Pattern mining for label ranking

Pinho Rebelo de Sá, C.F.

Chapter 6

Permutation Tests for Label Ranking

Cláudio Rebelo de Sá, Carlos Soares, Arno Knobbe

in local proceedings of the 27th Benelux Conference on Artificial Intelligence, 2015

Abstract

In recent years, many Label Ranking (LR) methods have been proposed, along with an increasing number of datasets. The validation of these algorithms has been done empirically, as is usual in Machine Learning, by testing them on a set of benchmark datasets. Due to the scarcity of real-world LR data, most of the experiments are based on LR datasets that were adapted from classification and regression datasets from the UCI repository and Statlog project. In this work, we want to test how strong is the relation between the target rankings and independent variables. In other supervised learning tasks, target swap randomization methods have been used to test it. We propose two target swap randomization approaches for LR and apply them on KEBI datasets. Our results show that there are meaningful relations between the independent variables and the target rankings and that the relative importance of each label in a ranking varies in some cases.

6.1 Introduction

The study of label ranking is receiving increased attention [27, 36, 28, 116, 64]. Label Ranking (LR) studies the problem of learning a mapping from instances to rankings over a finite number of predefined labels. It can be considered a variant of the conventional classification problem [26]. However, in contrast to a classification setting, where the objective is to assign examples to a specific class, in LR we are interested in predicting the (possibly incomplete) true preference order of the labels for every example. This means that the true ranking of the labels is available for the training examples.

Due to the lack of benchmark LR datasets, 16 semi-synthetic datasets were proposed in [26]. They are based on multi-class and regression datasets from the UCI repository and Statlog project. For multi-class problems, also referred as *type A*, the naive Bayes classifier was trained to give a probability score to each class, and the true ranking is based on that score. For the regression problems, *type B*, some numeric attributes were normalized and the true ranking was based on the relative order of their values.

Since then, this set of 16 datasets has been used by the majority and the most influential contributions in the Label Ranking field [28, 27, 116, 64]. However, it is unclear if the type B datasets contain any real relations between the target rankings and independent variables. While type A can be interpreted as the preferences of an agent, which in this case is the naive Bayes classifier, on type B, the relations is application-specific and it is unclear whether it exists or not. To test whether such a relation exist, we expect to find strong statistical validation of it.

In many data mining applications, Swap Randomizations techniques are used together with statistical tests to validate the significance of findings [62]. After swapping the position of the values along the attributes, the resulting models are compared with the ones obtained from the original data. Therefore, statistical significance tests can be used to validate the latter.

In this work, we investigate the usefulness of the type B datasets from the KEBI data repository by comparison to type A. For that purpose, we propose two swap randomization methods specific for the LR task. Our results show that both types of semi-synthetic data have relevant preference information.

The paper is organized as follows: Section 6.2 introduces the LR problem. Section 6.3 introduces the swap randomization and Section 6.4 describes the method proposed here. Section 6.5 presents the experimental setup and

discusses the results. Finally, Section 6.6 concludes this paper.

6.2 Label Ranking

The LR task is similar to classification. In classification, given an instance x from the instance space $\mathbb{X}$, the goal is to predict the label (or class) λ to which x belongs, from a predefined set $\mathcal{L} = \{\lambda_1, \ldots, \lambda_k\}$. In LR, the goal is to predict the ranking of the labels in $\mathcal{L}$ that are associated with x [74]. A ranking can be represented as a total order over $\mathcal{L}$ defined on the permutation space Ω. In other words, a total order can be seen as a permutation π of the set $\{1, \ldots, k\}$, such that $\pi(a)$ is the position of λ_a in π.

As in classification, we do not assume the existence of a deterministic $\mathbb{X} \to \Omega$ mapping. Instead, every instance is associated with a *probability distribution* over Ω [26]. This means that, for each $x \in \mathbb{X}$, there exists a probability distribution $\mathcal{P}(\cdot|x)$ such that, for every $\pi \in \Omega$, $\mathcal{P}(\pi|x)$ is the probability that π is the ranking associated with x. The goal in LR is to learn the mapping $\mathbb{X} \to \Omega$. The training data is a set of instances $D = \{\langle x_i, \pi_i \rangle\}, i = 1, \ldots, n$, where x_i is a vector containing the values $x_i^j, j = 1, \ldots, m$ of m independent variables describing instance i and π_i is the corresponding target ranking.

Given an instance x_i with label ranking π_i, and the ranking $\hat{\pi}_i$ predicted by an LR model, we evaluate the accuracy of the prediction with a loss function on Ω. One such function is the number of discordant label pairs,

$$\mathcal{D}(\pi, \hat{\pi}) = \#\{(a, b) | \pi(a) > \pi(b) \wedge \hat{\pi}(a) < \hat{\pi}(b)\}$$

If normalized to the interval $[-1, 1]$, this function is equivalent to Kendall's τ coefficient [85], which is a correlation measure where $\mathcal{D}(\pi, \pi) = 1$ and $\mathcal{D}(\pi, \pi^{-1}) = -1$ (π^{-1} denotes the inverse order of π).

The accuracy of a model can be estimated by averaging this function over a set of examples. This measure has been used for evaluation in recent LR studies [26, 40] and, thus, we will use it here as well. However, other correlation measures, like Spearman's rank correlation coefficient [118], can also be used.

6.2.1 IB-PL

Instance-Based Placket-Luce (IB-PL) is an highly competitive method in label ranking proposed in [24]. It is a local prediction method based on the nearest neighbor estimation principle. Given a new instance $\hat{x}$ it uses the $\{\pi_1, \ldots, \pi_K\}$ rankings associated with the K nearest neighbors to predict the ranking $\hat{\pi}$ associated with $\hat{x}$. The estimation of $\hat{\pi}$ is made using a Maximum Likelihood Estimation of the Plackett-Luce (PL) model which assumes that the rankings have been produced independently of each other.

6.2.2 APRIORI-LR

APRIORI-LR is an algorithm that generates *Label Ranking Association Rules* (LRAR) [36] which are a straightforward adaptation of Class Association Rules (CAR): $A \rightarrow \pi$ Where $A \subseteq desc(\mathbb{X})$ and $\pi \in \Omega$. Where $desc(\mathbb{X})$ is the set of descriptors of instances in $\mathbb{X}$, typically pairs $\langle attribute, value \rangle$. Similar to how predictions are made with CARs in CBA (Classification Based on Associations) [97], when an example matches the antecedent of the rule, $A \rightarrow \pi$, the predicted ranking is π.

6.2.3 Datasets

Even though Label Ranking potentially has a large number of practical applications [74], before the KEBI datasets, there were not many datasets available [26]:

- Meta-learning [17] on which we try to predict a total ranking of a set of algorithms accordingly to the best expected accuracy for each dataset.
- Microarray [74] which provides information of genes from Yeast on five different micro-array experiments (spo, heat, dtt, cold and diau).
- Image categorization [58] of landscape pictures from several categories (beach, sunset, field, fall foliage, mountain, urban).

To solve this problem, the KEBI Label Ranking data repository was created [26]. Data from the UCI repository and Statlog collection was transformed into Label Ranking data using the following two procedures:

type A The multi-label data is used in the training of a naive Bayes classifier. The predicted class probabilities are ranked by decreasing order

for each example, which will result in a label ranking. (To avoid incomplete rankings, the labels with lower indexes are ranked first in case of ties)

type B With the regression data, the process consisted of transforming some attributes into labels. A selected set of attributes are normalized and then ranked by descent order for each example. The remaining attributes will then be used to predict the rankings. As some of the attributes are correlated, this transformation is believed to keep a relation between predictors and rankings.

While the type A rankings can be interpreted as the preferences of a classifier, namely the naive Bayes, the interpretation on type B is not so clear. As mentioned in [26], type B datasets lead to more difficult learning problems. In this work, we analyze both data types.

6.3 Swap Randomization

Swap randomization consists of the creation of randomized datasets $\{D'_i\}_{i=1,\ldots,s}$ from a given dataset D to compare and validate the findings of data mining algorithms. We can maintain the margins of the attributes of D in all $\{D'_i\}_{i=1,\ldots,s}$ by swapping the position of the values per attribute (see Figure 6.1).

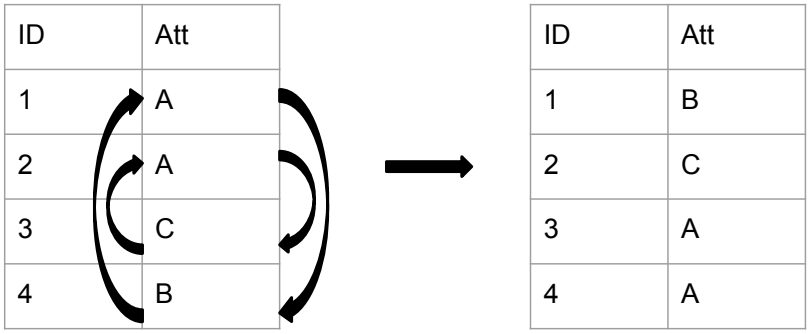

Figure 6.1: Illustration of a swap randomization per attribute.

Given an interest measure a_i, for example the accuracy of a learning method, an estimation of $\{a_i\}_{i=1,\ldots,s}$ can be obtained for $\{D'_i\}_{i=1,\ldots,s}$, respectively. Considering a_D as the estimation of a in the dataset D by a given method, if a_D deviates significantly from the distribution of $\{a_i\}_{i=1,\ldots,s}$ we can consider a_D to be significant, otherwise we do not consider it to be relevant [62].

The same concept has also been widely used in the classification task to validate classifiers [63, 103], and is commonly referred to as the *permutation test*. The *p-value* can be seen as the fraction of $\{D'_i\}_{i=1,\dots,s}$ where the classifiers obtained better results than in D. In other words, this procedure measures to what extent the accuracy of classifiers could have been due to chance [105]. The null hypothesis assumes that there is no relation between the independent variables and the targets.

If we reverse the interpretation, we can also use learning methods to assess the information contained in the datasets. By using more than one learner we can avoid the bias of the methods.

6.4 Validating ranking data with permutation tests

Swap Randomization is used to verify the significance of Data Mining discoveries from any given method [62]. If we use the same concept and randomly permute the position of the target attribute relatively to the independent variables, we should be able to verify if the relation attribute-target is also meaningful. While the target class has only one dimension, the target ranking has k dimensions. This property allows us to make partial permutations i.e. we permute the ranks of one label while leaving the remaining ranks unchanged. The different approaches are detailed below.

6.4.1 Random permutation of rankings

Randomly permuting the rankings is a natural adaptation of the methods used in classification like in [63]. By randomly permuting the target rankings, we want to test the *strength* of the relation $\mathbb{X} \to \Omega$ in the data, as exemplified in Table 6.2. After the permutation, since we break this relation, we can measure how the LR learners behave and compare with the results on the original data. If the differences are not significant, we can conclude that there is no real relation $\mathbb{X} \to \Omega$. Otherwise, we can statistically show that the attributes-ranking relations are meaningful.

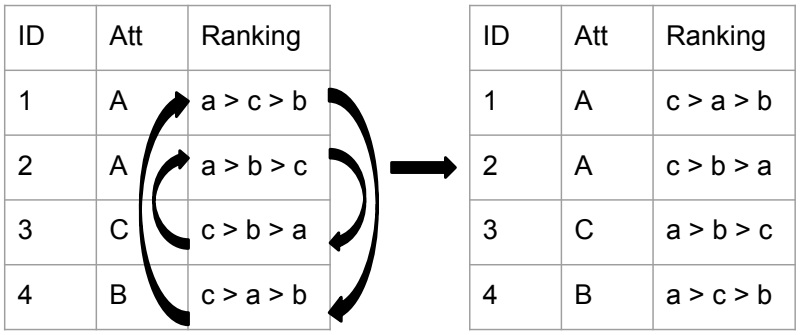

Figure 6.2: Illustration of a permutation of rankings.

6.4.2 Random permutation of labels

In LR the target can be seen as a multidimensional variable, from which both labelwise and pairwise levels of information can be extracted. We refer to a *labelwise permutation* when the ranks of a specific label are permuted.

By permuting one label at a time, we can assess the importance of each label by dataset. We can then compare the distributions of the permuted labels with the non-permuted results.

In [19], each attribute was permuted at a time to measure the impact of variables in prediction, in terms of misclassification rate. The results in [19] indicated that some variables did not contribute to increase the predictive power of the method used, while others were very important. Similarly, in our approach to labels, we test if similar conclusions can be drawn, but in terms of relevance of the labels of rankings.

We would like to note that this process will never lead to a completely different ranking from the original, since only the relation of one label versus the others is affected per ranking. This is exactly what we intend here, in order to test the relevance of a label at a time. The process is exemplified in Figure 6.3, where the label a is permuted within the rankings.

6.5 Experiments

We use two LR algorithms, APRIORI-LR [36] and IB-PL [24]. The performance of the methods is estimated using a ten-fold cross-validation in terms of Kendall's τ. The data for APRIORI-LR was discretized with *equal width*

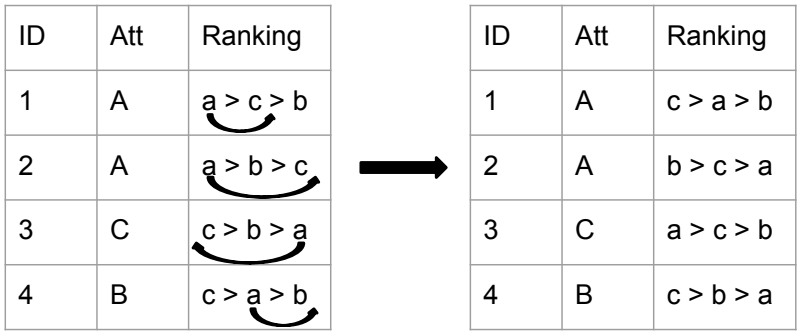

Figure 6.3: Illustration of a labelwise permutation of the label a.

discretization with 4 bins.

Table 6.1: Summary of the datasets.

Datasets	type	#examples	#labels	#attributes
bodyfat	B	252	7	7
calhousing	B	20,640	4	4
cpu-small	B	8,192	5	6
elevators	B	16,599	9	9
glass	A	214	6	9
housing	B	506	6	6
iris	A	150	3	4
segment	A	2310	7	18
stock	B	950	5	5
vehicle	A	846	4	18
vowel	A	528	11	10
wine	A	178	3	13
wisconsin	B	194	16	16

To compare the results, we used the *t.test* function from the *stats* package [113] with a confidence level of 95%. In Section 6.5.1 we use the standard *t-test* approach and in Section 6.5.2 a paired *t-test*. The p-values are mentioned below.

To check whether the mean accuracy in the original data is better or not, we use the following hypotheses:

H_0 The mean accuracy is equivalent in both original and permuted datasets.

H_1 The mean accuracy on the original datasets is better than the average accuracy on the permuted datasets.

If the p-value $< 5\%$, we reject H_0.

6.5.1 Ranking permutations

For each dataset, we performed 100 random permutations of the targets and measured the accuracy for APRIORI-LR and IB-PL. Then we compared with 10 repetitions on the original data. The distributions for IB-PL are shown in Figure 6.4.

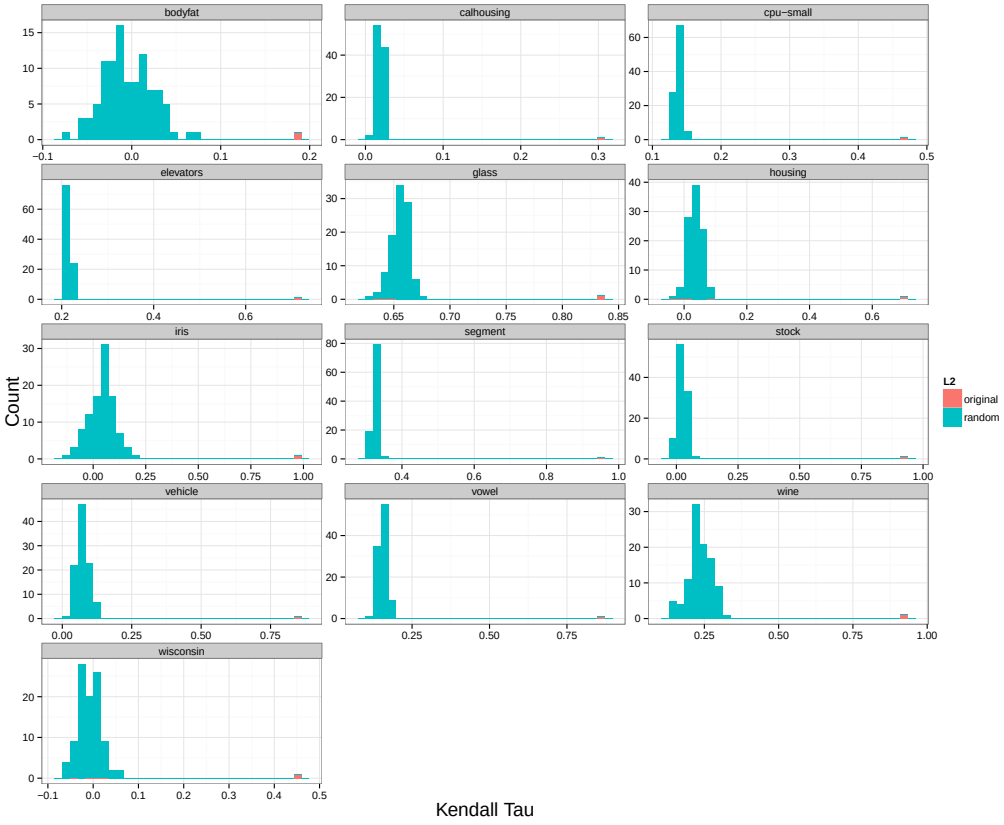

Figure 6.4: Distribution of the accuracy of IB-PL on randomized rankings (blue) and original data (red).

In Figure 6.4 it is clear that in most datasets the distribution of the accuracy on the permuted datasets is less than the accuracy with the original dataset. When the difference is big, it indicates that the algorithm is not being able to find relevant patterns in the randomized datasets. The statistical tests indicated that for all the cases, there is a significantly better mean accuracy

on the original datasets, with p-values $\ll 1\%$. Very identical results were obtained using the APRIORI-LR algorithm.

6.5.2 Labelwise permutations

In this part of the experiments, we permuted one label at a time with 10 repetitions. By comparing it to the 10 repetitions on the original data, we can statistically test whether the latter are better.

Even though we used two LR methods, if we can statistically show that at least one method yielded better results with the original data than with a label permuted, then we do not need to consider the other on that particular label.

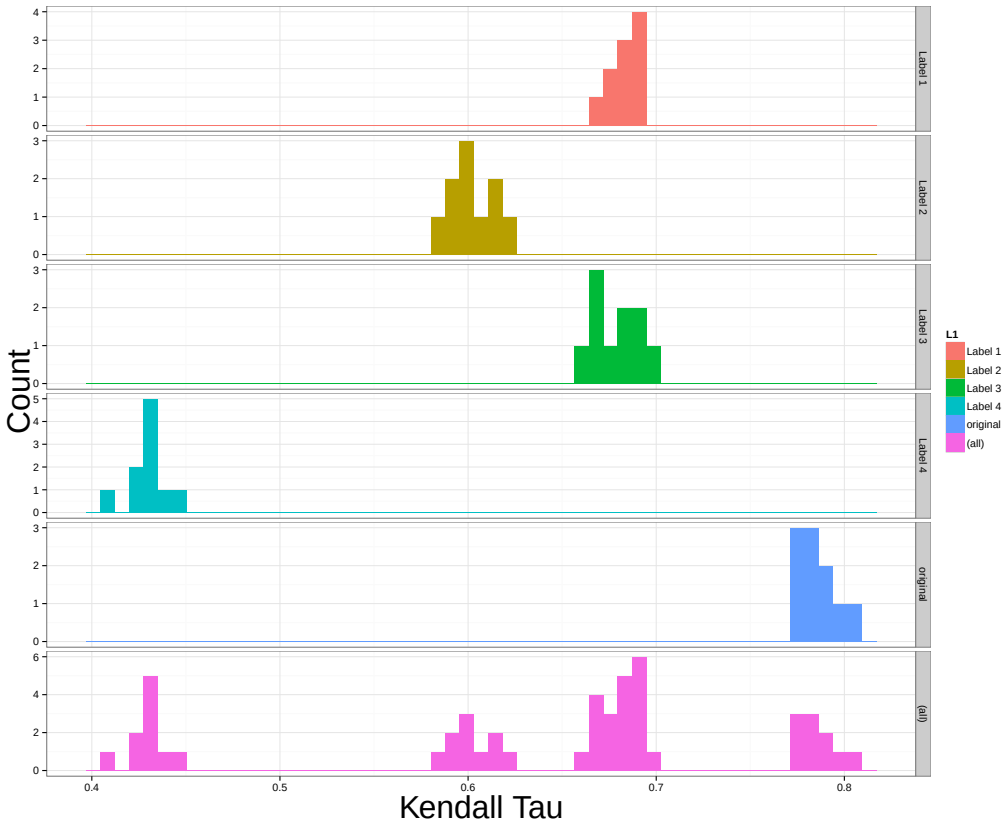

Figure 6.5: Distribution of the accuracy of APRIORI-LR on vehicle dataset per permuted label and original data.

In Figure 6.5 it is clear that the distribution of the accuracy with any label

permuted is less than the accuracy on the original dataset. Therefore it seems that there are no doubts about the importance of each label for the accuracy. Also, from Figure 6.5 it is clear how label 4, when randomized, affects the accuracy in a more extreme way than the remaining labels. Statistical tests confirm that with p-values $\ll 1\%$ using both APRIORI-LR and IB-PL.

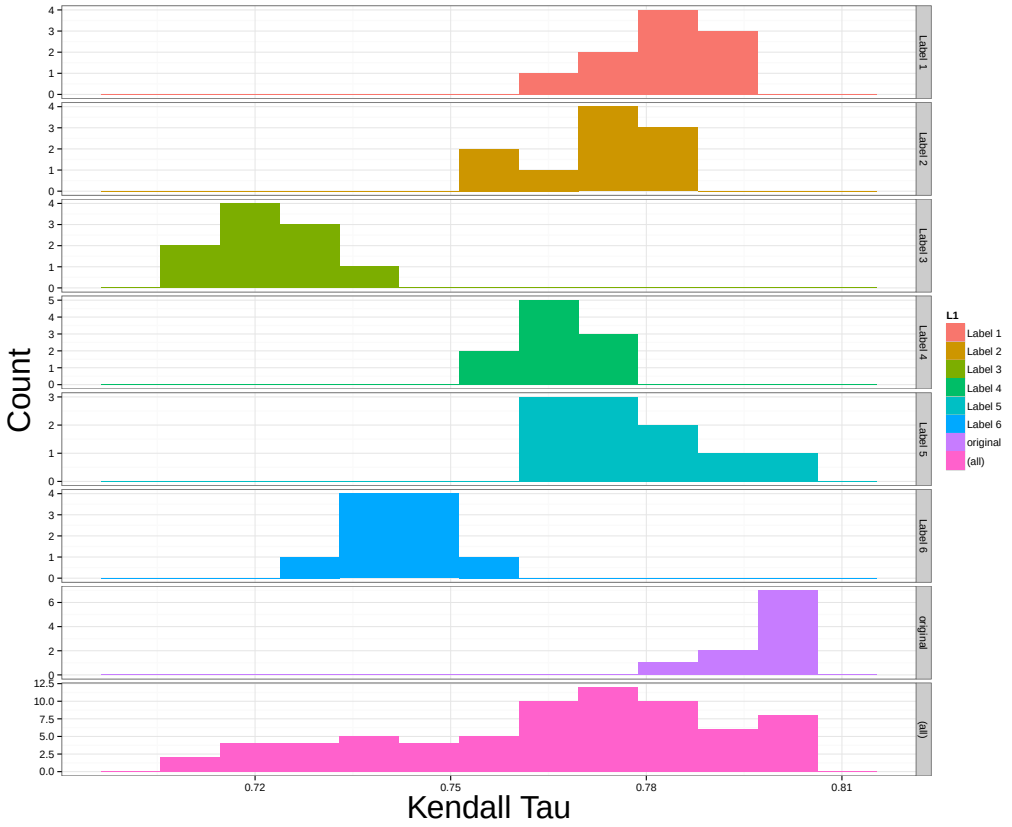

Figure 6.6: Distribution of the accuracy of APRIORI-LR on glass dataset per permuted labels and original data.

In the results obtained with the glass dataset, on Figure 6.6, the difference is not clear and the distribution of some randomized labels overlaps the distribution with the original data. However, statistical tests indicated that the distribution on the original data is significantly better.

On the other hand, it is also very interesting how labels 3 and 6 have a much higher impact on the accuracy for this model than the others. Similar to [19], we can suggest a level of relevance by label, using this approach.

Figure 6.7 gives the accuracy distribution of IB-PL per label permuted and with the original dataset. In this case, statistical tests obtained a p-value

$< 5\%$ for all the permuted labels except for label 6. Also APRIORI-LR failed to obtain a p-value $< 5\%$ for the same label.

This seems to indicate that label 6 does not have a very relevant relation with the other labels. This is somewhat expected from type B datasets rather than type A. Since in the former, some attributes were transformed into labels of a ranking, if these come from attributes that are not strongly related with the remaining, the label rank should also be measured as irrelevant.

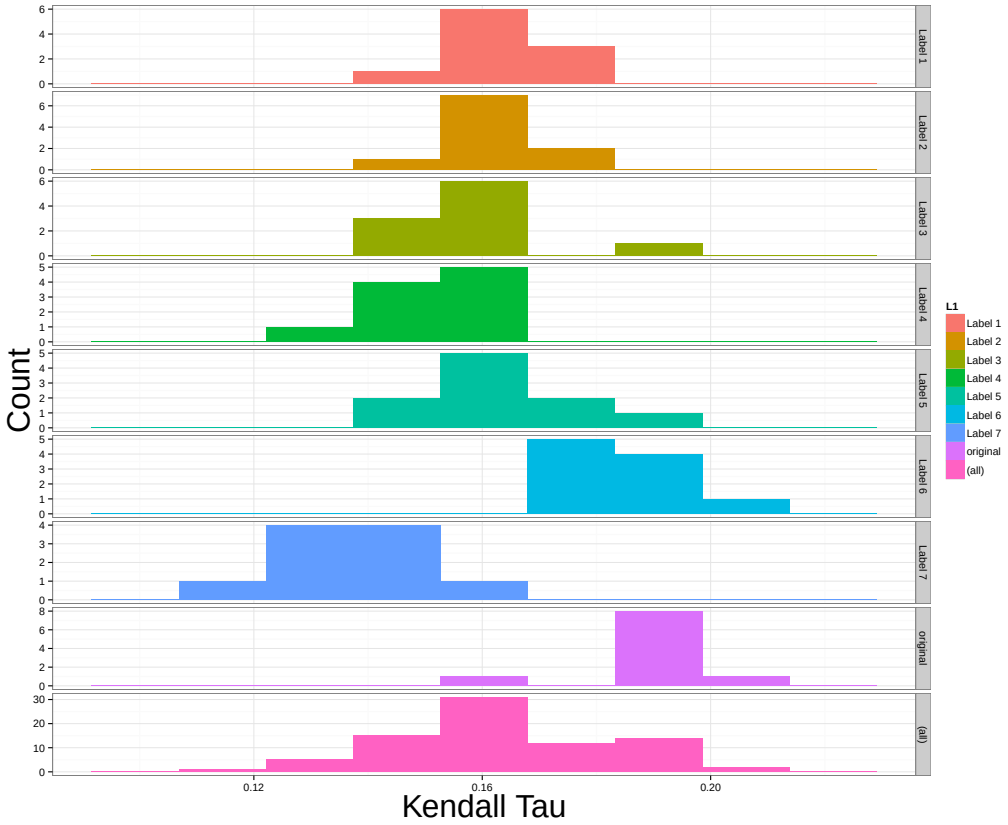

Figure 6.7: Distribution of the accuracy of IB-PL on the bodyfat dataset per permuted label and original data.

Due to space limitations we do not present results for all datasets, but instead we show the most relevant which are also representative of the others.

6.6 Conclusions

In this work, we show that, even though KEBI datasets have a semi-synthetic nature, they carry relevant preference information that can be learned by contemporary label rankers. In particular, there were no obvious differences between the type A and type B datasets. Statistical tests showed that the prediction models over this datasets are not due to chance.

This work also proposes a simple way to measure the relevance of each label on the prediction accuracy, based on the work of [19]. We also found out that some labels seem to affect the accuracy more than other, such as in the *glass* and *vehicle* dataset.

This methods can also be used on real world datasets too, in order to give a richer analysis. For example, by measuring the relative importance of each label or determining which algorithm is more resistant to *noise in rankings* [39]. In the future, we intend to propose a specific method to assess the relevance of ranking data with a proper statistical framework.