Multiple Hypothesis Testing in Pattern Discovery

Sami Hanhijärvi

Department of Information and Computer Science,
Aalto University, Finland
sami.hanhijarvi@aalto.fi

Abstract. The problem of multiple hypothesis testing arises when there are more than one hypothesis to be tested simultaneously for statistical significance. This is a very common situation in many data mining applications. For instance, assessing simultaneously the significance of all frequent itemsets of a single dataset entails a host of hypothesis, one for each itemset. A multiple hypothesis testing method is needed to control the number of false positives (Type I error). Our contribution in this paper is to extend the multiple hypothesis framework to be used in a generic data mining setting. We provide a method that provably controls the family-wise error rate (FWER, the probability of at least one false positive). We show the power of our solution on real data.

Keywords: multiple hypothesis testing, randomization, significance test, pattern mining.

1 Introduction

A plethora of data mining methods have been developed to find many different types of patterns with various criteria from a given dataset. The methods produce a collection of patterns for almost any dataset, even random ones. Basing future analysis and decisions on moot patterns is likely a futile effort that will waste time and money. Therefore, it is very important for the applicability of the results to identify the patterns that are a result of exceptional structure in the data.

Statistical significance testing provides theoretically founded framework for achieving this. It is based on a definition of a test statistic for a pattern, which can be perceived as a goodness measure of the pattern. A null hypothesis is then defined, which states how the test statistic is distributed if the original data were merely random. A significance testing method can be used to assess if the value of the test statistic in the observed dataset was drawn from this distribution, or was it drawn from some other unknown distribution.

The statistical significance testing problem is well understood when the hypotheses (patterns) to be tested are known in advance, before the data is observed, and the number of hypotheses is fixed (see [16]). However, this is not the case in data mining scenarios: before the data is observed, it is impossible to know which patterns the data mining algorithm will output from the set of all possible patterns. If instead hypotheses are assigned to each possible pattern, the large number of this set is likely to greatly reduce the power of the

T. Elomaa, J. Hollmén, and H. Mannila (Eds.): DS 2011, LNCS 6926, pp. 122–134, 2011.

significance test. If, on the other hand, the data is first mined and only the patterns that are output are tested for significance, the mining process needs to be accounted for in the significance testing. The process limits the possible values for the test statistic of a pattern, and therefore, may invalidate the assumptions about test statistic distribution in random data. Also, and more importantly, the varying number of patterns output for different datasets causes problems, that, if not handled correctly, may cause far too many patterns to be falsely declared significant.

A possible solution to overcome this problem consists of limiting the hypothesis space. For example, in frequent itemset mining, one could only consider all the itemsets of at most the given length [14]. If the search space can be trivially limited such that the portion of interesting patterns is relatively high, then these methods are expected to work well. However, if the limiting is not trivial or possible, such methods may fail to provide adequate results.

We propose a method that can be used for virtually any combination of data mining method, test statistic function, and a distribution of random data, as long as a very general assumption of subset pivotality holds. The method is based on drawing random samples of data sets, applying the data mining method to each random data set and comparing the values of the test statistic of the patterns found from random data to the test statistic values of the patterns in the original data. Therefore, we specifically take into consideration the data mining process and do not require to limit the search space in any way. The method builds on an existing one [16], which we extend to data mining settings and prove its validity.

2 Statistical Significance Testing in Data Mining

We consider the general case where we have a *data mining algorithm* A that, given an input dataset D, outputs a set of patterns P, or $A(D) = P$. The set P is a subset of a universe of all patterns $\mathcal{P}$. For different input datasets, the algorithm may output a different set of patterns, still from $\mathcal{P}$. We assume defined a *test statistic* $f(x, D) \in \mathbb{R}$, associated to an input pattern $x \in \mathcal{P}$ for the dataset D; small values of the statistic are assumed to be more interesting for the user. The choice of test statistic is arbitrary, but it should somehow express the goodness of a pattern to gain maximum power in the significance test.

The common definition for a statistical significance test in data mining [4,8,10,12,18] for a single pattern x is to test if the test statistic value of the pattern in the observed dataset D is an exceptionally small value among the test statistic values in random datasets. We assume defined a *null distribution* of datasets Π_0. The null hypothesis H_0^x for x states that the observed test statistic value $f(x, D)$ is not exceptionally small, *i.e.*, x is a *false pattern*. Conversely, the alternative hypothesis H_1^x states that $f(x, D)$ is exceptionally small, and therefore, x is a *true pattern*. The test for statistical significance is carried out by defining a p-value for the pattern x,

$$\begin{aligned}\pi_x(t) &= Pr_{D'}(f(x, D') \leq t)\\ p_x &= \pi_x(f(x, D))\,,\end{aligned} \tag{1}$$

where D' in the subscript of Pr means that D' is the random variable over which the probability is calculated. If not otherwise stated, a dataset D' in the subscript is sampled from Π_0.

A p-value represents the probability of observing equal or smaller test statistic value for x when H_0^x is true. If the p-value is very small, there is evidence to suggest that H_0^x is false, *i.e.*, x may be statistically significant and interesting. Conversely, if the p-value is large, there is no sufficient evidence to say that H_0^x is false, and therefore, x is not statistically significant. The p-value is either reported as is, or thresholded with a confidence threshold α, which defines the maximum accepted probability of falsely declaring the pattern significant. Declaring a false pattern significant, *i.e.*, rejecting the corresponding null hypothesis H_0^x when it is true, is called a false positive.

In data mining scenarios, the number of patterns, and therefore the number of null hypotheses to be tested, is often very large. This methodology can not be used for each pattern separately, since the number of false positives often increases when multiple null hypotheses are tested simultaneously (see [3,16] for a review and further references). As the basis of statistical significance test is to (most often) control the probability of falsely rejecting a null hypothesis, *multiple hypothesis testing* methods are developed to provide control for such probabilities in the presence of multiple hypotheses. Let V be the number of true null hypotheses that are falsely rejected, *i.e.*, false patterns that are falsely declared significant. One of the measures in multiple hypothesis testing is to control the Family-wise Error Rate (FWER), which is defined as the probability of falsely declaring any pattern significant, FWER$= Pr(V > 0)$. While there are other measures [2,3], we adopt this for its simplicity.

The multiple hypothesis testing methods are often defined in terms of *adjusted p-values*. The simplest and probably the best known multiple hypothesis testing method that controls the FWER is the Bonferroni test, that modifies the original, *unadjusted*, p-values and returns adjusted p-values given by

$$\tilde{p}_x^B = \min(1, m p_x)\,, \tag{2}$$

where m is the number of (null) hypotheses tested. A null hypothesis H_0^x is rejected if $\tilde{p}_x^B \leq \alpha$, and FWER is controlled at level α.

This method can be used if the hypotheses are known in advance. However, in data mining settings, this is not always the case as a data mining algorithm often outputs a different subset of patterns from the set of all possible patterns $\mathcal{P}$ for different datasets D. As all patterns in $\mathcal{P}$ have the possibility of being output by the algorithm, all of them should therefore be considered in statistical significance testing. However, assigning null hypotheses for all $x \in \mathcal{P}$ would reduce the power of the test as, in FWER, the probability of at least one false positive most likely increases as the number of hypotheses increases.

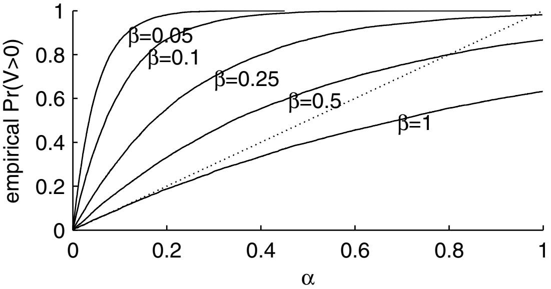


Fig. 1. Empirical probability of at least one false positive with respect to different confidence thresholds α. A subset of 1000 independent p-values are selected that are at most β, and these are then corrected with Bonferroni-method. The dotted line corresponds to $Pr(V > 0) = \alpha$, which should not be exceeded.

This problem has been discussed in the literature [1,10,15,13,14], and we call it the problem of varying set of hypotheses. If only the patterns that are output by the algorithm are tested for statistical significance, and the selection process is not taken into consideration, there is an elevated risk for false positives. This is illustrated by the following example.

Let the p-values be calculated for the patterns output by a data mining algorithm using Equation (1). Assume for clarity that the algorithm outputs patterns that have a p-value below some threshold β. This is close to a common scenario, where the test statistic is defined as the goodness measure of a pattern, and only patterns that have a small (or high) enough value in the measure are returned. Therefore, the algorithm only outputs patterns for which $p_x \leq \beta$, $A(D) = \{x \in \mathcal{P} | Pr_{D'}(f(x, D') \leq f(x, D)) \leq \beta\}$. Assume further that all patterns are independent and correspond to true null hypotheses (are null patterns). The p-values are adjusted with the Bonferroni-method in Equation 2, which is the most conservative of existing multiple hypothesis testing methods. Figure 1 illustrates the empirical probability of at least one false positive for different confidence thresholds α and p-value thresholds β. In the figure, the empirical FWER is larger than the accepted maximum α for many α and all $\beta < 1$. Therefore, FWER is not controlled even in the simple case where all patterns are null and independent, and hence, the common framework of the statistical tests can not be directly applied with existing multiple hypothesis correction methods in data mining scenarios.

3 Multiple Hypothesis Testing with Randomization

In this section, we introduce our contribution, which is to extend an existing resampling method [16] to the context of data mining and to prove the control of FWER.

Resampling methods are based on drawing samples of random datasets from a *null distribution* of datasets Π_0. We assume that we have at our disposal a randomization algorithm with which one can sample n datasets i.i.d. from Π_0. We denote the datasets sampled from the null distribution by D_i, $D_i \sim \Pi_0$, where $i \in [n]$, and $[n] = \{1, \ldots, n\}$.

The existing resampling based method [16] assumes the set of hypothesis is fixed, *i.e.,* $A(D) = \mathcal{P}$ for all $D \sim \Pi_0$, and a p-value is calculated with Equation (1) for each pattern for each random dataset. Therefore, it is assumed that no pattern is missing and that p_x can be easily calculated. The FWER-adjusted p-value for each pattern is then calculated as the fraction of random datasets for which the minimum p-value was equal or smaller than the p-value of the pattern x,

$$\tilde{p}_x^{Be} = \frac{|\{i \in [n+1] \,|\min_{y \in \mathcal{P}} \pi_y(f(y, D_i)) \leq \pi_x(f(x, D))\}|}{n+1}, \tag{3}$$

with $D_{n+1} = D$, *i.e.*, the original dataset is included in the calculation (see Section 3.2). The obtained adjusted p-values are essentially the empirical version of the Bonferroni-corrected p-values in Equation (2).

However, in data mining settings, the set of hypotheses is not fixed and the test statistics should not be restricted to p-values. We extend the method and calculate the adjusted p-values as follows.

Definition 1. *Let D be the original dataset, D_i, $i \in [n]$, be the datasets sampled from the null distribution Π_0 and $D_{n+1} = D$. Let $f(x, D_i)$ be the test statistic associated to an input pattern $x \in \mathcal{P}$ returned by algorithm A for dataset D_i. The FWER-adjusted p-values are defined as*

$$\tilde{p}_x = \frac{\left|\left\{i \in [n+1] \,\middle|\, (A(D_i) \neq \emptyset) \cap \left(\min_{y \in A(D_i)} f(y, D_i) \leq f(x, D)\right)\right\}\right|}{n+1}. \tag{4}$$

In other words, the FWER-adjusted p-value for a pattern is the fraction of random datasets that returned at least one pattern and any of the returned patterns had an equal or smaller test statistic value. The FWER-adjusted p-values provide a valid statistical significance test, in that they have the following property.

Theorem 1. *Given that subset pivotality holds, the null hypotheses H_x^0 of any pattern $x \in A(D)$ with $\tilde{p}_x \leq \alpha$ can be rejected with the certainty that FWER is controlled at level α.*

3.1 Proof of Theorem 1

Before we provide a proof for the theorem, we introduce two lemmas and the subset pivotality assumption. The first lemma shows that the expected and asymptotic forms of the adjusted p-values in Equation (4) are equal and correspond to the value of the cumulative distribution function of the minimum test statistic value at $f(x, D)$.

Lemma 1. *It holds for the adjusted p-values calculated using Equation (4) that*

$$\begin{aligned}\mathbb{E}[\tilde{p}_x] &= \lim_{n\to\infty} \tilde{p}_x \\ &= Pr_{D'\sim\Pi_0}\left((A(D') \neq \emptyset) \cap \left(\min_{y\in A(D')} f(y,D') \leq f(x,D)\right)\right) .\end{aligned}$$

The proof is a simple application of the law of large numbers with some algebra, and it is therefore omitted for brevity. Warranted by the properties in Lemma 1, we will ignore for the rest of the paper the sampling error due to the finite number of samples from the null distribution, that is, we assume that n is large enough.

The second lemma states a property of two identically distributed random variables.

Lemma 2. *For real valued random variables Y and X, that are distributed identically, and for any $q \in [0,1]$, $Pr_X(Pr_Y(Y \leq x) \leq q) = q$.*

The proof is again omitted for brevity.

Subset pivotality is an assumption about the dependency structure between the test statistics of the null patterns.

Definition 2. *(Subset pivotality) Let the original dataset be sampled from the unknown distribution Θ. Let $\mathcal{P}_0 \subseteq \mathcal{P}$ be the set of false patterns. The joint distribution of the test statistics of $x \in \mathcal{P}_0$ is identical between datasets sampled from Π_0 and datasets sampled from Θ.*

We require the assumption of subset pivotality, since we can not sample datasets from Θ, where the patterns correspond to both true and false null hypotheses. We can only sample from Π_0, where all null hypotheses are true. Therefore, as is common for all resampling methods, we assume that the distribution of test statistics for any subset of true null hypotheses is unaffected by the truth or falsehood of other null hypotheses. See [16] for discussion.

Proof. (of Theorem 1)

Assume the original dataset comes from the unknown distribution Θ. The set of all possible patterns $\mathcal{P}$ is divided to null patterns $\mathcal{P}_0$, which correspond to true null hypotheses, and to true patterns $\mathcal{P}_1$, for which the null hypotheses are false. The two sets are mutually exclusive and together cover $\mathcal{P}$, $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$ and $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$. The separation to $\mathcal{P}_0$ and $\mathcal{P}_1$ is unknown, since it is exactly what we are trying to find out. Let $\beta = Pr_{D\sim\Pi_0}(A(D) \neq \emptyset)$. First note that

$$\begin{aligned}&\min_{x\in A(D)} \tilde{p}_x \\ &= \min_{x\in A(D)} Pr_{D'}\left((A(D') \neq \emptyset) \cap \left(\min_{y\in A(D')} f(y,D') \leq f(x,D)\right)\right) \\ &= \min_{x\in A(D)} \beta Pr_{D'}\left(\min_{y\in A(D')} f(y,D') \leq f(x,D)\,|A(D') \neq \emptyset\right)\end{aligned}$$

$$
\begin{aligned}
&= \beta Pr_{D'}\left(\min_{y\in A(D')} f(y,D') \le \min_{x\in A(D)} f(x,D)\,|A(D') \neq \emptyset\right)\\
&= \beta\tau(D)\,,
\end{aligned}
$$

We assume $\alpha > \beta$ as otherwise FWER holds trivially. Assuming subset pivotality is satisfied, FWER is

$$
\begin{aligned}
&Pr(V>0)\\
&= Pr_{D\sim\Theta}\left((A(D)\cap\mathcal{P}_0 \neq \emptyset)\cap\left(\min_{x\in A(D)\cap\mathcal{P}_0} \tilde{p}_x \le \alpha\right)\right)\\
&= Pr_{D\sim\Pi_0}\left((A(D)\cap\mathcal{P}_0 \neq \emptyset)\cap\left(\min_{x\in A(D)\cap\mathcal{P}_0} \tilde{p}_x \le \alpha\right)\right)\\
&\le Pr_{D\sim\Pi_0}\left((A(D) \neq \emptyset)\cap\left(\min_{x\in A(D)} \tilde{p}_x \le \alpha\right)\right)\\
&= \beta Pr_D\left(\beta\tau(D) \le \alpha\,|A(D) \neq \emptyset\right)\\
&= \beta\frac{\alpha}{\beta} = \alpha\,,
\end{aligned}
$$

where we have used the subset pivotality between the second and third line, and Lemma 2 between the second to last and last line.

3.2 Empirical *p*-Values

Let us first remind that the original dataset is assumed to be drawn from Θ.

The definition in Equations (4) includes the original dataset in the calculations, following [11]. This may seem counter-intuitive, since we assume the dataset has been drawn from Θ but still consider it as to be drawn from Π_0. The reasons for doing this are two-fold: conservativeness and the assumption of true null hypothesis. When the original dataset is added as a random dataset, it is guaranteed that all p-values are strictly larger than 0. Otherwise we would claim that a pattern with a p-value of 0 could never be returned with random data, which we do not know. As for the assumption of true null hypothesis, it is initially assumed in hypothesis testing that each null hypothesis is true and that evidence is gathered that may result in rejecting the null hypothesis. Because of this, each pattern in the output with the original dataset is assumed to correspond to a true null hypothesis. Following this reasoning, we should use the test statistic of each of these patterns in the p-value calculations. A natural way of doing that is to include the original data in the calculation. For further discussion on empirical p-values, see [11].

3.3 Marginal Probabilities as Test Statistic

Any definition for a test statistic function can be used in the calculations in Equation (4). Furthermore, the identity of patterns need not be known in the equation, but only the smallest test statistic value. This makes the implementation of the method extremely simple, as it is sufficient to store only the smallest test statistic value for each random dataset. The indifference of identity is also beneficial in cases, where it is unreasonable to expect a pattern to be output by the algorithm for many of the random datasets.

Conversely, there are cases where the identity of a pattern is meaningful, such as frequent itemset mining. If the negative frequency of an itemset is used in Equation (4), as smaller values were considered more interesting, it is likely that small itemsets dominate the calculation because of the submodularity of itemsets. If a more equal comparison of patterns is sought, one can transform the original test statistics with an empirical variant of Equation (1)

$$f(x, D_j) = \frac{|\{i \in [n+1] \,|(x \in A(D_i)) \cap (g(x, D_i) \leq g(x, D_j))\}|}{n+1}, \tag{5}$$

where $g(x, D)$ is the original test statistic function. We call this the "threshold" transformation, since if a pattern is not output by the algorithm, it is assumed to have a larger test statistic value than with any dataset for which the pattern was output. In effect, the transformation acts as if the algorithm only outputs patterns with a test statistic value less than some threshold value, with a separate threshold for each pattern. This can be a reasonable transformation for example in frequent itemset mining if frequency is considered as a test statistic, since the mining algorithm truly outputs only patterns with a sufficiently good test statistic value, and any pattern that is not output for a dataset has a frequency less than the threshold in that dataset.

Another possibility is to define $f(x, D)$ as the marginal probability of the pattern having equal or smaller test statistic value with the condition that the pattern is output by the algorithm,

$$f(x, D_j) = \frac{|\{i \in [n+1] \,|(x \in A(D_i)) \cap (g(x, D_i) \leq g(x, D_j))\}|}{|\{i \in [n+1] \,|(x \in A(D_i))\}|}, \tag{6}$$

This transformation we call "conditional", since it is calculated using only the values that arise when the pattern is output by the algorithm. The interpretation is that in this case, the test statistic value of a missing pattern is unknown and nothing can be assumed about it. This can be reasonable in complex algorithms that do not have such a clear threshold as the frequency threshold of itemset mining. Notice that using the original test statistics or either of these transformations creates a different statistical test.

4 Related Work

Randomization has been studied in data mining scenarios [4,5,6,12,17] to assess the statistical significance of found results. These methods produce random versions of the original dataset while maintaining specific properties or structures in the dataset. These methods can be used to draw random datasets from the specified null distribution of datasets Π_0.

Multiple hypothesis testing has been studied in the context of data mining [1,8,9,10,15,13,14,18]. However, most of the methods are context dependent and can not be applied in general situations. These scenarios include assessing the significance of SQ-rules [18], association rules [10] and contrast sets [1]. Furthermore, some of the methods require to bootstrap the data [9], or to split it in half [15,13], both of which assume the data is a collection of samples from some distribution. If the data can not be bootstrapped or split in half, such as network or spatial data, theses methods can not be used.

The most comparable method is to use layered critical values [14]. In the method, the space of all possible patterns $\mathcal{P}$ is limited by stopping the data mining at a certain level. For example, in frequent itemset mining, only itemsets of length at most 8 are mined. This greatly reduces the size of $\mathcal{P}$ and hence increases the power of the multiple hypothesis test. While this method can be used in a variety of settings, it is still limited to level-wise searches and has power only if the interesting patterns are likely to be located on the lower levels.

5 Experiments

5.1 Frequent Itemsets

The first experiment was a common scenario in data mining, namely, mining frequent itemsets. The test statistic f was a variant of the lift:

$$f(x) = -\frac{\mathrm{freq}(x)}{\prod_{A \in x} \mathrm{freq}(A)}, \tag{7}$$

where x is an itemset, A is a single attribute of x, and $\mathrm{freq}(x) \in [0,1]$ is the relative frequency of itemset x. We used three different datasets: Courses, Paleo and Retail; all of which were used by [4].

Frequent itemsets were first mined from each dataset using the minimum support thresholds: 400, 7 and 200 for Courses, Paleo and Retail, respectively, and with minimum frequent set size of 2. We then used two different randomization methods [4]: Col, that randomizes the dataset while maintaining the column margins, and Swap, that additionally maintains the row margins. Each dataset was randomized 10000 times with both methods. Table 1 lists for different datasets basic properties of the datasets, the minimum support, number of frequent itemsets in the original data, and the mean and standard deviation of the number of frequent itemsets in the randomized datasets.

We calculated the number of patterns found significant for different controlled FWER levels α with the original f and the two transformations in Equations (5) and (6). We also used the layered critical values [14] when randomizing with Col, and calculated the unadjusted p-values using the frequency of an itemset with the binomial distribution with the product of the item frequencies as the success probability. This produces a similar null distribution for the frequencies of the itemsets to what is produced by the randomization. We set the maximum level to the actual maximum found from the data. This is a biased choice, as the level should be chosen before the data is seen. However, the amount of bias in these experiments are expected to be minimal. Notice that we could not use the

Table 1. Description of the datasets, mining parameters and statistics for frequent itemset mining. $|P|$ is the number of frequent itemsets with the original data, $|P_i^{\textsc{Col}}|$ the mean number of frequent itemsets with random data from Col, and $|P_i^{\textsc{Swap}}|$ the mean number of frequent itemsets with random data from Swap. Standard deviations are shown in parenthesis.

Dataset	# of rows	# of cols	# of 1's	density %	minsup	$\|P\|$	$\|P_i^{\textsc{Col}}\|$	$\|P_i^{\textsc{Swap}}\|$
Courses	2405	5021	65152	0.54	400	9678	146.4(2.8)	423.7(9.2)
Paleo	124	139	1978	11.48	7	2828	221.7(11.3)	266.8(14.9)
Retail	88162	16470	908576	0.06	200	1384	860.2(6.9)	1399.6(4.7)

layered critical values with Swap as no analytical distribution for the itemset frequencies is known from which the unadjusted p-values could be derived.

Figure 2 depicts the results. They show that the swap randomization is in general more restricted and, as expected, less patterns were found significant in comparison with Col. An interesting result can be seen with Retail and Swap randomization. A single pattern in the randomization always obtains an extremely small test statistic value and no original test statistic value is less than that value. Therefore, no pattern is found significant when the original test statistic is used. However, with transformed test statistics, some patterns could be found significant with as high a FWER value as 0.1. The layered critical values had equal power to the proposed method with the original and thresholded test statistic in Courses, and less power than with the original test statistic in Paleo and Retail.

5.2 Frequent Subgraphs

In the second experiment, we mined for frequent subgraphs from a collection of graphs. We used the FSG algorithm by [7], which is a part of Pafi[1], to mine subgraphs from a dataset of different compounds[2], which has 340 different graphs and the largest graph has 214 nodes. We mined the graph with minimum support level 40, and obtained 140 frequent subgraphs. We calculated the test statistic f for each subgraph x as

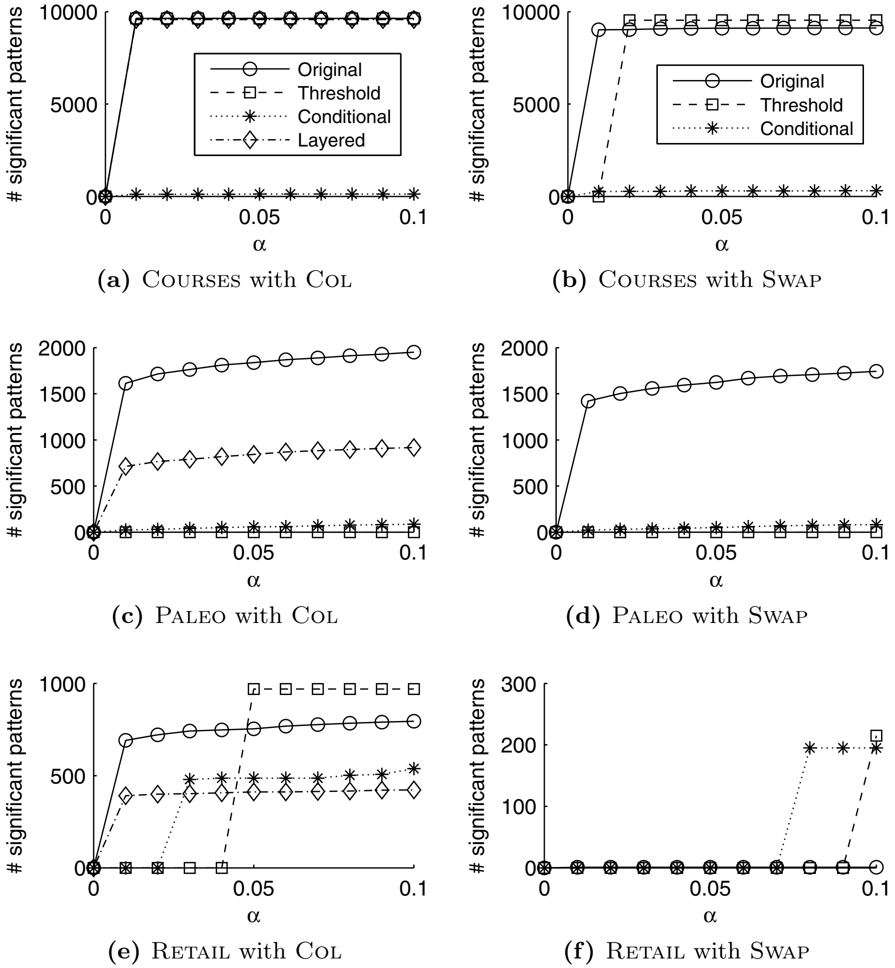
$$f(x) = -\mathrm{freq}(x)\log(\#\,\mathrm{nodes\,in}\,x)\,.$$

The logarithm term is to weight larger subgraphs slightly more, because they are considered more interesting than small ones.

We randomized the graphs preserving the node degrees while creating a completely different topology for the graph [5]. Since the dataset is a set of graphs, we randomized each graph individually by attempting 500 swaps, and combined the randomized graphs back to a transactional dataset. We used 10000 random datasets at support level 40, which resulted in mean number of subgraphs 191.7 with a standard deviation 13.4.

[1] `http://glaros.dtc.umn.edu/gkhome/pafi/overview`

[2] `http://www.doc.ic.ac.uk/~shm/Software/Datasets/carcinogenesis/progol/carcinogenesis.tar.Z`



(a) COURSES with COL

(b) COURSES with SWAP

(c) PALEO with COL

(d) PALEO with SWAP

(e) RETAIL with COL

(f) RETAIL with SWAP

Fig. 2. Frequent set mining results. The number of patterns found significant for different controlled FWER levels (α) for both randomization methods, the original and transformed test statistics, the layered critical values when applicable, and all datasets.

Figure 3 depicts the number of frequent subgraphs found statistically significant for different α levels with the original and transformed test statistics. Notice that the layered critical values were not be used, as a distribution for the subgraph frequency under the randomization is unknown. Again, some subgraphs had very small test statistic values in random data, and therefore, the adjusted p-values for the original frequent subgraphs were high. Conversely, if the test statistics are transformed, more subgraphs are found significant.

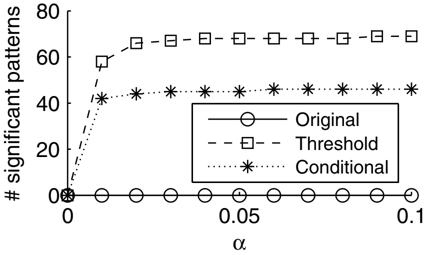


Fig. 3. Frequent subgraph mining results with Compound dataset. The lines illustrate the number of patterns found significant for different controlled FWER levels.

6 Discussion and Conclusions

As shown by the recent interest in randomization methods, there is a clear need for new significance testing methods in data mining applications. Especially within the framework of multiple hypothesis testing, the significance tests for data mining results have been lacking.

In this paper, we have extended an existing method to test the significance of patterns found by a generic data mining algorithm. The method is based on comparing the goodness of the patterns of the original data to the ones found when mining random data. The method works with any algorithm, test statistic and null distribution of datasets. And, unlike much of the previous work, we do make only a very general assumption about the combination of all of these three, and no assumptions about the data. Hence, our approach is suitable in many data mining scenarios where the significance test is based on a null distribution of datasets.

Acknowledgments. We thank Kai Puolamäki and Gemma C. Garriga for discussions and suggestions.

References

1. Bay, S.D., Pazzani, M.J.: Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery 5(3), 213–246 (2001)
2. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 57(1), 289–300 (1995)
3. Dudoit, S., Shaffer, J.P., Boldrick, J.C.: Multiple hypothesis testing in microarray experiments. Statistical Science 18(1), 71–103 (2003)
4. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. ACM Transactions on Knowledge Discovery from Data 1(3) (2007)
5. Hanhijärvi, S., Garriga, G.C., Puolamäki, K.: Randomization techniques for graphs. In: Proceedings of the Ninth SIAM International Conference on Data Mining, SDM 2009 (2009)

6. Hanhijärvi, S., Ojala, M., Vuokko, N., Puolamäki, K., Tatti, N., Mannila, H.: Tell me something i don't know: randomization strategies for iterative data mining. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 379–388. ACM, New York (2009)
7. Kuramochi, M., Karypis, G.: An efficient algorithm for discovering frequent subgraphs. IEEE Transactions on Knowledge and Data Engineering 16(9), 1038–1051 (2004)
8. Lallich, S., Teytaud, O., Prudhomme, E.: Association rule interestingness: measure and statistical validation. Quality Measures in Data Mining, 251–275 (2006)
9. Lallich, S., Teytaud, O., Prudhomme, E.: Statistical inference and data mining: false discoveries control. In: 17th COMPSTAT Symposium of the IASC, La Sapienza, Rome, pp. 325–336 (2006)
10. Megiddo, N., Srikant, R.: Discovering predictive association rules. In: Knowledge Discovery and Data Mining, pp. 274–278 (1998)
11. North, B.V., Curtis, D., Sham, P.C.: A note on the calculation of empirical P values from Monte Carlo procedures. The American Journal of Human Genetics 71(2), 439–441 (2002)
12. Ojala, M., Vuokko, N., Kallio, A., Haiminen, N., Mannila, H.: Assessing data analysis results on real-valued matrices. Statistical Analysis and Data Mining 2, 209–230 (2009)
13. Webb, G.: Discovering significant patterns. Machine Learning 68, 1–33 (2007)
14. Webb, G.: Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. Machine Learning 71, 307–323 (2008)
15. Webb, G.I.: Discovering significant rules. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2006, pp. 434–443. ACM, New York (2006)
16. Westfall, P.H., Young, S.S.: Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley, Chichester (1993)
17. Ying, X., Wu, X.: Graph generation with predescribed feature constraints. In: Proceedings of the Ninth SIAM International Conference on Data Mining, SDM 2009 (2009)
18. Zhang, H., Padmanabhan, B., Tuzhilin, A.: On the discovery of significant statistical quantitative rules. In: KDD 2004: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 374–383. ACM, New York (2004)