Multi-Label Dataset Analyzer
User Guide

Contents

1 Introduction **1**

2 Metrics **5**
2.1 Notation 5
2.2 Dimensionality metrics 5
2.2.1 Attributes 5
2.2.2 Labels 5
2.2.3 Instances 5
2.2.4 Labelsets 5
2.2.5 LxIxF 7
2.2.6 Ratio of number of instances to the number of attributes 7
2.3 Labels distribution metrics 7
2.3.1 Cardinality 7
2.3.2 Density 7
2.3.3 Frequency 7
2.3.4 Standard deviation of label cardinality 8
2.3.5 Minimal, maximal and mean of entropy of labels . . 8
2.4 Imbalance metrics 8
2.4.1 Mean of IR inter-class 8
2.4.2 Max of IR inter-class 9
2.4.3 CVIR inter-class 9
2.4.4 Mean of IR intra-class 9
2.4.5 Max of IR intra-class 9
2.4.6 Mean of standard deviation of IR intra-class 10
2.4.7 Mean of IR per labelset 10
2.4.8 Max of IR per labelset 10
2.4.9 Kurtosis cardinality 10
2.4.10 Skewness cardinality 11
2.4.11 Proportion of maxim label combination (PMax) . . . 11
2.4.12 Proportion of unique label combination (PUniq) . . . 12
2.5 Labels relationship metrics 12
2.5.1 Bound 12
2.5.2 Diversity 12
2.5.3 SCUMBLE 12
2.5.4 Proportion of distinct labelsets 12
2.5.5 Number of labelsets up to n examples 13
2.5.6 Ratio of labelsets up to n examples 13
2.5.7 Average examples per labelset 13
2.5.8 Standard deviation of examples per labelset 13
2.5.9 Number of unique labelsets 13
2.5.10 Ratio of labelsets with number of examples less than half of the attributes 13

2.5.11 Number of unconditionally dependent label pairs by chi-square test . . . 14
2.5.12 Ratio of unconditionally dependent label pairs by chi-square test . . . 14
2.5.13 Average of unconditionally dependent label pairs by chi-square test . . . 14
2.6 Attributes metrics . . . 15
2.6.1 Number of binary attributes . . . 15
2.6.2 Number of nominal attributes . . . 15
2.6.3 Number of numeric attributes . . . 15
2.6.4 Proportion of binary attributes . . . 15
2.6.5 Proportion of nominal attributes . . . 15
2.6.6 Proportion of numeric attributes . . . 15
2.6.7 Proportion of numeric attributes with outliers . . . 16
2.6.8 Mean of entropies of nominal attributes . . . 16
2.6.9 Mean of mean of numeric attributes . . . 16
2.6.10 Mean of standard deviation of numeric attributes . . . 17
2.6.11 Average gain ratio . . . 17
2.6.12 Average absolute correlation between numeric attributes 17
2.6.13 Mean of kurtosis . . . 18
2.6.14 Mean of skewness of numeric attributes . . . 18

3 MLDA user guide 21
3.1 Downloading and executing . . . 21
3.2 Format of the datasets . . . 21
3.2.1 Multi-label datasets . . . 21
3.2.2 Multi-view multi-label datasets . . . 23
3.3 Summary tab . . . 23
3.4 Preprocess . . . 25
3.4.1 Instance selection . . . 25
3.4.2 Feature selection . . . 26
3.4.3 Data partitioning . . . 27
3.4.4 Dataset format conversion . . . 28
3.5 Transformation . . . 29
3.6 Labels tab . . . 31
3.7 Attributes tab . . . 38
3.8 Dependences tab . . . 39
3.8.1 Chi and Phi coefficients . . . 39
3.8.2 Co-occurrence values and co-ocurrence graphs . . . 40
3.8.3 Conditional probabilities and heatmap graphs . . . 42
3.9 Multiple datasets . . . 44
3.10 MVML . . . 46

4 API 49
4.1 Metrics . . . 49
4.2 Structure . . . 49
4.3 Usage . . . 50
4.4 Extending the API with new metrics . . . 53

References 55

Chapter 1

Introduction

The objective of data mining is to extract relevant and potentially useful knowledge under large amounts of information. Classification is one of the main tasks of data mining. Given a set of objects, each one labeled with a pre-defined class, for each new object whose class is unknown, tries to predict which is its class.

An example of classic classification is the Iris dataset [9]. In this problem, depending on the size of petals and sepals of Iris plant, they are classified in one of its three types: setosa, virginica or versicolor. Each instance can be classified only in one class, since each Iris plant can be only of one type (Figure 1.1).

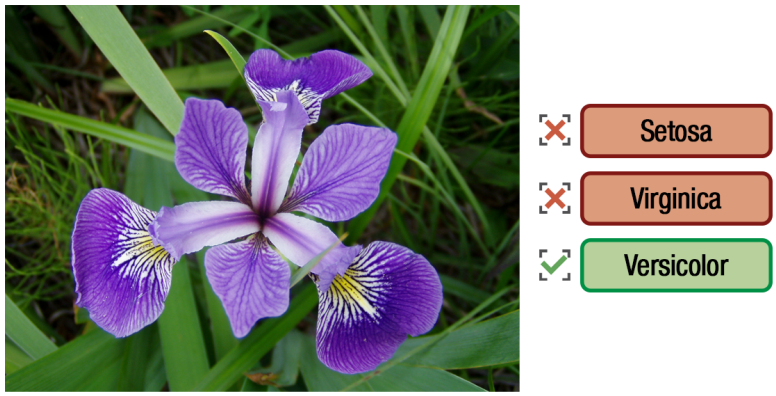


FIGURE 1.1: Classification example with Iris dataset

This definition corresponds to the classic version of classification, which implies the restriction of one class per pattern. However, there are a lot of real-world problems where objects can be associated with more than one class simultaneously. For example, in image annotation, as in Figure 1.2, having 4 labels previously defined, the image could be associated to the "Mountain", "Forest" and "River" labels. In classic classification, for the same example, the image could only be classified in one class or label, so the output would not be complete.

Thereby, they appear new paradigms in machine learning and data mining, which represents the information in a more flexible way. In Multi-Label Learning (MLL), unlike classic learning, an instance could have several classes (labels) simultaneously associated. This kind of learning has experimented a big growth in the last years, due to its success in problems as text categorization [15], social networks mining [25], medical diagnosis [23] or direct marketing [31].

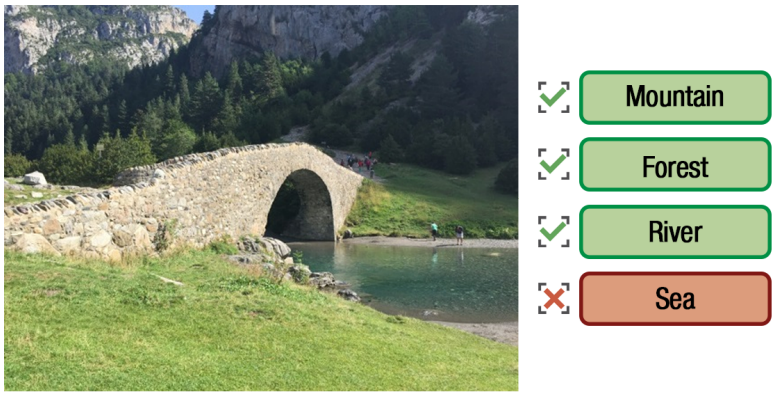


FIGURE 1.2: Example of object with several labels

On the other hand, in many fields of application, an object has different representations and is described by several sets of features (views), which have been obtained from different data sources or feature extractors. As a result, the descriptor variables (input space) can be partitioned. For instance, in image annotation (Figure 1.3), many times a single representation is not able to characterize all the classes. With different representations, each view would characterize a different concept (label). For instance, colour information (RGB), shape cue (SIFT) or global structure (GIST) could represent natural substances (e.g. sky or clouds), man-made objects (e.g. plane or motorbike) or scenes (e.g. seaside) respectively [16].

FIGURE 1.3: Example of object with multiple views

Despite the fact that many real-world data are are stored in a distributed and partitioned way and with an heterogeneous organization in multiple data sources, traditional algorithms infer models from a single homogeneous data set, which combines or merges different set of features. This methodology does not perform properly in every problem and may complicate the learning process. Multi-View Learning (MVL) tries to solve this type of problems, combining heterogeneous and complementary information from distinct views [30].

As a result of the hybridization of this two techniques, the Multi-View Multi-Label learning paradigm (MVML) appears. This more flexible learning approach allows to each pattern being represented by several sets of attributes and also being associated with several labels. Figure 1.4 shows an example of the image as an object with multiple views and multiple labels.

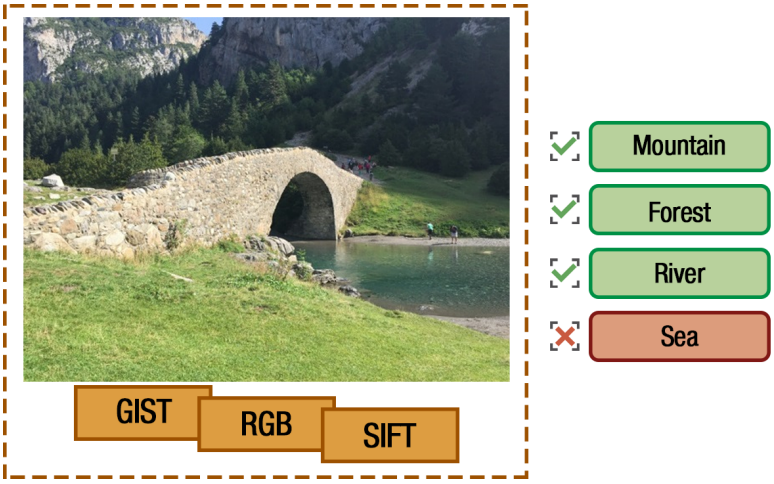


FIGURE 1.4: Example of object with multiple views and multiple labels

In MLL and MVML problems, many datasets have characteristics as imbalance (some labels or label combinations are very frequent while other are barely present), high dimensionality (in number of instances, attributes and labels) or relationship among labels [10]. Also, these characteristics could have direct effect on algorithms performance, hence the importance on characterizing and analyzing data [7]. This has led some authors to go further from previous exploratory analysis, developing meta-learning models based on these metrics in order to determine which algorithm is more appropriate according to the characteristics of each dataset. Furthermore, there are two libraries for multi-label learning, Mulan [29] and Meka [17], which implements different algorithms and also have different dataset formats. For that, it could be interesting to convert the format of the datasets. By last, for MVML datasets, as input space is partitioned, it is interesting to characterize each view separately or filter some of them.

The main objective of MLDA (Multi-Label and multi-view Datasets Analyzer) is to offer a tool for data exploration and analysis of multi-label and multi-view multi-label datasets, which allows to represent its more representative characteristics and save its values for further analysis. Moreover, it allows to perform an analysis of imbalance and relationship among labels. MLDA also includes preprocessing tasks as data partitioning, feature and instance selection, transformations and dataset format conversion. The tool allows to load some datasets in order to calculate its characteristics simultaneously. In case of MVML datasets, it also allows to get some specific metrics for multi-view multi-label learning, shows the attribute list of each view and let save a new dataset with a set of user-selected views. Finally, it is implemented a Java API for multi-label dataset characterization, including the full set of metrics of MLDA. All this functionality allows to analyze

problems in order to adapt better the technique selected for its solution, as well as transform and preprocess them in an easy way.

Chapter 2

Metrics

A taxonomy of metrics for multi-label datasets characterization is defined in [6]. Starting from this taxonomy, the basic metrics from Mulan and Meka are also included. Finally, the set of metrics proposed in [7] have been categorized into the previous taxonomy, updating it with a new group of attribute metrics. Therefore, metrics are categorized into the following groups: dimensionality, labels distribution, relationship among labels, labels imbalance and attribute metrics. Figure 2.1 shows the taxonomy and metrics composing each group.

2.1 Notation

Formally, let $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_q\}$ be the set of labels of the dataset. Each subset of $\mathcal{L}$ is also called labelset. $\mathcal{T} = \{t_1, t_2, \ldots t_n\}$ is the set of input features. $\mathcal{D} = \{(T_1, l_1), (T_2, l_2), \ldots (T_m, l_m)\}$ is the multi-label dataset, composed by a set of m instances, being T_i the feature vector associated with the instance i and $l_i \subseteq \mathcal{L}$ the subset of labels associated with the instance.

2.2 Dimensionality metrics

This kind of metrics indicates the size or dimensionality of the dataset, either in number of instances, attributes or labels.

2.2.1 Attributes

This metric indicates the number of attributes or input features $|\mathcal{T}|$ of the dataset.

2.2.2 Labels

It indicates the number of labels or output features $|\mathcal{L}|$ of the dataset.

2.2.3 Instances

This metric indicates the number of instances $|\mathcal{D}|$ of the dataset.

2.2.4 Labelsets

It indicates the number of label combinations appearing in the dataset. Each combination can be represented as a binary vector, where the i^{th} position represents with an 1 or a 0 if the i^{th} label is present or not respectively. It is also called $DistinctLabelsets(\mathcal{D})$ or $DL(\mathcal{D})$

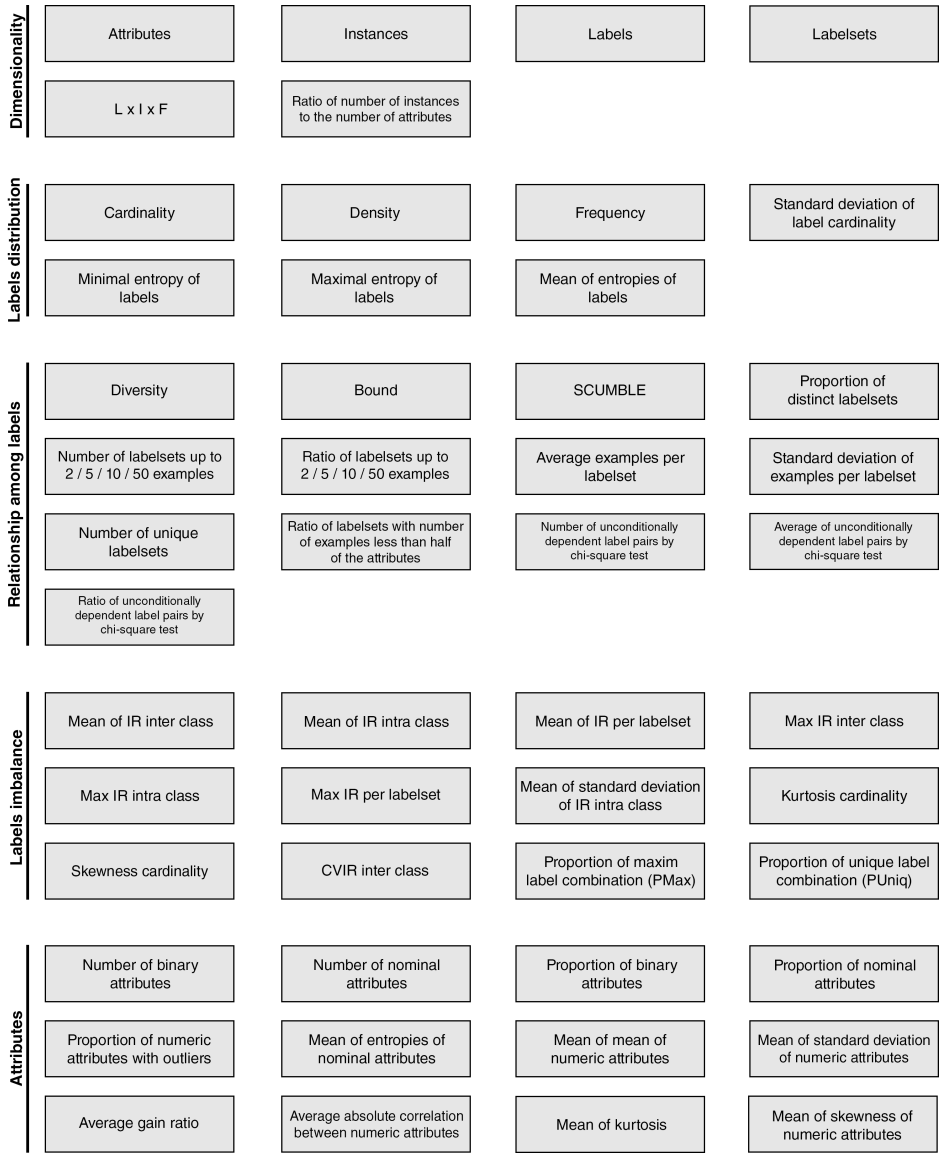


FIGURE 2.1: Taxonomy of metrics for MLL datasets characterization

2.2.5 LxIxF

LxIxF or *Labels* $\times$ *Instances* $\times$ *Features* is a metric proposed in [19], useful for measuring the complexity of the dataset regarding to the number of labels, instances and attributes.

$$LxIxF(\mathcal{D}) = |\mathcal{L}| \times |\mathcal{D}| \times |\mathcal{T}| \tag{2.1}$$

2.2.6 Ratio of number of instances to the number of attributes

It calculates the proportion between the number of instances and the number of attributes [7], useful to know the complexity of the dataset. Thus, if the metric value is under 0, it indicates that the number of attributes is greater than the number of instances, while if it is greater than 0, the number of instances is greater than the number of attributes.

$$Ratio_{inst-attr}(\mathcal{D}) = \frac{|\mathcal{D}|}{|\mathcal{T}|} \tag{2.2}$$

2.3 Labels distribution metrics

This kind of metrics indicates the distribution of labels in the dataset.

2.3.1 Cardinality

Cardinality is defined as the mean number of labels associated for instance. This metric was introduced in [27].

$$Card(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} |l_i| \tag{2.3}$$

2.3.2 Density

Density is defined as cardinality divided by the number of labels. Thus, it relates the mean number of labels associated with each instance with the total number of labels. This metric was introduced in [27]. The density is a good metric to have an idea about the frequency of labels [19].

$$Dens(\mathcal{D}) = \frac{Card(\mathcal{D})}{|\mathcal{L}|} \tag{2.4}$$

2.3.3 Frequency

The frequency of a label is defined as the number of appearances of this label divided by the total number of instances.

$$Freq(\lambda) = \frac{|\lambda|}{|\mathcal{D}|} \tag{2.5}$$

2.3.4 Standard deviation of label cardinality

It calculates the standard deviation of the number of labels associated with each instance [7]. Being $Card(\mathcal{D})$ the mean number of labels associated with each instance, the metric is defined as:

$$StdvCard(\mathcal{D}) = \sqrt{\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}(|l_i| - Card(\mathcal{D}))^2} \tag{2.6}$$

2.3.5 Minimal, maximal and mean of entropy of labels

This metrics returns the minimum, maximum and mean of entropies of labels [7, 26]. The entropy of a set of nominal values is defined as $H(X) = -\sum_i p(x_i)log_2p(x_i)$. As each label can take 1 and 0 values, the frequency of the label is taken as the probability of being 1, and 1 minus the frequency the probability of being 0. Thus, the entropy for a label is defined as:

$$H(\lambda) = -\left[Freq(\lambda)\cdot log_2(Freq(\lambda)) + (1 - Freq(\lambda))\cdot log_2(1 - Freq(\lambda))\right] \tag{2.7}$$

The greater the entropy is, the more uniform the distribution of label values is. So, the metrics of minimal, maximal and mean of entropies of labels are defined as:

$$MinEntropy_{Labels}(\mathcal{D}) = \arg\min_{\lambda\in\mathcal{L}}(H(\lambda)) \tag{2.8}$$

$$MaxEntropy_{Labels}(\mathcal{D}) = \arg\max_{\lambda\in\mathcal{L}}(H(\lambda)) \tag{2.9}$$

$$MeanEntropy_{Labels}(\mathcal{D}) = \left(\frac{1}{|\mathcal{L}|}\sum_{i=1}^{|\mathcal{L}|}H(\lambda_i)\right) \tag{2.10}$$

2.4 Imbalance metrics

This kind of metrics shows the existing imbalance in the dataset. In many cases, the imbalance indicates that there are labels or labelsets that appears with more frequency than others.

2.4.1 Mean of IR inter-class

The IR (Imbalance Ratio) inter-class measures the degree of imbalance among labels [4]. IR inter-class for a specific label is obtained dividing the number of positive examples of most frequent label by the number of positive examples of the current label. Thus, the mean IR inter-class is defined as the average value for all labels.

$$IR_{InterClass}(\lambda) = \frac{\arg\max_{\lambda'\in\mathcal{L}}(Freq(\lambda'))}{Freq(\lambda)} \tag{2.11}$$

$$MeanIR_{InterClass}(\mathcal{D}) = \frac{1}{|\mathcal{L}|}\sum_{i=1}^{|\mathcal{L}|} IR_{InterClass}(\lambda_i) \tag{2.12}$$

2.4.2 Max of IR inter-class

This metric returns the maximum value of IR inter-class, which corresponds to the most imbalanced label.

$$MaxIR_{InterClass}(\mathcal{D}) = \arg\max_{\lambda\in\mathcal{L}}(IR_{InterClass}(\lambda)) \tag{2.13}$$

2.4.3 CVIR inter-class

It is defined as the coefficient of variation of IR inter-class [4].

Being:

$$\sigma_{IR_{InterClass}}(\mathcal{D}) = \sqrt{\sum_{i=1}^{|\mathcal{L}|} \frac{(IR_{InterClass}(\lambda_i) - MeanIR_{InterClass}(\mathcal{D}))^2}{|\mathcal{L}| - 1}} \tag{2.14}$$

the metric is defined as:

$$CVIR_{InterClass}(\mathcal{D}) = \frac{\sigma_{IR_{InterClass}}(\mathcal{D})}{MeanIR_{InterClass}(\mathcal{D})} \tag{2.15}$$

2.4.4 Mean of IR intra-class

IR intra-class measures the degree of imbalance inside the label [24]. For a specific label, IR intra-class is calculated dividing the number of examples of the majority value of the class by the number of examples of the minority value. Thereby, the mean of IR intra-class is defined as the average value for all the labels. Being $Max(\lambda) = \arg\max(Count(\lambda), |\mathcal{D}| - Count(\lambda))$ a function returning the number of instances of the majority value of the label, either the number of instances that have associated the label or the number of instances that not have it, and $Min(\lambda) = \arg\min(Count(\lambda), |\mathcal{D}| - Count(\lambda))$ the number of instances of the minority value.

$$IR_{IntraClass}(\lambda) = \frac{Max(\lambda)}{Min(\lambda)} \tag{2.16}$$

$$MeanIR_{IntraClass}(\mathcal{D}) = \frac{1}{|\mathcal{L}|}\sum_{i=1}^{|\mathcal{L}|} IR_{IntraClass}(\lambda_i) \tag{2.17}$$

2.4.5 Max of IR intra-class

This metric returns the maximum value of IR intra-class, which corresponds to the IR value of the most imbalanced label.

$$MaxIR_{IntraClass}(\mathcal{D}) = \arg\max_{\lambda\in\mathcal{L}}(IR_{IntraClass}(\lambda)) \tag{2.18}$$

2.4.6 Mean of standard deviation of IR intra-class

It computes the mean of the standard deviations intra-class of each label. The standard deviation intra-class is defined taking into account the number of positive instances for the label (instances that have the label associated, L^p) and negative instances (instances that have not the label associated, L^n).

$$Stdv_{IntraClass}(\lambda) = \sqrt{\frac{(L_i^p - nInstances/2)^2 + (L_i^n - nInstances/2)^2}{2}} \tag{2.19}$$

$$MeanStdv_{IntraClass}(\mathcal{D}) = \frac{1}{|\mathcal{L}|}\sum_{i=1}^{|\mathcal{L}|} Stdv_{IntraClass}(\lambda_i) \tag{2.20}$$

2.4.7 Mean of IR per labelset

The IR per labelset is defined in the same way that IR inter-class, but instead of for each label, it is calculated for each labelset. It is calculated dividing the number of instances associated with the most frequent labelset by the number of instances of the current labelset.

$$IR_{labelset}(l) = \frac{\arg\max_{l' \in DL(\mathcal{D})}(Freq(l'))}{Freq(l)} \tag{2.21}$$

$$MeanIR_{labelset}(\mathcal{D}) = \frac{1}{DL(\mathcal{D})}\sum_{l \in DL(\mathcal{D})} IR_{labelset}(l) \tag{2.22}$$

2.4.8 Max of IR per labelset

This metric returns the maximum value of IR per labelset, which corresponds to the IR value of the most imbalanced labelset.

$$MaxIR_{labelset}(\mathcal{D}) = \arg\max_{l \in DL(\mathcal{D})}(IR_{labelset}(l)) \tag{2.23}$$

2.4.9 Kurtosis cardinality

This metric applies the Kurtosis formula [20] to the number of associated labels of each instance. The higher Kurtosis (ku) value, the higher the concentration of values near the mean of the distribution. If $ku > 0$, values are near to the mean, while if $ku < 0$ values are distant from the mean.

Kurtosis is defined as:

$$Kurtosis = \frac{(n-1)}{(n-2)(n-3)}((n+1)g_{ku} + 6) \tag{2.24}$$

being $g_{ku} = \frac{m_4}{(m_2)^2} - 3$, where m_2 is the second moment of the sample with respect to the mean $m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ and m_4 is the fourth moment of the sample $m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4$.

In this case, n is the total number of instancesc, x_i the number of labels associated with the instance i and $\overline{x}$ is the dataset cardinality. So, m_2 y m_4 are defined as:

$$m_2(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}(|l_i| - Card(\mathcal{D}))^2 \tag{2.25}$$

$$m_4(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}(|l_i| - Card(\mathcal{D}))^4 \tag{2.26}$$

Therefore, g_{ku} is defined as $g_{ku}(\mathcal{D}) = \frac{m_4(\mathcal{D})}{(m_2(\mathcal{D}))^2} - 3$. Finally, Kurtosis cardinality is defined as:

$$Kurtosis(\mathcal{D}) = \frac{(|\mathcal{D}| - 1)}{(|\mathcal{D}| - 2)(|\mathcal{D}| - 3)}((|\mathcal{D}| + 1)g_{ku}(\mathcal{D}) + 6) \tag{2.27}$$

2.4.10 Skewness cardinality

It measures the skewness of the number of labels associated with each instance, using the skewness metric defined in [7]:

$$Skewness = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\frac{(x_i - \overline{x})^3}{\sigma^3} \tag{2.28}$$

If skewness value is greater than 0, values are biased to the right, while if the value is negative, they are biased to the left. In this case, n is the number of instances, x_i is the number of associated labels to the instance i, $\overline{x}$ is the dataset cardinality and σ^3 is the deviation of the number of associated labels per instance, expressed as:

$$\sigma^3 = \left[\frac{1}{|\mathcal{D}| - 1}\sum_{i=1}^{|\mathcal{D}|}(|l_i| - Card(\mathcal{D}))^3\right]^{\frac{3}{2}} \tag{2.29}$$

So, finally, skewness cardinality is defined as:

$$Skewness_{Card}(\mathcal{D}) = \frac{|\mathcal{D}|}{(|\mathcal{D}| - 1)(|\mathcal{D}| - 2)}\sum_{i=1}^{|\mathcal{D}|}\frac{(|l_i| - Card(\mathcal{D}))^3}{\sigma^3} \tag{2.30}$$

2.4.11 Proportion of maxim label combination (PMax)

This metric calculates the proportion of instances associated with the most frequent labelset. According to [19], if this metric has high values, it is considered *label skew*, i.e. there is a high number of instances that have associated the most frequent labelset.

$$PMax(\mathcal{D}) = \arg\max_{l \in DL(\mathcal{D})}\frac{|l|}{|\mathcal{D}|} \tag{2.31}$$

2.4.12 Proportion of unique label combination (PUniq)

It measures the proportion of labelsets that are only associated with an instance, divided by the number of instances of the dataset [19].

$$PUniq(\mathcal{D}) = \frac{|l \subseteq DL(\mathcal{D}) : |l| = 1|}{|\mathcal{D}|} \tag{2.32}$$

2.5 Labels relationship metrics

This kind of metrics measures the relationship among labels.

2.5.1 Bound

It represents the maximum number of labelsets that may exist in the dataset [27].

$$Bound(\mathcal{D}) = 2^{|\mathcal{L}|} \tag{2.33}$$

2.5.2 Diversity

It represents the percentage of labelsets present in the dataset divided by the number of possible labelsets.

$$Diversity(\mathcal{D}) = \frac{DL(\mathcal{D})}{Bound(\mathcal{D})} \tag{2.34}$$

2.5.3 SCUMBLE

This metric defined in [5] aims to quantify the variation of the imbalance among labels of each instance. It is based on Atkinson index [2] and *IR inter-class*. While the former is a metric to evaluate inequalities among individuals in a population, the latter measures the degree of imbalance of the dataset. To calculate the Atkinson index, each instance is considered as a population and each active label as an individual. If label λ_i is active in the instance d, then $IR_{d\lambda_i} = IR_{InterClass}(\lambda_i)$, and 0 otherwise. Being $\overline{IR_d}$ the average value of $IR_{InterClass}(\lambda_i)$ of the labels associated with the instance d, the metric is defined as:

$$SCUMBLE(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{d=1}^{|\mathcal{D}|}\left[1 - \frac{1}{\overline{IR_d}}\left(\prod_{i=1}^{|\mathcal{L}|} IR_{d\lambda_i}\right)^{\frac{1}{|\mathcal{L}|}}\right] \tag{2.35}$$

2.5.4 Proportion of distinct labelsets

This metric calculates the proportion between the number of distinct labelsets and the number of instances of the dataset [27].

$$ProportionDistinctLabelsets(\mathcal{D}) = \frac{DL(\mathcal{D})}{|\mathcal{D}|} \tag{2.36}$$

2.5.5 Number of labelsets up to *n* examples

This metric returns the number of labelsets appearing the up to n times in the dataset. It is defined in [7], and it uses as default values of n: 2, 5, 10 and 50.

$$NumLabelsetsUpToN(\mathcal{D}, n) = |l \subseteq \mathcal{L} : |l| \leq n| \tag{2.37}$$

2.5.6 Ratio of labelsets up to *n* examples

It returns the ratio of labelsets appearing the up to n times in the dataset. It is defined in [7], and it uses as default values of n: 2, 5, 10 and 50.

$$RatioLabelsetsUpToN(\mathcal{D}, n) = \frac{NumLabelsetsUpToN(\mathcal{D}, n)}{DistinctLabelsets(\mathcal{D})} \tag{2.38}$$

2.5.7 Average examples per labelset

This metric is calculated as the average number of instances associated with each labelset [7]. It is defined as:

$$AvgExamplesLabelset(\mathcal{D}) = \frac{|\mathcal{D}|}{DL(\mathcal{D})} \tag{2.39}$$

2.5.8 Standard deviation of examples per labelset

It calculates the standard deviation of the number of examples per labelset [7], defined as:

$$StdvExamplesLabelset(\mathcal{D}) = \\ = \sqrt{\frac{1}{DL(\mathcal{D})} \sum_{i=1}^{|DL(\mathcal{D})|} (|l_i| - MeanExamplesLabelsets(\mathcal{D}))^2} \tag{2.40}$$

2.5.9 Number of unique labelsets

It returns the number of labelsets that are only associated with an instance of the dataset.

$$UniqueLabelsets(\mathcal{D}) = |l \subseteq DL(\mathcal{D}) : |l| = 1| \tag{2.41}$$

2.5.10 Ratio of labelsets with number of examples less than half of the attributes

This metric calculates the ratio of labelsets which number of appearances is less or equal to the half of the number of attributes [7].

$$RatioLabelsetsHalfAttributes(\mathcal{D}) = \frac{|l \subseteq \mathcal{L} : |l| \leq \frac{|\mathcal{T}|}{2}|}{DL(\mathcal{D})} \tag{2.42}$$

2.5.11 Number of unconditionally dependent label pairs by chi-square test

Coefficients Chi (χ^2) [11] and Phi (ϕ) [8] identify the unconditionally relation between pairs of labels. Chi coefficient identifies the unconditionally relation between pairs of labels using Chi-square test between each pair of labels. Thus, if Chi value for a pair is greater than 6.635, labels are considered dependent at 99% confidence [11]. Given two labels λ_i and λ_j, and a contingency table for both labels as in Table 2.1, Chi coefficient is calculated as:

$$\chi^2 = \frac{(AD - BC)^2(A + B + C + D)}{(A + B)(C + D)(A + C)(B + D)} \tag{2.43}$$

TABLE 2.1: Contingency table for two labels

	λ_j	$\neg\lambda_j$
λ_i	A	B
$\neg\lambda_i$	C	D

Also, given the previous contingency table, Phi coefficient is calculated as:

$$\phi = \frac{AD - BC}{\sqrt{(A + B)(C + D)(A + C)(B + D)}} \tag{2.44}$$

Chi and Phi coefficients have a relation $\chi^2 = n \cdot \phi^2$, being n the number of instances.

Thereby, this metric represents the number of pairs of labels unconditionally dependent that reject the null hypotesis of the Chi-square test at 99% confidence [7, 11]. It is defined as:

$$DependentLabelPairsChiSquare(\mathcal{D}) = \{Count(\lambda_i, \lambda_j) : \chi^2(\lambda_i, \lambda_j) > 6.635\} \tag{2.45}$$

2.5.12 Ratio of unconditionally dependent label pairs by chi-square test

It returns the proportion of pairs of labels dependent at 99% confidence divided by the number of existing pairs [7].

$$\begin{aligned} &RatioDependentLabelPairsChiSquare(\mathcal{D}) = \\ &= DependentLabelPairsChiSquare(\mathcal{D}) \left(\frac{|\mathcal{L}||\mathcal{L} - 1|}{2} \right)^{-1} \end{aligned} \tag{2.46}$$

2.5.13 Average of unconditionally dependent label pairs by chi-square test

This metric returns the average value of chi-square test for the pairs of labels that reject the null hypotesis [7]. Being *depPairs* the set of dependent pairs

of labels according to the chi-square test at 99% confidence, the metric is defined as:

$$AvgDependentLabelPairsChiSquare(\mathcal{D}) = \sum_{pair}^{|depPairs|} \chi^2(pair) \tag{2.47}$$

2.6 Attributes metrics

This kind of metrics characterize different aspects of the dataset related to the attributes.

2.6.1 Number of binary attributes

It calculates the number of binary attributes $|T^b|$ of the dataset [7].

2.6.2 Number of nominal attributes

It calculates the number of nominal or categorical attributes $|T^c|$ of the dataset [7].

2.6.3 Number of numeric attributes

It calculates the number of numeric attributes $|T^n|$ of the dataset [7].

2.6.4 Proportion of binary attributes

It calculates the ratio of binary attributes regarding to the total number of attributes [7].

$$Proportion_{binary}(\mathcal{D}) = \frac{|T^b|}{|\mathcal{T}|} \tag{2.48}$$

2.6.5 Proportion of nominal attributes

It calculates the ratio of nominal attributes regarding to the total number of attributes [7].

$$Proportion_{nominal}(\mathcal{D}) = \frac{|T^c|}{|\mathcal{T}|} \tag{2.49}$$

2.6.6 Proportion of numeric attributes

It calculates the ratio of numeric attributes regarding to the total number of attributes [7].

$$Proportion_{numeric}(\mathcal{D}) = \frac{|T^n|}{|\mathcal{T}|} \tag{2.50}$$

2.6.7 Proportion of numeric attributes with outliers

It represents the proportion of numeric attributes with outliers. The outliers are values that behave in a different way from the majority. This value is defined in $[0, 1)$ interval, and the higher value, the greater the number of outliers [7, 18, 21].

For an attribute t^n, a list of numeric values sorted in ascending order $t^n = t^n_1, t^n_2, \ldots, t^n_m; 1 \leq m \leq |t^n|$ is defined. From this list, a new list $r^n \subseteq t^n$ is created, removing extreme values. Being the number of values to remove in each extreme $nTrim = \frac{\alpha}{2}$ and using a value of $\alpha = 0,05$, the new list r^n is obtained as:

$$r^n = t^n_{nTrim}, t^n_{nTrim+1}, \ldots, t^n_{|T^n|-nTrim} \tag{2.51}$$

Then, the variance of both sets is calculated, and as $V(r^n) < V(t^n)$, the ratio is calculated as:

$$Ratio(t^n) = \frac{V(r^n)}{V(t^n)} \tag{2.52}$$

If the ratio is less than 0.7, t^n is considered to have outliers [7]. To calculate the proportion of attributes with outliers, first the number of numeric attributes with outliers is calculated, and then it is divided by the total number of numeric attributes.

$$Count_{outliers}(T^n) = |t^n : Ratio(t^n) < 0,7| \tag{2.53}$$

$$Ratio_{outliers}(T^n) = \frac{Count_{outliers}(T^n)}{|T^n|} \tag{2.54}$$

2.6.8 Mean of entropies of nominal attributes

It calculates the entropy of each nominal attribute [26]. For that, it is considered that an attribute t^c has the values $t^c_1, t^c_2, \ldots, t^c_m$. Thus, $H(t^c) = -\sum_{i=1}^{m} p(t^c_i) log_2 p(t^c_i)$ returns the entropy of a nominal attribute, where t^c_i is each one of the values that t^c can take [7]. So, the mean entropy is defined as:

$$MeanEntropy(T^c) = \frac{1}{|T^c|} \sum_{i=1}^{|T^c|} H(t^c_i) \tag{2.55}$$

2.6.9 Mean of mean of numeric attributes

This metric calculates the mean of means of all numeric attributes [7]. Being $Mean(t^n) = \frac{1}{|T^n|} \sum_{j=1}^{|t^n|} t^n_j$ the mean for a numeric attribute, the metric is defined as:

$$MeanOfMean_{numeric}(\mathcal{D}) = \frac{1}{|T^n|} \sum_{i=1}^{|T^n|} Mean(t^n_i) \tag{2.56}$$

2.6.10 Mean of standard deviation of numeric attributes

This metric calculates the average of standard deviations of all numeric attributes [7]. Being $\sigma(t^n)$ the standard deviation of the values of a numeric attribute, the metric is defined as:

$$MeanOfStdv_{numeric}(\mathcal{D}) = \frac{1}{|T^n|}\sum_{i=1}^{|T^n|}\sigma(t_i^n) \tag{2.57}$$

2.6.11 Average gain ratio

Average Gain Ratio obtains the average value of Gain Ratio values for all attributes and each label [1]. Being $H()$ the entropy, the Gain Ratio is defined as:

$$GainRatio(\lambda, t) = \frac{(H(\lambda) - H(\lambda|t))}{H(t)} \tag{2.58}$$

So, the average is obtained as:

$$AvgGainRatio(\mathcal{D}) = \frac{1}{|\mathcal{L}|}\frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{L}|}\sum_{j=1}^{|\mathcal{T}|}GainRatio(\lambda_i, t_j) \tag{2.59}$$

2.6.12 Average absolute correlation between numeric attributes

It calculates the average of the correlation coefficient, r, among all pairs of numeric attributes. Taking x and y as numeric value sets, the correlation coefficient r is defined as:

$$r = \sigma_{xy} \cdot (\sigma_x \sigma_y)^{-1} \tag{2.60}$$

Being σ_{xy} and σ_x the covariance and standard deviation respectively, they are defined as:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i \cdot y_i) \tag{2.61}$$

$$\sigma_x = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i)^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2} \tag{2.62}$$

Correlation coefficient r takes values between $(-1, 1)$. A value near to -1 indicates a strong inverse correlation, a value near to 1 indicates a strong direct correlation, and a value near to 0 indicates a weak correlation [13, 14].

In this case, the calculation of correlation between pairs of numeric attributes refers to $r(t_i^n, t_j^n); (t_i^n, t_j^n) \subseteq T^n, i \neq j$. With these conditions, the correlation between a pair of numeric attributes is defined as:

$$r(t_i^n, t_j^n) = \frac{\sigma(t_i^n, t_j^n)}{\sigma(t_i^n) \cdot \sigma(t_j)^n} \tag{2.63}$$

$$\sigma(t_i^n, t_j^n) = \frac{1}{|T^n|} \sum_{w=1}^{|t^n|} \left[(t_i^n)_w \cdot (t_j^n)_w \right] \tag{2.64}$$

$$\sigma(t^n) = \sqrt{\left(\frac{1}{|t^n|} \sum_{i=1}^{|t^n|} (t_i^n)^2 \right) - \left(\frac{1}{|t^n|} \sum_{i=1}^{|t^n|} t_i^n \right)^2} \tag{2.65}$$

$$MeanCorrelation_{numeric}(T^n) = \left[\frac{(|T^n| \cdot (|T^n| - 1))}{2} \right]^{-1} \left[\sum_{i=1}^{|T^n|} \sum_{j=i+1}^{|T^n|} r(t_i^n, t_j^n) \right] \tag{2.66}$$

2.6.13 Mean of kurtosis

This metric calculates the Kurtosis for each numeric attribute and then, it calculates the average [7]. A numeric attribute t^n contains $t_1^n, t_2^n, \ldots, t_m^n; 1 \leq m \leq |t^n|$ numeric values, where its mean is $Mean(t^n)$. Thus, Kurtosis is defined as:

$$Kurtosis(t^n) = \frac{(|t^n| - 1)}{(|t^n| - 2)(|t^n| - 3)} ((|t^n| + 1) g_{ku}(t^n) + 6) \tag{2.67}$$

$$g_{ku}(t^n) = \frac{m_4(t^n)}{(m_2(t^n))^2} - 3 \tag{2.68}$$

$$m_2(t^n) = \frac{1}{|t^n|} \sum_{i=1}^{|t^n|} (t_i^n - Mean(t^n))^2 \tag{2.69}$$

$$m_4(t^n) = \frac{1}{|t^n|} \sum_{i=1}^{|t^n|} (t_i^n - Mean(t^n))^4 \tag{2.70}$$

So, the mean of Kurtosis for numeric attributes T^n is defined as:

$$MeanKurtosis(T^n) = \frac{1}{|T^n|} \sum_{i=1}^{|T^n|} Kurtosis(t_i^n) \tag{2.71}$$

2.6.14 Mean of skewness of numeric attributes

It calculates the mean of skewness of numeric attributes [7]. The skewness for numeric attributes is defined as:

$$Skewness(t^n) = \frac{|t^n|}{(|t^n| - 1)(|t^n| - 2)} \sum_{i=1}^{|t^n|} \frac{(t_i^n - Mean(t^n))^3}{\sigma^3(t^n)} \tag{2.72}$$

$$\sigma^3(t^n) = \left[\frac{1}{|t^n| - 1} \sum_{i=1}^{|t^n|} (t_i^n - Mean(t^n)^3 \right]^{\frac{3}{2}} \tag{2.73}$$

Thus, the mean of skewness of numeric attributes is defined as:

$$MeanSkewness(t^n) = \frac{1}{|T^n|} \sum_{i=1}^{|T^n|} Skewness(t_i^n) \tag{2.74}$$

Chapter 3

MLDA user guide

3.1 Downloading and executing

Before downloading the tool, it is necessary to have Java version 1.8 or higher installed. In order to download and install Java, you have to access to `https://www.java.com/es/download/`, download the version which corresponds to your operative system and install it.

The tool is released via GitHub, which last version v1.2.2 is available at `https://github.com/i02momuj/MLDA/releases/tag/1.2.2`. The file *MLDA_GUI_v1_2_2.zip* stores the GUI tool. Once downloaded, the file have to be decompressed, and it contains, among others, the executable *.jar* file. To execute the tool, you have to open a command-line interface, access to the path where the *.jar* file is located and type the command `java -jar MLDA.jar`, or double click to the file. The downloaded file also includes both javadoc and pdf documentation, as well as some ML and MVML datasets.

3.2 Format of the datasets

3.2.1 Multi-label datasets

The tool is able to read two multi-label dataset formats, which are Mulan [29] and Meka [17] formats. Both of them are based on the *.arff* format from Weka [12]. Weka's *.arff* format has the following characteristics:

- The relation name goes in the first line, following the statement "*@relation*". If the relation name has spaces, it must be between quotes.
- Each attribute must be in a different line, and its syntax is: "*@attribute name type*" .The attribute name must be between quotes if it has spaces. The attribute type could be:
  - numeric: integer and real are treated as numeric.
  - <nominal-values>: between braces and separated by commas all the possible values.
  - string.
  - date[<format>].
- Instances start with the statement "*@data*", and each one will be in a different line. Each attribute value is separated by commas, and they will be in the same order they were declared. Also there is a shortly way to write the instances, where attributes with zero value are not

included. In this case, instances will be in braces, and separated by commas each pair composed by attribute and value.

- Line comments are inserted with the character "%".

Mulan and Meka use this format but they are different in how they define the labels. Mulan uses two different files: an *.arff* file and a *.xml* file. In the *.arff* file the full set of attributes and labels are exposed, without distinguishing among them. The only constraint is that labels may be binary attributes (values 0 and 1). The *.xml* file indicates which are the attributes that act like labels in the dataset. This *.xml* file is necessary because nowhere in the *.arff* file is indicated which the labels are. Figures 3.1 and 3.2 shows the *.arff* and *.xml* files respectively from a dataset in Mulan format.

```
@relation emotions_test

@attribute att1 numeric
@attribute att2 numeric
@attribute att3 numeric
   ...
@attribute att72 numeric
@attribute amazed-suprised {0,1}
@attribute happy-pleased {0,1}
@attribute relaxing-calm {0,1}
@attribute quiet-still {0,1}
@attribute sad-lonely {0,1}
@attribute angry-aggresive {0,1}

%Starting with the data
@data
0.094829,0.204498,0.082824, ..., 0.335371,1,0,0,0,1,1
0.065248,0.117975,0.08597, ..., 0.442898,0,0,0,1,0,0
0.101287,0.23254,0.078028, ..., 1.183461,1,1,0,0,0,0
   ...
0.172427,0.378696,0.081777, ..., 1.294949,1,1,1,0,0,0
```

FIGURE 3.1: Mulan *.arff* file

```
<?xml version="1.0" encoding="utf-8"?>
  <labels xmlns="http://mulan.sourceforge.net/labels">
   <label name="amazed-suprised"></label>
   <label name="happy-pleased"></label>
   <label name="relaxing-calm"></label>
   <label name="quiet-still"></label>
   <label name="sad-lonely"></label>
   <label name="angry-aggresive"></label>
</labels>
```

FIGURE 3.2: Mulan *.xml* file

On the other hand, Meka format only needs the *.arff* file. In this case, the separation between features and labels is done in the *.arff* file. Meka's *.arff* format is similar to Mulan's *.arff*, excepting the relation name line, where Meka indicates which attributes are the labels. Following the relation name, and separated by a colon, it is indicated with parameter "-C" and an integer positive number if the first q attributes are labels, or with an integer negative number if the labels are the last q attributes. Figure 3.3 shows the *.arff* file from a Meka dataset.

```
@relation "emotions: -C -6"

@attribute att1 numeric
@attribute att2 numeric
@attribute att3 numeric
   ...
@attribute att72 numeric
@attribute amazed-suprised {0,1}
@attribute happy-pleased {0,1}
@attribute relaxing-calm {0,1}
@attribute quiet-still {0,1}
@attribute sad-lonely {0,1}
@attribute angry-aggresive {0,1}

%Starting with the data
@data
0.094829,0.204498,0.082824, ..., 0.335371,1,0,0,0,1,1
0.065248,0.117975,0.08597, ..., 0.442898,0,0,0,1,0,0
0.101287,0.23254,0.078028, ..., 1.183461,1,1,0,0,0,0
   ...
0.172427,0.378696,0.081777, ..., 1.294949,1,1,1,0,0,0
```

FIGURE 3.3: Meka *.arff* file

As seen, both formats are similar, and while Meka's format is simpler, Mulan's format is more powerful because it does not need to define all the labels at the beginning or the end of the attributes set.

3.2.2 Multi-view multi-label datasets

The format to define the multiple views consists only in adding this information to the header of the *.arff* file. To define the views, inside the relation name the parameter "*-V*" is included. It is followed by a colon ":" and the set of views. Each view is defined with an interval with lower and higher attribute indices of the view, both included, and separated by a hyphen "-". Also, it is allowed to include attributes separated by commas "," if they are not consecutive. Different views are separated by an exclamation mark "*!*". Figure 3.4 shows an example of a multi-view multi-label dataset, where two views are defined, one from attribute 0 to 63 and other from attribute 64 to 71. This declaration of views is applicable to both Mulan and Meka formats.

3.3 Summary tab

Once started, the first tab is the summary, where main metrics for dataset characterization can be calculated. For loading a dataset, you have to click on the button in the upper right corner (Figure 3.5), and select an *.arff* file in Mulan or Meka formats. As it can be seen, this tab shows in its top a set of basic metrics for characterizing datasets, while in the bottom it shows a larger set of metrics, from which you can select a subset of them, and calculate with the *Calculate* button (Figure 3.6).

In order to select the metrics, there also exist four extra buttons: *All*, to select all the metrics, *None*, to unselect all of them, *Invert*, to invert the selection, and *Clear*, which unselect and clear all the calculated values. The

```
@relation "emotions -V:0-63!64-71"

@attribute att1 numeric
@attribute att2 numeric
@attribute att3 numeric
   ...
@attribute att72 numeric
@attribute amazed-suprised {0,1}
@attribute happy-pleased {0,1}
@attribute relaxing-calm {0,1}
@attribute quiet-still {0,1}
@attribute sad-lonely {0,1}
@attribute angry-aggresive {0,1}

%Starting with the data
@data
0.094829,0.204498,0.082824, ..., 0.335371,1,0,0,0,1,1
0.065248,0.117975,0.08597, ..., 0.442898,0,0,0,1,0,0
0.101287,0.23254,0.078028, ..., 1.183461,1,1,0,0,0,0
   ...
0.172427,0.378696,0.081777, ..., 1.294949,1,1,1,0,0,0
```

FIGURE 3.4: MVML dataset *.arff* file

Save button allows to save all the metric values in four formats: *.txt*, *.csv*, *.arff* y *.tex*.

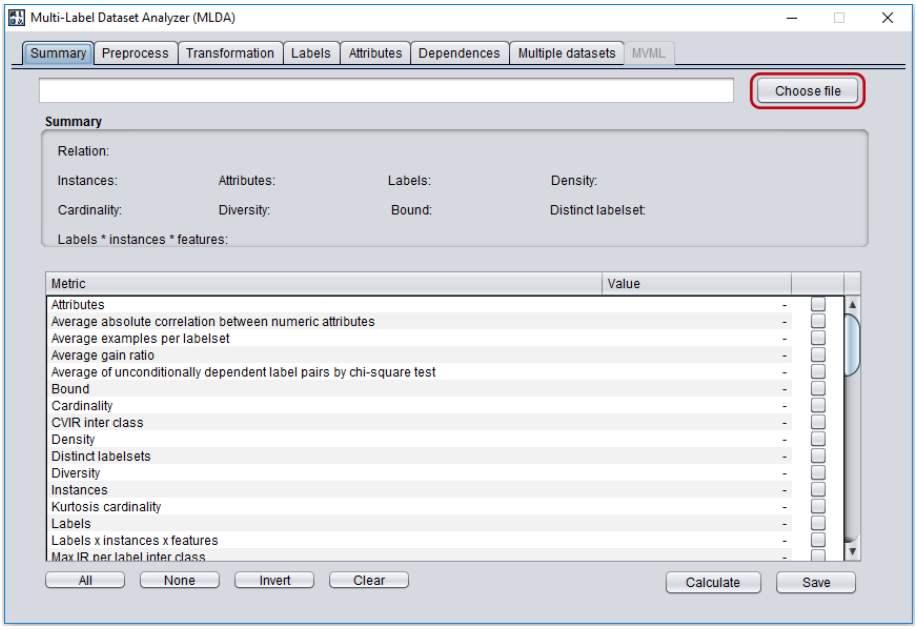


FIGURE 3.5: Selecting a dataset in summary tab

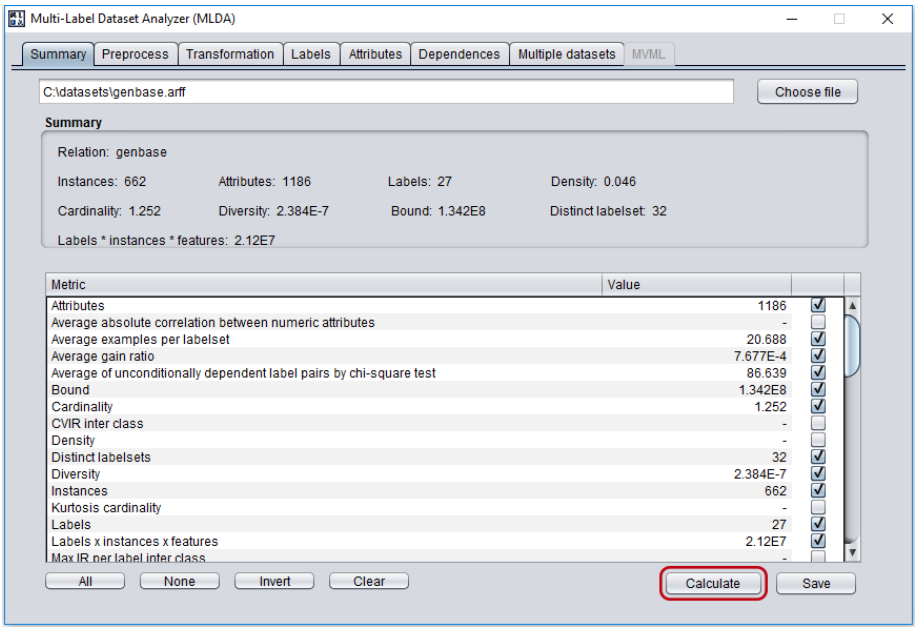


FIGURE 3.6: Calculating metrics in summary tab

3.4 Preprocess

The preprocess tab allows to perform preprocess tasks over the multi-label datasets. It includes instance selection, feature selection, data partitioning and format conversion. All the preprocess types could be done separately or together, where preprocessing is performed in the following: instance selection, feature selection, data partitioning and format conversion. That is to say, in case of selecting all the preprocessing types, first it will be made the instance selection, then, over the modified dataset, it will be made the feature selection, and last the splitting. Finally, the dataset could be saved in Mulan or Meka format.

3.4.1 Instance selection

To perform the random instance selection, only the number of instances to select has to be indicated, which must be less than the number of instances in the original dataset (Figure 3.7).

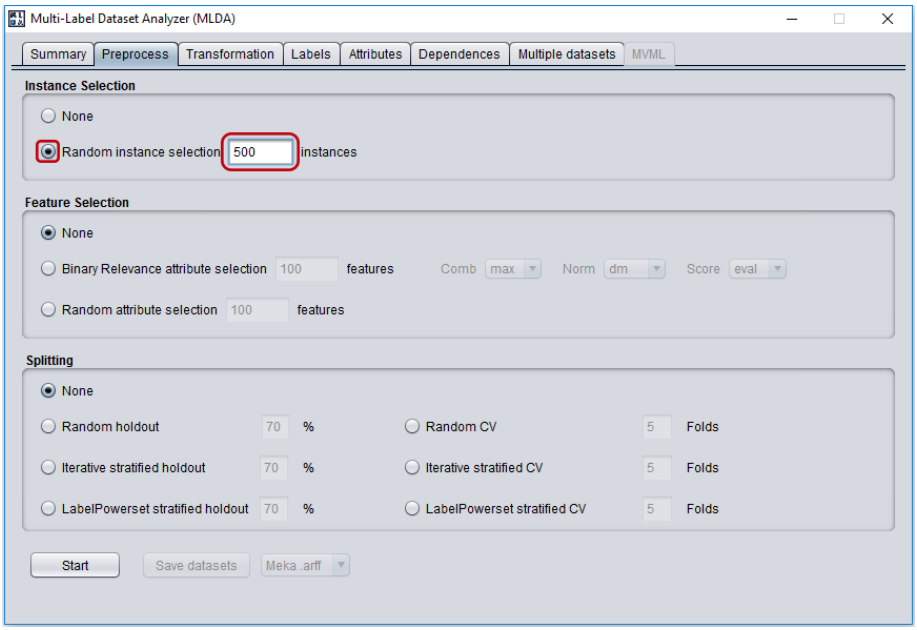


FIGURE 3.7: Random instance selection

3.4.2 Feature selection

For feature selection, the tool includes two methods: BR based [28] and random. For both types, the number of features must be indicated, always less than the number of features in the original dataset. As shown in Figure 3.8, the BR based method, which is specific for multi-label learning, needs three parameters: the combination approach method (which may be by maximum *max*, average *avg* or minimum *min*), normalization mode (which may be dividing by length *dl*, dividing by maximum *dm* or no normalization *none*), and score mode (which may be by evaluation score *eval* or by ranking score *rank*).

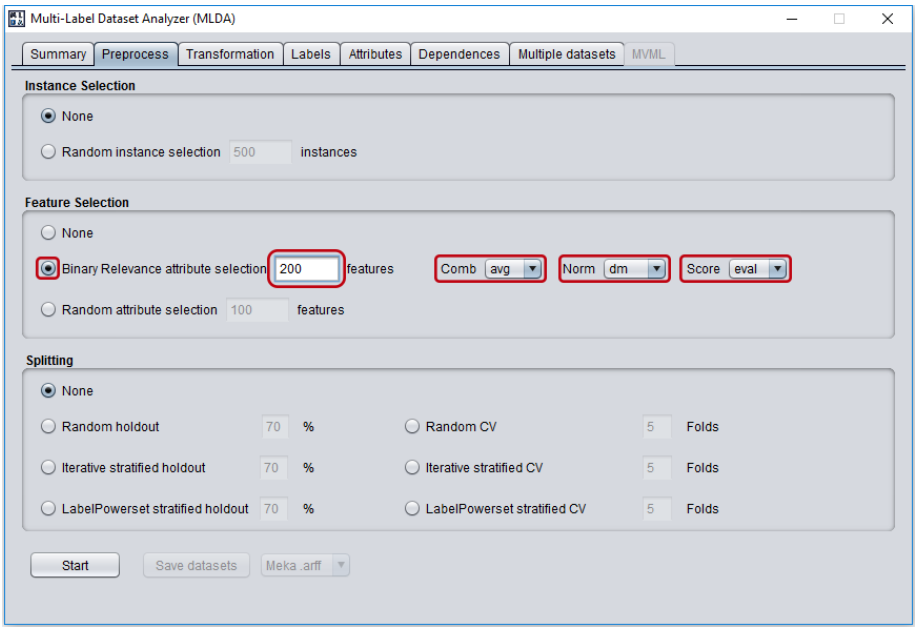


FIGURE 3.8: BR based feature selection

3.4.3 Data partitioning

Multi-label datasets allows splitting to be carried out in two different ways: holdout and k-folds cross-validation (cv) (Figure 3.9). For holdout partitioning, the percentage of instances for the train file must be specified, while for k-folds cv, the number of folds to split the original dataset has to be indicated. Holdout partitioning saves two new dataset files, a train file and a test file, while the k-folds cv saves k train files and k test files (Figure 3.10). For both types, there exists three methods for splitting: random, iterative stratified and LP stratified [22]. Iterative and LP stratification methods are specific for multi-label learning.

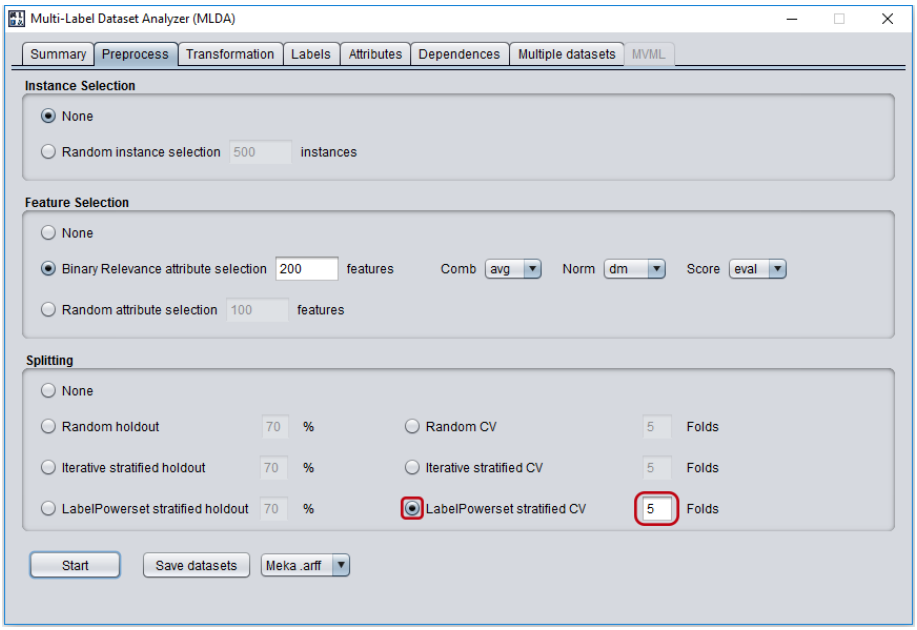


FIGURE 3.9: LP *k*-folds cv dataset splitting

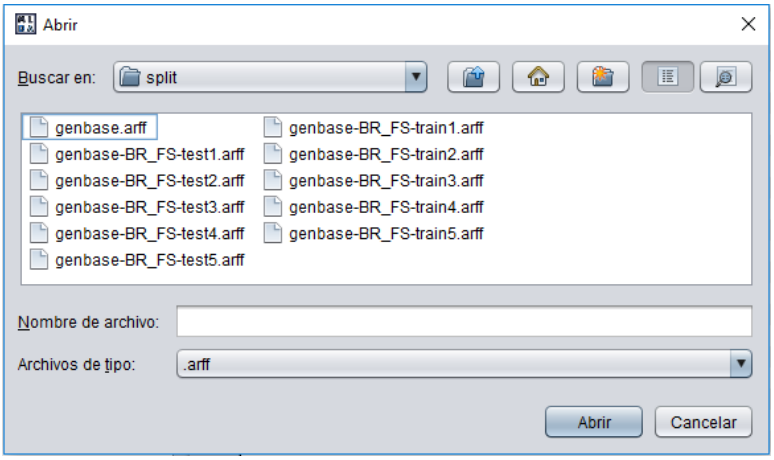


FIGURE 3.10: Result of a 5-folds cv splitting

3.4.4 Dataset format conversion

Since the tool is able to load both Mulan and Meka dataset formats and also allows to save the preprocessed datasets in both formats, if no preprocessing type is selected (Figure 3.11), the dataset format could be converted without modifying data. In order to convert the format of the dataset, *Start* button has to be clicked, and *Save datasets* to save the dataset with the selected format.

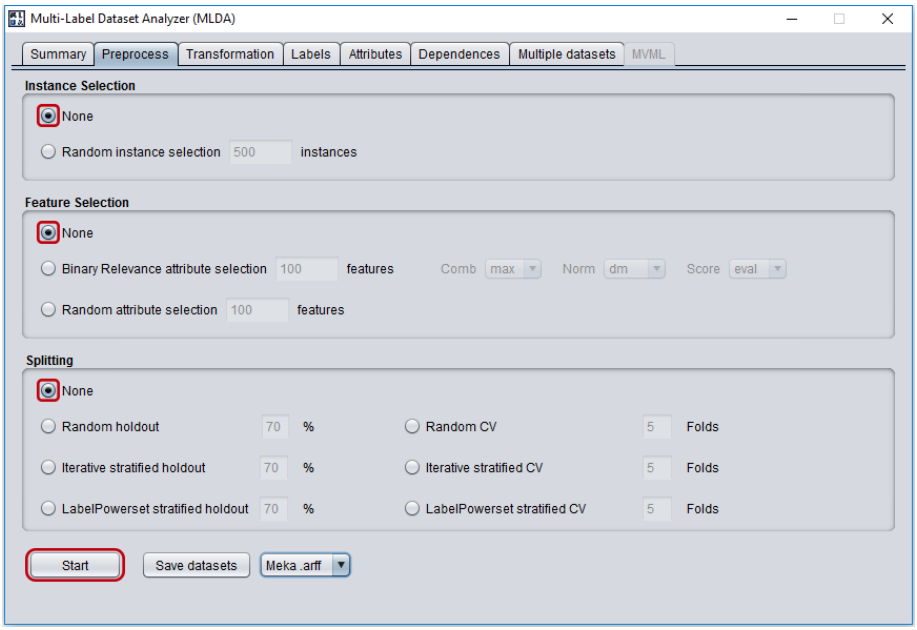


FIGURE 3.11: Dataset format conversion

3.5 Transformation

Transformation tab includes four transformation methods for multi-label datasets into single-label datasets. It includes: Binary Relevance transformation [28], Label Powerset transformation [3], include labels transformation and remove all labels transformation. Binary Relevance method generates q binary datasets, one for each existing label (Figure 3.12). Label Powerset generates one multi-class dataset, where the classes are the labelsets of the original dataset (Figure 3.13). Include labels transformation generates a binary dataset. For that, each instance is replicated q times, where each new instance is extended with the label name and the class is the label value of this instance (Figure 3.14). Finally, remove all labels transformation generates a new dataset removing all the label attributes (Figure 3.15).

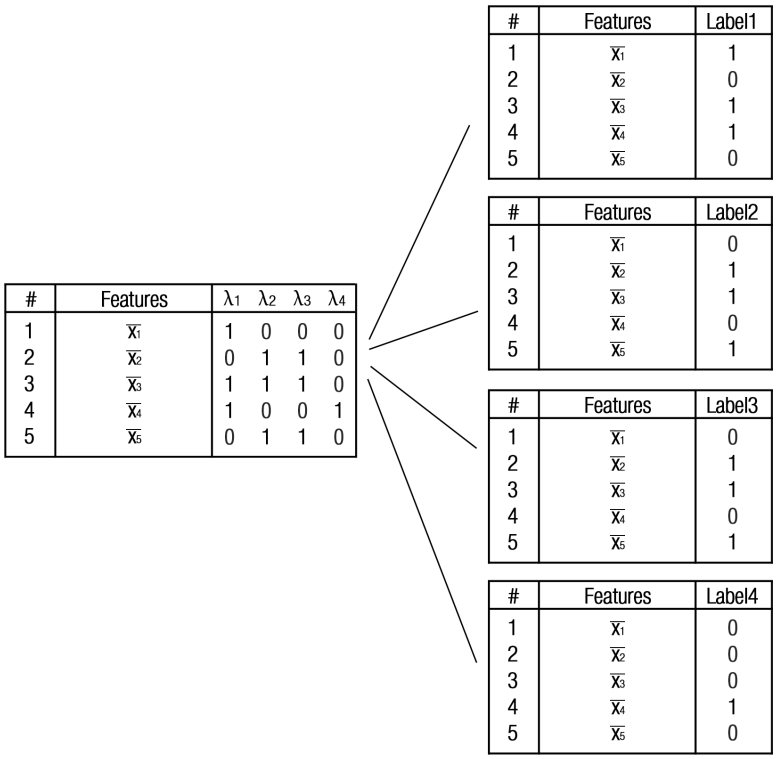


FIGURE 3.12: BR transformation

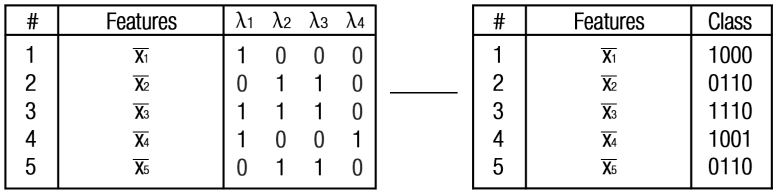


FIGURE 3.13: LP transformation

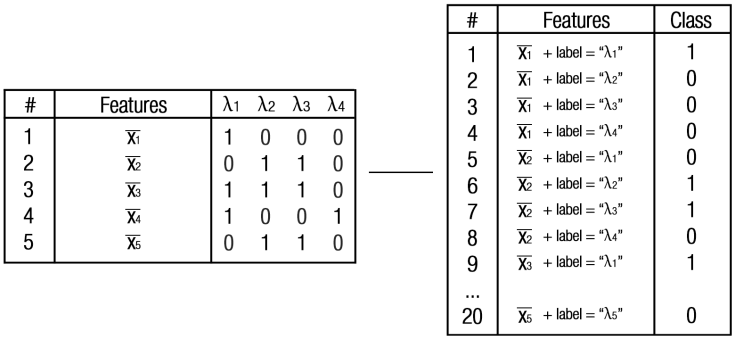


FIGURE 3.14: Include labels transformation

#	Features	λ_1	λ_2	λ_3	λ_4
1	$\overline{x_1}$	1	0	0	0
2	$\overline{x_2}$	0	1	1	0
3	$\overline{x_3}$	1	1	1	0
4	$\overline{x_4}$	1	0	0	1
5	$\overline{x_5}$	0	1	1	0

———

#	Features
1	$\overline{x_1}$
2	$\overline{x_2}$
3	$\overline{x_3}$
4	$\overline{x_4}$
5	$\overline{x_5}$

FIGURE 3.15: Remove all labels transformation

To transform the current dataset, we just have to select the desired method, click on the *Transform* button, and then save it with the *Save* button, as shown in Figure 3.16.

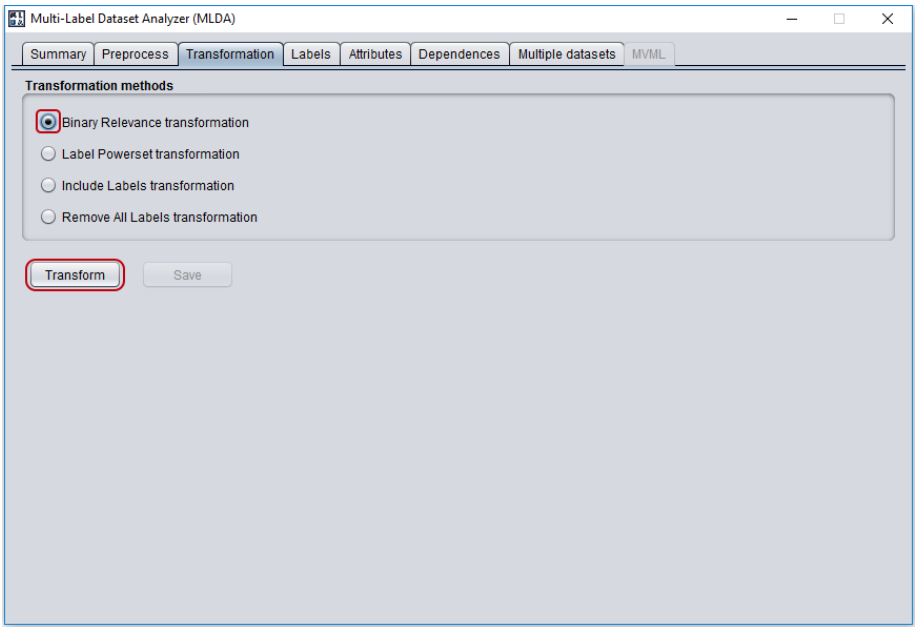


FIGURE 3.16: Transformation tab

3.6 Labels tab

Labels tab allows to obtain information and charts about the distribution and imbalance of labels. On the top appears a dropdown menu (Figure 3.17) where the graphic type can be selected. In all cases, on the left sidea table with the information appears, which also can be saved in *.csv* format by clicking on the *Save* button. The chart appears on the right side. By clicking in the chart with the right mouse button (Figure 3.18), some options are provided: zoom in, zoom out, auto range, print, properties, or saving the chart as a *.png* file in the *Save as* option (Figure 3.19) .

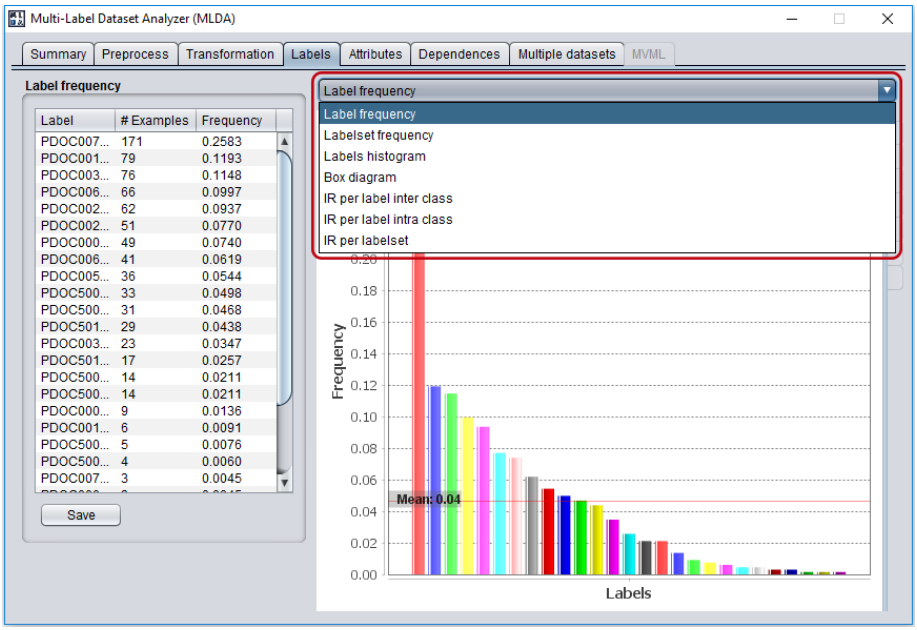


FIGURE 3.17: Dropdown menu on labels tab

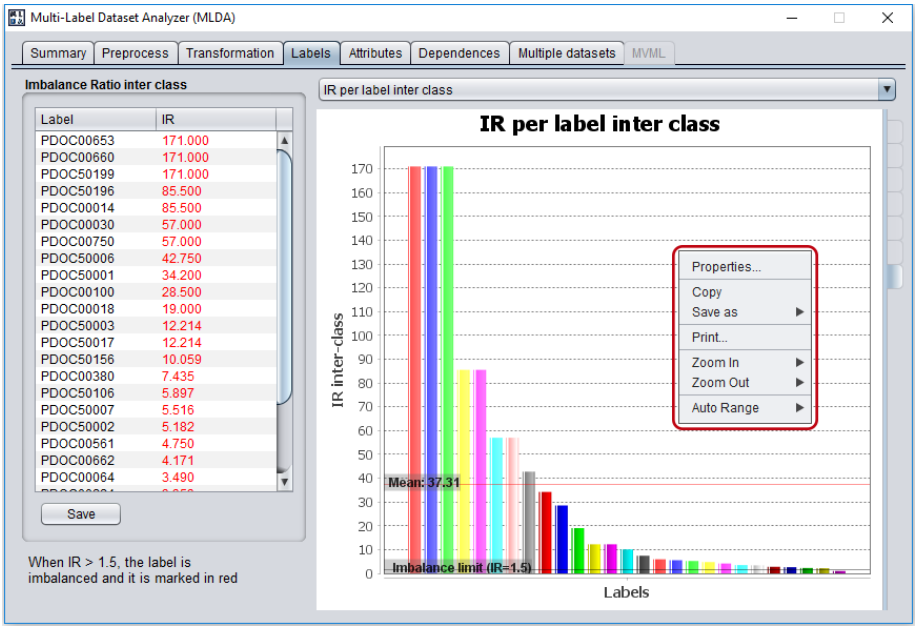


FIGURE 3.18: Chart options

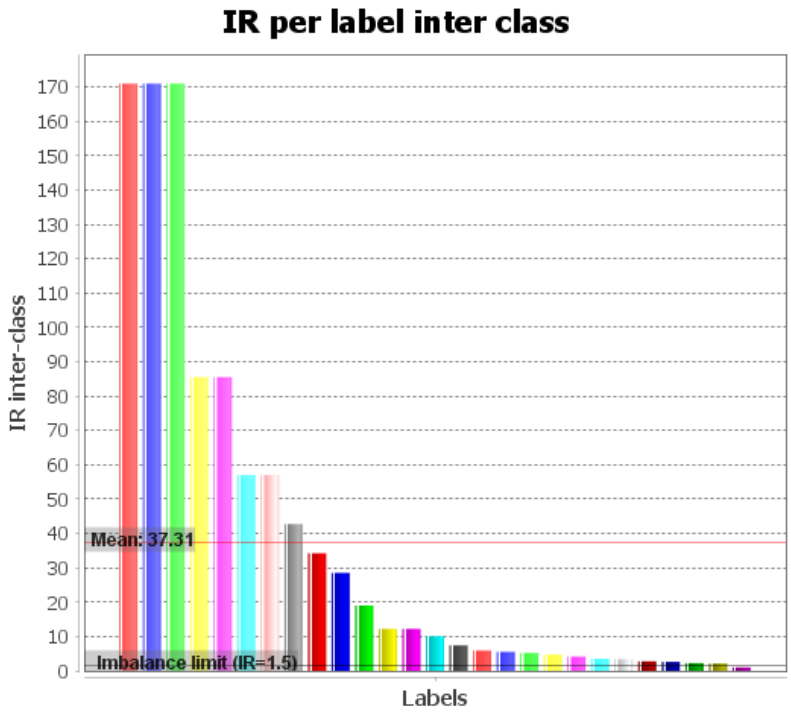


FIGURE 3.19: Example of chart saved in *.png* format

In the dropdown menu, the following options are provided:

- **Label frequency:** on the left side it shows a table with the labels, the number of examples to which it is associated and the frequency of the label. On the right side it shows a bar chart where each bar corresponds to a label and its height indicates the frequency. Besides, it shows the average frequency value as a horizontal line. If the mouse is over a bar, a tooltip message with the label name and frequency is showed (Figure 3.20).

- **Labelset frequency:** this tab shows the frequency of each labelset. So, the left table shows, ordered by frequency, the number of instances associated to each labelset and its frequency. When clicking on each labelset in the table, it appears information about the labelset, showing the number of labels that compose it, the binary coded labelset (1s and 0s shows if each label is associated or not with the labelset respectively), the labelset frequency and the name and frequency of each label included in the labelset (Figure 3.21). In the bar chart of the right side, each bar corresponds to a labelset and the height shows its frequency.

- **Labels histogram:** in this case, the table shows how many examples are associated with a certain number of labels and its frequency. That is to say, it shows how many examples have only one label associated, how many have two labels, how many have three, and so on. The bar chart indicates the frequency of the previous table. If the mouse is over a bar, a tooltip message with the number of associated labels and their frequency is showed (Figure 3.22).

- **Box diagram:** it shows two possible boxplot charts: distribution of the number of examples by label (Figure 3.23) or distribution of number of examples by labelset (Figure 3.24).

- **IR inter class:** the IR inter class shows the imbalance ratio between labels [4]. The table on the left side shows each label name and its IR inter class value, which are sorted by descending IR value. The bar chart shows the IR inter class value for each label. Also it shows a line for the mean IR inter class value and other line for the imbalance bound (Figure 3.25). This bound is fixed to 1.5, and each label whose IR is greater than this bound, is considered to be imbalanced. The IR values in the table are shown in red if they are greater than 1.5. As in previous charts, if the mouse is over a bar, it shows a tooltip message with the label name and its IR inter class value.

- **IR intra class:** IR intra class shows the imbalance ratio inside a label [24]. Left side table shows each label name and its IR intra class value. The labels in the table are sorted by descending IR value. The bar chart shows the IR intra class value for each label. Also it shows a line for the mean IR intra class value and other line for the imbalance limit (Figure 3.26). This limit is also fixed to 1.5. IR values in the table are shown in red if they are greater than 1.5. As in previous charts, if the mouse is over a bar, it shows a tooltip message with the label name and its IR intra class value.

- **IR per labelset:** IR per labelset is similar to IR inter class, but it shows the imbalance ratio between labelsets. Left side table shows each labelset and its IR value. The labelsets in the table are sorted by descending IR value. The bar chart shows the IR value for each labelset, and also it shows a line for the mean IR value (Figure 3.27). IR values in the table are shown in red if they are greater than the imbalance limit fixed to 1.5. As in previous charts, if the mouse is over a bar, it shows a tooltip message with its IR value.

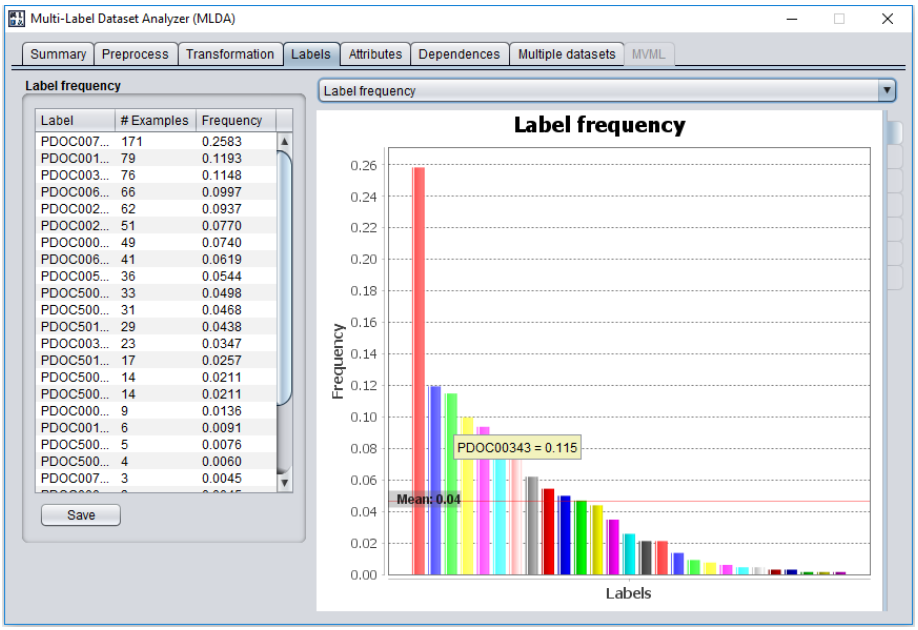


FIGURE 3.20: Example of label frequency chart

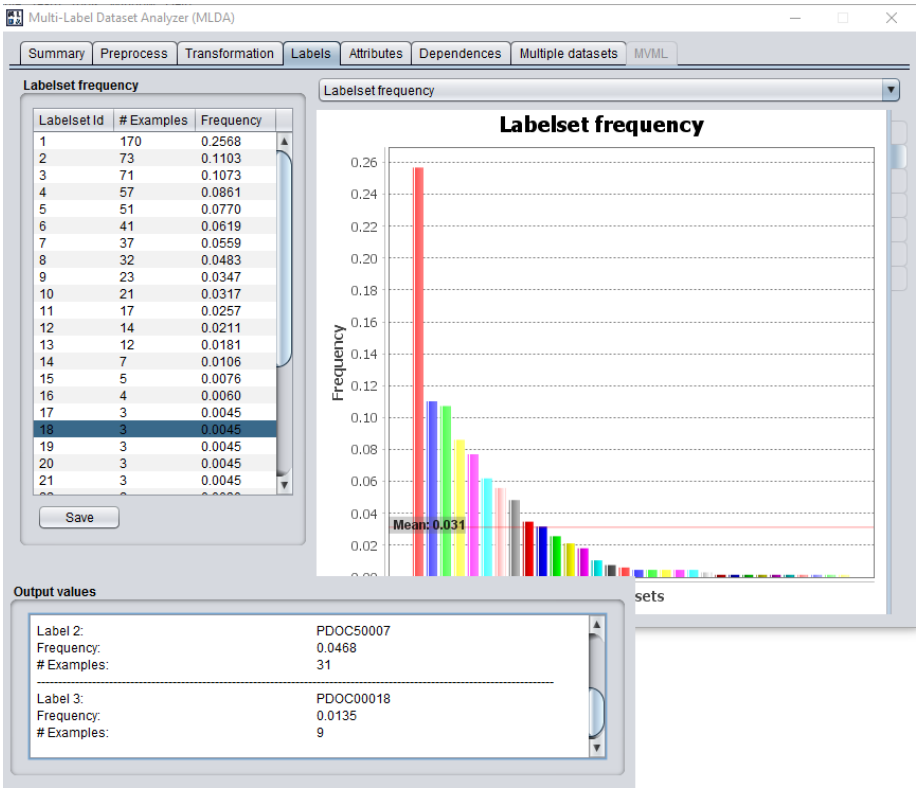


FIGURE 3.21: Example of labelset frequency chart

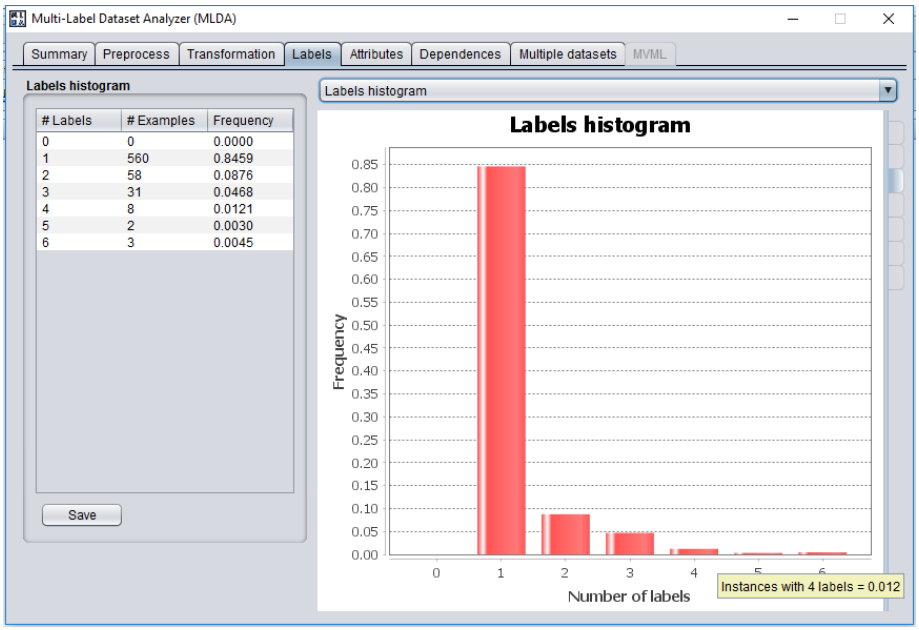


FIGURE 3.22: Example of labels histogram chart

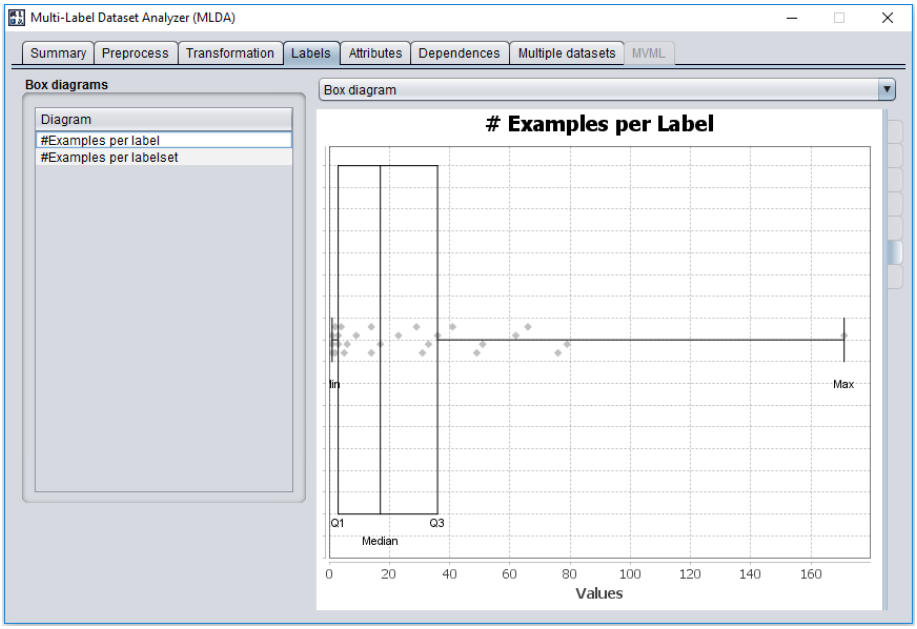


FIGURE 3.23: Example of boxplot for number of examples by label

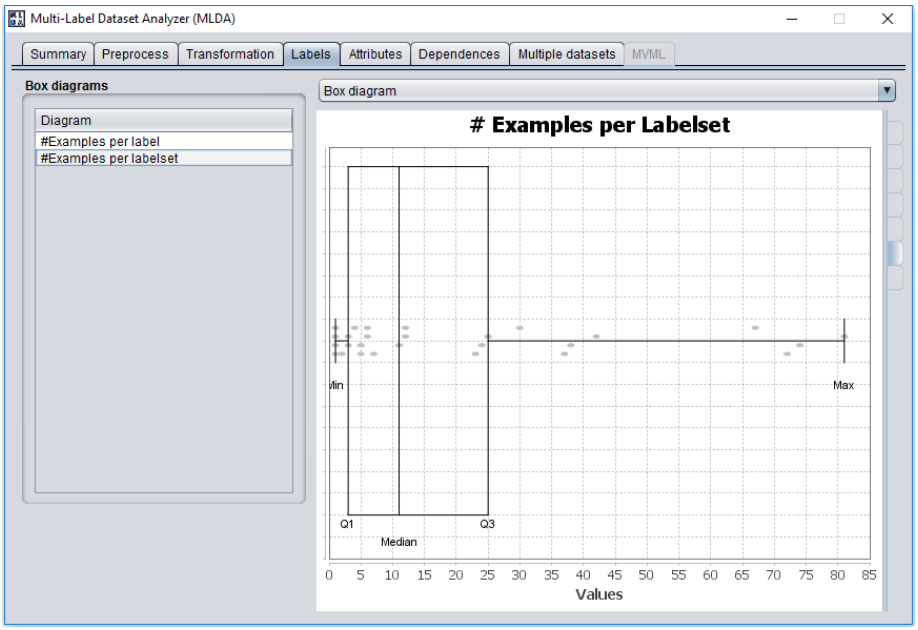


FIGURE 3.24: Example of boxplot for number of examples by labelset

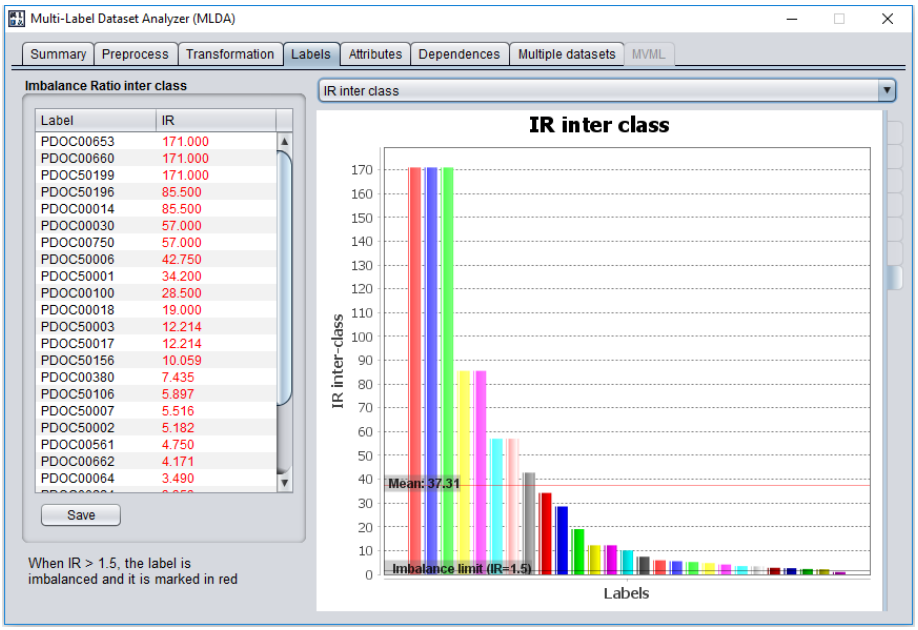


FIGURE 3.25: Example of IR inter class chart

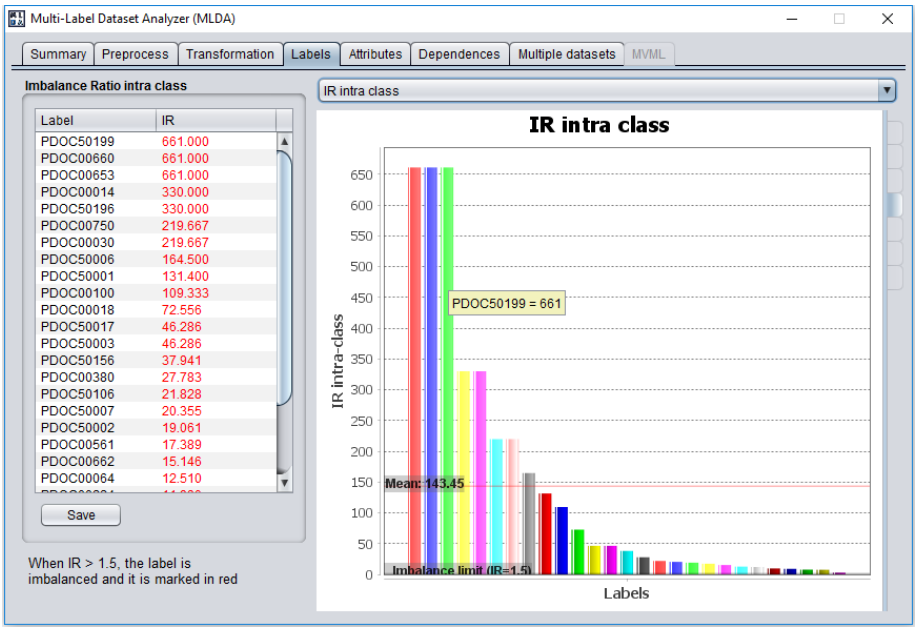


FIGURE 3.26: Example of IR intra class chart

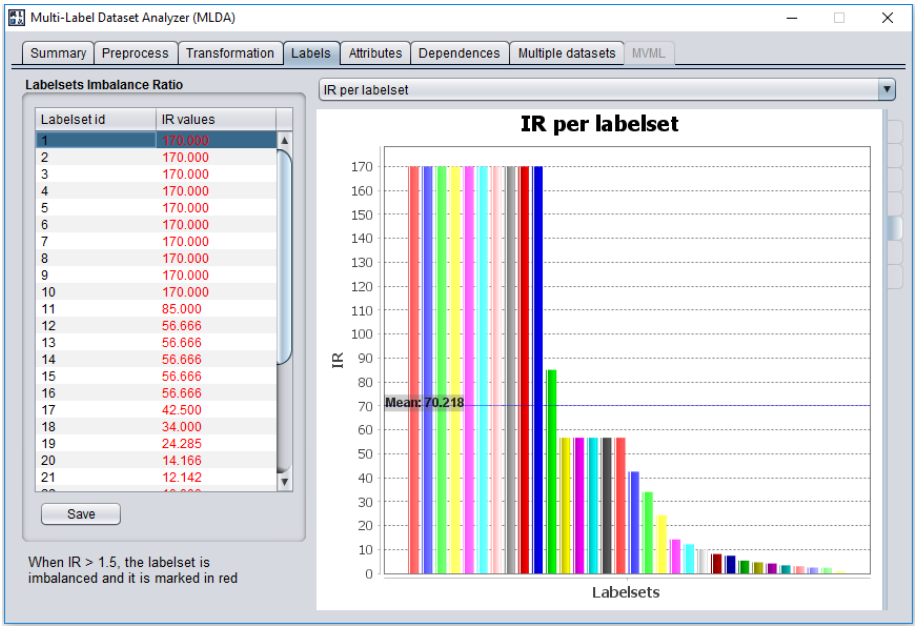


FIGURE 3.27: Example of IR per labelset chart

3.7 Attributes tab

The attributes tab shows boxblot charts for each numeric attribute. As shown in Figure 3.28, on the left side there is a table listing all the numeric attributes, and on the right, there is a boxplot chart showing the values of the selected numeric attribute. When an attribute is selected in the list, a boxplot with its values is showed.

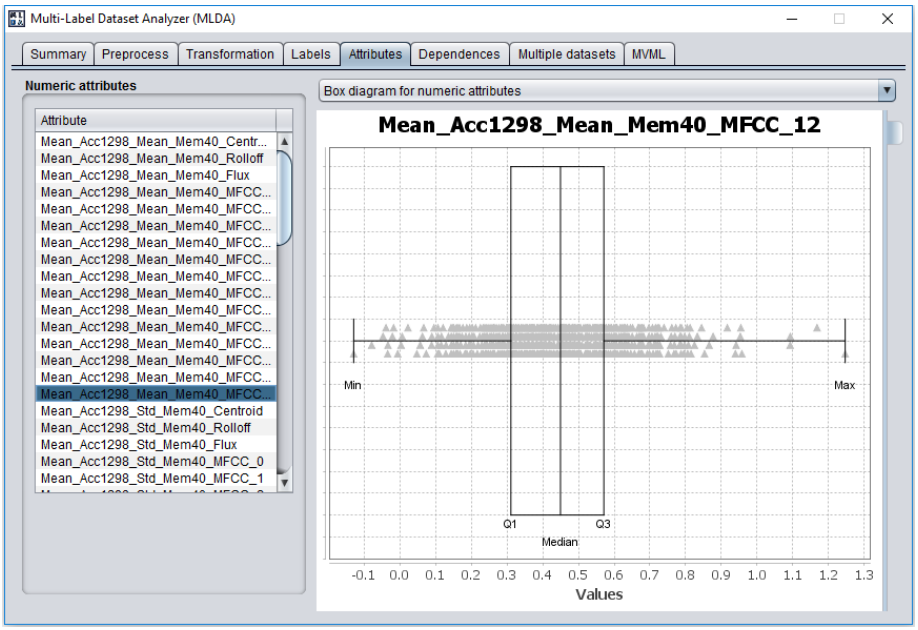


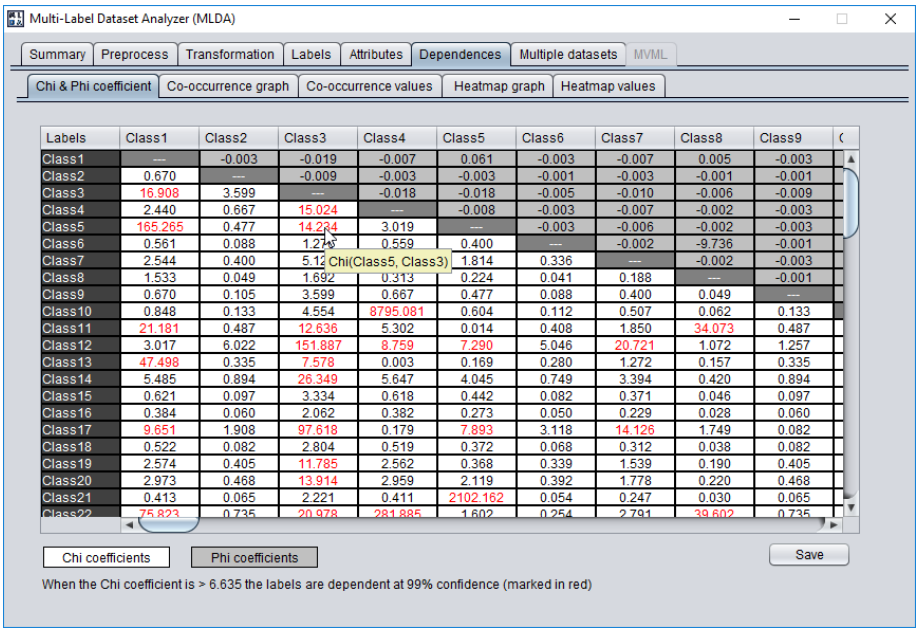
FIGURE 3.28: Example of numeric attributes boxplot

3.8 Dependences tab

The dependences tab shows information about the dependences or relationship between labels. This tab is divided in five sub-tabs, including three dependence types: chi and phi coefficients, co-occurrence and heatmap.

3.8.1 Chi and Phi coefficients

Chi coefficient (χ^2) identifies the unconditional relationship between label pairs, using Chi-square tests between each label pair. If Chi value is greater than 6.635, labels will be considered dependent at 99% confidence [11]. Phi (ϕ) coefficient [8] can be calculated from Chi as $\chi^2 = n \cdot \phi^2$, being n the number of instances. The *Chi & Phi coefficient* sub-tab shows a table with the Chi and Phi values among each pair of labels. As shown in Figure 3.29, Chi values are in white cells and Phi values in gray cells. Also, Chi values which are greater than 6.635 are marked in red, in order to show the dependence between these labels. If the mouse is over a cell, a tooltip message with the label names to which correspond the value is shown. Finally, the table can be saved in *.csv* format by clicking on the *Save* button. It will store a file with two tables, one for each coefficient.



Labels	Class1	Class2	Class3	Class4	Class5	Class6	Class7	Class8	Class9
Class1	---	-0.003	-0.019	-0.007	0.061	-0.003	-0.007	0.005	-0.003
Class2	0.670	---	-0.009	-0.003	-0.003	-0.001	-0.003	-0.001	-0.001
Class3	16.908	3.599	---	-0.018	-0.018	-0.005	-0.010	-0.006	-0.009
Class4	2.440	0.667	15.024	---	-0.008	-0.003	-0.007	-0.002	-0.003
Class5	165.265	0.477	14.284	3.019	---	-0.003	-0.006	-0.002	-0.003
Class6	0.561	0.088	1.27	0.559	0.400	---	-0.002	-9.736	-0.001
Class7	2.544	0.400	5.12	Chi(Class5, Class3)	1.814	0.336	---	-0.002	-0.003
Class8	1.533	0.049	1.692	0.313	0.224	0.041	0.188	---	-0.001
Class9	0.670	0.105	3.599	0.667	0.477	0.088	0.400	0.049	---
Class10	0.848	0.133	4.554	8795.081	0.604	0.112	0.507	0.062	0.133
Class11	21.181	0.487	12.636	5.302	0.014	0.408	1.850	34.073	0.487
Class12	3.017	6.022	151.887	8.759	7.290	5.046	20.721	1.072	1.257
Class13	47.498	0.335	7.578	0.003	0.169	0.280	1.272	0.157	0.335
Class14	5.485	0.894	26.349	5.647	4.045	0.749	3.394	0.420	0.894
Class15	0.621	0.097	3.334	0.618	0.442	0.082	0.371	0.046	0.097
Class16	0.384	0.060	2.062	0.382	0.273	0.050	0.229	0.028	0.060
Class17	9.651	1.908	97.618	0.179	7.893	3.118	14.126	1.749	0.082
Class18	0.522	0.082	2.804	0.519	0.372	0.068	0.312	0.038	0.082
Class19	2.574	0.405	11.785	2.562	0.368	0.339	1.539	0.190	0.405
Class20	2.973	0.468	13.914	2.959	2.119	0.392	1.778	0.220	0.468
Class21	0.413	0.065	2.221	0.411	2102.162	0.054	0.247	0.030	0.065
Class22	75.823	0.735	20.978	281.885	1.602	0.254	2.791	39.602	0.735

FIGURE 3.29: Chi and Phi coefficients table in Dependences tab

3.8.2 Co-occurrence values and co-ocurrence graphs

The co-occurrence measures how many times a label appears with another. The sub-tab *Co-occurrence values* show the values of co-occurrence between each pair of labels, as shown in Figure 3.30. This table is triangular, because co-occurrence between i and j is the same as between j and i. If the mouse is over a cell, a tooltip message with the labels that correspond to this value is shown. White cells with no value means no co-occurrence, the same as a 0 value. This table could be saved by the *Save* button.

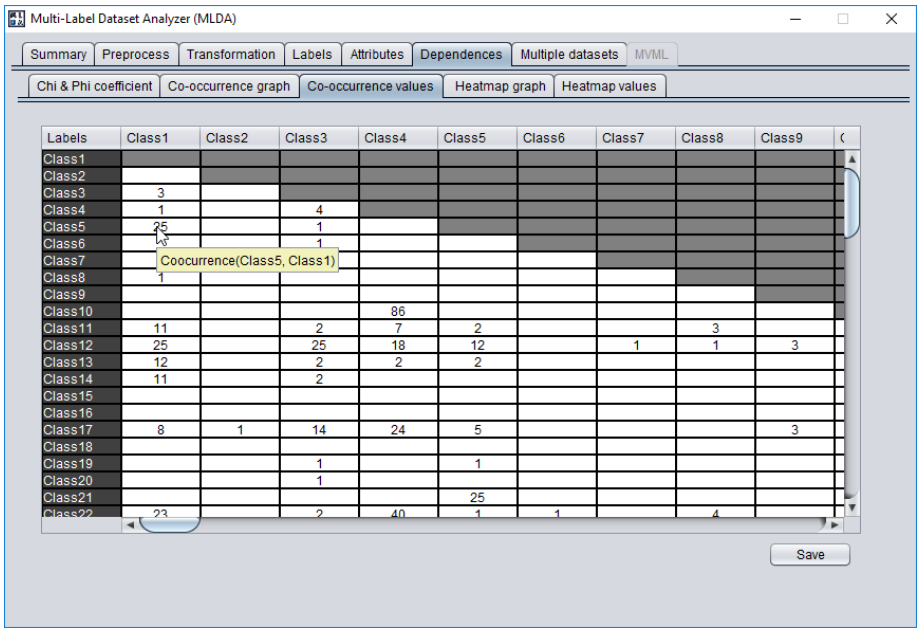


FIGURE 3.30: Co-occurrence values table in Dependences tab

The *Co-occurrence graph* sub-tab shows a graph representing the previous table. In this graph, the label thickness represents the *priori* probability of the current label $P(\lambda_i)$ and the link thickness represents the co-occurrence probability between two labels $P(\lambda_i \wedge \lambda_j)$. On the left side, it is shown a table with the label names and frequencies, which is ordered by descending frequency. Under the table, there are four different options:

- Show selected: creates a graph with the selected labels in the table.
- Show most frequent: creates a graph with the n most frequent labels. The value of n must be less than the number of labels.
- Show most related: creates a graph with the n most related labels. For that, it selects the pairs of labels with higher value of co-occurrence between them. The value of n must be less than the number of labels. This graph is shown by default with the 10 most related labels.
- Show most frequent $\cup$ most related: creates a graph with the union of the n most frequent labels and the n most related labels. The number of labels in the graph may vary between n and $2n$, because some labels could be in both most frequent and most related sets. The value of n must be less than the number of labels.

Figure 3.31 shows an example of co-occurrence graph, selecting the 10 most related labels. As it can be seen, after clicking on the *Show most related* button, the table automatically selects the rows of the labels in the graph. The graph may be modified, moving the nodes or resizing them. Then, the *Save* button allows us to save the graph in *.png* format.

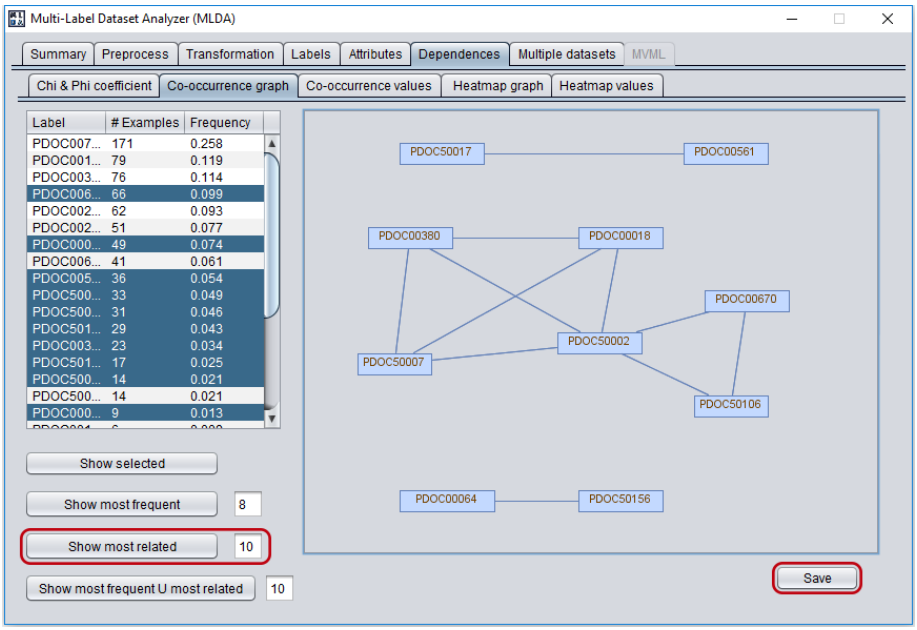


FIGURE 3.31: Co-occurrence graph for 10 most related labels in Dependences tab

3.8.3 Conditional probabilities and heatmap graphs

Heatmap table in *Heatmap values* sub-tab also represents the relationship among labels. Each cell in the table or matrix $\mathbf{M}$ represents the conditional probability, being $\mathbf{M}_{ij} = P(\lambda_i|\lambda_j) = P(\lambda_i \wedge \lambda_j)/P(\lambda_j)$, where i is a row and j a column. The diagonal (gray background) represents the *a priori* probabilities $\mathbf{M}_{jj} = P(\lambda_j)$. White cells with no value represents a probability of 0.0. And as in previous cases, if the mouse is over a cell, it shows a tooltip message with the labels that correspond to this value. This table could be saved by the *Save* button (Figure 3.32).

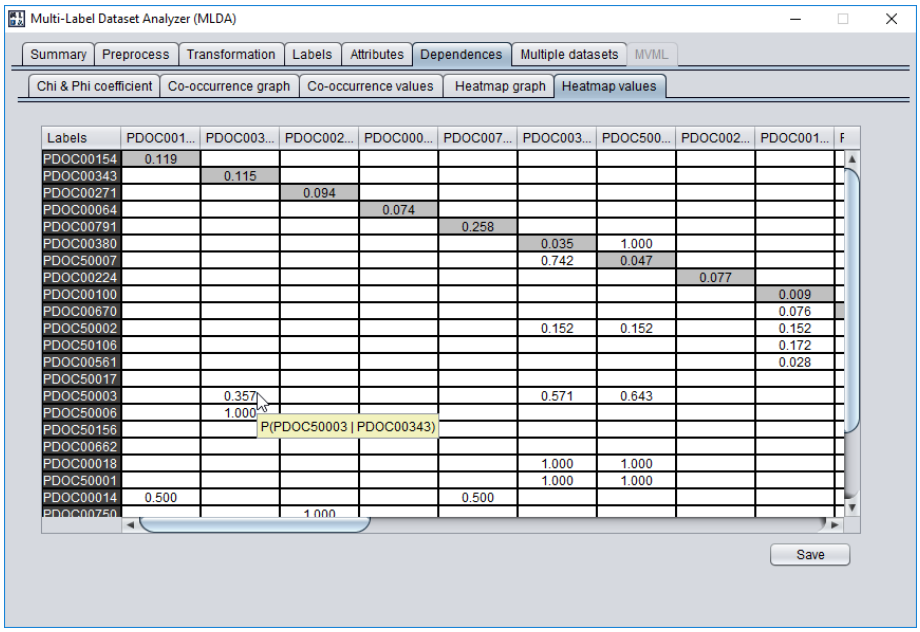


FIGURE 3.32: Heatmap values table in Dependences tab

The *Heatmap graph* sub-tab shows a graph representing the previous heatmap matrix. In this graph, probabilities are represented in grayscale, where black is a conditional probability of 0.0 and white is a conditional probability of 1.0, being the gray colors values between these. The diagonal running from the bottom left to the upper right corner represents the *prior* probabilities. On the left side, it is shown a table with the label names and frequencies, which is ordered by descending frequency. Under the table, there are four different options:

- Show selected: creates a graph with the selected labels in the table.
- Show most frequent: creates a graph with the n most frequent labels. The value of n must be less than the number of labels.
- Show most related: creates a graph with the n most related labels. For that, it selects the pairs of labels with higher value of conditional probability. The value of n must be less than the number of labels.
- Show most frequent $\cup$ most related: creates a graph with the union of the n most frequent labels and the n most related labels. The number of labels in the graph may vary between n and $2n$, because some labels could be in both most frequent and most related sets. The value of n must be less than the number of labels.

Figure 3.33 shows an example of heatmap. It has been created with all the labels (default), in order to prove that this type of graph is better when the number of labels is high. Finally, the *Save* button allows to save the graph in *.png* format.

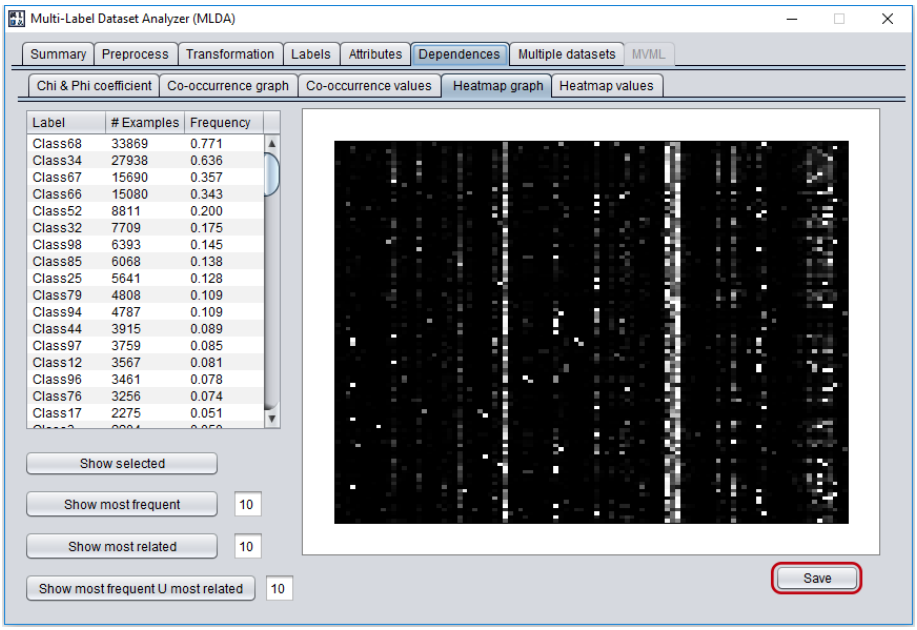


FIGURE 3.33: Heatmap graph in Dependences tab

3.9 Multiple datasets

The multiple datasets tab allows to load several datasets simultaneously, in order to calculate a set of metrics for all of them. As shown in Figure 3.34, on the left side there is an empty list and two buttons: *Add* and *Remove*. To add datasets to the list, we have to click on the *Add* button and select one or more datasets (Figure 3.35). We can add as many datasets as we want. Also, with the *Remove* button, the current selected dataset could be removed from the list.

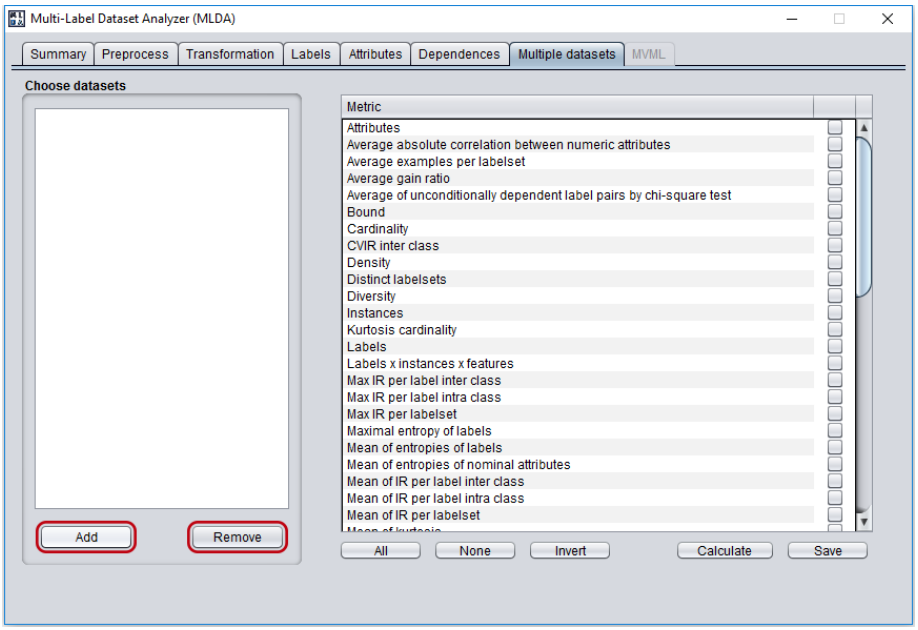


FIGURE 3.34: Add/remove datasets in multiple datasets tab

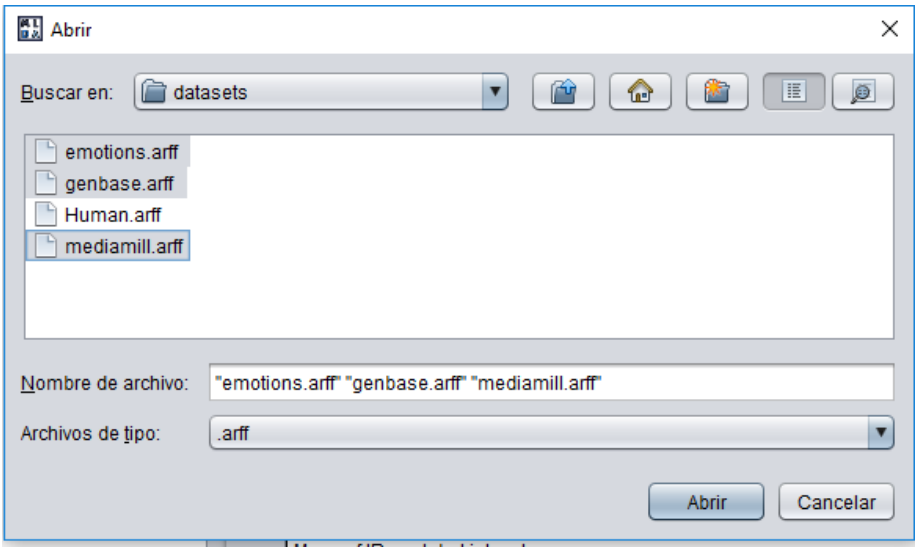


FIGURE 3.35: Adding datasets in multiple datasets tab

Once the datasets have been added, on the right side there is a table with metrics as in the summary tab. A set of metrics can be selected and calculated clicking on the *Calculate* button. Then, the metric values can be saved with the *Save* button in *.csv*, *.arff*, *.txt* and *.tex* formats (Figure 3.36).

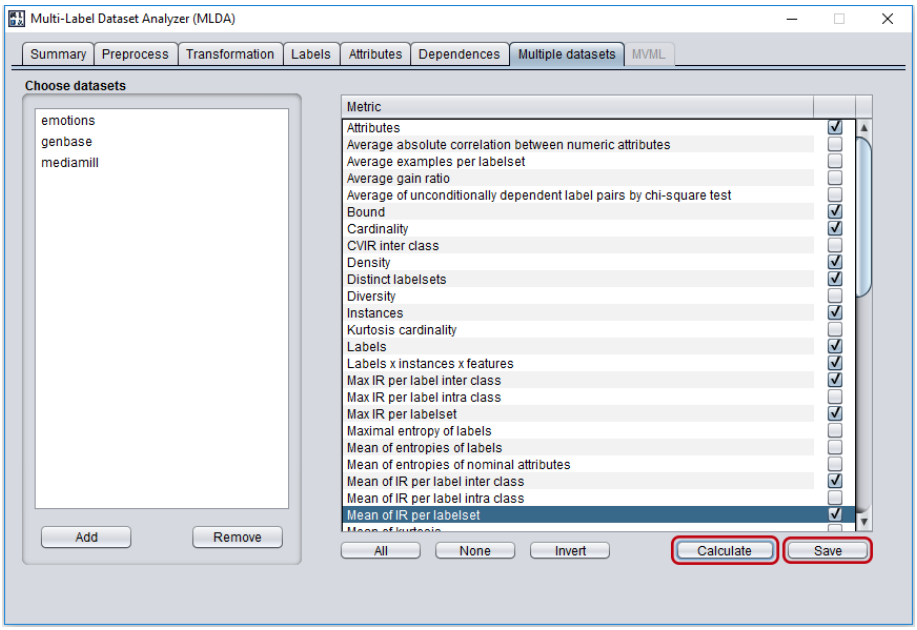


FIGURE 3.36: Calculating metrics for multiple datasets

3.10 MVML

The MVML tab gives an extra utility to datasets which are multi-view multi-label (MVML). Multi-view datasets are those in which the input space is partitioned in some views. This tab is disabled if the dataset is not MVML.

The difference in format between a multi-label dataset and a MVML dataset is in its header. The *@relation* clause in the *.arff* file includes a `-V` parameter indicating the views. For example, the header of a MVML *.arff* file could be: `@relation 'relationName -V:0-31,71!32-63!64-70'`. The format for MVML dataset was described in Section 3.2.2.

As shown in Figure 3.37, this tab at the top shows some basic metrics for multi-view datasets, for instance the number of views and the minimum, maximum and average number of attributes per view. On the bottom, the tool shows a table with the different views, giving for each one the number of attributes and some metrics like LxIxF, ratio of number of instances to the number of attributes or average gain ratio. By selecting views on the table, it can be seen a list with the attribute names of the selected views. This tab allows to save both the table with the metrics for each view and the list of attributes for each selected view by clicking on the *Save table* button. Finally, the tool also allows to save a new multi-view multi-label dataset only with the selected views in the table. For that, we only have to select the format to save (Mulan or Meka) and click on *Save views* (Figure 3.37).

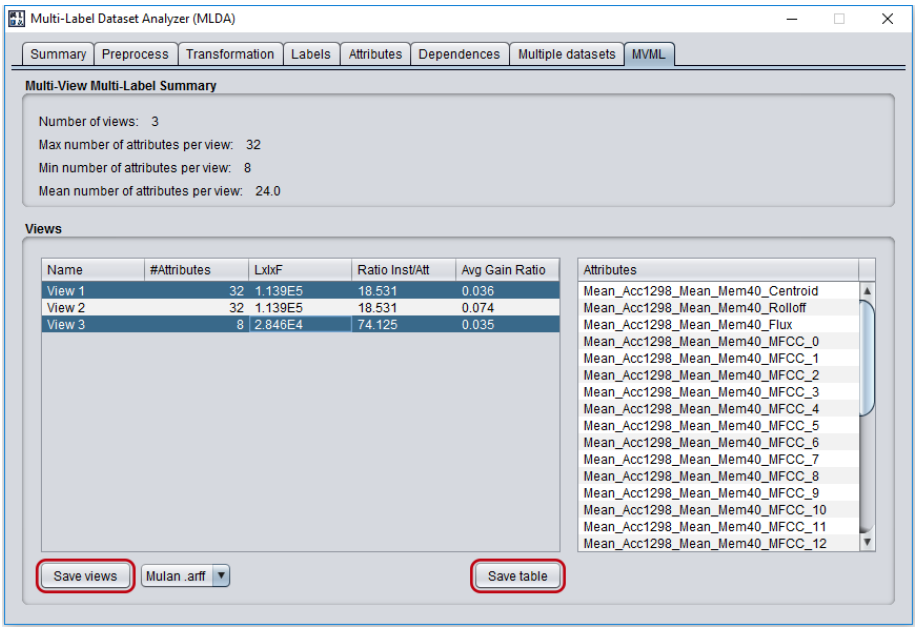


FIGURE 3.37: Saving views and table in MVML tab

Chapter 4

API

4.1 Metrics

As indicated previously, the set of metrics included in the API has been selected starting with the taxonomy proposed in [6], being extended with the metrics of both Mulan and Meka, and also the proposed metrics in [7]. Thus, the taxonomy is composed of 5 groups of metrics: dimensionality, labels distribution, labels imbalance, relationship among labels and attributes. The taxonomy could be seen in Figure 2.1.

4.2 Structure

The API have been implemented in Java, built over Mulan and Weka libraries. It is divided into the following packages:

- **base:** it includes *MLDataMetric* and *MLDataCharacterization*, the main classes for dataset characterization. The former is the base implementation for any implemented metric, including the *calculate()* method, which calculates the metric value. The latter is useful to calculate a set of metrics instead only one.
- **attributes:** it includes the implementations of attribute metrics.
- **dimensionality:** it includes the implementations of dimensionality metrics.
- **imbalance:** it includes the implementations of imbalance metrics.
- **labelsDistribution:** it includes the implementations of labels distribution metrics.
- **labelsRelation:** it includes the implementations of labels relationship metrics.
- **metricsName:** it includes the names of all implemented metrics.
- **util:** it includes some necessary methods for metrics calculation.

Figure 4.1 shows the API class diagram. The classes that implements the metrics inherits from *MLDataMetric* class, from *base* package. For simplicity, all the metrics are not included in the diagram, but they are included the necessary classes to understand the structure. The *labelsRelation* package has a generic class *LabelsetsUpToNExamples*, from which inherits 4 predefined classes for $n = 2$, 5, 10 and 50. On the other hand, *imbalance* package

includes the *ImbalanceDataMetric* class, inheriting from *MLDataMetric* and from which inherits all the imbalance metrics. *ImbalanceDataMetric* includes a new field of type *ImbalancedFeature* in order to store the characteristics of imbalanced features. Table 4.1 shows the correspondence between each metric and the class that implements it.

4.3 Usage

The API can be downloaded from `https://github.com/i02momuj/MLDA/releases/tag/1.2.2`. Once downloaded, it has to be included in the Java project. This API has two main goals: calculate one characterization metric for a multi-label dataset or calculate a set of metrics for the same dataset.

To calculate one metric, an object of the desired metric have to be created. After creating the metric object, the *calculate()* method have to be called with the multi-label dataset as parameter. The metric value can be obtained in two ways: getting the returned value of the *calculate()* method, or accessing to the metric value with the *getValue()* method. Figure 4.2 shows an example for calculating one metric.

```
1 //Creating the object corresponding to the metric
2 Density density = new Density();
3
4 //Calculating metric value
5 double value = density.calculate(mlData);
6
7 //Other way to get the metric value
8   //After calling calculate() method
9 double value2 = density.getValue();
```

FIGURE 4.2: Calculating one metric

On the other hand, to calculate a set of metrics instead of only one, the API includes the *MLDataCharacterization* class. To create a *MLDataCharacterization* object it just need the multi-label dataset as parameter. Then, the metrics are added with *addMetric()* or *addMetrics()* methods, passing as parameter a metric or a list of metrics respectively. Once calculated with *calculateMetrics()* method, the *getMetric()* method returns a metric of the list identified by its name. The *getAvailableMetrics()* method returns a set with the names of all available metrics. Figure 4.3 shows an example of how to calculate some metrics for a dataset.

TABLE 4.1: Correspondence between metrics and the class that implements it

Metric	Class
Dimensionality	
Attributes	`Attributes`
Labels	`Labels`
Instances	`Instances`
Labelsets	`DistinctLabelsets`
LxIxF	`LxIxF`
Ratio of number of instances to the number of attributes	`RatioInstancesToAttributes`
Labels distribution	
Cardinality	`Cardinality`
Density	`Density`
Maximal entropy of labels	`MaxEntropy`
Mean of entropies of labels	`MeanEntropy`
Minimal entropy of labels	`MinEntropy`
Standard deviaton of label cardinality	`StdvCardinality`
Relationship among labels	
Average examples per labelset	`AvgExamplesPerLabelset`
Average of unconditionally dependent label pairs by chi-square test	`AvgUnconditionalDependentLabelPairsByChiSquare`
Bound	`Bound`
Diversity	`Diversity`
Number of labelsets up to 2 examples	`LabelsetsUpTo2Examples`
Number of labelsets up to 5 examples	`LabelsetsUpTo5Examples`
Number of labelsets up to 10 examples	`LabelsetsUpTo10Examples`
Number of labelsets up to 50 examples	`LabelsetsUpTo50Examples`
Number of labelsets up to N examples	`LabelsetsUpToNExamples`
Number of unconditionally dependent label pairs by chi-square test	`NumUnconditionalDependentLabelPairsByChiSquare`
Proportion of distinct labelsets	`ProportionDistinctLabelsets`
Ratio of labelsets up to 2 examples	`RatioLabelsetsUpTo2Examples`
Ratio of labelsets up to 5 examples	`RatioLabelsetsUpTo5Examples`
Ratio of labelsets up to 10 examples	`RatioLabelsetsUpTo10Examples`
Ratio of labelsets up to 50 examples	`RatioLabelsetsUpTo50Examples`
Ratio of labelsets up to N examples	`RatioLabelsetsUpToNExamples`
Ratio of labelsets with number of examples less than half of the attributes	`RatioLabelsetsWithExamplesLessThanHalfAttributes`
Ratio of unconditionally dependent label pairs by chi-square test	`RatioUnconditionalDependentLabelPairsByChiSquare`
SCUMBLE	`SCUMBLE`
Standard deviation of examples per labelset	`StdvExamplesPerLabelset`
Number of unique labelsets	`UniqueLabelsets`
Labels imbalance	
CVIR inter class	`CVIRInterClass`
Kurtosis Cardinality	`KurtosisCardinality`
Max IR inter class	`MaxIRInterClass`
Max IR intra class	`MaxIRIntraClass`
Max IR per labelset	`MaxIRLabelset`
Mean of IR inter class	`MeanIRInterClass`
Mean of IR intra class	`MeanIRIntraClass`
Mean of IR per labelset	`MeanIRLabelset`
Mean of standard deviation of IR intra class	`MeanStdvIRIntraClass`
Proportion of maxim label combination (PMax)	`PMax`
Proportion of unique label combination (Puniq)	`Puniq`
Skewness cardinality	`SkewnessCardinality`
Attributes	
Average absolute correlation between numeric attributes	`AvgAbsoluteCorrelationBetweenNumericAttributes`
Average gain ratio	`AvgGainRatio`
Number of binary attributes	`BinaryAttributes`
Number of nominal attributes	`NominalAttributes`
Number of numeric attributes	`NumericAttributes`
Mean of entropies of nominal attributes	`MeanEntropiesNominalAttributes`
Mean of mean of numeric attributes	`MeanOfMeanOfNumericAttributes`
Mean of standard deviation of numeric attributes	`MeanStdvNumericAttributes`
Proportion of binary attributes	`ProportionBinaryAttributes`
Proportion of nominal attributes	`ProportionNominalAttributes`
Proportion of numeric attributes	`ProportionNumericAttributes`
Proportion of numeric attributes with outliers	`ProportionNumericAttributesWithOutliers`
Mean of kurtosis	`MeanKurtosis`
Mean of skewness of numeric attributes	`MeanSkewnessNumericAttributes`

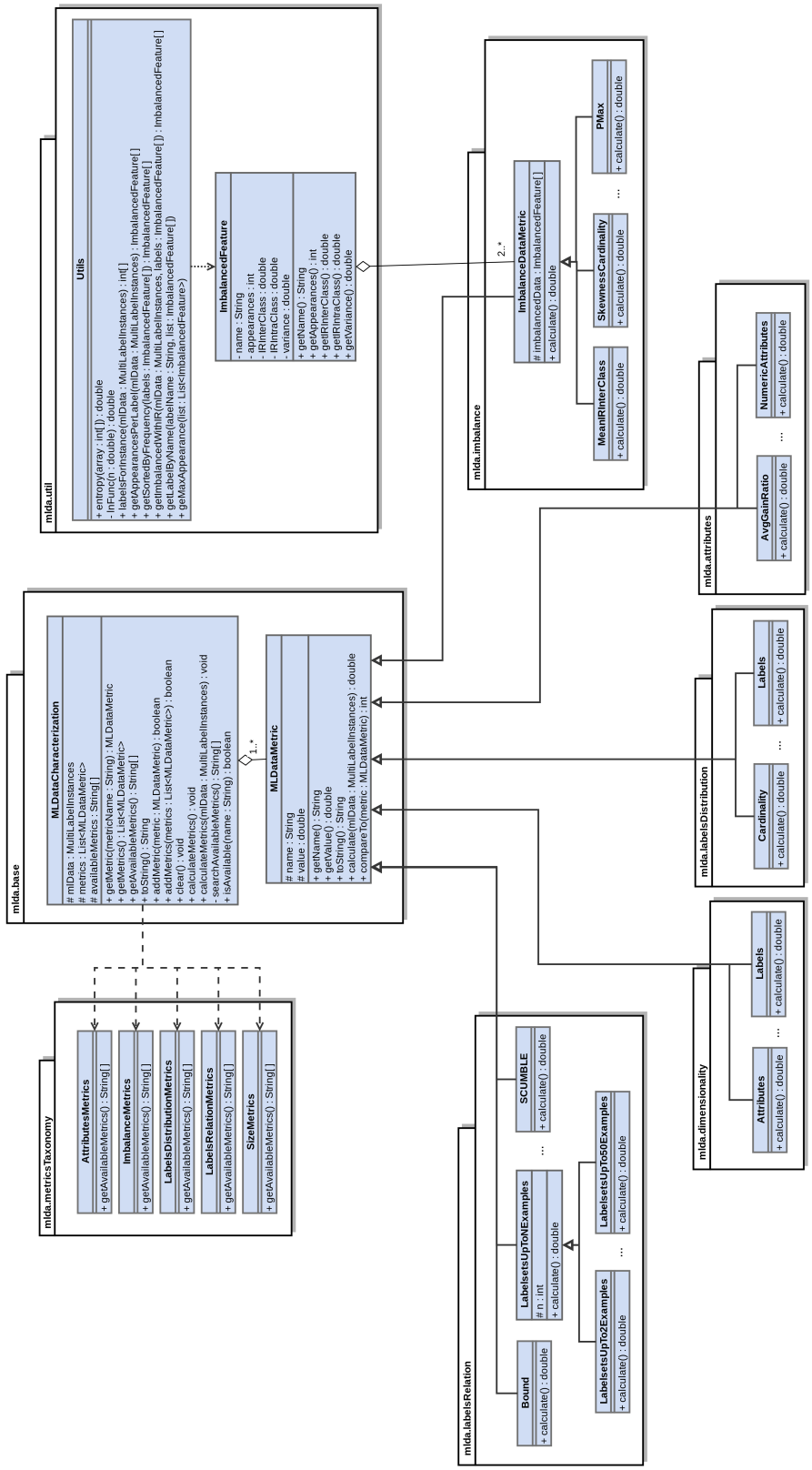


FIGURE 4.1: API class diagram

```
1  //Creating object MLDataCharacterization
2  MLDataCharacterization mldc = new MLDataCharacterization(mlData);
3
4  //Including metrics with addMetrics method
5  ArrayList<MLDataMetric> m = new ArrayList<>();
6  m.add(new Attributes());
7  m.add(new Labels());
8  m.add(new Instances());
9  mldc.addMetrics(m);
10
11 //Including metrics with addMetric method
12 mldc.addMetric(new Cardinality());
13 mldc.addMetric(new Density());
14
15 //Calculating
16 mldc.calculateMetrics();
17
18 //Getting values
19 double attributes = mldc.getMetrc("Attributes").getValue();
20 double labels = mldc.getMetric("Labels").getValue();
21 double instances = mldc.getMetric("Instances").getValue();
22 double cardinality = mldc.getMetric("Cardinality").getValue();
23 double density = mldc.getMetric("Density").getValue();
```

FIGURE 4.3: Calculating some metrics

4.4 Extending the API with new metrics

As mentioned in Section 4.2, each metric inherits from *MLDataMetric* class. Thus, to expand the API with new metrics, a new class extending *MLDataMetric* class is required (Figure 4.4). This new class must implement two methods: a constructor and a *calculate()* method. The constructor has not parameters and must call the super constructor with the name of the metric as parameter (line 4). Then, in order to include the metric in the list of available metrics of a specific type, the *addMetric* method of the corresponding class must be called (line 5). On the other hand, the *calculate* method receives as parameter the multi-label dataset and must return the value of the metric (lines 8-14).

```
1  public class Metric extends MLDataMetric{
2
3    public Metric() {
4      super("Name of the metric");
5      AttributesMetrics.addMetric("Name of the metric");
6    }
7
8    public double calculate(MultiLabelInstances mlData){
9      double value = 0;
10
11     /* Calculate the value of the metric */
12
13     return value;
14   }
15 }
```

FIGURE 4.4: Extending the API with a new metric

References

[1] Devi C. Akalya, B. Surendiran, and K. E. Kannammal. *Software Fault Prediction: A Software Fault Prediction Model by Hybrid Feature Selection and Hybrid Classifier Approach*. LAP Lambert Academic Publishing, 2012. ISBN: 3659144819, 9783659144813.

[2] A.B. Atkinson. "On the measurement of inequality". In: *Journal of Economic Theory* 2.3 (1970), pp. 244–263. DOI: 10.1016/0022-0531(70)90039-6.

[3] M Boutell et al. "Learning multi-label scene classification". In: *Pattern recognition* 37 (2004), pp. 1757–1771.

[4] F. Charte et al. "A first approach to deal with imbalance in multi-label datasets". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8073 LNAI (2013), pp. 150–160. DOI: 10.1007/978-3-642-40846-5_16.

[5] F. Charte et al. "Concurrence among imbalanced labels and its influence on multilabel resampling algorithms". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8480 LNAI (2014), pp. 110–121. DOI: 10.1007/978-3-319-07617-1_10.

[6] Francisco Charte and David Charte. "Working with Multilabel Datasets in R: The mldr Package". In: *The R Journal* 7.2 (2015), pp. 149–162.

[7] L Chekina, L Rokach, and B Shapira. "Meta-learning for selecting a multi-label classification algorithm". In: 2011, pp. 220–227. DOI: 10.1109/ICDMW.2011.118.

[8] Jacob Cohen et al. *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences*. 2002.

[9] R. A. Fisher. "The use of multiple measurements in taxonomic problems". In: *Annals of Eugenics* 7 (1936), pp. 179–188.

[10] Eva Gibaja and Sebastian Ventura. "A tutorial on multilabel learning". In: *ACM Computing Surveys* 47.3 (2015). DOI: 10.1145/2716262.

[11] Priscilia E Greenwood and Michael S Nikulin. "A guide to chi-squared testing". In: *Wiley-Interscience* 280 (1996).

[12] Mark Hall et al. "The WEKA Data Mining Software: An Update". In: *SIGKDD Explor. Newsl.* 11.1 (2009), pp. 10–18. ISSN: 1931-0145. DOI: 10.1145/1656274.1656278.

[13] Charles J. Kowalski. "On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 21.1 (1972), pp. 1–12.

[14] J. Lee Rodgers and W. Alan Nice Wander. "Thirteen ways to look at the correlation coefficient". In: *American Statistician* 42.1 (1988), pp. 59–66. DOI: 10.1080/00031305.1988.10475524.

[15] E Loza and J Fürnkranz. "Efficient multilabel classification algorithms for large-scale problems in the legal domain". In: *Semantic Processing of Legal Texts*. Vol. 6036. 2010, pp. 192–215.

[16] Y Luo et al. "Multiview vector-valued manifold regularization for multilabel image classification". In: *Neural Networks and Learning Systems, IEEE Transactions on*. Vol. 24(5). 2013, pp. 709–722.

[17] *MEKA: A Multi-label Extension to WEKA*. http://meka.sourceforge.net/. Último acceso: 21-04-2016.

[18] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. "Efficient Algorithms for Mining Outliers from Large Data Sets". In: *SIGMOD Rec.* 29.2 (May 2000), pp. 427–438. DOI: 10.1145/335191.335437.

[19] Jesse Read. "Scalable Multi-label Classification". In: *PhD Thesis, University of Waikato* (2010).

[20] Glen Meeden Richard A. Groeneveld. "Measuring Skewness and Kurtosis". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 33.4 (1984), pp. 391–399.

[21] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Statistics. Wiley, 2005. ISBN: 9780471725374.

[22] K. Sechidis, G. Tsoumakas, and I. Vlahavas. "On the stratification of multi-label data". In: *Lecture Notes in Computer Science* 6913 LNAI.PART 3 (2011), pp. 145–158. DOI: 10.1007/978-3-642-23808-6_10.

[23] H Shao et al. "Symptom selection for multi-label data of inquiry diagnosis in traditional Chinese medicine". In: *Sci China Ser F-Info Sci* 1 (2010), pp. 1–13.

[24] E Stamatatos. "Author identification: Using text sampling to handle the class imbalance problem". In: *Information Processing and Management* 44.2 (2008), pp. 790–799. DOI: 10.1016/j.ipm.2007.05.012.

[25] L Tang and H Liu. "Scalable learning of collective behavior based on sparse social dimensions". In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. New York, NY, USA, 2009, pp. 1107–1116.

[26] C. Tsallis. "Possible generalization of Boltzmann-Gibbs statistics". In: *Journal of Statistical Physics* 52.1-2 (1988), pp. 479–487. DOI: 10.1007/BF01016429.

[27] G. Tsoumakas and I. Katakis. "Multi-label classification: An overview". In: *International Journal of Data Warehousing and Mining* 3.3 (2007), pp. 1–13.

[28] G Tsoumakas, I Katakis, and I Vlahavas. "Data Mining and Knowledge Discovery Handbook, Part 6". In: Springer, 2010. Chap. Mining Multi-label Data, pp. 667–685.

[29] G. Tsoumakas et al. "Mulan: A Java Library for Multi-Label Learning". In: *Journal of Machine Learning Research* 12 (2011), pp. 2411–2414.

[30] Chang Xu, Dacheng Tao, and Chao Xu. "A Survey on Multi-view Learning". In: *CoRR* abs/1304.5634 (2013). URL: http://arxiv.org/abs/1304.5634.

[31] Y Zhang et al. "Ensemble pruning via semi-definite programming". In: *Journal of Machine Learning Research* 7 (2006), pp. 1315–1338.