A Network Perspective on Stratification of Multi-Label Data

Piotr Szymański* PIOTR.SZYMANSKI@PWR.EDU.PL
Tomasz Kajdanowicz TOMASZ.KAJDANOWICZ@PWR.EDU.PL
Department of Computational Intelligence, Wrocław University of Science and Technology, Wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław, Poland

Editors: Luís Torgo, Bartosz Krawczyk, Paula Branco and Nuno Moniz.

Abstract

We present a new approach to stratifying multi-label data for classification purposes based on the iterative stratification approach proposed by Sechidis et. al. in an ECML PKDD 2011 paper. Our method extends the iterative approach to take into account second-order relationships between labels. Obtained results are evaluated using statistical properties of obtained strata as presented by Sechidis. We also propose new statistical measures relevant to second-order quality: label pairs distribution, the percentage of label pairs without positive evidence in folds and label pair - fold pairs that have no positive evidence for the label pair. We verify the impact of new methods on classification performance of Binary Relevance, Label Powerset and a fast greedy community detection based label space partitioning classifier. The proposed approach lowers the variance of classification quality, improves label pair oriented measures and example distribution while maintaining a competitive quality in label-oriented measures. We also witness an increase in stability of network characteristics.

Keywords: multi-label classificaiton, multi-label stratification, label space clustering, data-driven classification

1. Introduction

In the recent years, we have witnessed the development of multi-label classification (MLC) methods which utilize the structure of the label space in a divide and conquer approach to improve classification performance and allow large data sets to be classified efficiently. Without taking into account the structure of label relationships, binary and label powerset transformations can be used, the first one is not efficient due to a large number of classifiers, the other due to underfitting problems described by Tsoumakas and Vlahavas (2007). Yet most of the available data sets have been provided in train/test splits that did not account for maintaining a distribution of higher-order relationships between labels among splits or folds. As a result, classification methods are prone to make mistakes in generalization from data that was not stratified properly. This is especially relevant to problem transformation based approaches to MLC, which convert the problem to a set of binary or multi-class problems. However in such problems coming up from transformed MLC accounting for evidence distribution per class is even more important, because - as Charte et al. (2013) note - *the imbalance level in multi-label datasets is much larger than in binary or multi-class*

* The authors thank Ljupco Todorovski, Sašo Džeroski and Dragi Kocev for their feedback on the seminar at Josef Stefan Institute, where the idea for this paper was born.

datasets. Very often the imbalance ration is equal to the number of samples in the majority class after transformation, as the minority class has singular evidence.

The problem of label distribution in MLC in the problem transformation approach has two important elements: dealing with class imbalance after transformation and providing evidence for each class. The class can be understood in many ways, as Zhang et al. (2015) note, transformation exploit different orders of label relationships to form classes: *first-order approaches which assume independence among class labels, second-order approaches which consider correlations between a pair of class labels, and high-order approaches which consider correlations among all the class labels or subsets of class labels.*

Problem adaptation approaches to MLC tackle the label imbalance problems by incorporating modifications into the learning method, which is a task called imbalance learning, examples of such methods are multi-label Charte et al. (2015b) variant of SMOTE (Torgo et al. (2013)), COCOA by Zhang et al. (2015), a review of perspectives was written by Krawczyk (2016). Label imbalance is usually coped with by performing resampling procedures like Charte et al. (2015a) or generating synthetic data.

While Krawczyk (2016) notes *little attention was paid to imbalanced learning in the multi-label classification context, despite the fact that area suffers from it*, even less attention has been paid to stratification of multi-label data without resampling or synthesizing data. Resampling rare classes is important in multi-class classification, however, the existence of very rare classes in labelset-to-class transformations (as in label powerset method) is often a problem of underfitting/overfitting due to the exponentiality of the transformed space. This can be overcome by dividing the label space into smaller subspaces of labels that are more related to each other and their evidence distribution among labels is less skewed. Clustering approaches often depend on second-order approaches Szymanski et al. (2016). Resampling is also not viable for non-labelset transformations such as Classifier Chains by Read et al. (2009), which aim to order create a chain of single-class classifiers dependent on each other in a Bayes chain-rule fashion. Most effective of the classifier chains variants (ex. Read et al. (2014)) explore second-order label relations to construct the ordering of the chain before training the classifiers. A similar case may be argued for other second-order MLC methods such as Calibrated Label Ranking by Fürnkranz et al. (2008) and similar by Zhang and Zhang (2010).

Sechidis et al. (2011) published the only paper on multi-label stratification to date where they provide an iterative algorithm to maintain the availability of evidence per label. However, their approach is directed at maintaining first-order label evidence distributed across the strata as they have come across the lack of evidence problem upon Binary Relevance transformation. A stratification approach has also been proposed by Charte et al. (2016) alongside with a theoretical measure of dataset complexity. A genetic algorithm approach has been evaluated in a Ph.D. research of Fernández del Pozo et al., yet it has not been published, neither the paper describing the procedure, nor the code.

In this paper, we propose an extended version of Iterative Stratification approach, which we call the Second-Order Iterative Stratification, which takes the desirability of label pairs and not just single labels into account when performing stratification while maintaining a graceful fallback to IS when second-order relationships are not well exhibited in the data. We compare it to IS, stratified and traditional k-fold approaches. We evaluate a 10 fold

moving window cross validation division - relevant to parameter estimation for MLC base classifiers.

2. Proposed Method

Algorithm 1: Second Order Iterative Stratification (SOIS)
Input: Set of samples D, labels L, number of folds k, list of desired proportions per fold r
$\Lambda \leftarrow \{\{\lambda_i, \lambda_j\} : (\exists (x, Y) \in D)(\{\lambda_i, \lambda_j\} \subset Y)\}$; **foreach** $e \in \Lambda$ **do**
  $D^e \leftarrow \{(x, Y) : Y \cap e \neq \emptyset\}$;
end
for $j = 1..k$ **do**
  $c_j \leftarrow |D| * r_j$; **forall** $e \in \Lambda$ **do**
    $c_j^e \leftarrow |D^e| * r_j$;
  end
end
return *DistributeOverFolds*(D, Λ, c)

Algorithm 2: Iterative distribution of samples into folds (DistributeOverFolds)
Input: Set of samples D, set of edges with samples Λ, percentages of desired samplsing from a given edge per fold c
while $|\{(x, Y) \in D : Y \neq \emptyset\}| > 0$ **do**
  foreach $\lambda_i \in \Lambda$ **do**
    $D^i \leftarrow \{(x, Y) : Y \cap \lambda_i \neq \emptyset\}$;
  end
  $l \leftarrow \arg\min_{i, D^i \neq \emptyset} |D^i|$; **forall** $(x, Y) \in D^l$ **do**
    $M \leftarrow \arg\max_{j=1...|L|} c_j^l$; **if** $|M| == 0$ **then**
      $m \leftarrow onlyElement(M)$;
    end
    else
      $M' \leftarrow \arg\max_{j \in M} c_j$; **if** $|M'| == 0$ **then**
        $m \leftarrow onlyElementOf(M')$;
      end
      else
        $m \leftarrow randomElementOf(M')$;
      end
    end
    $S_m \leftarrow S_m \cup (x, Y)$; $D \leftarrow D \setminus (x, Y)$; $c_m^l \leftarrow c_m^l - 1$; $c_m \leftarrow c_m - 1$;
  end
end
foreach $(x, Y) \in D$ **do**
  $M \leftarrow \arg\max_{i=1...k} c_i$; $m \leftarrow randomElementOf(M)$; $S_m \leftarrow S_m \cup (x, Y)$; $c_m \leftarrow c_m - 1$;
end
return $S_1, \ldots, S_k$

We propose an extended version of the Iterative Stratification (IS) algorithm from Sechidis et. al. so that takes into account the second-order label relations (we call it SOIS).

In order to introduce the algorithm we start with the following notations. Let X denote the input space, L - the set of labels, $D \subset X \times 2^L$ - the data set, k - the number of desired folds, and $r_i|_1^k$ the desired proportion of labels in each of the folds ($\sum_{i=1}^{k} r_i = 1$). In a typical 10-fold CV scenario: $k = 10$, and $r_i = \frac{1}{10}$. Let E denote the set of all pairs of labels that occur together in D:

$$E = \big\{\{\lambda_i, \lambda_j\} : \big(\exists(\bar{x}, \Lambda) \in D\big)\big(\lambda_i \in \Lambda \wedge \lambda_j \in \Lambda\big)\big\}$$

The proposed algorithm - Second Order Iterative Stratification (Algorithm 1) first calculates the desired number of samples for each label pair, per fold. In the second part Algorithm 2 is iterating over label pairs from E, it selects the label pair with the least samples available, iterates over all samples with this label pair assigned, assigning the sample to the fold that desires the label pair the most, randomly breaking the ties. The relevant counters of label pair availability and per fold sample, label and label pair desirability are updated, and the internal loop over samples progresses, once all samples evidencing the selected label pair are used up it continues with another iteration of the outer loop.

Once all label pairs are distributed, the same algorithm is employed to distribute labels from L in a similar manner - which is the graceful fallback to the IS algorithm once all the label pair evidence has been distributed. For each output set (label, label pair or label set), we define the positive evidence to be the set of all samples labeled with a given output set, while negative evidence consists of the samples not labels with that output set. Once all positive evidence of labels is distributed, negative evidence is randomly distributed as to satisfy sample desirability in each of the folds. SOIS includes both label pairs and labels (represented as i, i pairs) in E for consideration. Negative evidence is distributed as in SOIS once all the positive evidence has been distributed.

2.1. Experimental Setup

We perform experimental evaluation of presented stratification approaches on 16 benchmark data sets that were available in the MULAN repository MULAN (2016): Corel5k, bibtex, delicious (not used in network approaches as calculations did not finish), emotions, enron, genbase, mediamill, medical, rcv1subset1, rcv1subset2, rcv1subset3, rcv1subset4, rcv1subset5, scene, tmc2007-500, yeast. Experiments were performed using the scikit-multilearn library by Szymański (2016).

Stratification methods were evaluated by analyzing the characteristics of each fold in terms of statistical measures, classification quality measures with classification performed using Binary Relevance, Label Powerset and Data-Driven Label Space Partitioning with Label Powerset.

Additionaly we evaluate the impact of stratification methods on a models' ability to perform generalization approaches using classification and label ranking quality metrics provided by the scikit-learn library by Pedregosa et al. (2011). Evaluated models include Binary Relevance, Label Powerset, and data-driven label space partitioning following our previous research.

3. Results

We evaluate the considered stratification methods in terms of three types of properties. First we are interested in the quality of sample distribution over folds in terms of the statistical properties of output spaces that the model will work on in a cross-validation setting. Next we evaluate what is the impact of generalization quality in two baseline approaches - Binary Relevance which should depend on how well each of the label is evidenced and counterevidenced in each fold, and Label Powerset which should be more prone to higher-order relation misstratification. Finally we look into the stratification methods' impact on label co-occurrence graphs, the detected communities, their stability, the obtained modularities and generalization quality of under the partitioned scheme.

3.1. Statistical properties of folds

In this section we compare sampling approaches using statistical properties of obtained data subsets using the properties from Sechidis et. al.'s paper and also their second-order label relations equivalents. We follow the notation from previous paragraphs to define the measures used in this section.

Label Distribution (LD) is a measure that evaluates how the proportion of positive evidence for a label to the negative evidence for a label deviates from the same proportion in the entire data set, averaged over all folds and labels. In the following notation S_j^i and D^i are the sets of samples that have the i-th label from L assigned in the j-th fold and the entire data set, respectively:

$$LD = \frac{1}{|L|}\sum_{i=1}^{|L|}\Big(\frac{1}{k}\sum_{j=1}^{k}\Big|\frac{|S_j^i|}{|S_j| - |S_j^i|} - \frac{|D^i|}{|D| - |D^i|}\Big|\Big)$$

Label Pair Distribution (LPD) is an extension of the LD measure that operates on positive and negative subsets of label pairs instead of labels. In the following definition S_j^i and D^i are the sets of samples that have the i-th label pair from E assigned in the j-th fold and the entire data set, respectively:

$$LPD = \frac{1}{|E|}\sum_{i=1}^{|E|}\Big(\frac{1}{k}\sum_{j=1}^{k}\Big|\frac{|S_j^i|}{|S_j| - |S_j^i|} - \frac{|D^i|}{|D| - |D^i|}\Big|\Big)$$

Examples Distribution (ED) is a measure of how much a given fold's size deviates from the desired number of samples in each of the folds:

$$ED = \frac{1}{k}\sum_{j=1}^{k}\Big||S_j| - c_j\Big|$$

In a cross-validation setting we are also interested in how, we thus define: **FZ** - the number of folds that contain at least one label with no positive examples, **FLZ** - the number of fold-label pairs with no positive examples, **FLPZ** - a second-order extension of FLZ - the number of fold - label pair pairs with no positive examples. In the case of FLPZ, as it happens that label pairs do not have enough evidence to split over the evaluated 10 folds,

we only count these label pair - fold pairs that had more folds without positive examples, than the inevitable minimum value corresponding to the number of folds minus the number of available samples with a label pair.

	kfold		labelset		SOIS		IS	
	mean	std	mean	std	mean	std	mean	std
Corel5k	0.828	0.04	0.820	0.28	**0.699**	0.01	0.709	0.01
bibtex	0.694	0.03	0.851	0.29	**0.662**	0.02	0.687	0.02
delicious	0.592	0.00	0.887	0.30	**0.582**	0.00	0.584	0.00
emotions	0.285	0.11	0.256	0.14	**0.161**	0.04	0.251	0.09
enron	0.649	0.07	0.806	0.28	**0.578**	0.02	0.602	0.02
genbase	0.686	0.15	0.601	0.31	**0.487**	0.16	0.494	0.14
mediamill	0.491	0.03	0.596	0.23	**0.324**	0.01	0.364	0.01
medical	0.762	0.06	0.762	0.30	**0.736**	0.03	0.751	0.04
rcv1subset1	0.712	0.02	0.729	0.26	**0.581**	0.01	0.606	0.02
rcv1subset2	0.712	0.05	0.727	0.26	**0.574**	0.01	0.598	0.02
rcv1subset3	0.721	0.04	0.731	0.26	**0.583**	0.01	0.606	0.02
rcv1subset4	0.720	0.08	0.709	0.26	**0.574**	0.01	0.600	0.02
rcv1subset5	0.714	0.03	0.732	0.26	**0.584**	0.02	0.603	0.02
scene	0.711	0.10	0.277	0.11	**0.276**	0.05	0.312	0.14
tmc2007-500	0.218	0.02	0.347	0.17	**0.159**	0.01	0.207	0.03
yeast	0.078	0.03	0.095	0.04	**0.062**	0.01	0.064	0.02

Table 1: Percentage of label pairs without positive evidence, averaged over 10 folds, with standard deviation. The lesser the better. The best performing division method in bold. Methods with smallest variance are underlined.

As there is little reason to generalizing FZ to label pairs as an integer measure, because all folds miss at least one label pair, we generalize it as a measure of percentage of label pairs that are not present in each of the folds. We provide average percentages per data set per method alongside with standard deviations in Table 1.

The best method for multi-label stratification should provide folds that have a small Example, Label and Label Pair Distribution scores, as such a stratification remains well balanced both in terms of evidence and in terms of size. It should also yield small number of folds that miss evidence for labels and label pairs and preferably if a miss happens it should be as small as possible, thus FZ, FLZ and FLZP should be as small as possible. Similarly the percentage of label pairs not evidenced per fold should be both small on average, but also stable. Let us look at Figure 1 to see how the evaluated methods rank on average from the statistical properties perspective.

The k-fold approach is a clear winner when it comes to lowest deviation fold sizes (ED) which does not surprise us, as the only criterion of the traditional k-fold division is the number of examples. While simplest, available in practically all multi-label classification libraries and thus most often used - it remains the worst ranked in FLZ, LD, LPD. It also ranks second worst in terms of FLZP and the percentage of label pairs not evidenced per fold

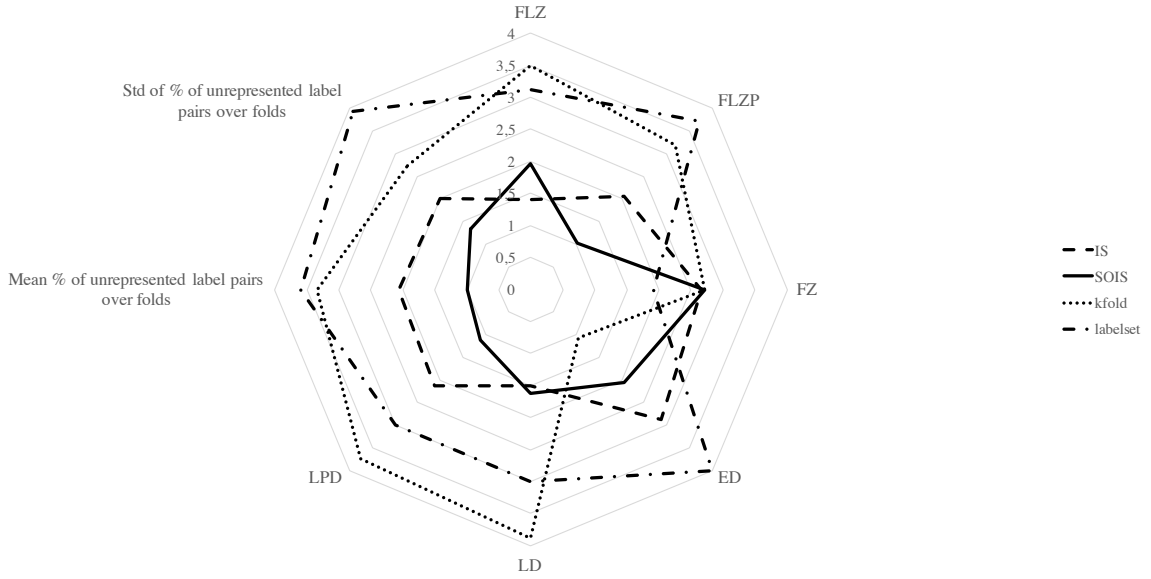


Figure 1: Average ranks of proposed stratification approaches with regard to statistical properties of generated strata.

both on average and in the scale of standard deviation of the percentage of label pairs not evidenced per fold. It is noteworthy that in most of the evaluated data sets this approach generates folds in which, on average, lack positive samples for 70-80% of label pairs. k-fold does not provide folds that maintain a distribution of labels or label pairs. Clearly this measure should only be used when the data set authors have taken other precautions concerning label and label pair distributions before performing division.

The stratified labelset approach ranks on par with the best in FZ, worst in ED, FLZP and coverage of label percentages with positive evidence - both in average and standard deviation. It ranks second worse in other measures. In practice what we have observed is that there this approach creates the most informed fold first. That fold contains positive evidence for as many label combinations (classes) as possible, leaving few samples to serve as such evidence in other folds. Such an approach yields large deviation of percentages of unevidenced label pairs and also creates a disproportion in fold sizes. It is succesful in minimizing FZ as the first fold is always sure to be well-evidenced. This method should only be used in the case when there is little to no imbalance of positive evidence distribution among labelsets, in practice - never.

The Iterative Stratification approach ranks best in terms of FLZ and LD, and on par with the best in terms of FZ. It performs second best in label pair measures, but it ranked visibly worse than SOIS. This approach performs best stratification when it comes to making

sure that all labels have positive evidence in all folds, but underperforms when it comes to positive evidence for label pairs. It also ranks second worse in ED losing only to stratified labelset approach.

Second-Order versions of IS (SOIS) perform best in measures related to label pairs and is better in ED than other non-kfold approaches, while also performing second best in all other measures, ranking closely to the best performers.

Out of the two methods scoring well in label and label pair measures, SOIS is clearly a better choice as the gain in stability of label pair measures is larger than the loss in FLZ. In other single-label measures SOIS ranks closely to IS. The method successfuly finds a compromise between evidencing labels and label pairs while maintaining small deviations of sample sizes per fold.

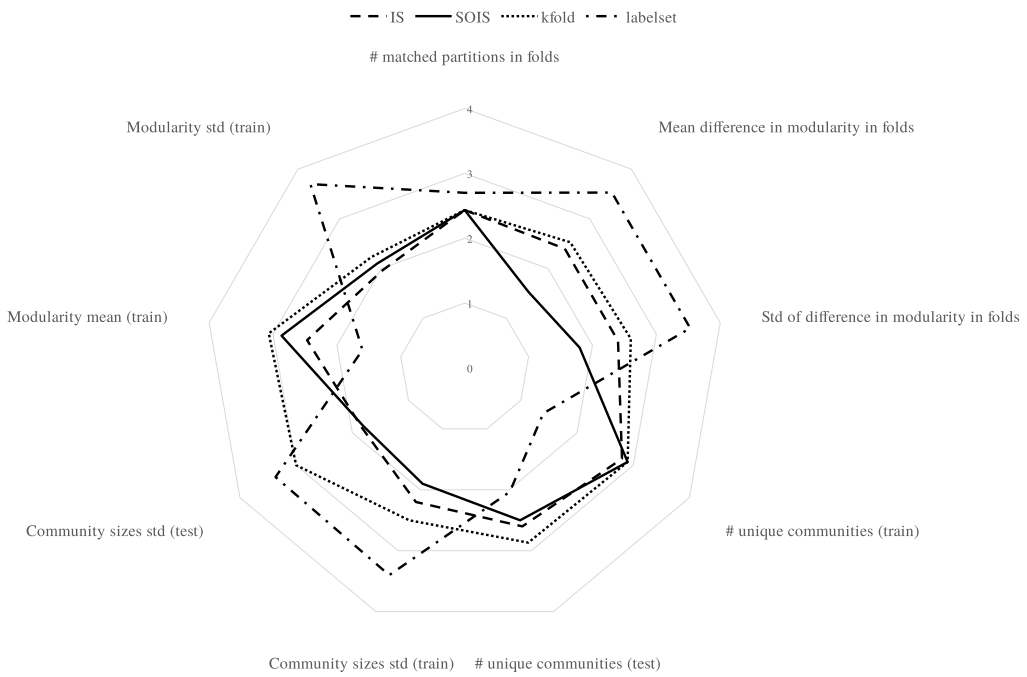
3.2. Stability of Network Characteristics

From the Network Perspective it is important that a stratification methods provides stability in obtained modularity scores both over the training folds and between train and test folds of a given strata. In the perfect case the stratification algorithm should provide data that allow constructing graphs similar enough that the community detection algorithm would find exactly the same community in all folds, and exactly the same community in every train/test fold pair per stratum.

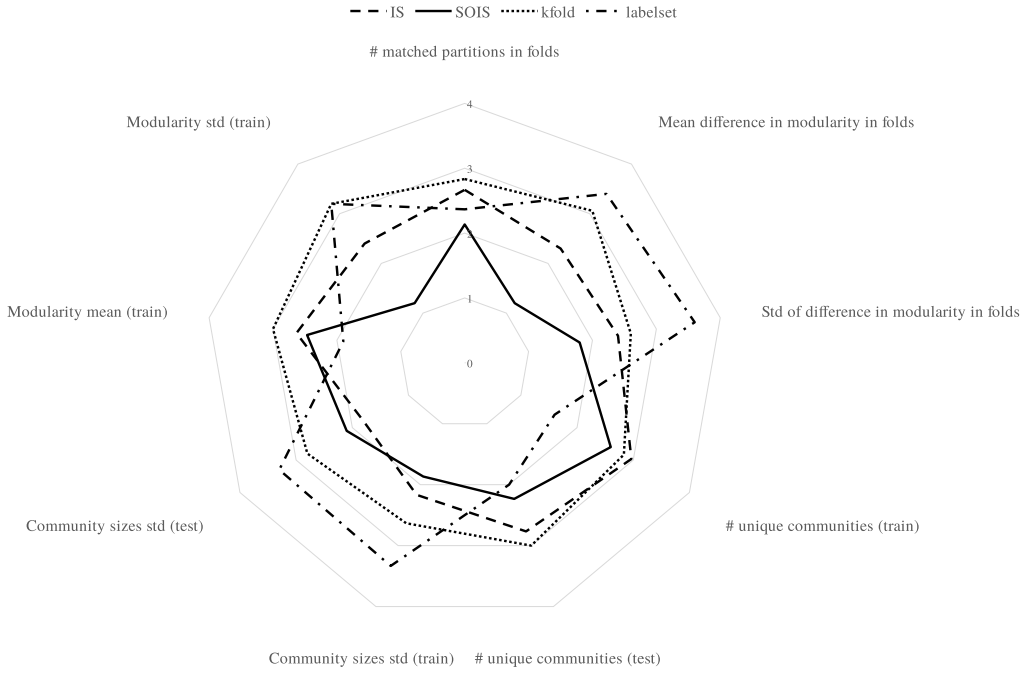
We used the fast greedy modularity maximization scheme provided by the igraph Csardi and Nepusz (2006) library to detect communities on label graphs constructed from training and test examples in each of the folds. We constructed both the unweighted and weighted graphs and performed community detection on both of them. We review the case of each of the graphs separately.

We evaluated the following Network Characteristics: the mean and standard deviation of modularity scores over training folds, the stability (i.e. the standard deviation) of the number of sizes of communities detected in each train and test fold and the number of unique communities. We also count the number of partitions that were exactly matched per train-test subsets of every fold and the mean and standard deviation of modularity differences between train-test subsets of every fold. The results are illustrated in Figure 2 for the unweighted graph case and in Figure 2 for the weighted case.

The labelset stratification approach ranks best when it comes to obtained modularity mean on train examples, yet worst when it comes to standard deviation of the modularity score. It is like this because the first fold is always provided with as complete evidence as possible, which makes any mean score higher, while other folds do not include rare data and become different problems - yielding a very high standard deviation. Similar case happens with unique communities, where the problems with less evidence become more similar, yet simpler, yielding less communities due to lack of edges - as in this case edges are binary indications of existence of samples labeled with a given label pair. In all other measures the labelset approach performs worst and, as was in the statistical measures case, should not be used in practice.



a. fast-greedy unweighted (FG)



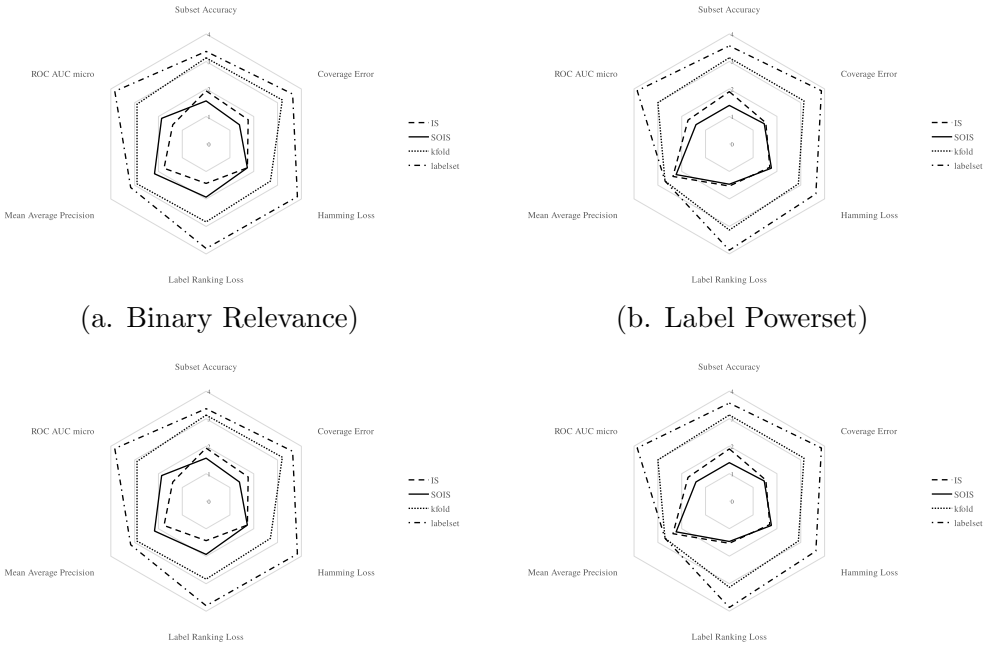
b. fast-greedy weighted (FGW)

Figure 2: Average ranks of proposed stratification approaches with regard to different characteristics of the weighted label co-occurence graph constructed on stratified folds.

SOIS ranks higher than IS in every measure apart from the standard deviation of obtained community sizes in test sets and is only slightly higher ranked in terms of mean obtained modularity on train data. We see that SOIS ranks consistently better in better matching of partitions between relevant train-test pairs and yielding lower and more stable modularity differences among these pairs.

We observe that SOIS is closer to realizing the ideal scenario than IS in most of the measures on weighted graphs, where the number and not just the presence of samples is most important. SOIS also maintains and advantage in terms of unweighted graphs, but the difference with IS is less significant. Similarly as in the case of statistical measures we note that SOIS is a better choice than IS when it comes to maintaining network characteristics across folds.

3.3. Variance of generalization quality



(a. Binary Relevance) (b. Label Powerset)

(c. fast-greedy unweighted (FG)) (d. fast-greedy weighted (FGW))

Figure 3: Average ranks of proposed stratification approaches with regard to standard deviation of scores in evaluated generalization measures when classification was performed using a given classification method (underneath) over stratified folds.

In terms of generalization quality one would expect for stratification methods to allow comparable generalization perspectives to the model in each of the folds, while not compromising the average generalization quality. For evaluation purposes we take two standard approaches to classification - Binary Relevance (BR, Figure 3a) and Label Powerset (LP, Figure 3b), and two wariants of the data-driven label space clustering using fast greedy

modularity maximization on unweighted (FG, Figure 3c) and weighted (FGW, Figure 3d) label co-occurrence graphs. We do not compare them to each other, instead we compare the standard deviation of their generalization quality over folds generated by each stratification method. We recall the original measures presented in Sechidis et. al.'s work: Subset Accuracy, Coverage Error, Hamming Loss, Label Ranking Loss, Mean Average Precision (also known as macro-averaged precision), micro-averaged Receiver Operating Characteristic Area Under Curve.

The Binary Relevance case is fairly evident, with label powerset yielding the highest standard deviations, followed by k-fold and IS, while SOIS ranks best with most stable generalization. The Label Powerset provides similar worst-performing picture of labelset and kfold approaches. In this case however the distances between the best ranked IS and SOIS are small and what SOIS gains in Mean Average Precision or Coverage Error or Subset Accuracy it loses in Label Ranking Loss. From a practical point of view the methods perform equally well in this case.

Similarly to the case of network characteristics when it comes to unweighted fast greedy case, the standard deviation of generalization scores is similar between IS and SOIS, while the other methods rank last and second last. In this case again IS ranks better in Label Ranking Loss, Mean Average Precision and ROC AUC micro, while SOIS ranks better in Subset Accuracy and Coverage Error. The distances between the ranks are small and what one method gains in one measure's stability, it loses in the another one.

In the case of weighted label co-occurence graphs we observe that, consistently with other experimental results, kfold and stratification approaches rank worst when it comes to standard deviation of generalization measures. In this case we also observe, what is compatible with the network characteristics results for the weighted graph, that SOIS ranks better or on par than IS becoming a stratification method of choice.

We observe that kfold and stratification methods perform worst in classification stability across all evaluated cases. The two algorithms that perform best: IS and SOIS provide similar generalization stability with Label Powerset or unweighted fast-greedy scheme. When Binary Relevance or weighted fast-greedy approach are used, SOIS performs better.

We did not provide the result tables in print as all of the data and result tables are available in the Github repository associated with this paper[1] to maintain the standards of reproducibility. In print the tables would span multiple pages and would not serve the purpose of illustrating the results in an understandable fashion. The result tables, notebooks and code are thus provided in the repository and can be browsed digitally to allow comfortable review.

4. Conclusions and future research

Our experiments show that the stratification based on label powerset transformation results in distributing as much positive evidence as available in the first fold(s) and running out of positive evidence for other folds. Thus the data set actually becomes divided into completely different problems - a more complicated one based on the super fold, and the easier in the other folds. While such a division allows better scoring due to lack of hard test samples in most of the folds, it is far from providing data that allow a stable generalization. The other

1. https://github.com/niedakh/multilabelstratification

traditional method for data set division - kfold - ranks consistently worst in terms of the scale of standard deviation obtained in evaluated measures.

We discourage the use of k-folding and label powerset transformation based stratification and instead propose to use an iterative approach that takes second-order relationships into account and provides folds that exhibit more stability in terms of statistical measures, network characteristics and generalization quality. We recommend using SOIS instead of IS as the stability increase yielded by SOIS is usually a greater advantage then the rare cases of SOIS ranking lower that IS, noting the small distances in ranks in these cases.

Future research into the topic should examine the impact on other community detection and clustering methods for example k-means, infomap, etc, a larger number of data sets and stronger theoretical considerations. It would also be very interesting to evaluate other scenarios such as 5x2 standard train-test divisions and validating the methods on controlled artificial data sets. We also plan to compare evaluated solutions with the stratification approach proposed by Charte et al. (2016). We plan to evaluated the impact of stratification approaches on multi-label classification when algorithm adaptation methods are used instead of problem transformation, for example the multi-label decision trees by Vens et al. (2008).

Acknowledgments

The work was partially supported by The National Science Centre the research projects no. 2016/21/N/ST6/02382 and 2016/21/D/ST6/02948 and by the Faculty of Computer Science and Management, Wrocław University of Science and Technology statutory funds. Research financed under the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 691152 (RENOIR); the Polish Ministry of Science and Higher Education fund for supporting internationally co-financed projects in 2016-2019 (agreement no. 3628/H2020/2016/2)

References

Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. *A First Approach to Deal with Imbalance in Multi-label Datasets*, pages 150–160. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-40846-5. doi: 10.1007/978-3-642-40846-5_16. URL http://dx.doi.org/10.1007/978-3-642-40846-5_16.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3 – 16, 2015a. ISSN 0925-2312. doi: http://dx.doi.org/10.1016/j.neucom.2014.08.091. URL http://www.sciencedirect.com/science/article/pii/S0925231215004269. Recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World Problems Progress in Intelligent Systems Mining Humanistic Data.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385 – 397, 2015b. ISSN 0950-7051. doi: http://dx.

doi.org/10.1016/j.knosys.2015.07.019. URL http://www.sciencedirect.com/science/article/pii/S0950705115002737.

Francisco Charte, Antonio Rivera, María José del Jesus, and Francisco Herrera. On the impact of dataset complexity and sampling strategy in multilabel classifiers performance. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 500–511. Springer, 2016.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL http://igraph.org.

J. A. Fernández del Pozo, P. Larrañaga, and C. Bielza. Stratified cross-validation in multilabel classification using genetic algorithms. http://leo.ugr.es/MD-PGMs/ficheros_presentaciones/albacete3/SCVMLCGA.pdf. Accessed: 2017-06-30.

Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine learning*, 73(2):133–153, 2008.

Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

MULAN. Multilabel datasets. http://mulan.sourceforge.net/datasets-mlc.html, January 2016.

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

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning and Knowledge Discovery in Databases*, pages 254–269, 2009.

Jesse Read, Luca Martino, and David Luengo. Efficient monte carlo methods for multi-dimensional learning with classifier chains. *Pattern Recognition*, 47(3):1535–1546, 2014.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, 2011.

Piotr Szymanski, Tomasz Kajdanowicz, and Kristian Kersting. How is a data-driven approach better than random choice in label space division for multi-label classification? *CoRR*, abs/1606.02346, 2016. URL http://arxiv.org/abs/1606.02346.

P. Szymański. Scikit-multilearn: Enhancing multi-label classification in python. http://scikit-multilearn.github.io/, January 2016.

Luís Torgo, Rita P Ribeiro, Bernhard Pfahringer, and Paula Branco. Smote for regression. In *EPIA*, volume 8154, pages 378–389, 2013.

Grigorios Tsoumakas and Ioannis Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. *Machine learning: ECML 2007*, pages 406–417, 2007.

Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.

Min-Ling Zhang and Kun Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 999–1008. ACM, 2010.

Min-Ling Zhang, Yu-Kun Li, and Xu-Ying Liu. Towards class-imbalance aware multi-label learning. In *IJCAI*, pages 4041–4047, 2015.