Synthetic dataset generator for multi-label classification with control of label correlation

Maria Luiza S. de Sousa[1], **Marcelo de Souza Lauretto** [1], **Helena Brentani** [2], **Fátima L. S. Nunes** [1], **Ariane Machado-Lima**[1]

[1]Escola de Artes Ciências e Humanidades – Universidade de São Paulo
São Paulo – SP – Brasil

[2]Faculdade de Medicina – Universidade de São Paulo
São Paulo – SP – Brasil

marialuiza.silva@usp.br, marcelolauretto@usp.br

helena.brentani@gmail.com, fatima.nunes@usp.br, ariane.machado@usp.br

Abstract. *In Data Science, synthetic datasets allow comparing algorithms under controlled conditions. In multi-label classification, current dataset generators ignore label associations, which is relevant in areas such as medicine to represent comorbidities, for instance. This work proposes a generator that explicitly incorporates these associations. Each instance is created by two mechanisms: (i) label generation using conditional co-occurrence probabilities; (ii) feature generation with normal distributions whose mean depends on the present labels. Its differential is enabling algorithm analysis and comparison while simultaneously controlling label correlation and their separability in the feature space.*

1. Introduction

Multi-label classification (MLC) is a machine learning paradigm in which each instance can be simultaneously associated with multiple labels, in contrast to traditional single-label classification [Zhang and Zhou 2014]. This type of problem naturally arises in several real-world applications, such as medicine, text categorization, image annotation, bioinformatics, and recommender systems, where objects often exhibit multiple characteristics or belong to multiple classes at the same time [Tsoumakas and Katakis 2007].

One of the main characteristics that distinguishes multi-label classification is the presence of dependencies between labels. Unlike independent scenarios, labels in MLC often exhibit complex correlations that may carry relevant information for the learning process [Zhang and Zhou 2020]. Properly modeling these dependencies is a central challenge and is considered a critical factor for the performance of algorithms in this context.

Despite recent advances, the evaluation of multi-label classification methods still faces important challenges, particularly related to the availability and quality of datasets. In many cases, real-world datasets present limitations such as imbalance, noise, uncontrolled correlations, and lack of structural interpretability [Madjarov et al. 2012]. Furthermore, obtaining annotated data can be costly and, in some domains, impractical.

In this scenario, the use of synthetic data becomes a relevant alternative. The generation of artificial datasets allows the creation of controlled environments in which

specific data properties — such as distribution, separability, and label dependency — can be explicitly manipulated [Tomás et al. 2014]. This is particularly useful for the experimental analysis of algorithms, enabling systematic investigation of model behavior under different conditions. However, generating synthetic data for multi-label classification introduces additional challenges compared to the traditional case, as it is necessary to simultaneously control: (i) the probabilistic structure of the labels, including their dependencies, and (ii) the geometric structure of the features given the labels, which defines how the data are distributed in the feature space.

Existing approaches often address only one of these aspects. In particular, many methods either assume independence between labels or do not explicitly model their dependencies, despite the fact that label co-occurrence plays a central role in multi-label problems. While some approaches focus on the geometric generation of instances in the feature space [Tomás et al. 2014], others rely on general-purpose machine learning tools such as scikit-learn [Pedregosa et al. 2011], while some focus on resampling and balancing techniques, without providing explicit control over label dependency structures [Charte et al. 2015]. As a consequence, there is a gap in the literature for approaches that integrate both probabilistic dependencies between labels and the geometric structure of the feature space in a unified and interpretable manner.

To address this limitation, this work proposes a synthetic dataset generator for multi-label classification that combines: (i) an explicit probabilistic model to control label dependencies, and (ii) a geometric model based on theoretical centroids to control the distribution of features in the feature space. Label generation is performed using parameterized conditional distributions, allowing explicit definition of relationships such as $P(L_2 \mid L_1)$, where L_1 and L_2 are two labels to be present in the dataset. Feature generation, in turn, is based on sampling theoretical centroids in a vector space, with distance constraints that control class separability and data dispersion.

This approach enables the construction of datasets with interpretable and controllable properties, providing a suitable environment for experimental analysis and validation of multi-label classification methods.

2. Methods

The proposed model for synthetic data generation was designed to explicitly control both the probabilistic properties of the labels and the geometric structure of the features. To achieve this, the generation process is organized into two main stages: (i) label generation and (ii) feature generation conditioned on the labels. For simplicity, this work considers two binary labels, denoted by L_1 and L_2, where each label can take the value 1 (presence) or 0 (absence).

In the first stage, labels are assigned to each instance based on prior probabilities and conditional dependencies. In the second stage, feature vectors are generated in the feature space under four multivariate normal distributions defined by the presence/absence of labels, ensuring consistency with the assigned labels.

2.1. Label Generation

The label generation process follows a probabilistic approach that allows explicit modeling of dependencies between labels.

The label L_1 is sampled from a Bernoulli distribution with user-defined probability $P(L_1 = 1)$. After sampling L_1, the label L_2 is generated conditionally on the value of L_1. Two conditional probabilities are specified as model parameters:

$$P(L_2 = 1 \mid L_1 = 0) \quad \text{and} \quad P(L_2 = 1 \mid L_1 = 1).$$

This formulation allows direct control over the relationship between labels. For example, when $P(L_2 = 1 \mid L_1 = 1) > P(L_2 = 1 \mid L_1 = 0)$, the presence of L_1 increases the probability of occurrence of L_2, characterizing a positive dependency between labels. In this way, the model enables explicit control over label dependencies, allowing the construction of multi-label classification scenarios with different levels of correlation.

2.2. Feature Generation

We consider that each feature vector $\mathbf{x} \in \mathbb{R}^{\mathbf{k}}$ follows a multivariate Normal distribution, as follows. Initially, two vectors $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are generated in a k-dimensional space, corresponding to the theoretical centroids of labels L_1 and L_2, respectively. These vectors are sampled with a random direction and controlled magnitude, such that their distances to the origin lie within a predefined interval (Equation 1), while the minimum distance between them is also constrained (Equation 2). These constraints allow explicit control over the positioning of the theoretical centroids for each label and the degree of separation between them in the feature space (Figure 1).

$$d_{\text{min\_origin}} \leq \|\boldsymbol{\mu_i}\| \leq d_{\text{max\_origin}}, \; i = 1, 2; \tag{1}$$

$$\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\| \geq d_{\text{between}} \tag{2}$$

Based on these generated $\boldsymbol{\mu}$ vectors, a theoretical centroid C is defined for each label combination $L_1 \times L_2$:

- $\mathrm{C}(0, 0) = \mathbf{0^k}$;
- $\mathrm{C}(1, 0) = \boldsymbol{\mu_1}$;
- $\mathrm{C}(0, 1) = \boldsymbol{\mu_2}$;
- $\mathrm{C}(1, 1) = (\boldsymbol{\mu_1} + \boldsymbol{\mu_2})/2$.

These points are referred to as *theoretical centroids*, as they represent the ideal positions of the groups in the feature space.

The feature vector $\mathbf{x_i}$ of each instance i is then generated from these theoretical centroids by adding Gaussian noise:

$$\mathbf{x_i} = \boldsymbol{\mu} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ is the center associated with the instance's label combination and $\boldsymbol{\Sigma}$ is a user-defined covariance matrix.

Figure 1 illustrates the conceptual model used for feature generation, highlighting the main geometric elements and constraints imposed by the method. The origin of the feature space is used as a central reference point, and the positions of the theoretical centroids associated with the labels are restricted to an annular region defined by the

distances $d_{\min\_origin}$ and $d_{\max\_origin}$ from the origin. This region represents the set of valid positions for the theoretical centroids, ensuring control over their magnitude.

Within this region, the theoretical centroids $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$, associated with labels L_1 and L_2, are positioned in random and unbiased directions. Additionally, a constraint on the distance between them, represented by d_{between}, directly controls the separability between groups, with larger distances leading to greater separation in the feature space.

Around each centroid, data points are generated from normal distributions, resulting in dispersion regions that represent intra-class variability and are controlled by the noise introduced during instance generation. Furthermore, when multiple labels are present in an instance, the corresponding generating center is defined as the mean of the individual centroids, producing intermediate regions that reflect the combined influence of the labels. Overall, this conceptual representation provides an intuitive understanding of how the model parameters influence the geometry of the generated data, including cluster positions, separation, and dispersion.

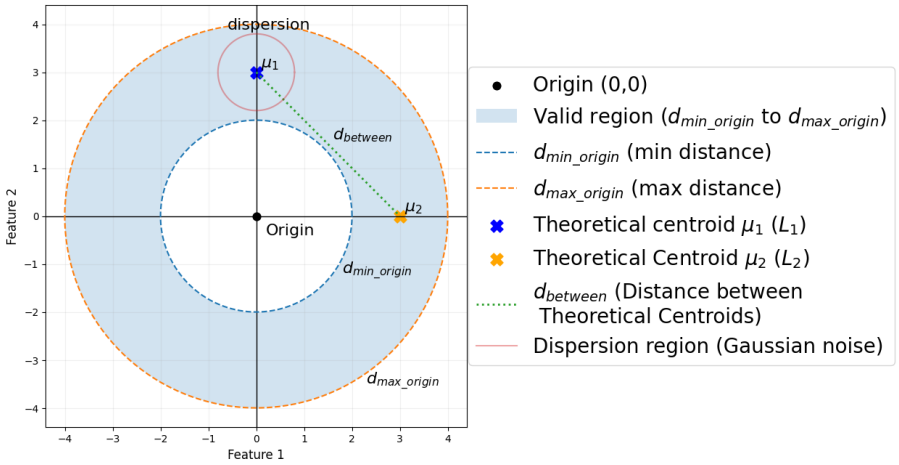

Figure 1. Conceptual model of feature generation in two dimensions. Centroids are placed within an annular region ($d_{\min\_origin}$, $d_{\max\_origin}$), with separation controlled by d_{between}. Points are sampled around them, forming clusters.

2.3. Experimental Setup

To evaluate the behavior of the proposed model, controlled experiments were conducted using different configurations of geometric parameters given a particular setup of the probabilistic structure of the labels.

In all experiments, datasets containing 500 instances were generated, with 10 independent datasets created for each configuration. In Section 3, Figures 2, 3, and 4 correspond to a representative example selected from these executions for each configuration. The remaining datasets exhibited similar behavior, consistent with the proposed model. This approach allows analyzing result variability and the stability of the generation process.

The probabilistic parameters were kept constant across all experiments and defined as follows:

- $P(L_1 = 1) = 0.3$

- $P(L_2 = 1 \mid L_1 = 0) = 0.7$
- $P(L_2 = 1 \mid L_1 = 1) = 0.4$

Additionally, to allow direct visualization in two-dimensional space, the experiments used feature vectors $\mathbf{x} \in \mathbb{R}^2$. For simplicity, we set $\boldsymbol{\Sigma} = \boldsymbol{I}$, which implies that the features are uncorrelated and each has unit variance.

Three experimental configurations were considered, varying the geometric parameters:

Table 1. Experimental configurations

Experiment	$d_{\min\_origin}$	$d_{\max\_origin}$	d_{between}
1	4	5	3
2	2	3	4
3	4	5	1

These configurations were designed to analyze the direct impact of geometric constraints on the data distribution, particularly regarding group separability and centroid positioning. The results were analyzed through visual inspection of the feature space, allowing the evaluation of cluster behavior under different generation conditions. The empirical centroids were also calculated to inspect their match to the theoretical centroids.

3. Results and Discussion

This section presents the results obtained in the synthetic dataset generation under the experimental configurations described in Section 2.3, focusing on a qualitative analysis of the data distribution in the feature space. For each experiment, the two-dimensional features distributions and their labels are presented, allowing direct visualization of the spatial organization of the data and the separation between groups associated with different label combinations.

Figures 2, 3, and 4 illustrate the distribution of instances for different configurations of the parameters $d_{\min\_origin}$, $d_{\max\_origin}$, and d_{between} defined in Table 1. The analysis focuses on the impact of these distances on group separability and centroid positioning.

3.1. Experiment 1

In Experiment 1, the theoretical centroids are positioned farther from the origin, according to the interval $[4, 5]$. This results in more dispersed clusters in the feature space. The constraint $d_{\text{between}} = 3$ ensures a moderate separation between centroids associated with labels L_1 and L_2.

Visually, it can be observed that groups (10)[1] and (01) are distinguishable, while group (11) appears in an intermediate position, as expected. Some overlap between clusters is still present due to the dispersion introduced by Gaussian noise, which is also expected in real datasets.

[1]Group (xy) corresponds to instances containing labels L_1 = x and L_2 = y.

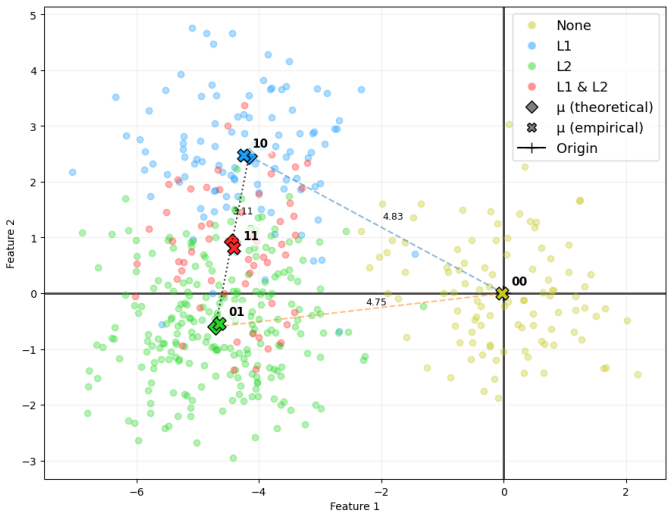

Figure 2. Two-dimensional feature distribution for a representative dataset. Theoretical and empirical centroids are shown, along with the Euclidean distance between each pair of centroids. The parameters d_{min_origin} and d_{max_origin} define the valid region for centroid placement.

3.2. Experiment 2

In Experiment 2, the theoretical centroids are positioned closer to the origin, according to the interval $[2, 3]$. However, increasing the parameter d_{between} to 4 imposes greater separation between them.

As a result, groups (10) and (01) become more distant from each other, significantly reducing cluster overlap. Even though the centroids are closer to the origin, the increased inter-centroid distance dominates the geometry of the problem, leading to improved separability.

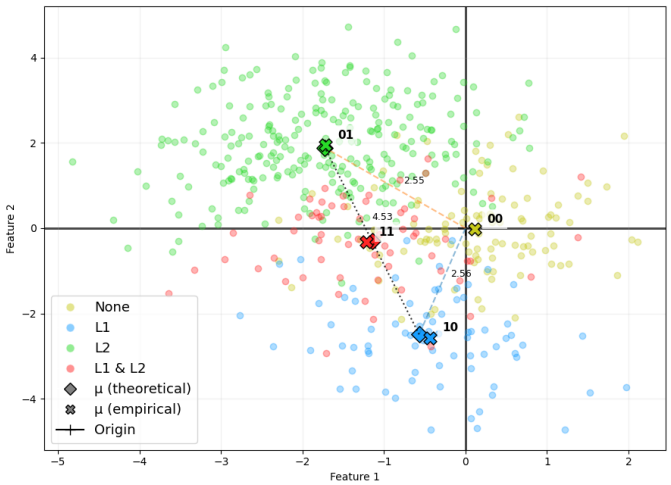

Figure 3. Dataset generated with $d_{\min\_origin} = 2$, $d_{\max\_origin} = 3$, and $d_{\text{between}} = 4$. Theoretical and empirical centroids are shown, along with the Euclidean distance between each pair of centroids. The centroids are closer to the origin, but more separated from each other.

3.3. Experiment 3

In Experiment 3, the centroids are again positioned far from the origin, but with a very small distance between them ($d_{\text{between}} = 1$).

This configuration may produce highly overlapping clusters, significantly increasing the overlap between groups. As a consequence, it becomes more difficult to distinguish the different label combinations in the feature space. In the example presented in Figure 4, group (11), being located near the mean of the centroids, shows strong overlap with the others, highlighting the direct impact of reducing centroid distance on data separability.

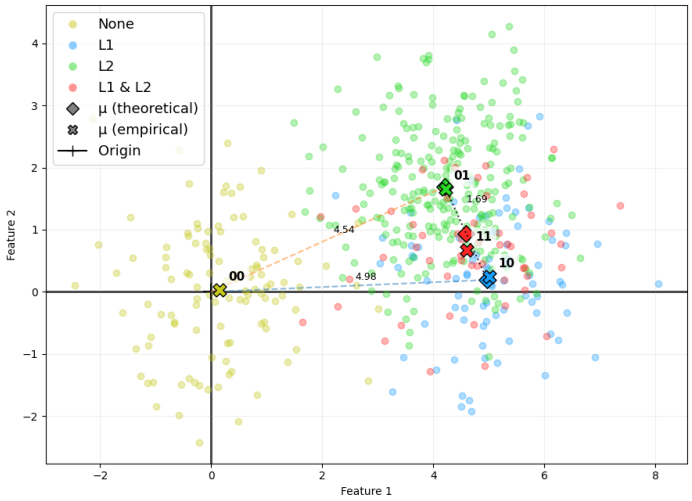

Figure 4. Dataset generated with $d_{\min\_origin} = 4$, $d_{\max\_origin} = 5$, and $d_{\text{between}} = 1$. Theoretical and empirical centroids are shown, along with the Euclidean distance between each pair of centroids. The centroids are far from the origin but very close to each other, resulting in high overlap.

3.4. General discussion

The obtained results demonstrate that the proposed model allows explicit control over the geometric properties of the generated data. The parameter d_{between} plays a central role in defining the separability between groups, while $d_{\min\_origin}$ and $d_{\max\_origin}$ control the global positioning of centroids in the space.

In addition, it is observed that the empirical centroids remain close to the theoretical centroids in all scenarios, indicating that the feature generation process is consistent with the defined model.

Overall, the experiments confirm that the generator is capable of producing controlled and interpretable geometric structures, enabling the analysis of classification algorithms under different levels of class separation.

4. Conclusion

This work presented a synthetic dataset generator for multi-label classification that explicitly integrates probabilistic and geometric aspects in data construction. The proposed approach allows control over both label dependencies, through conditional probabilities,

and feature distribution in the feature space, through a centroid-based model with distance constraints.

Experimental results showed that the defined geometric parameters directly influence data separability, enabling the generation of scenarios with different levels of class overlap. Furthermore, the proximity between theoretical and empirical centroids indicates the consistency of the generation process.

As the main contribution, the model provides a controlled and interpretable environment for evaluating multi-label classification algorithms, allowing the analysis of how specific data properties, such as label correlations, affect model performance.

Future work includes extending the model to a larger number of labels, exploring alternative centroid combination strategies, and quantitatively evaluating classification algorithms on the generated datasets.

Acknowledgments

This work was supported by the National Council for Scientific and Technological Development (CNPq), under grant number 307710/2022-0, and by São Paulo Research Foundation (FAPESP) grants #2025/03097-2, #2025/07136-2, #2025/07394-1 and #2022/03698-8.

Declaration of AI use

The authors used Writefull, an AI-based language tool, to refine the English expression and grammar of this manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Charte, F., Rivera, A. J., del Jesus, M. J., and Herrera, F. (2015). MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397.

Madjarov, G., Kocev, D., Gjorgjevikj, D., and Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Tomás, J. T., Spolaôr, N., Cherman, E. A., and Monard, M. C. (2014). A framework to generate synthetic multi-label datasets. *Electronic Notes in Theoretical Computer Science*, 302:155–176.

Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: An overview. In *International Journal of Data Warehousing and Mining*.

Zhang, M.-L. and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.

Zhang, M.-L. and Zhou, Z.-H. (2020). Multi-label learning: From problem transformation to algorithm adaptation. *IEEE Transactions on Knowledge and Data Engineering*.