: International Conference on Hybrid Artificial Intelligence Systems. pp. 487–499. Springer (2016)
7. Doquire, G., Verleysen, M.: Feature selection for multi-label classification problems. In: International work-conference on artificial neural networks. pp. 9–16. Springer (2011)
8. García, S., Luengo, J., Herrera, F.: Data preprocessing in data mining. Springer (2015)
9. Garcia, S., Luengo, J., Sáez, J.A., Lopez, V., Herrera, F.: A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering **25**(4), 734–750 (2012)
10. Herrera, F., Charte, F., Rivera, A.J., Del Jesus, M.J.: Multilabel classification. In: Multilabel Classification, pp. 17–31. Springer (2016)
11. Liu, H., Motoda, H.: Feature extraction, construction and selection: A data mining perspective, vol. 453. Springer Science & Business Media (1998)
12. Liu, H., Motoda, H.: On issues of instance selection. Data Mining and Knowledge Discovery **6**(2), 115–130 (2002)
13. Liu, H., Motoda, H.: Computational methods of feature selection. CRC Press (2007)
14. Pawlak, Z.: Rough sets. International journal of computer & information sciences **11**(5), 341–356 (1982)
15. Pereira, R.B., Plastino, A., Zadrozny, B., Merschmann, L.H.: Correlation analysis of performance measures for multi-label classification. Information Processing & Management **54**(3), 359–369 (2018)
16. Triguero, I., Derrac, J., Garcia, S., Herrera, F.: A taxonomy and experimental study on prototype generation for nearest neighbor classification. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**(1), 86–100 (2011)
17. Tsoumakas, G., Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan multi-label dataset repository. URL¡ http://mulan. sourceforge. net/datasets. html (2014)
18. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition **40**(7), 2038–2048 (2007)