Exploratory Data Analysis of Multi-label Classification Tasks with Formal *Context* Analysis

Francisco J. Valverde-Albacete[1] *, Carmen Peláez-Moreno[1], Inma P. Cabrera[2], Pablo Cordero[2], and Manuel Ojeda-Aciego[2]

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

[2] Dpt. Matemática Aplicada, Univ. de Málaga, Málaga, Spain
ipcabrera@uma.es pcordero@uma.es aciego@uma.es

Abstract. We introduce a new framework, Formal Context Analysis (FxA), for the exploratory analysis of data tasks cast in the guise of formal contexts. FxA gathers a number of results from Formal Concept Analysis, Formal Independence Analysis and Formal Equivalence Analysis to enhance the establishment and processing of hypothesis about data. We apply this framework to the study of the Multi-label Classification (MLC) task and obtain a number of results of technical nature about how the induction mechanism for MLC classifiers should proceed. The application is based on an analysis of multilabel classification from the standpoint of FxA.

1 Introduction

Exploratory Data Analysis (EDA) was introduced by Tukey as a complement to Confirmatory Data Analysis, or Predictive Modeling, as is often called nowadays [1]. It preconizes the analysis of data prior to issuing hypotheses, that may lead to a better model for it, the premise being that by investigating the data and gaining intuitions about it, better informed hypotheses and models may arise.

Formal Concept Analysis (FCA) is eminently suited for EDA of binary table data—collected in *formal contexts*—, as proposed by Wille in his "Landscapes of Knowledge" (LofK) metaphor [2], but concentrates in the hierarchical type of knowledge explicited by the *concept lattice* and, especially, *attribute implications* [3]. An extensive analysis of LofK affordances vis-à-vis other metaphors in Data Mining can be found in [4, § 4.1].

Ever since Wille himself cautioned against *only* reading hierarchical knowledge from it, there have been attempts at "other readings" from the information collected in a formal context, e.g. [5, 6]. More recently, Formal Independence Analysis (FIA) [7, 8] and Formal Equivalence Analysis (FEA) [9] have tried to

[*] Corresponding author.

Published in Francisco J. Valverde-Albacete, Martin Trnecka (Eds.): Proceedings of the 15th International Conference on Concept Lattices and Their Applications, CLA 2020, pp. 171–183, 2020.

cast representation theorems for formal contexts as lattices of anti-chains of a poset and partitions of a base set, respectively, in the guise of FCA-like theorems. Being a representation in terms of anti-chains, FIA describes the independence between parts of the formal context, while FEA takes a closer look into the refinements of the standard congruences on objects and attributes imposed by the polars of the context. These results are collected in Section 2 for future reference.

In this position-paper-with-application we propose *Formal Context Analysis* (FxA) as an umbrella term to gather the results and affordances of FCA, FIA, FEA and other types of formal analysis of tabular data bound to emerge. Specifically, to show the affordances of such a framework, we interweave it with the process of designing a valid strategy for carrying out a multi-label classification (MLC) task.

1.1 The Multi-label Classification Task

MLC is a relatively recently-formalized task in Machine Learning [10] with applications in *text categorization*, *gene expression analysis*, etc. A recent tutorial explains the progress in methods and concerns [11], while a more up-to-date exposition with special emphasis on software tools is [12]. We describe here a variant of the MLC task setting to accommodate feature transformations, as depicted in Fig. 1.

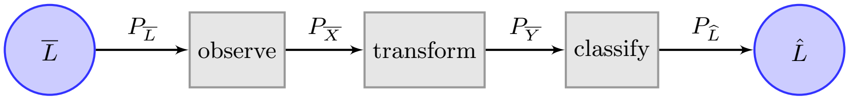

Fig. 1: Basic scheme for multi-label classification

The following way of solving the problem is based on the theory of Statistical Learning: consider a binary multivariate source $\overline{L}$, emitting binary label vectors or *labelsets* from a label space $\mathcal{L} \equiv 2^{m_L}$, by virtue of the isomorphism between sets and their characteristic vectors[3]. 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* or *observation* in a feature space $\mathcal{X} \equiv \mathbb{R}^{m_X}$. *The multi-label classification problem is to tag any (feature) vector $x \in \mathcal{X}$, with a labelset $l \in \mathcal{L}$.*

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

1. Model the source of labelsets as random label vectors $\overline{L} \sim P_{\overline{L}}$ and that of the observations as the feature vectors $\overline{X} \sim P_{\overline{X}}$ over their respective spaces.
2. Collect a sample or *task data*, $\mathcal{D} = \{(x^j, l^j)\}_{j=1}^{n}$ of labelsets and their observed features so you can induce a classifier from observations to labelsets $h \colon \overline{X} \to \overline{L}$.

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

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 *sampling* of the data into a set of *training examples* $\mathcal{D}_T = \{(x^j, l^j)\}_{j=1}^{n_T}$ and a set of *test examples* $\mathcal{D}_E = \{(x^k, l^k)\}_{k=1}^{n_E}$ so that the training data are used to induce the classifiers and the training data to test them, and *validate* these results, e.g. using *n-fold validation.*

If we were to tackle the tagging of observed features in the considerably harder *unsupervised setting*, we might first have to discern what the set of possible tags would be—e.g. the *clustering* problem [13]—but even if we knew the labels, probabilistic methods would have to be invoked to reliably relate them to observations.

The development of Machine Learning has proven that the intricacies of the above mentioned process can best be undertaken with Statistical Learning [14]. For instance, several issues have to be dealt with in an informed way when trying to solve the task:

- *Data preparation.* Due to limitations of classifier (theory and) technology it is better to obtain from the observations another representation better suited for classification by defining a transformation $g \colon \overline{X} \to \overline{Y}$ to improve classifier performance.
- *Classifier type selection.* Several competing classifier architectures need to be chosen and tested in parallel, since demonstrably, no single type of classifier can best the rest in all possible tasks.
- *Performance measure selection.* Not all measures are made equal, nor do they measure the same issues of a classifier, hence it is healthy to collect several of them in parallel so as not to bias the analysis.
- *Classifier construction and evaluation.* Again, demonstrably, unless you carry out *validation* in a repeatable manner your classifier(s) *will be cheating on the data.*

In the present paper we show how some of these problems can be tackled with the use of FxA. For this purpose, in Section 2 we present the basis of FEA and FIA, while on Section 3 we apply FxA—including FCA, FIA and FEA—to the MLC task. We end with a discussion and some avenues of future research.

2 Preliminary notions and theoretical basis

2.1 Basic Methods in MLC

Since MLC can be considered a strict generalization of the *binary and multi-class classification* tasks in that instances may have more than one label (class) assigned to them [10], most of the techniques for classifier selection and construction have been imported therefrom—albeit with a limited success, e.g k-Nearest Neighbors [15]. This is the basis of the **Algorithm Adaptation** approach [12].

On the other hand, performance measure selection, data preparation and classifier evaluation have required extensions to cater for the peculiarities of MLC: since the theory of machine learning is firmly-grounded only for the mutually-exclusive labelling cases, dealing with labelsets poses a challenge usually solved by means of **Problem Transformation**. Two extreme cases of these transformations are [16]:

- **Binary relevance (BR)** [10], a method that learns L binary classifiers—one for each different label in $\mathcal{L}$—and then transforms the original data set into L data sets $D_{l_j}; j = 1 \dots L$ that contain all examples of the original data set, labeled positively if the labelset of the original example contained l_j and negatively otherwise. To classify a new instance BR outputs the union of the labels l_j that are positively predicted by the L classifiers.
- **Label Powerset (LP)**, a simple but effective problem transformation method that considers each unique set of labels in a multi-label training set as one of the classes of a new single-label classification task. Given a new instance, the single-label classifier of LP outputs the most probable class, which is actually a set of labels.
- **Modeling the dependency between labels** by adding some labels to the set of predictors when predicting some other labels and further proceed by using either BR or LP. The initial and best example of this are Classifier Chains (CC) [17].

2.2 FIA: Formal Independence Analysis

For a poset $\mathbb{P} = \langle P, \leq \rangle$, a set of pairwise incomparable elements of P is called an *anti-chain* [18]. We denote the set of anti-chains of a poset as $A(\mathbb{P})$. FIA is based on an analysis of incomparability in terms of anti-chains.

Definition 1. *Given $\langle P, \leq \rangle$ a poset, it is possible to lift the ordering structure to the powerset 2^P by defining*

$$X \mathrel{{}_\star\!\preccurlyeq} Y \iff \textit{for all } x \in X \textit{ there exists } y \in Y \textit{ such that } x \leq y \tag{1}$$

$$X \preccurlyeq^\star Y \iff \textit{for all } y \in Y \textit{ there exists } x \in X \textit{ such that } x \leq y \tag{2}$$

$$X \preccurlyeq Y \iff X \mathrel{{}_\star\!\preccurlyeq} Y \textit{ and } X \preccurlyeq^\star Y \tag{3}$$

In general ${}_\star\!\preccurlyeq$ and $\preccurlyeq^\star$ are both simply preordering relations in $\langle 2^P, \subseteq \rangle$. Observe that in the set of anti-chains $A(\mathbb{P})$, the relations ${}_\star\!\preccurlyeq$ and $\preccurlyeq^\star$ are also anti-symmetric.

There exists a relationship with the inclusion ordering of ideals and filters since given $S, T \subseteq P$ then $S \mathrel{{}_\star\!\preccurlyeq} T \iff \downarrow S \subseteq \downarrow T$ and $S \preccurlyeq^\star T \iff \uparrow T \subseteq \uparrow S$. Because of these equivalences, we will call ${}_\star\!\preccurlyeq$ the *ideal containment relation* and $\preccurlyeq^\star$ the *filter containment relation.*

Definition 2. *For a poset $\langle P, \leq \rangle$, an anti-chain $\gamma \in A(\mathbb{P})$ is said to be* maximal *if every element of P is comparable to some element of γ.*

For any subset $Q \subseteq P$, the set of elements of P which are comparable to some element of Q is called the *neighborhood* of Q and it is denoted by $\updownarrow Q = \uparrow Q \cup \downarrow Q$. An anti-chain γ is maximal if and only if $\updownarrow \gamma = P$. The set of maximal anti-chains of a set is denoted as $MA(\mathbb{P})$, $MA(\mathbb{P}) = \{\gamma \in A(\mathbb{P}) \mid \updownarrow \gamma = P\}$.

It is worth noting that the orderings ${}_{\star}\!\preccurlyeq$ and $\preccurlyeq^{\star}$ coincide in $MA(\mathbb{P})$.

Proposition 1 ([19]). *If $\mathbb{P}$ is a finite poset then $\langle MA(\mathbb{P}), {}_{\star}\!\preccurlyeq \rangle$ is a lattice.*

Given an anti-chain A of P, the completion of A to a maximal anti-chain is not unique but there exists a unique lowest completion [20].

Definition 3. *For a finite partial order $\langle P, \leq \rangle$, and $A, B \in 2^P$ we define the* highest anti-chain complement of A, denoted A^-, *and the* lowest anti-chain complement of B, denoted B_-, *as*

$$\begin{aligned} &\cdot^- : 2^P \to 2^P && \cdot_- : 2^P \to 2^P \\ &A \mapsto A^- = \max(P \smallsetminus \updownarrow A) && B \mapsto B_- = \min(P \smallsetminus \updownarrow B)\,. \end{aligned} \tag{4}$$

Due to the universal complete lattice representation capabilities of FCA, we expect the lattices of anti-chains to be describable as the concept lattice of a context. In fact, when focusing on maximal anti-chains, we have the following isomorphism credited by Reuter [20] to Behrendt [19] and Wille [21].

Proposition 2. *Let $\mathbb{P} = \langle P, \leq \rangle$ be a poset. Then $\langle MA(\mathbb{P}), \preccurlyeq \rangle \cong \underline{\mathfrak{B}}(P, P, \not\geq)$.*

The proposition above states that maximal anti-chains can be obtained as a concept lattice for a certain context. On the other hand, Behrendt's theorem [19] is a universal representation theorem for lattices in terms of maximal anti-chains.

Theorem 1 (Behrendt). *Let $\mathbb{L} = \langle L, \leq \rangle$ be a finite lattice. Then there exists a poset $\mathbb{P} = \langle P, \leq_P \rangle$ such that $|P| = 2|L|$, where any chain has at most 2 elements and such that $\mathbb{L} \cong MA(\mathbb{P})$, i.e., $\mathbb{L}$ is isomorphic to the lattice of maximal anti-chains of $(P, \leq_P)$.*

It is straightforward that the three following structures are equivalent: formal contexts, bipartite graphs, and posets without chains with length higher than 2.

Definition 4. *Given a context (G, M, I), for all $\alpha \subseteq G$ and $\beta \subseteq M$ we define*

$$\alpha^{\sim} = M \smallsetminus \bigcup_{g \in \alpha} I(g, \cdot) \qquad\qquad \beta_{\sim} = G \smallsetminus \bigcup_{m \in \beta} I(\cdot, m) \tag{5}$$

Then a formal tomos *is a pair $(\alpha, \beta) \in 2^G \times 2^M$, such that $\alpha^{\sim} = \beta$ and $\beta_{\sim} = \alpha$. The set of formal tomoi of the context (G, M, I) will be denoted by $\mathfrak{A}(G, M, I)$.*

It is not difficult to see that there is a bijection between tomoi and maximal anti-chains in the formal context interpreted as a poset of height 2. That is, the set of formal tomoi with the supset-subset hierarchical ordering, denoted

$\underline{\mathfrak{A}}(G, M, I)$, is isomorphic to the corresponding lattice of maximal anti-chains. Moreover, since

$$\alpha^{\sim} = \{m \in M \mid g \not I\, m \text{ for all } g \in \alpha\} \quad \beta_{\sim} = \{g \in G \mid g \not I\, m \text{ for all } m \in \beta\} \quad (6)$$

it turns out that $\underline{\mathfrak{A}}(G, M, I) = \underline{\mathfrak{B}}(G, M, \not I)^d$, which is the "translation" into FCA terms. These operators reflect the underlying philosophy of FIA and, in this terminology, we can obtain the following alternative statement of Behrendt's theorem in terms of tomoi: *every finite lattice is isomorphic to a lattice of tomoi.*

Now, we can state an analogue for tomoi of the basic theorem of FCA as follows:

Theorem 2 (Basic theorem of formal independence analysis).

1. **Analysis phase.** *Given a formal context $\mathbb{K} = (G, M, I)$,*
   - *(a) The operators $\cdot^{\sim} : 2^G \to 2^M$ and $\cdot_{\sim} : 2^M \to 2^G$ of (5) form a right-Galois connection $(\cdot^{\sim}, \cdot_{\sim}) : (2^G, \subseteq) \leftharpoondown\!\!\rightharpoondown (2^M, \subseteq)$ whose* formal *tomoi are the pairs (α, β) such that $\alpha^{\sim} = \beta$ and $\alpha = \beta_{\sim}$.*
   - *(b) The set of formal tomoi $\mathfrak{A}(G, M, I)$ with the relation*

     $$(\alpha_1, \beta_1) \leq (\alpha_2, \beta_2) \ \text{ iff } \ \alpha_1 \supseteq \alpha_2 \ \text{ iff } \ \beta_1 \subseteq \beta_2$$

     is a complete lattice, which is called the tomoi lattice of (G, M, I) *and denoted $\underline{\mathfrak{A}}(G, M, I)$, where infima and suprema are given by:*

     $$\bigwedge_{t\in T} (\alpha_t, \beta_t) = \left(\bigcup_{t\in T} \alpha_t, \left(\bigcap_{t\in T} \beta_t \right)_{\sim}{}^{\sim} \right) \quad \bigvee_{t\in T} (\alpha_t, \beta_t) = \left(\left(\bigcap_{t\in T} \alpha_t \right)^{\sim}{}_{\sim}, \bigcup_{t\in T} \beta_t \right)$$

   - *(c) The mappings $\overline{\gamma} : G \to \underline{\mathfrak{A}}(G, M, I)$ and $\overline{\mu} : M \to \underline{\mathfrak{A}}(G, M, I)$*

     $$g \mapsto \overline{\gamma}(g) = (\{g\}^{\sim}{}_{\sim}, \{g\}^{\sim}) \qquad m \mapsto \overline{\mu}(m) = (\{m\}_{\sim}, \{m\}_{\sim}{}^{\sim})$$

     are such that $\overline{\gamma}(G)$ is infimum-dense in $\underline{\mathfrak{A}}(G, M, I)$, $\overline{\mu}(M)$ is supremum-dense in $\underline{\mathfrak{A}}(G, M, I)$.
2. **Synthesis phase.** *Given a complete lattice $\mathbb{L} = \langle L, \leq \rangle$,*
   - *(a) $\mathbb{L}$ is isomorphic to $\underline{\mathfrak{A}}(G, M, I)$ if and only if there are mappings $\overline{\gamma} : G \to L$ and $\overline{\mu} : M \to L$ such that*
     - *$\overline{\gamma}(G)$ is infimum-dense in $\mathbb{L}$, $\overline{\mu}(M)$ is supremum-dense in $\mathbb{L}$, and*
     - *$g\ I\ m$ is equivalent to $\overline{\gamma}(g) \not\geq \overline{\mu}(m)$ for all $g \in G$ and all $m \in M$.*
   - *(b) In particular, $\mathbb{L} \cong \underline{\mathfrak{A}}(L, L, \not\geq)$ and, if L is finite, $\mathbb{L} \cong \underline{\mathfrak{A}}(M(\mathbb{L}), J(\mathbb{L}), \not\geq)$ where $M(\mathbb{L})$ and $J(\mathbb{L})$ are the sets of meet- and join-irreducibles, resp., of $\mathbb{L}$.*

2.3 FEA: Formal Equivalence Analysis

Given (G, M, I) define operators $(\cdot)_I^{\exists}\colon 2^G \to 2^M$ and $(\cdot)_I^{\forall}\colon 2^M \to 2^G$ as follows

$$X_I^{\exists} = \{m \in M \mid m^{\downarrow} \cap X \neq \varnothing\} \qquad Y_I^{\forall} = \{\, g \in G \mid g^{\uparrow} \subseteq Y\} \tag{7}$$

These are the *span* of a set of objects $X_I^{\exists}$ as the set of attributes related to some object $g \in X$, and the *content* of a set of attributes $Y_I^{\forall}$ as the set of objects which can be completely described by the attributes in Y [5].

Based on the operators above, we can define a pair of partition-forming operators, respectively, on the sets of objects and attributes.

Let us introduce the operators $\kappa_G(\cdot) : 2^G \to \Pi(G)$ and $\kappa_M(\cdot) : 2^M \to \Pi(M)$ defined as follows:

$$g_1 \equiv g_2(\kappa_G(X)) \quad \text{iff} \quad \begin{cases} g_1 = g_2 \text{ or} \\ (g_1)_I^{\exists} = (g_2)_I^{\exists} \text{ and } g_1 \in X \end{cases}$$

$$m_1 \equiv m_2(\kappa_M(Y)) \quad \text{iff} \quad \begin{cases} m_1 = m_2 \text{ or} \\ (\overline{m}_1)_I^{\forall} = (\overline{m}_2)_I^{\forall} \text{ and } m_1 \in Y \end{cases} \tag{8}$$

where we have used the notation $\overline{m}$ to refer to the set $M \setminus \{m\}$.

Now, we can define mappings $\overline{\gamma}_{\Pi}(\cdot)\colon G \to \Pi(G) \times \Pi(M)$ and $\overline{\mu}_{\Pi}(\cdot)\colon M \to \Pi(G) \times \Pi(M)$ which assign to each object, respectively, each attribute, a pair of partitions as follows:

$$\overline{\gamma}_{\Pi}(g) = \left(\kappa_G\left((g_I^{\exists})_I^{\forall}\right), \kappa_M\left(g_I^{\exists}\right)\right) \qquad \overline{\mu}_{\Pi}(m) = \left(\kappa_G\left(\overline{m}_I^{\forall}\right), \kappa_M\left((\overline{m}_I^{\forall})_I^{\exists}\right)\right) \tag{9}$$

Now, finally, for partitions of objects G and attributes M let us define the mappings $(\cdot)_{\Pi}^{\exists}\colon \Pi(G) \to \Pi(M)$ and $(\cdot)_{\Pi}^{\forall} : \Pi(M) \to \Pi(G)$:

$$\pi_{\Pi}^{\exists} = \bigvee\{\sigma_c \mid c \in \overline{\mu}_{\Pi}(M), \pi \geq \pi_c\} \qquad \sigma_{\Pi}^{\forall} = \bigwedge\{\pi_c \mid c \in \overline{\gamma}_{\Pi}(G), \sigma \leq \sigma_c\} \tag{10}$$

We can finally state the:

Theorem 3. (Basic theorem of formal equivalence analysis)

1. **Analysis phase.** *Given a formal context $\mathbb{K} = (G, M, I)$,*
   (a) *The operators $(\cdot)_{\Pi}^{\exists}$ and $(\cdot)_{\Pi}^{\forall}$ form a left adjunction whose fixpoints, the* formal partitions*, are the pairs (π, σ) such that $\pi_{\Pi}^{\exists} = \sigma$ and $\sigma_{\Pi}^{\forall} = \pi$.*
   (b) *The set of formal partitions, denoted $\mathfrak{P}(\mathbb{K})$, with the relation $(\pi_1, \sigma_1) \leq (\pi_2, \sigma_2)$ iff $\pi_1 \leq \pi_2$ iff $\sigma_1 \leq \sigma_2$ is a complete lattice, which is called the* partition lattice of $\mathbb{K}$ *and denoted $\underline{\mathfrak{P}}(\mathbb{K})$, with infima and suprema given by:*

$$\bigwedge_{t \in T}(\pi_t, \sigma_t) = \left(\bigwedge_{t \in T} \pi_t, \left[\left(\bigwedge_{t \in T} \sigma_t\right)_{\Pi}^{\forall}\right]_{\Pi}^{\exists}\right)$$

$$\bigvee_{t \in T}(\pi_t, \sigma_t) = \left(\left[\left(\bigvee_{t \in T} \pi_t\right)_{\Pi}^{\exists}\right]_{\Pi}^{\forall}, \bigvee_{t \in T} \sigma_t\right)$$

   (c) The mappings in (9) $\overline{\gamma}_{\Pi}(\cdot) : G \to \mathfrak{P}(\mathbb{K})$ *and* $\overline{\mu}_{\Pi}(\cdot) : M \to \mathfrak{P}(\mathbb{K})$ *are such that* $\overline{\gamma}_{\Pi}(G)$ *is* $\vee$*-dense in* $\mathfrak{P}(\mathbb{K})$ *,* $\overline{\mu}_{\Pi}(M)$ *is* $\wedge$*-dense in* $\mathfrak{P}(\mathbb{K})$*.*

2. **Synthesis phase.** *Given a complete lattice* $\mathbb{L} = \langle L, \leq \rangle$*,*

   (a) $\mathbb{L}$ *is isomorphic to* $\mathfrak{P}(G, M, I)$ *if and only if there are mappings* $\overline{\gamma} : G \to L$ *and* $\overline{\mu} : M \to L$ *such that*
   - $\overline{\gamma}(G)$ *is* $\vee$*-dense in* $\mathbb{L}$ *,* $\overline{\mu}(M)$ *is* $\wedge$*-dense in* $\mathbb{L}$*, and*
   - $g\ I\ m$ *is equivalent to* $\overline{\gamma}(g) \not\geq \overline{\mu}(m)$ *for all* $g \in G$ *and all* $m \in M$*.*

   (b) In particular, $\mathbb{L} \cong \mathfrak{P}(L, L, \not\geq)$ *and, if* L *is finite,* $\mathbb{L} \cong \mathfrak{P}(J(\mathbb{L}), M(\mathbb{L}), \not\geq)$ *where* $M(\mathbb{L})$ *and* $J(\mathbb{L})$ *are the sets of meet- and join-irreducibles, resp., of* $\mathbb{L}$*.*

3 Exploratory Data Analysis of Multilabel Classification with FxA

General setup. Consider a generic dataset $\mathcal{D} = \{(x^i, l^i)\}_{i=1}^n$ capturing the essence of the MLC task described in Section 1.1 and set the indices over the samples as formal objects $G = \{1, \ldots, n\}$. Next build a formal context from it using the set of labels L as attributes and considering, for each object $i \in G$, its labelset as a row of the incidence matrix $I_{i\cdot} = l_i$, whence $\mathbb{D}_L = (G, L, I)$ is the *binary formal context of samples and their labels.*

For observation vectors $\{x^j\}_{j=1}^n$ their context $\mathbb{D}_F = (G, F, R)$—with F the set of features—will not be binary, in general, but many-valued $R_{i\cdot} = x_i$. This is the turf of *traditional* machine-learning techniques [14], but can be also tackled with generalizations of FCA for incidences with entries in an idempotent semiring, e.g. FCA in a fuzzy context [22] or $\mathcal{K}$-FCA where $\mathcal{K}$ is a semifield [23].

With the previous modeling relevant concepts in MLC correspond to relevant concepts in FxA. For instance, *labelsets are object intents,* and they can be found through the polars $l_i = \{i\}^{\uparrow}$. We would like to use the affordances of FxA (FCA + FIA + FEA) on this context $\mathbb{D} = \mathbb{D}_L \mid \mathbb{D}_F$ to help in solving the MLC task as described in Sections 1.1 and 2.1. We next introduce a number of possible ways to do so.

3.1 Task Subdivision with FIA

From a purely Machine Learning perspective, it is important to reduce the cardinality of $\mathcal{L}$ for LP. Furthermore, using BR only makes sense if the labels exhibit a great degree of independence. In fact, an extreme case of independence between labels is when the variation of one of them is obtained when the other is missing altogether and viceversa. This is the kind of situation that FIA detects [7]:

Proposition 3. *If FIA can be carried successfully in the context to obtain different subcontexts, then* $\mathbb{D}_L$ *admits a decomposition of the problem into as many subtasks of reduced complexity as subcontexts detected by FIA.*

$\mathbb{D}_L$	a	b	c	d	e	g
1	×					
2		×	×			
3a			×			
3b			×			
4				×	×	×
5					×	×
6				×		×

(a) Tabular representation of $\mathbb{D}_L$

$$\mathbb{D}_L = \frac{\mathbb{D}_1 \,\big|\, \varnothing}{\varnothing \,\big|\, \mathbb{D}_2}$$

(b) Schematic of $\mathbb{D}_L$

Fig. 2: **FIA reordered representations of an example label context $\mathbb{D}_L$.**

Proof. (Sketch.) Suppose that FIA of $\mathbb{D}_L$ is capable of providing the evidence that allows to split the data context into—without loss of generality—two subcontexts as in the toy example of Figure 2, that is, tomoi $(\{4, 5, 6\}, \{a, b, c\})$ and $(\{1, 2, 3a, 3b\}, \{d, e, g\})$. Notice that detecting and reordering the rows of $\mathbb{D}_L$ would also require the same reordering of the rows in the feature subcontext $\mathbb{D}_F$. Then we would be tempted to try to solve the subtasks built of $\mathbb{D}_1$ and $\mathbb{D}_2$ and their respective rows of observations, thereby reducing the complexity of the MLC induction process.

However, from the point of view of downstream processing such blind FIA is risky: we simplify the task by creating independent subtasks, but if we only use the subcontexts $\mathbb{D}_1$ and $\mathbb{D}_2$ for solving the subtasks we restrict the evidence for the negative cases of the labels. To avoid discarding data, then we should use the subpositions $\mathbb{D}_1/\varnothing$ and $\mathbb{D}_2/\varnothing$ for training—again, with proper reordering of rows. □

Note that FIA does not detect *statistical independence*, though, and there would also be merit in detecting such type of independence for BR, specially when inducing probability-based classifiers with $\mathcal{D}_F$ [14]. For an example of how to detect subcontexts with FIA, see [8].

3.2 Stratified sampling and FEA

In general, each labelset *present in the database* will be assigned to many different formal objects. The concept-forming function $\overline{\gamma}$ induces a partition on the set of objects $\ker \overline{\gamma}$ on G by equality of object-extents, equivalently labelsets: $(i_1, i_2) \in \ker \overline{\gamma} \iff \{i_1\}^{\uparrow} = \{i_2\}^{\uparrow} = l_{i_1}$ [3, § 1.5]. On the other hand, we expect the sampling to be good enough $m_L \ll n$ so it is safe to suppose that no two labels are predicated of the same set of objects, otherwise, they would be indistinguishable in the sample D and one of them excised. Therefore we expect the partition on labels induced by $\overline{\mu}$ to be the identity $\ker \overline{\mu} = \iota_L$.

Example 1. Table 1 shows a selection of low- to middle-complexity datasets used as testbeds for the MLC task. For instance, the `emotions` MLC dataset [15] has $n = 593$ instances and $m_L = 6$ labels, but only 27 of the $2^6 = 64$ labelsets

Table 1: **A selection of multi-label classification databases (from the *mldr* package [24])**, *distinct* and *single* refer to distinct and unitary labelsets, respectively. *max.freq* is the maximum absolute frequency for a label.

Name	n	m_F	m_L	*distinct*	*single*	*max.freq*
`emotions`	593	72	6	27	4	81
`birds`	645	260	19	133	73	294
`ng20`	19 300	1 006	20	55	17	997

have been realized. As many as 4 of them are *hapaxes*, that is, they occur only once, while at least one of the labelsets has appeared 81 times. This wildly imbalanced behaviour is typical of MLC datasets [24]. As expected, distinct labels have distinct extensions. □

The pair of partitions of objects and attributes $(\ker\overline{\gamma}, \ker\overline{\mu}) \in \Pi(G)\times\Pi(M)$ are the concern of FEA [9]. In the present case, since no refinement of the label partition is feasible, we concentrate on describing how FEA of the object partition guides the resampling process of the learning algorithms.

The MLC induction and assessment procedures demand that we generate train and test resamplings of the original data. i.e. splitting the original context $\mathbb{D}$ into two subposed subcontexts of training $\mathbb{D}_T$ and testing $\mathbb{D}_E$ data so that $\mathbb{D} = \mathbb{D}_T/\mathbb{D}_E$. Note that:

1. Since the samples are supposed to be independent and identically distributed, the order of these contexts in the subposition, as indeed the reordering of the rows in the incidence, is irrelevant.
2. The resampling of the labelset context is tied to the resampling of the observations: we decide on the labelset information and this carries over to the observations.

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 we state the following:

Proposition 4. *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 lattice of all of these must be the same:*

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

Proof. Due to the identification of object intents and labelsets, we know that to respect the complexity of the labelset samples in each subcontext, one sufficient condition is that one of the labelsets associated to each block in the partition $\ker\overline{\gamma}$ is accorded to each of the subcontexts.

If this is the case, then the sampled subcontexts being join- and meet-dense, will generate isomorphic concept lattices. Since they each are a clarification of the original context $\mathbb{D}$, their concept lattices are all isomorphic. □

However, if we only retained the meet- and join-irreducibles to obtain these concept lattices, then the labelsets of reducible attributes would be lost and this would change the relative importance of the samples (both labels and observations, remember), which will therefore impact the induction scheme of the classifiers. Hence *not only the labelsets but also their frequencies of occurrence are important.*

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 *FEA allows us to spot possible problems with the classifier induction and validation schemes using resampling.*

Example 2 (Continued). Fig. 3 represents the labelset histogram of the `emotions` dataset. Recall there are 4 hapaxes and this should appear in any resampling of the data, so it is subject to the problems detected above, that are ubiquitous. Table 1 lists the number of hapaxes or single samples for other MLC datasets.

□

3.3 Label Dependence Modelling with FCA

Actually, whether BR or LP will outperform the other is presently acknowledged to depend on the degree of dependence on labels among themselves: if labels are mostly non-dependent, then the BR method is superior to LP, while the contrary is expected to hold when dependence between labels is commonplace [17, 16]. This line of work has evolved from the Classifier Chain approach [17, Chap. 7] as more and more solutions to MLC try to model explicitly such dependencies. Furthermore, it has also been recently hinted that performance measures also presuppose one model of dependence or other [25], hence explicit modeling of dependences has become an issue to understand the task.

A crucial step in these solutions is the determination of an order to model dependencies between attributes adequately. In this context, we believe the ability of FCA to obtain implications between attributes to be specially relevant, particularly implications with a single consequent. This study will be left for future work.

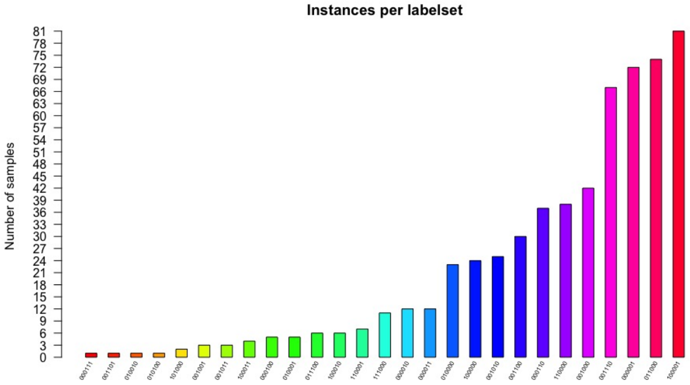

Fig. 3: (Color online) Label powerset histogram for `emotions`.

4 Summary and Discussion

We have sketched in this paper the basics of Formal Context Analysis (FxA), a framework to exploit the findings of FCA, FIA and FEA under a common umbrella, in order to help carry out the Exploratory Data Analysis of MLC tasks.

We have shown how labelset modelling is straightforward in this framework, and how FIA helps in the subdivision of the task into easier subtasks, and how FEA settles the table to understand how "stratified sampling" of MLC tasks can be made to respect the underlying formal conceptual model. We have also sketched how FCA could help in the actual design of classifiers that incorporate the dependencies between labels in the classifier induction step.

This is work in progress. Next steps will be to provide quantitative support for this framework by using the recently developed package `fcaR`, as well as filling in the details about how to use the probability mass distribution associated to $\ker \overline{\gamma}$, the partition of samples due to labelsets, in conjunction with the purely FxA methods.

References

[1] Tukey, J.W.: Exploratory data analysis. Addison-Wesley (1977)

[2] Wille, R.: Conceptual landscapes of knowledge: a pragmatic paradigm for knowledge processing. In Mineau, G., Fall, A., eds.: Proceedings of the Second International Symposium on Knowledge Retrieval, Use and Storage for Efficiency, Vancouver (1997) 2–13

[3] Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin, Heidelberg (1999)

[4] Valverde-Albacete, F.J., González-Calabozo, J.M., Peñas, A., Peláez-Moreno, C.: Supporting scientific knowledge discovery with extended, generalized formal concept analysis. Expert Systems with Applications **44** (2016) 198 – 216

[5] Düntsch, I., Gediga, G.: Modal-style operators in qualitative data analysis. In: Proc. IEEE International Conference on Data Mining, ICDM 2002. (2002) 155–162

[6] Dubois, D., Prade, H.: Possibility theory and formal concept analysis: Characterizing independent sub-contexts. Fuzzy Sets And Systems (2012)

[7] Valverde-Albacete, F., Peláez-Moreno, C., Cabrera, I., Cordero, P., Ojeda-Aciego, M.: Formal independence analysis. CCIS **853** (2018) 596–608

[8] Valverde-Albacete, F.J., Peláez-Moreno, C., Cabrera, I.P., Cordero, P., Ojeda-Aciego, M.: A Data Analysis Application of Formal Independence Analysis. In: Concept Lattices and their Applications (CLA'18). (2018) 1–12

[9] Valverde-Albacete, F.J., Peláez-Moreno, C., Cordero, P., Ojeda-Aciego, M.: Formal equivalence analysis. In: Proc. Conf. Internat. Fuzzy Syst. Assoc. and European Soc. Fuzzy Logic and Technol. (EUSFLAT 2019), Atlantis Press (2019)

[10] Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recognition **37**(9) (2004) 1757–1771

[11] Gibaja, E., Ventura, S.: A Tutorial on Multilabel Learning. ACM Computing Surveys **47**(3) (2015) 1–38

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

[13] Mirkin, B.: Clustering for Data Mining. A Data Recovery Approach. Computer Science and Data Analysis Series. Chapman & Hall (2005)

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

[15] 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

[16] Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In Maimon, O., Rokach, L., eds.: Data Mining and Knowledge Discovery Handbook. (2010) 667–685

[17] Read, J.: Scalable Multi-label Classification. PhD thesis, Univ. Waikato, Australia (2010)

[18] Davey, B., Priestley, H.: Introduction to lattices and order. 2nd edn. Cambridge University Press, Cambridge, UK (2002)

[19] Behrendt, G.: Maximal antichains in partially ordered sets. Ars Combinatoria **25**(C) (1988) 149–157

[20] Reuter, K.: The jump number and the lattice of maximal antichains. Discrete Mathematics **88**(2-3) (1991) 289–307

[21] Wille, R.: Finite distributive lattices as concept lattices. Atti Inc. Logica Mathematica **2** (1985) 635–648

[22] Bělohlávek, R.: Fuzzy Relational Systems. Foundations and Principles. Volume 20 of IFSR International Series on Systems Science and Engineering. Kluwer Academic (2002)

[23] Valverde-Albacete, F.J., Peláez-Moreno, C.: Extending conceptualisation modes for generalised Formal Concept Analysis. Information Sciences **181** (2011) 1888–1909

[24] Charte, F., Charte, F.D.: Working with multilabel datasets in R: The mldr package. The R journal **7**(5) (2015) 149–162

[25] Dembczyński, K., Waegeman, W., Cheng, W., Hüllermeier, E.: Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. In: Proc. European Conf. on Machine Learning, (ECML PKDD 2010). (2010) 280–295