Local sets for multi-label instance selection*

Álvar Arnaiz-González
Dpto. Ingeniería Civil
Universidad de Burgos
Burgos, Spain
alvarag@ubu.es

José-Francisco Díez-Pastor
Dpto. Ingeniería Civil
Universidad de Burgos
Burgos, Spain
jfdpastor@ubu.es

Juan J Rodríguez
Dpto. Ingeniería Civil
Universidad de Burgos
Burgos, Spain
jjrodriguez@ubu.es

César García-Osorio
Dpto. Ingeniería Civil
Universidad de Burgos
Burgos, Spain
cgosorio@ubu.es

***Abstract*—This is a summary of our article published in Applied Soft Computing [1], presented to the Multi-Conference CAEPIA'18 KeyWorks.**

***Index Terms*—multi-label classification, data reduction, instance selection, nearest neighbor, local set**

I. Summary

Single-label classification is a predictive data mining task that consists of assigning a label to an instance for which the label is unknown. Multi-label classification presents a similar task, although the difference is that the instances have a collection of labels, known as a labelset, rather than only one label. The maximum size of the labelset is determined by the number of different labels in the data set. The aforementioned labelset concept can also be considered as a sequence of binary output attributes (as many attributes as there are labels in the whole data set). Each attribute indicates whether the corresponding label is applicable to the instance. Only one of the attributes is active in single-label problems, while several attributes may be active in multi-label problems [5]. In other words, the labels in multi-label learning are not mutually exclusive [8]. This feature implies a much harder and more challenging problem, due to the high relevance of the relations between the different labels [10].

Despite the well-established usefulness of single-label instance selection, there are still very few methods for multi-label classification. To the best of our knowledge, only two instance selection methods for multi-label have been developed. Since both algorithms are based on Wilson Editing (ENN) [9], to avoid any confusion with the acronyms, we refer to them by the initials of their authors: the KADT method [6] and the CRJH method [3]. In this paper, we have attempted to fill that gap by proposing a new technique for computing local sets in multi-label data sets. This new proposal was used to adapt two single-label instance selection methods, LSSm and LSBo, for multi-label problems. The adaptation was tested against the few instance selection methods existing for multi-label learning and against the classifiers (MLkNN [11] and IBLR-ML [4]) trained on the whole data sets.

The main contributions of the paper were:

We would like to thank the *Ministerio de Economía y Competitividad* of the Spanish Government for financing the project TIN2015-67534-P (MINECO/FEDER, UE) and the *Junta de Castilla y León* for financing the project BU085P17 (JCyL/FEDER, UE) both cofinanced from European Union FEDER funds.

- The definition of the local set concept in the context of multi-label data sets.
- The proposal that defines two new instance selection methods, based on the adaptation of single-label classification algorithms to multi-label learning: LSBo and LSSm [7].
- The experimental evaluation of the new algorithms. The new methods were compared with the few existing algorithms.

Instance selection methods usually focus on boundaries between classes. Boundaries are the keystone of the predictive process, because they define whether an instance belongs to one class or another. The simplest classification problem is a binary class data set: there is only one class, thus one instance may or may not belong to it (in practice, this task is similar to determining one of two categories to which the instance belongs). In multi-class classification, more classes are present but, as in the previous case, each instance can only belong to one. The challenge that emerges in multi-label data sets is that instances can belong to more than one class at the same time, which blurs the boundaries (because different labels overlap).

The concept of local set has been used for designing several instance selection algorithms for single-label data sets [2], [7]. Local sets are defined by the nearest enemy, which is straightforward to compute in single-label data sets. The problem with multi-label data sets is how the nearest enemy is defined: it is no trivial task, because every single instance has a set of labels, rather than only one, as in single-label classification. An intuitive solution would be to consider each labelset (the vector of labels of an instance) as a class in itself. However, the results of several experiments have demonstrated that this approach is of little or no use, due to the large amount of different labelsets that multi-label data sets usually have. For example, for a data set with three different classes, the number of different labelsets could be up to $2^3 = 8$; if a data set has nine labels, the number of labelsets could reach $2^9 = 512$. The number of possible labelsets therefore increases exponentially with the number of labels. Hence, local sets calculated in this way will be too small (many of them only made up of a single instance) and, therefore, the algorithms based on local sets would not work properly.

The proposal that was presented in the paper was to use the Hamming loss (calculated over labelsets) to measure the

degree of difference in the labelsets[1]. If the Hamming loss between the labelsets of two instances is greater than a predefined threshold, the instances are considered to be of different '*classes*'. This concept of *class* can be seen as a '*soft-class*' in the same sense as in regression data sets. The Hamming distance is computed as follows:

$$Hamming\ distance(\mathbf{a}, \mathbf{b}) = |\omega_{\mathbf{a}} \triangle \omega_{\mathbf{b}}| \quad (1)$$

where, $\omega_{\mathbf{a}}$ and $\omega_{\mathbf{b}}$ are the labelsets of instances $\mathbf{a}$ and $\mathbf{b}$, respectively, and $\triangle$ is the symmetric difference between two labelsets[2].

The Hamming distance according to the previous definition is a whole number. The Hamming loss value is commonly used in multi-label learning HL $\in [0, 1]$.

$$Hamming\ loss(\mathbf{a}, \mathbf{b}) = \frac{1}{|\Omega|}|\omega_{\mathbf{a}} \triangle \omega_{\mathbf{b}}| \quad (2)$$

Pseudocode 1 shows the proposed method for local set calculation in multi-label data sets. It has two inputs: the multi-label data set and the value of the Hamming loss threshold that determines when two labelsets are considered distinct. The function has two outputs: an array of local sets and an array of nearest enemies. Every single instance has its local set (made of one or more instances) and its nearest enemy.

Algorithm 1: Function `computeLocalSets`: computes the local sets of a multi-label data set.

Input: A training set $X = \{(\mathbf{x}_1, \omega_1), ..., (\mathbf{x}_n, \omega_n)\}$, a threshold θ

Output: The local sets $LSS = \{\mathbf{lss}_1, ..., \mathbf{lss}_n\}$, the nearest enemy of each instance $NE = \{ne_1, ..., ne_n\}$

```
1  for i ∈ {1...n} do
2      lss_i ← ∅
3      dist_ne_i ← ∞
       /* Find the nearest enemy of x_i      */
4      for j ∈ {1...n} do
5          d ← EuclideanDistance(x_i, x_j)
6          if HammingLoss(ω_i, ω_j) > θ and d < dist_ne_i
           then
7              ne_i ← x_j
8              dist_ne_i ← d
       /* Compute the local set of x_i       */
9      for j ∈ {1...n} do
10         if EuclideanDistance(x_i, x_j) < dist_ne_i then
11             lss_i ← lss_i ∪ {x_j}
```

After the calculation of local sets, any local set-based algorithm can be used without changes. In the experimental study, we considered LSSm and LSBo, because their use of local sets is more robust than the use of local sets in ICF (the heuristic used in ICF has fundamental problems that were reported in [7]).

The experimental study used a broad range of data sets from different domains, several multi-label measures and statistical tests. The results revealed the two main benefits of our proposal: *i*) HDLSSm, as an edition algorithm, is not only capable of outperforming the other instance selection methods in terms of its results, but it also capable of outperforming the classifier trained with the whole data set; *ii*) HDLSBo, as a condensed algorithm, achieved a remarkable compression, while maintaining a statistically equivalent performance to the performance of the other methods. Furthermore, the existence of a threshold for controlling local set sizes implies an adaptable and versatile proposal.

[1]We decided to use Hamming loss, because its computation is fast and it is a commonly used measure in multi-label learning.

[2]The symmetric difference is the exclusive disjunction (XOR) of two sets, that is the set of all elements that are in one set, but not in the other set.

References

[1] Álvar Arnaiz-González, José F. Díez-Pastor, Juan J. Rodríguez, and César García-Osorio. Local sets for multi-label instance selection. *Applied Soft Computing*, 68:651 – 666, 2018.

[2] Henry Brighton and Chris Mellish. *On the Consistency of Information Filters for Lazy Learning Algorithms*, pages 283–288. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[3] Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. MLeNN: A first approach to heuristic multilabel undersampling. In *Intelligent Data Engineering and Automated Learning – IDEAL 2014: 15th International Conference, Salamanca, Spain, September 10-12, 2014. Proceedings*, pages 1–9, Cham, 2014. Springer International Publishing.

[4] Weiwei Cheng and Eyke Hüllermeier. Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2):211–225, Sep 2009.

[5] Francisco Herrera, Francisco Charte, Antonio J. Rivera, and María J. del Jesus. *Multilabel Classification: Problem Analysis, Metrics and Techniques*. Springer Publishing Company, Incorporated, 2016.

[6] Sawsan Kanj, Fahed Abdallah, Thierry Denœux, and Kifah Tout. Editing training data for multi-label classification with the k-nearest neighbor rule. *Pattern Analysis and Applications*, 19(1):145–161, 2016.

[7] Enrique Leyva, Antonio González, and Raúl Pérez. Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective. *Pattern Recognition*, 48(4):1523 – 1537, 2015.

[8] Newton Spolaôr, Maria Carolina Monard, Grigorios Tsoumakas, and Huei Diana Lee. A systematic review of multi-label feature selection and a new method based on label construction. *Neurocomputing*, 180:3 – 15, 2016. Progress in Intelligent Systems Design Selected papers from the 4th Brazilian Conference on Intelligent Systems (BRACIS 2014).

[9] Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-2(3):408–421, July 1972.

[10] Zoulficar Younes, Fahed Abdallah, Thierry Denœux, and Hichem Snoussi. A dependent multilabel classification method derived from the k-nearest neighbor rule. *EURASIP Journal on Advances in Signal Processing*, 2011(1):1–14, 2011.

[11] Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038 – 2048, 2007.