Leveraging Formal Concept Analysis to Improve n-fold validation in Multilabel Classification

Francisco J. Valverde-Albacete[1] ⋆ and Carmen Peláez-Moreno[1]

Depto. Teoría de Señal y Comunicaciones, Univ. Carlos III de Madrid, Madrid, Spain,
fva@tsc.uc3m.es carmen@tsc.uc3m.es

Abstract. In this application note we revisit previous work on the exploratory analysis of Multilabel Classification (MLC) tasks in Machine Learning. We combine Information Theory and Formal Concept Analysis (FCA) to formalize the intuition that inference of classifiers in MLC tasks can only proceed when the training and testing data concerning the labels define the same Concept Lattice. We instantiate our procedure on the `emotions` dataset, but the procedure is independent of the dataset being explored. An R language interactive notebook carrying out the procedure is available upon request.

Keywords: Multilabel Classification · Entropy-based assessment · Data preprocessing · n-fold validation · Formal Context Analysis.

1 Introduction and Motivation

1.1 The Multilabel Classification Task

MLC is a relatively recently-formalized task in Machine Learning [1] with applications in *text categorization*, *gene expression analysis*, etc. A recent tutorial explains the progress in methods and concerns [2], while a more up-to-date exposition with special emphasis on software tools is [3]. We describe here a variant of the MLC task setting to accommodate feature transformations, as depicted in Fig. 1.

Fig. 1: Basic scheme for multi-label classification

The following way of solving the problem is based on the theory of Statistical Learning [4]: consider a binary multivariate source $\overline{Y}$, emitting binary label vectors or *labelsets* from a label space $\mathcal{Y} \equiv 2^{m_Y}$, by virtue of the isomorphism between sets and their characteristic vectors[1]. Suppose that the labelsets are *hidden* and we can only access the result of an *observation mechanism* of the labelsets in terms of *visible instance*

⋆ Corresponding author.

[1] We assign to each of the labels a certain "meaning" but this is out of this mathematical model.

RealDataFCA'2021: Analyzing Real Data with Formal Concept Analysis, June 29, 2021, Strasbourg, France

or *observation* in a feature space $\mathcal{X} \equiv \mathbb{R}^{m_X}$. *The multi-label classification problem is to tag any (feature) vector $\vec{x} \in \mathcal{X}$, with a labelset $\vec{y} \in \mathcal{Y}$ [3].*

The usual way to solve the problem in the simpler, *supervised setting* is:

1. Model the *source* of labelsets as random label vectors $\overline{Y} \sim P_{\overline{Y}}$ and their *observations* as the feature vectors $\overline{X} \sim P_{\overline{X}}$ over their respective spaces.
2. Collect a *(task) dataset* with s samples, $\mathcal{D} = \{(\vec{x}^i, \vec{y}^i)\}_{j=1}^{s}$ of labelsets and their observed features to induce a classifier from observations to labelsets $h\colon \mathcal{X} \to \mathcal{Y}$.
3. Choose the *classifier type* and induction scheme for it.
4. In order to assess the classifiers, choose an adequate measure of performance, and implement (any of a number of) schemes of iterated *re-sampling* of the data into a set of *training examples* $\mathcal{D}_T = \{(\vec{x}^j, \vec{y}^j)\}_{j=1}^{s_T}$ and a set of *test examples* $\mathcal{D}_E = \{(\vec{x}^k, \vec{y}^k)\}_{k=1}^{s_E}$ so that the training data are used to induce the classifiers and the testing data to test them, and *validate* these results, e.g. using k*-fold validation* [4].

*Example 1 (**`emotions`**[5]).* A traditional dataset to test new ideas in MLC, `emotions` captures $m_X = 72$ features and the emotions they produce—as $m_Y = 6$ labels like `amazed-surprised`, `angry-aggressive`, etc.—for a total of $s = 593$ song tracks. As expected, distinct labels have distinct expressions in samples.

Note that only 27 of the $2^6 = 64$ possible labelsets occur, and as many as 4 of them are *hapaxes*—they occur only once—while at least one of the labelsets appears 81 times. This wildly imbalanced behaviour is typical of MLC datasets [6]. □

1.2 An FCA-based Interpretation of MLC

Note that in the standard approach above FCA is all but absent. However, it is easy to see that FCA and its variants are relevant in modelling this ML task.

First, with the labelsets $\{\vec{y}^i\}_{i=1}^{s}$ of the task dataset we can build a formal context using the set of labels L as attributes, with $|L| = m_Y$, and taking for each sample—considered as a formal object $i \in G$ with $|G| = s$—its labelset as a row of the incidence matrix $I_{i\cdot} = \vec{y}_i$, whence $\mathbb{D}_L = (G, L, I)$ is the *binary formal context of samples and their labels*. This context captures the information in the stochastic source $\overline{Y}$.

Secondly, for observation vectors $\{\vec{x}^i\}_{i=1}^{s}$ their context $\mathbb{D}_F = (G, F, R)$—with F the set of features—will not be binary, in general, but many-valued $R_{i\cdot} = \vec{x}_i$. This context captures the information in the observations $\overline{X}$ [2].

With the previous modelling, relevant notions in MLC correspond to relevant notions in FCA. For instance, *labelsets are object intents,* and they can be found through the polars $\vec{y}_i = \{i\}^{\uparrow}$. In this way, **we can reason about the sampling of the stochastic variables $\overline{Y}$ and $\overline{X}$—the dataset—in terms of the contexts above, and *vice-versa***.

For instance, each labelset present in the dataset—that is, each object-extent—may appear with different samples (formal objects). Recall that the concept-forming function $\overline{\gamma}$ induces a partition $\ker \overline{\gamma}$ on G by equality of labelsets: $(i_1, i_2) \in \ker \overline{\gamma} \iff \{i_1\}^{\uparrow} = \{i_2\}^{\uparrow} = \vec{y}_{i_1}$. *This partition is crucial for defining the entropy of the label source* (see § 2). In the other direction, we expect the sampling to be good enough $m_Y \ll s$ so it is

[2] Note that Fig. 1 suggests another context formed by the *transformed observations* captured by the random vector $\overline{Z}$, but this will not be dealt with here.

safe to suppose that no two labels are predicated of the same set of objects. Therefore we expect the partition on labels induced by $\overline{\mu}$ to be the identity $\ker \overline{\mu} = \iota_L$, which *holds on standard testing datasets* (see e.g. those in [3]).

1.3 The Problem and the FCA-induced Solution

The following problem and its theoretical solution were proposed in [7] as an example of the above-suggested operation: The MLC classifier induction and assessment procedures demand that we generate train and test resamplings of the original data [8, 4]. This amounts to splitting the original context $\mathbb{D}_L$ into two subposed subcontexts of training $\mathbb{D}_T$ and testing $\mathbb{D}_E$ data: $\mathbb{D} = \mathbb{D}_T / \mathbb{D}_E$. Note that

1. Since the samples are supposed to be independent and identically distributed, the order of these contexts in the subposition—indeed the row order of I—is irrelevant.
2. The resampling of the labelset context is tied to that of the observations: we decide on the labelset context and this carries over to the observation context.

Since the data are a formal context we know that an important part of the information contained in it comes from the concept lattice, hence:

Hypothesis 1. *1. (FCA intuition) A necessary condition for the resampling of the data $\mathcal{D}$ into training part $\mathcal{D}_T$ and testing part $\mathcal{D}_E$ to be meaningful for the MLC task, is that the concept lattices of all of these be isomorphic:*

$$\underline{\mathfrak{B}}(\mathbb{D}) \cong \underline{\mathfrak{B}}(\mathbb{D}_T) \cong \underline{\mathfrak{B}}(\mathbb{D}_E)$$

2. (ML intuition) Not only the labelsets but also their ocurrence frequencies as quantified by the cardinalities of the blocks of $\ker \gamma$ are important.

The last consideration follows because if we only retained the meet- and join-irreducibles to obtain these concept lattices, then the labelsets of reducible attributes would be lost so the relative importance of the samples—both labels and observations—would change, impacting the induction scheme of the classifiers.

The above proposition suggests that the analogue of *stratified sampling* in MLC is a procedure in which the stratification must proceed on a block-by-block basis with respect to $\ker \overline{\gamma}$. However this comes at a price, when there are hapaxes in the data. If we choose, for instance, to maintain 80% of the data for training and 20% for testing, regardless of these proportions, stratified sampling will force us to include all hapaxes with the following deleterious consequences:

- The relative frequency of the hapaxes will be distorted wrt to other labelsets.
- We will be using some data (the hapaxes) both for training and testing, which is known to obtain too optimistic performance results in whichever measure of it.

Furthermore, if we use, e.g. k-fold validation we have to repeat this procedure and ensure that the resamplings are somehow different. A usual procedure is to distribute the original dataset into k blocks in order to aggregate $k-1$ of them into the training dataset $\mathbb{D}_T$ and use the leftover as the testing dataset $\mathbb{D}_E$. It is common to use $k = 5$ or $k = 10$. This can only compound the previous problem, therefore *FCA allows us to spot possible problems with the classifier induction and validation schemes using resampling.*

In this paper we will use side-by-side FCA-based and Information Theory-based devices to support the claim that they improve the understanding of the task when used in synergy. For that purpose we first review an Exploratory Data Analysis (EDA) technique in Sec. 2 to help in interpreting the experimental results in Sec. 3.

2 Methods: Source Multivariate Entropy Triangles

The Source Multivariate Entropy Triangle (SMET, [9]) is a visual tool based on Information Theory to explore the informational content of multivariate sources. It is better suited to discrete sources—be they binary or multiclass—and optimally suited for analyzing the set of label assignments of MLC tasks.

Due to space limitations, we will only describe the graphic approach to the SMET. We refer the reader to [9] for its theoretical bases. In brief, the information content of a discrete multivariate source can be depicted in the simplex of Fig. 2, where the vertices describe extreme behaviours of the labels. If a label appears in the lowest vertex, is not stochastic. If it appears in the right vertex, its information can be predicted from that of the rest of the labels, whereas if it appears in the left vertex, it has a lot of private information. The opposite sides to these vertices represent their opposite behaviours. In general, the behaviour of labels is a mix of these three types (see Fig. 3).

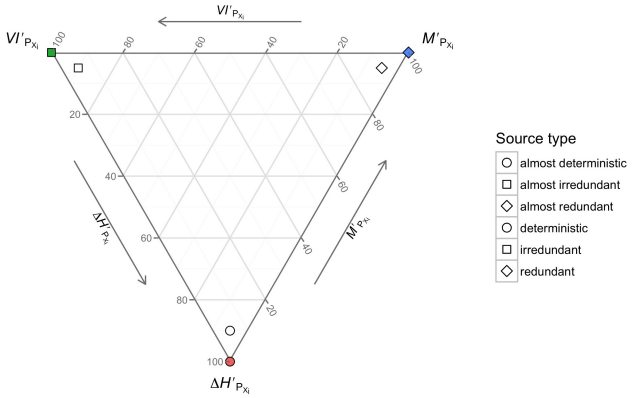

Fig. 2: General Source Multivariate Entropy Triangle from [9]

To assess the effectiveness of stratified sampling schemes for MLC, this paper proposes to use the visualization of the information balance of a formal context of labels $\mathbb{D}_L$ using the SMET.

3 Results: Example Data Analysis for an MLC Dataset

SW resources. The following analysis is carried out on the Example 1 `emotions` dataset [5], as pre-processed and presented by the *mldr* R package [6]. The interactive R notebook embodying the analysis is available from the authors upon request.

Basic EDA of the labels. Since we are only considering the set of labels $\overline{Y}$, we extracted the histogram of the labelsets $\{\vec{y}_j, n_j\}_{j \in J}$ from the dataset and considered a set of minimal frequencies of occurrence $n_T \in \{0, 1, 4, 9, 16, 25\}$ acting as thresholds based on it. The case $n_T = 0$ actually represents the original dataset and Fig. 3 shows the information balance of the six labels of `emotions` as well as the average balance for them all. We see that most labels are rather random, with `relaxing-calm` completely so. No label is completely specified by the rest of them, nor is any totally independent. This in essence means that the dataset is truly multilabel.

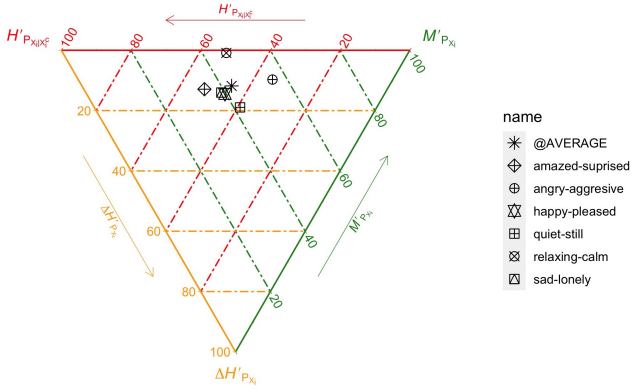

Fig. 3: SMET for the original `emotions` dataset, e.g. with $n_T = 0$

Disposing of hapaxes to improve stratified sampling. Previous analyses of the histogram of labelsets made us realize that this dataset is not adequate for resampling due to hapaxes and in general low-counts of many labelsets [7]. This applies to most MLC datasets used at present [3].

To dispose of hapaxes without disposing of samples we must re-assign each to a more frequent labelset. The rationale for this decision is because we consider hapaxes errors in label codification, and assume that the "real" labelset is the closest non-hapax in Hamming distance[3]. However, this re-assignment changes the histogram of labelsets what results in an impoverishment of the information balance of the labels and the dataset in general.

To explore this trade-off, at each threshold n_T, a labelset $\vec{y}$ was considered a *generalized hapax* if $n_{\vec{y}} < n_T$. For each threshold n_T we calculated the Hamming distance between each generalized hapax $\vec{y}_{n_T}$ and the non-hapaxes, and found the set of those closest to it. Then we re-assigned $\vec{y}_{n_T}$ to one of them uniformly at random (allowing for repetitions) [4].

[3] Recall that the Hamming distance between two sequences of bits of identical length is the number of positions in which they differ.

[4] Note that an alternative strategy would have been a scheme considering the original frequencies in the histogram, to simulate a rich-get-richer phenomenon. But such a procedure would decrease the source entropy more than the one we have chosen.

This reassignment defined a new dataset whose information balance was represented by the SMET, as suggested in Section 2, whence Fig. 4 ensued. What we can

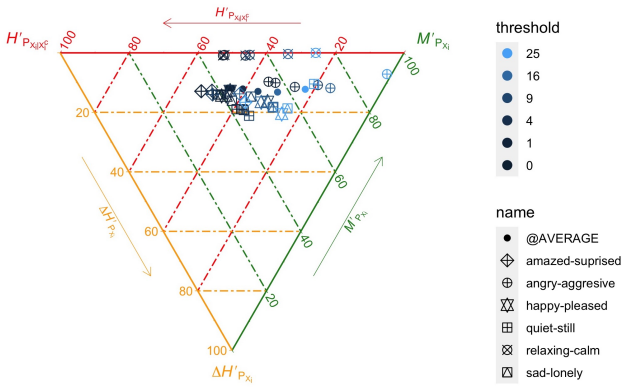

Fig. 4: SMET for `emotions` in several thresholds.

see is a general tendency to the increment of the total correlation as the thresholds increase. But this entails that the individual distinctiveness of each label is diminished. See, for instance, the case for `angry-aggressive` that can actually be predicted from the other labels when $n = 25$, confirming that too aggressive a threshold will substantively change the relative information content of the labels in the dataset.

Choosing the adequate threshold. Note that a threshold of n is needed to request an $(n+1)$-fold cross-validation of any magnitude about the dataset, since all labelset will have at least $(n+1)$ representatives for the stratified sampling requested by the cross validation procedure. Is it possible to balance the identical sampling property on train and test, yet avoid too much loss of information content?

Figure 5 depicst a choice of thresholds typically used in validation—1, 4 and 9, corresponding to 2-, 5- and 10-fold validation—for three differently behaving labels—`angry-aggressive`, `quiet-still` and `relaxing-calm`—and the average of the dataset, both for the ensembles of training and testing folds.

- As applied to the estimation of the entropies, the $(n+1)$-fold validation yields the same result in train and test, the sought-for result.
- We can see the general drift towards increased correlation in all labels, but much more in, say, `angry-aggressive` than in `quiet-still`.
- For this particular dataset, a threshold of $n_T = 4$ with 5-fold validation seems to be a good compromise for attaining statistical validity vs. dataset fidelity.

FCA confirmation. To strengthen the validity of the last two conclusions, we calculated the number of concepts of all of the train and test label contexts using the `fcaR` package [10]. This is possible since such datasets are just binary tables. After creating the contexts, we clarified and obtained the lists of concepts, then we compared the cardinality of the training and test concept lattices both for the unsplit dataset—after reassigning the generalized hapaxes, when needed—and the $(n+1)$-cross validated versions. The results are shown in Fig. 6.

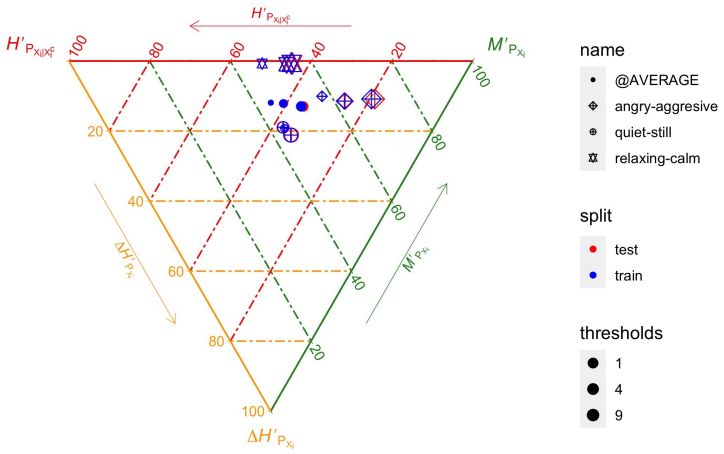

Fig. 5: SMET for `emotions` with cross-validated entropies (some labels not shown for clarity).

As expected, for $n_T = 0$ the difference in number of concepts between the non-sampled and sampled versions of the dataset make it non-adequate for appropriate sampling[5]. The training and test splits had the same number of concepts for every other threshold. For $n_T \in \{1, 4, 16\}$, the number of concepts was constant among folds, but due to the randomness inherent in sampling for $n_T \in \{9, 25\}$ one of the folds was different (not shown explicitly).

4 Summary and Discussion

We have tested and strengthened a data hypothesis put forward in previous work [7]. This hypothesis demands that the training and testing splits for solving a MLC task have the same concept lattice associated with their multilabel contexts, and also similar entropies in the induced partition of the sample set.

We first explored the influence of carrying out different degrees of smoothing of the labelset frequencies and found a balance between it and the increase in total correlation of the labels, leading to a loss of their individual information content.

We also checked that the recipe for defining an appropriate threshold n to carry out an $(n+1)$-fold cross validation actually works by comparing the informational balances of training and testing splits of the dataset as random estimates of their actual values.

Further work should start applying this rationale to solving of the MLC task. Preliminary work on using both FCA- and information theoretic-approaches suggest that present-day solutions are far from satisfactory in this respect. In the future, we plan to use more FCA-induced techniques from [7] to improve on this.

[5] It is a fluke of the dataset that both the training and test subcontexts have the same number of concepts as some of the hapaxes are singletons

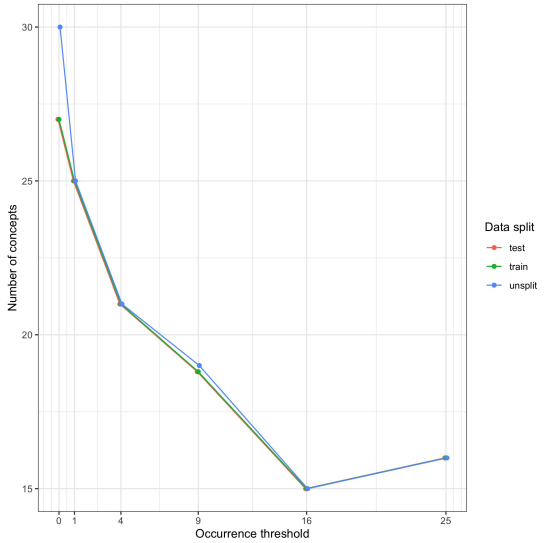

Fig. 6: Number of concepts vs. threshold for different n and splits of the dataset

References

[1] Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition **37** (2004) 1757–1771

[2] Gibaja, E., Ventura, S.: A Tutorial on Multilabel Learning. ACM Computing Surveys **47** (2015) 1–38

[3] Herrera, F., Charte, F., Rivera, A.J., del Jesus, M.J.: Multilabel Classification. Problem Analysis, Metrics and Techniques. Springer (2016)

[4] Murphy, K.P.: Machine Learning. A Probabilistic Perspective. MIT Press (2012)

[5] Wieczorkowska, A., Synak, P., Raś, Z.W.: Multi-label classification of emotions in music. In Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K., eds.: Intelligent Information Processing and Web Mining, Springer Berlin Heidelberg (2006) 307–315

[6] Charte, F., Charte, F.D.: Working with multilabel datasets in R: The mldr package. The R journal **7** (2015) 149–162

[7] Valverde Albacete, F.J., Pelaez-Moreno, C., Cabrera, I.P., Cordero, P., Ojeda-Aciego, M.: Exploratory Data Analysis of Multi-Label Classification Tasks with Formal Context Analysis. In Trnecka, M., Valverde Albacete, F.J., eds.: Concept Lattices and their Applications CLA, Tallinn University of Technology (2020) 171–183

[8] Sechidis, K., Tsoumakas, G., Vlahavas, I.P.: On the Stratification of Multi-label Data. In: Lecture Notes in Computer Science. Volume 6913., Berlin, Heidelberg, Springer Berlin Heidelberg (2011) 145–158

[9] Valverde-Albacete, F.J., Peláez-Moreno, C.: The evaluation of data sources using multivariate entropy tools. Expert Systems with Applications **78** (2017) 145–157

[10] Lopez Rodriguez, D., Mora, A., Dominguez, J., Villalon, A.: fcaR: Formal Concept Analysis. (2020) R package version 1.0.7.