Maximum Entropy Based Significance of Itemsets

Nikolaj Tatti
HIIT Basic Research Unit, Department of Computer Science
Helsinki University of Technology, Helsinki, Finland

E-mail: ntatti@cc.hut.fi

Abstract

We consider the problem of defining the significance of an itemset. We say that the itemset is significant if we are surprised by its frequency when compared to the frequencies of its sub-itemsets. In other words, we estimate the frequency of the itemset from the frequencies of its sub-itemsets and compute the deviation between the real value and the estimate. For the estimation we use Maximum Entropy and for measuring the deviation we use Kullback-Leibler divergence.

A major advantage compared to the previous methods is that we are able to use richer models whereas the previous approaches only measure the deviation from the independence model.

We show that our measure of significance goes to zero for derivable itemsets and that we can use the rank as a statistical test. Our empirical results demonstrate that for our real datasets the independence assumption is too strong but applying more flexible models leads to good results.

1 Introduction

How significant is a given itemset? Itemsets are popular and well-studied patterns in binary data mining. The major drawback is that, given a dataset, there are exponential number of itemsets. Hence, we need to rank itemsets in order to prune the uninteresting ones.

Traditionally, the frequency of an itemset is used as a rank measure. The higher the frequency, the more significant is the itemset. Frequency has many virtues: It is easy to interpret and because of its property of anti-monotonicity there exist efficient algorithms for finding all frequent itemsets [2, 3]. There are, however, major drawbacks. First, a frequent itemset may be insignificant: An itemset AB may be frequent just because itemsets A and B are frequent. Second, an infrequent itemset may be significant: If itemsets A and B are frequent, the infrequency of AB is interesting information.

Alternative methods for ranking itemsets are suggested in [1, 6, 12]. These methods are discussed in more detail in Section 4. A common feature to these methods is that they compare the frequency of an itemset to an estimate obtained from the independence model. That is, the more the itemset deviates from the independence model, the more surprising, and thus the more significant, the itemset is.

Our proposal for ranking itemsets resembles the aforementioned approaches. We estimate the frequency of a given itemset from the frequencies of some selected sub-itemsets. Namely, we use Maximum Entropy for the estimation. This approach is more flexible than the independence model, since the independence model uses only the margins (the frequencies of itemsets of size 1) for prediction whereas our approach allows to use the information available from the itemsets of larger size. While our ranking method is based on well-known tools, no similar framework has been suggested previously.

Unlike the frequency, our measure is not decreasing with respect to set inclusion. Hence we cannot mine significant itemsets in a level-wise fashion. However, it turns out that in some cases we can prune a large set of uninteresting itemsets (w.r.t. the measure). Namely, if the itemset is derivable [7], then the measure is equal to 0. We also point out that can be used as a statistical test, thus providing a clear interpretation for the measure.

The rest of the paper is organized as follows: Preliminaries are given in Section 2. The definition and the properties of the measure are given in Section 3. We present related work in Section 4. Section 5 is devoted to experiments and finally we provide conclusions in Section 6.

1550-4786/07 $25.00 © 2007 IEEE
DOI 10.1109/ICDM.2007.43

2 Preliminaries and Notation

In this section we review briefly theory of itemsets and also introduce some notation that will be used later on.

A *binary dataset* D is a collection of M binary vectors, *transactions*, having length K. Such dataset can be naturally represented as a matrix of size $M \times K$. We denote the number of transactions by $|D| = M$. To each column of the matrix we assign an *attribute* a_i. Let $A = \{a_1, \ldots, a_K\}$ be the collection of all attributes. An itemset $X \subseteq A$ is a set of attributes.

We say that a transaction (binary vector) ω *covers* an itemset X if $a_i \in X$ implies $\omega_i = 1$. Given a dataset D, a *frequency* of an itemset X is a proportion of the transaction in D covering X. Note that if an itemset Y is a subset of X, then the frequency of Y is larger than or equal to the frequency of X. In other words, frequency is decreasing with respect to set inclusion.

A sample space Ω is the set of all binary vectors of length K. We take a simplistic approach in defining distributions: A distribution $p : \Omega \to [0, 1]$ is a function from a sample space Ω to a real number between 0 and 1 such that $\sum_{\omega \in \Omega} p(\omega) = 1$. Given an itemset X, a frequency of X calculated from a distribution p is the probability of binary vector covering X. We denote this by

$$p(X = 1) = p(\omega \text{ covers } X).$$

A family of itemsets $\mathcal{F}$ is called *anti-monotonic* or *downward closed* if every subset of each member of $\mathcal{F}$ is also a member of $\mathcal{F}$. Note that a collection of σ-frequent itemsets, that is, itemsets having frequency larger than some given threshold σ, is downward closed. We are interested in three particular families:

- $\mathcal{I}$, the family containing only itemsets of size 1.
- $\mathcal{C}$, the family containing itemsets of size 1 and 2.
- $\mathcal{A}$, the family containing all itemsets.

Given a dataset D, we say that an itemset X is *derivable* if by knowing the frequencies (calculated from D) of each proper subset of X we can deduce the frequency of X. For example, if some subset of X has a frequency 0, then we know that X must also have frequency 0. Thus, in this case, X is derivable. An itemset that is not derivable is called *non-derivable.* A family of all non-derivable itemsets is downward closed [7].

3 Maximum Entropy Ranking

In this section we introduce our ranking method and discuss its theoretical properties. The fundamental idea behind our approach is to measure how surprising an itemset is compared to its subsets. In other words, we estimate the itemset frequency by using the frequencies of its subsets and compare how close is our estimation to the actual value. The estimation is done using Maximum Entropy method and the comparison is done using Kullback-Leibler divergence.

3.1 Definition

Let D be a binary dataset and let $\{a_1, \ldots, a_K\}$ be its attributes. The number of columns in D is K. Assume that we are given G, an itemset we wish to rank. We define a projected dataset D_G by keeping only the attributes included in G.

Let $\Omega_G = \{0, 1\}^{|G|}$ be a space of binary vectors of length $|G|$. We define an *empirical distribution* $q_G : \Omega_G \to [0, 1]$ to be

$$q_G(\omega) = \frac{\text{Number of samples in } D_G \text{ equal to } \omega}{|D_G|}.$$

Our goal is to compare the distribution q_G to a distribution obtained by using Maximum Entropy [20], a method that we will describe next.

Assume now that we are given a family of itemsets $\mathcal{F} \subseteq \mathcal{A}$ and let θ_X be the frequency of $X \in \mathcal{F}$ calculated from D. Our next step is to define an approximative distribution using only the itemsets in $\mathcal{F}$. In defining q_G we projected out the attributes outside G. Similarly, we are only interested in subsets of G. Hence we define a *projected family* $\mathcal{F}_G$ to be

$$\mathcal{F}_G = \{X \in \mathcal{F} \mid X \subset G, X \neq G, X \neq \emptyset\}.$$

Note that $\mathcal{F}_G$ may contain $2^{|G|} - 2$ itemsets, at maximum. This is the case if $\mathcal{F} = \mathcal{A}$.

We say that a distribution $p : \Omega_G \to [0, 1]$ *satisfies the itemsets* $\mathcal{F}_G$ if for each itemset $X \in \mathcal{F}_G$ and its frequency θ_X we have

$$p(X = 1) = \theta_X.$$

Let $\mathbb{P}$ be the set of all distributions satisfying the itemsets $\mathcal{F}_G$. This set is not empty since $q_G \in \mathbb{P}$. We select the distribution from $\mathbb{P}$ maximizing the entropy

$$-\sum_{\omega \in \Omega_G} p(\omega) \log p(\omega).$$

We denote this distribution by p^{ME}. Note that p^{ME} depends on G, $\mathcal{F}$, and θ but we have omitted these variables from the notation for the sake of clarity.

We define the rank measure $r\left(G; \mathcal{F}, D\right)$ to be the divergence between q_G and p^{ME}, that is,

$$r\left(G; \mathcal{F}, D\right) = \sum_{\omega \in \Omega_G} q_G(\omega) \log \frac{q_G(\omega)}{p^{ME}(\omega)}.$$

We omit D from the notation when the dataset is clear from the context.

Example 1. *Assume the simplest case where $G = a$ is an itemset of size 1. Let θ_G be the frequency of G. Note that $\mathcal{F}_G = \emptyset$, hence there are no constraints on selecting p^{ME}. This means that p^{ME} is the uniform distribution, that is, $p^{ME}(0) = p^{ME}(1) = 1/2$. In this case the measure is*

$$r\left(a; \mathcal{F}\right) = (1 - \theta_G)\log\left(2(1 - \theta_G)\right) + \theta_G \log\left(2\theta_G\right)$$

obtaining its minimum when $\theta_G = 1/2$ and is at its maximum when $\theta_G = 0$ or $\theta_G = 1$.

We are mainly interested in three kinds of measures: The first is $r\left(G; \mathcal{I}\right)$ in which $\mathcal{I}$ is the family of itemsets of size 1. In this case the Maximum Entropy distribution is equal to the independence model.

The second case is $r\left(G; \mathcal{C}\right)$, where $\mathcal{C}$ contains the itemsets of size 1 and 2. We can show that there exists a matrix B such that for the non-zero entries of p^{ME} we have

$$p^{ME}(\omega) \propto \exp\left(\omega^T B \omega\right).$$

Hence, $r\left(G; \mathcal{C}\right)$ can be seen as the measure of the deviation from the discrete Gaussian model.

Our third type of measure is $r\left(G; \mathcal{A}\right)$ in which p^{ME} is predicted from all the proper sub-itemsets of G. In this case we can prove that for a certain set of real numbers r_i we have for the non-zero entries of p^{ME}

$$p^{ME}(\omega) \propto \prod_{X_i \in \mathcal{A}_G} \exp\left(r_i I\left(\omega \text{ covers } X_i\right)\right),$$

where I is the indicator function. We discuss the evaluation of our approach in Section 3.3.

3.2 Properties

In this section we discuss various properties of $r\left(G\right)$. We will first point the connection between $r\left(G\right)$ and derivable itemsets and then discuss the use of $r\left(G\right)$ as a statistical test.

Theorem 2. *Let G be a derivable itemset. Then*

$$r\left(G; \mathcal{A}\right) = 0.$$

Proof. We can argue that if we know the frequencies of all sub-itemsets of G, we can derive the distribution q_G and vice versa. This implies that there is one-to-one correspondence between the distribution $p \in \mathbb{P}$ satisfying the itemsets $\mathcal{A}_G$ and the frequency $p(G = 1)$. Since we can derive the frequency of G from $\mathcal{A}_G$, it follows that $\mathbb{P} = \{q_G\}$, and hence $p^{ME} = q_G$. □

We can reformulate the previous theorem in a stronger form by pointing out that we need to know only non-derivable itemsets.

Theorem 3. *Let $\mathcal{F}$ be a family of all non-derivable itemsets. Let G be outside of $\mathcal{F}$. Then $r\left(G; \mathcal{F}\right) = 0$.*

Proof. Since all unknown sub-itemsets of G are derivable from $\mathcal{F}_G$, the argument of Theorem 2 holds. □

The following theorem provides the interpretation to the value of $r\left(G\right)$ and points out that we can use $r\left(G\right)$ as a statistical test.

Theorem 4. *Let G be a non-derivable itemset. Under the 0-hypothesis that G is distributed according to p^{ME}, the quantity $2\left|D\right| r\left(G; \mathcal{A}\right)$ is distributed asymptotically as χ^2 with degree 1 of freedom.*

Theorem 4 is a special case of the following more general statement.

Theorem 5. *Let G be a non-derivable itemset and let $\mathcal{F}$ be an itemset family. Define $\mathcal{H}$ to be*

$$\mathcal{H} = \left\{X \in \mathcal{A} \mid X \subseteq G, X \neq \emptyset, X \notin \mathcal{F}_G\right\},$$

that is, $\mathcal{H}$ is a family of sub-itemsets of G not belonging to $\mathcal{F}_G$. Under the 0-hypothesis that the itemsets in $\mathcal{H}$ are distributed according to p^{ME}, the quantity $2\left|D\right| r\left(G; \mathcal{F}\right)$ is distributed asymptotically as χ^2 with degree $\left|\mathcal{H}\right| = 2^{|G|} - 1 - \left|\mathcal{F}_G\right|$ of freedom.

Theorem 5 is stated (but not proven) in a more general form in [20]. A rather technical proof is provided in Appendix A.

3.3 Computing Rank

Evaluating the measure requires computing p^{ME} distribution and comparing it to the empirical distribution. Both distributions have $\left|\Omega_G\right| = 2^{|G|}$ entries. Solving p^{ME} distribution can be done using Iterative Scaling algorithm [10, 18]. The algorithm consists of consecutive steps. One such step requires $O\left(\left|\Omega_G\right|\right) = O\left(2^{|G|}\right)$ time. Hence computing the measure requires exponential time but it is doable for itemsets of reasonable size.

Note that in defining the measure we only use itemsets that are subsets of the query itemset G. This pruning guarantees that the number of entries in the distributions is $2^{|G|}$ and not, at worst, 2^K, where K is the number of columns in the dataset. We can show that in the general case solving p^{ME} is an **NP**-complete problem [25, 8]; hence pruning attributes is essential.

4 Related Work

Our work resembles approach of [6] in which the authors defined the significance of an itemset by comparing the distribution q_G against the independence model. The authors used χ^2 statistical test as a measure, that is, if p is the distribution related to the independence model, the rank measure is

$$r_b\left(G\right) = \sum_{\omega \in \Omega_G} \frac{\left(q_G(\omega) - p(\omega)\right)^2}{p(\omega)}. \qquad (1)$$

In [12] the authors also compare the frequency of an itemset against the independence model but in addition they use Bayes screening to smooth the values. Also, in [1] the authors proposed the collective strength as a measure of significance. To be more specific, we say that a transaction $\omega \in \Omega_G$ is *good* if it contains only 0s or only 1s. Let p be the distribution related to the independence model. Then the measure is

$$r_{cs}\left(G\right) = \frac{q_G\left(\omega \text{ is good}\right)}{p\left(\omega \text{ is good}\right)} \frac{p\left(\omega \text{ is bad}\right)}{q_G\left(\omega \text{ is bad}\right)}. \qquad (2)$$

This measure obtains small values when data obeys the independence model. In a related work presented in [11] the authors define an itemset to be interesting if its frequency increases significantly from one dataset to another. In [15] the authors order itemsets based on their p-values. In [17] the authors used entropy of tree models for ranking itemsets.

The authors in [24] showed empirically that Maximum Entropy model provides excellent estimates for itemsets. Rank can be used for pruning a large family of itemsets by picking the itemsets having the largest rank. Other pruning methods are proposed in [4, 7, 23]. The authors in [27] suggest a generic framework for discovering significant rules. In addition, a relevant framework is described in [21]; the authors define a pattern ordering given an estimation algorithm and a loss function. In [22] the authors use information component analysis to find patterns in a drug safety database.

5 Experiments

In this section we present our empirical results. In the first 3 sections we explain the datasets and the setup. In our experiments we investigate the significance of itemsets, how different measures are related to each other, and the monotonicity of the ranks.

5.1 Synthetic Datasets

For the testing purposes we created two synthetic datasets. Each dataset contained 100 attributes and 5000 rows. The first dataset, *gen-ind*, was generated such that the attributes were independent. The margins were sampled uniformly from $[0, 1]$. In the second dataset, *gen-copy*, each column was a copy of the previous column corrupted by the symmetric white noise. The amount of noise, that is the probability

$$p\left(a_i = 1 \mid a_{i-1} = 0\right) = p\left(a_i = 0 \mid a_{i-1} = 1\right),$$

was selected uniformly from $[0, 1]$ for each column a_i, individually. The first column was generated by a coin flip. Our expectations are that in *gen-ind* the itemsets of size 1 are significant and that in *gen-copy* the itemsets of size 2 are significant.

5.2 Real Datasets

In our experiments we used the following real-world datasets. Data in *Accidents*[1] were obtained from the Belgian "Analysis Form for Traffic Accidents" forms that is filled out by a police officer for each traffic accident that occurs with injured or deadly wounded casualties on a public road in Belgium. In total, 340 183 traffic accident records are included in the dataset [16]. The datasets *POS*[2], *WebView-1*[3] and *WebView-2*[4] were contributed by Blue Martini Software as the KDD Cup 2000 data [19]. *POS* contains several years worth of point-of-sale data from a large electronics retailer. *WebView-1* and *WebView-2* contain several months worth of click-stream data from two e-commerce web sites. *Kosarak*[5] consists of (anonymized) click-stream data of a Hungarian on-line news portal. *Retail*[6] is a retail market basket data supplied by an anonymous Belgian retail supermarket store [5]. The dataset *Paleo*[7] contains information of species fossils found in specific paleontological sites in Europe [13], preprocessed as in [14].

5.3 Setup for the Experiments

In this section we will describe how we conducted our experiments. We reduced the largest datasets by selecting the first 10000 rows and 200 most frequent attributes. From each dataset we computed all *almost non-derivable* itemsets. By almost non-derivable we mean that the difference between the upper bound and

[1] `http://fimi.cs.helsinki.fi/data/accidents.dat.gz`

[2] `http://www.ecn.purdue.edu/KDDCUP/data/BMS-POS.dat.gz`

[3] `http://www.ecn.purdue.edu/KDDCUP/data/BMS-WebView-1.dat.gz`

[4] `http://www.ecn.purdue.edu/KDDCUP/data/BMS-WebView-2.dat.gz`

[5] `http://fimi.cs.helsinki.fi/data/kosarak.dat.gz`

[6] `http://fimi.cs.helsinki.fi/data/retail.dat.gz`

[7] NOW public release 030717 available from [13].

the lower bound of a given itemset, say G, is at least n transactions. In other words, if we know the frequencies of all sub-itemsets of G, then we cannot predict the frequency of G within n transactions. If $n = 0$, then an itemset is non-derivable. It is known that the family of almost non-derivable itemsets is anti-monotonic [7, Lemma 3.1]. A reason to use almost non-derivable itemsets instead of frequent itemsets is the statement of Theorem 3, that is, $r(G; \mathcal{A}) = 0$ if the itemset is derivable. The other reason is that we want to study how the measure behaves for infrequent itemsets.

To keep the sizes of the obtained families within reasonable bounds we used different thresholds for different datasets: For *gen-ind*, *Retail* and *WebView-2* we set $n = 5$. For *POS* the threshold n was set to 10 and for *gen-copy* and *Accidents* n was set to 100. For the rest of the datasets we set $n = 0$, that is, we mined all non-derivable itemsets from these datasets.

For each itemset from the obtained itemsets we queried the following measures:

- Frequency.
- Rank measures $r(G; \mathcal{I})$, $r(G; \mathcal{C})$, $r(G; \mathcal{A})$. We normalized these measures by applying Theorem 5.
- Measures discussed in Section 4: A χ^2 test $r_b(G)$ defined in Eq. 1 and a collective strength $r_{cs}(G)$ defined in Eq. 2.

The evaluation times and the sizes of the query families are given in Table 1.

Data	n	# of queries	max $\lvert G \rvert$	Time
gen-ind	5	156699	6	414s
gen-copy	100	111487	4	29s
Accidents	100	354399	6	316s
Kosarak	5	223734	5	8s
Paleo	0	166903	5	23s
POS	10	246640	6	20s
Retail	0	818813	6	32s
WebView-1	5	226313	5	7s
WebView-2	0	715398	6	58s

Table 1. The evaluation times and the sizes of the query families. The second column is the threshold used in mining almost non-derivable itemsets. The fourth column is the maximal size of a query itemset. The evaluation time includes the calculation of all measures but not the mining process itself.

5.4 Significant Itemsets

Our first experiment is to study how many of the itemsets are significant. We did this by comparing the P-values of our measures with risk level 0.05. The results are given in Tables 2–4. We also provide a typical example of box plots in Figure 1.

	itemset size						
Data	1	2	3	4	5	6	All
gen-ind	.92	.05	.04	.03	.02	.01	.03
gen-copy	.08	.14	.24	.03	–	–	.07
Accidents	.99	.60	.95	1	1	1	.97
Kosarak	1	.62	.99	1	1	–	.96
Paleo	1	.30	.81	.99	1	–	.88
POS	1	.45	.99	1	1	1	.95
Retail	1	.14	.30	.93	1	1	.45
WebView-1	1	.70	1	1	1	–	.97
WebView-2	1	.20	.69	1	1	1	.85

Table 2. The percentages of significant itemsets according to $r(G; \mathcal{I})$. Each entry is a fraction of itemsets of specific size calculated from a specific dataset. Significance is measured using χ^2 distribution with 0.05 risk level.

	itemset size						
Data	1	2	3	4	5	6	All
gen-ind	.92	.05	.06	.05	.04	.03	.05
gen-copy	.08	.14	.06	.03	–	–	.03
Accidents	.99	.60	.21	.45	.62	.60	.45
Kosarak	1	.62	.32	.50	.38	–	.37
Paleo	1	.30	.12	.15	.21	–	.15
POS	1	.45	.09	.21	.43	.66	.17
Retail	1	.14	.04	.08	.12	.38	.05
WebView-1	1	.70	.48	.32	.52	–	.48
WebView-2	1	.20	.11	.20	.88	1	.17

Table 3. The percentages of significant itemsets according to $r(G; \mathcal{C})$. Each entry is a fraction of itemsets of specific size calculated from a specific dataset. Significance is measured using χ^2 distribution with 0.05 risk level.

Let us first study *gen-ind*, a synthetic dataset with independent columns. We see from Table 2 that according to $r(G; \mathcal{I})$ a large portion of itemsets of size

	itemset size						
Data	1	2	3	4	5	6	All
gen-ind	.92	.05	.06	.06	.06	.07	.06
gen-copy	.08	.14	.06	.05	–	–	.05
Accidents	.99	.60	.21	.07	.06	.11	.12
Kosarak	1	.62	.32	.10	.06	–	.33
Paleo	1	.30	.12	.21	.64	–	.18
POS	1	.45	.09	.06	.08	.41	.11
Retail	1	.14	.04	.49	.61	.75	.15
WebView-1	1	.70	.48	.10	.26	–	.45
WebView-2	1	.20	.11	.55	.79	1	.36

Table 4. The percentages of significant itemsets according to $r(G;\mathcal{A})$. Each entry is a fraction of itemsets of specific size calculated from a specific dataset. Significance is measured using χ^2 distribution with 0.05 risk level.

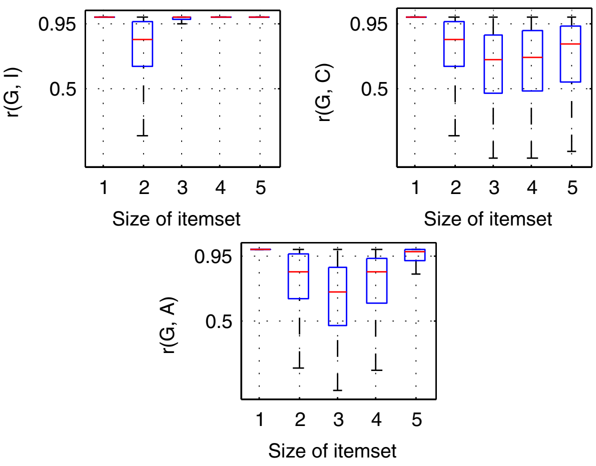

Figure 1. Box plots of the rank measures computed from *Paleo*.

1 are significant but only a small portion of itemsets having size larger than 1 is significant. This is an expected result since the frequencies obey the independence model. In Tables 3 and 4 we have similar results for $r(G;\mathcal{C})$ and for $r(G;\mathcal{A})$. However, the values of $r(G;\mathcal{C})$ and for $r(G;\mathcal{A})$ tend to be larger than the values of $r(G;\mathcal{I})$. The reason for this is a type of overlearning: Since the frequencies of itemsets are calculated from the datasets, they are imprecise. Hence, the itemsets with larger size mislead us during prediction, because the resulting Maximum Entropy distribution is not an independent model (although close to one).

Let us continue by studying *gen-copy*, a synthetic data in which an attribute is a noisy copy of the previous attribute. We see that $r(G;\mathcal{C})$ tends to have smaller ranks than $r(G;\mathcal{I})$ when G has size 3. The reason for this is that, unlike with *gen-ind*, the independence model cannot explain the dataset. However, when we predict using also the itemsets of size 2, the prediction becomes more accurate.

We turn our attention to real datasets. We see that for these datasets the independence model is too strict: According to $r(G;\mathcal{I})$ almost all itemsets are significant: The results change drastically, when we use richer models. According to $r(G;\mathcal{C})$ or $r(G;\mathcal{A})$ only 5%–50% of the itemsets are significant, depending on the dataset. Similar overfitting that occurred with *gen-ind* also occurs in some but not all real datasets (see Figure 1). For instance, in *Retail* $r(G;\mathcal{A})$ tends to produce higher values than $r(G;\mathcal{C})$ but not in *POS*.

5.5 The Effect of the Known Itemsets

We continued our experiments by comparing the measures $r(G;\mathcal{I})$, $r(G;\mathcal{C})$, and $r(G;\mathcal{A})$ against each other. This was done by calculating the correlations between the rank measures. The results are given in Table 5.

Data	$r(G;\mathcal{I})$ vs. $r(G;\mathcal{C})$	$r(G;\mathcal{I})$ vs. $r(G;\mathcal{A})$	$r(G;\mathcal{C})$ vs. $r(G;\mathcal{A})$
gen-ind	0.74	0.26	0.40
gen-copy	0.52	0.28	0.53
Accidents	0.17	0.09	0.43
Kosarak	0.15	0.14	0.92
Paleo	0.16	0.22	0.67
POS	0.12	0.10	0.77
Retail	0.62	0.68	0.86
WebView-1	0.15	0.15	0.92
WebView-2	0.43	0.58	0.61

Table 5. Correlations between the measures $r(G;\mathcal{I})$, $r(G;\mathcal{C})$, and $r(G;\mathcal{A})$.

From the results we see that all correlations are positive. For the real datasets the correlations between $r(G;\mathcal{C})$ and $r(G;\mathcal{A})$ are systematically higher than the correlations between $r(G;\mathcal{I})$ and $r(G;\mathcal{A})$ or between $r(G;\mathcal{C})$ and $r(G;\mathcal{A})$. This suggests that $r(G;\mathcal{I})$ produces different ranks whereas $r(G;\mathcal{C})$ and $r(G;\mathcal{A})$ are more similar. This supports the behavior we have seen in Section 5.4.

5.6 Rank vs. Other Methods

We compared our measures against the other ranking methods described in Section 5.3. Namely, we calculated the correlations of $r\left(G;\mathcal{I}\right)$, $r\left(G;\mathcal{C}\right)$, and $r\left(G;\mathcal{A}\right)$ against the frequency of G, $r_b\left(G\right)$, the χ^2 test for independency, and $r_{cs}\left(G\right)$, the collective strength of the itemset G. The results are presented in Table 6. We also studied the relationships by plotting our measures as functions of the aforementioned approaches and such examples are given in Figure 2.

Our first observation is that $r\left(G;\mathcal{I}\right)$ correlates strongly with $r_b\left(G\right)$. This is an expected result since both test the independency of attributes inside the itemsets and also because $r\left(G;\mathcal{I}\right)$ is asymptotically a χ^2 test (see Theorem 5). There is some correlation between $r_b\left(G\right)$ and $r\left(G;\mathcal{C}\right)$ and $r\left(G;\mathcal{A}\right)$ although this correlation is much weaker compared to $r\left(G;\mathcal{I}\right)$.

Apart from *WebView-2*, there is little correlation between the measures and the frequency.

The correlation between the measures and the collective strength $r_{cs}\left(G\right)$ exists but varies depending on the method and the dataset. The strongest correlations are obtained when $r_{cs}\left(G\right)$ is compared against $r\left(G;\mathcal{I}\right)$. This is a natural result since $r_{cs}\left(G\right)$ produces small values when attributes are independent.

5.7 Monotonicity of Rank

In this section we investigate the relationship between the rank of an itemset and the ranks of its sub-itemsets. Namely, we tested whether the measures are monotonic, that is, whether $r\left(G;\mathcal{F}\right) \geq r\left(H;\mathcal{F}\right)$ for all $H \subset G$. We deliberately ignored sub-itemsets having size 1 since they all have very high rank. We also tested whether the measures are anti-monotonic, that is, decreasing w.r.t. set inclusion.

From the results given in Tables 7–8 our first observation is that $r\left(G;\mathcal{I}\right)$ are increasing for real datasets but not for the synthetic datasets. The raw values of $r\left(G;\mathcal{I}\right)$ are indeed increasing but this does not hold for the P-values since the number of degrees varies.

On the contrary, $r\left(G;\mathcal{C}\right)$ and $r\left(G;\mathcal{A}\right)$ are increasing for extremely few itemsets. Table 8 suggests that $r\left(G;\mathcal{C}\right)$ and $r\left(G;\mathcal{A}\right)$ satisfies the anti-monotonicity to some degree. Measures $r\left(G;\mathcal{C}\right)$ and $r\left(G;\mathcal{A}\right)$ are anti-monotonic for relatively high percentage of itemsets of size 3. Among itemsets of size 4, $r\left(G;\mathcal{A}\right)$ satisfies the property of anti-monotonicity for a slightly larger portion of itemsets than $r\left(G;\mathcal{C}\right)$.

6 Conclusions

We have given a definition of a measure for ranking itemsets. The idea is to predict the frequency of an itemset from the frequencies of its sub-itemsets and measure the deviation between the actual frequency and the prediction. The more the itemset deviates from the prediction, the more it is significant. We estimated the frequencies using Maximum Entropy and we used Kullback-Leibler divergence to measure the deviation. The measure can be computed in $O(2^{|G|})$ time, where $|G|$ is the size of the itemset needed to be ranked.

A clear advantage of our approach to the previous methods is that the previous solutions calculate the deviation from the independence model whereas we are able to use the information available from the itemsets of larger size, and thus use more flexible models.

Our empirical results for real data show that the independence is too strict assumption: Almost all itemsets were significant according to $r\left(G;\mathcal{I}\right)$. The results changed when we applied the more flexible models, $r\left(G;\mathcal{C}\right)$ and $r\left(G;\mathcal{A}\right)$. We also observed an interesting type of overfitting: In some cases we obtain a better prediction if we do not use all the available information.

We showed that there is a little correlation between our measures and the other approaches. For instance, infrequent itemset may be significant and frequent itemset may be insignificant. We also observed that $r\left(G;\mathcal{I}\right)$ is monotonic for a large portion of itemsets, whereas $r\left(G;\mathcal{C}\right)$ and $r\left(G;\mathcal{A}\right)$ are anti-monotonic for a significant portion of itemsets.

Acknowledgments

The author wishes to thank Gemma Garriga, Heikki Mannila, and Robert Gwadera for their comments.

References

[1] C. C. Aggarwal and P. S. Yu. A new framework for itemset generation. In *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 18–24. ACM Press, 1998.

[2] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

	$r(G;\mathcal{I})$ vs.			$r(G;\mathcal{C})$ vs.			$r(G;\mathcal{A})$ vs.		
Data	freq.	$r_b(G)$	$r_{cs}(G)$	freq.	$r_b(G)$	$r_{cs}(G)$	freq.	$r_b(G)$	$r_{cs}(G)$
gen-ind	0.06	0.99	−0.01	0.03	0.72	−0.01	0	0.25	−0.01
gen-copy	0.15	1	0.02	0.07	0.52	0.02	0	0.27	0.01
Accidents	0.01	1	0.02	−0.01	0.17	0.05	0.04	0.08	0.01
Kosarak	0.01	0.98	0.20	0.01	0.15	0.28	0	0.13	0.23
Paleo	0.18	0.95	0.39	0.01	0.15	0.10	−0.03	0.20	0.03
POS	0.05	0.99	0.22	0.09	0.13	0.20	0.07	0.10	0.01
Retail	0.04	0.97	0.31	0.05	0.56	0.17	0.05	0.62	0.28
WebView-1	0.06	0.98	0.19	0.07	0.16	−0.28	0.06	0.15	−0.30
WebView-2	0.12	0.96	0.33	0.17	0.36	0.39	0.15	0.49	0.35

Table 6. Correlations between the ranking methods $r(G;\mathcal{I})$, $r(G;\mathcal{C})$, and $r(G;\mathcal{A})$ and the base measures: the frequency of G, $r_b(G)$, the χ^2 test for independency, and $r_{cs}(G)$, the collective strength of the itemset G.

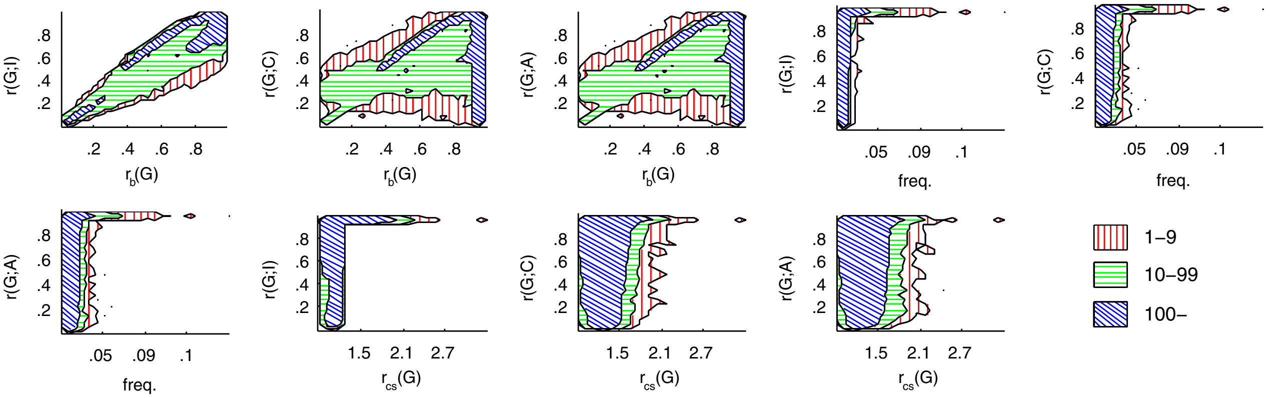

Figure 2. Ranks as functions of the base measures. The plots are calculated from *Paleo* dataset.

	$r(G;\mathcal{I})$					$r(G;\mathcal{C})$					$r(G;\mathcal{A})$				
Data	3	4	5	6	All	3	4	5	6	All	3	4	5	6	All
gen-ind	.09	.02	.01	0	.03	.27	.03	.01	0	.05	.27	.11	.06	.03	.11
gen-copy	.15	.01	–	–	.04	.20	.02	–	–	.05	.20	.10	–	–	.12
Accidents	.78	.92	.97	.99	.90	.01	.02	.02	0	.02	.01	0	0	0	0
Kosarak	.93	.98	1	–	.93	0	0	0	–	0	0	0	0	–	0
Paleo	.40	.61	.84	–	.51	.04	0	0	–	.02	.04	0	0	–	.02
POS	.87	1	1	1	.92	0	0	.01	0	0	0	0	0	0	0
Retail	.11	.42	.92	1	.19	.04	0	0	0	.03	.04	.11	0	0	.06
WebView-1	.98	1	1	–	.98	.04	0	0	–	.04	.04	0	0	–	.04
WebView-2	.39	.88	1	1	.67	.04	0	.09	1	.02	.04	.02	0	0	.03

Table 7. Percentages of itemsets satisfying the property of monotonicity. The itemset G satisfies the property if $r(G;\mathcal{F}) \geq r(H;\mathcal{F})$ for all $H \subset G$ such that $|H| \geq 2$.

[3] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances*

	$r\left(G;\mathcal{I}\right)$					$r\left(G;\mathcal{C}\right)$					$r\left(G;\mathcal{A}\right)$				
Data	3	4	5	6	All	3	4	5	6	All	3	4	5	6	All
gen-ind	.21	.07	.03	.02	.07	.25	.08	.03	.01	.08	.25	.07	0	0	.06
gen-copy	.15	.06	–	–	.08	.25	.08	–	–	.11	.25	.07	–	–	.10
Accidents	.03	0	0	0	.01	.62	.05	.01	0	.16	.62	.20	0	0	.24
Kosarak	.02	.06	0	–	.02	.93	.03	.01	–	.83	.93	.29	0	–	.86
Paleo	.02	0	0	–	.01	.43	.04	0	–	.22	.43	.07	0	–	.23
POS	.01	.01	.04	.23	.01	.87	.09	.01	0	.58	.87	.21	0	0	.62
Retail	.17	0	0	0	.13	.38	.05	.01	0	.30	.38	.01	0	0	.29
WebView-1	0	0	0	–	0	.69	.12	0	–	.62	.69	.24	0	–	.64
WebView-2	.07	.01	.14	.96	.04	.48	.06	0	0	.25	.48	.03	0	0	.23

Table 8. Percentages of itemsets satisfying the property of anti-monotonicity. The itemset G satisfies the property if $r\left(G;\mathcal{F}\right) \leq r\left(H;\mathcal{F}\right)$ for all $H \subset G$ such that $|H| \geq 2$.

in Knowledge Discovery and Data Mining, pages 307–328. AAAI Press/The MIT Press, 1996.

[4] J-F. Boulicaut, A. Bykowski, and C. Rigotti. Approximation of frequency queries by means of free-sets. In *Principles of Data Mining and Knowledge Discovery*, pages 75–85, 2000.

[5] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets. Using association rules for product assortment decisions: A case study. In *Knowledge Discovery and Data Mining*, pages 254–260. ACM, 1999.

[6] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In Joan Peckham, editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, pages 265–276. ACM Press, May 1997.

[7] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2002.

[8] G. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2–3):393–405, Mar. 1990.

[9] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, Feb. 1975.

[10] J. Darroch and D. Ratchli. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[11] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Knowledge Discovery and Data Mining*, pages 43–52, 1999.

[12] W. DuMouchel and D. Pregibon. Empirical bayes screening for multi-item associations. In *Knowledge Discovery and Data Mining*, pages 67–76, 2001.

[13] M. Fortelius. Neogene of the old world database of fossil mammals (NOW). University of Helsinki, `http://www.helsinki.fi/science/now/`, 2005.

[14] M. Fortelius, A. Gionis, J. Jernvall, and H. Mannila. Spectral ordering and biochronology of european fossil mammals. paleobiology. *Paleobiology*, 32(2):206–214, 2006.

[15] A. Gallo, T. De Bie, and N. Christianini. Mini: Mining informative non-redundant itemsets. In *11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 438–445, 2007.

[16] K. Geurts, G. Wets, T. Brijs, and K. Vanhoof. Profiling high frequency accident locations using association rules. In *Proceedings of the 82nd Annual Transportation Research Board, Washington DC. (USA), January 12-16*, 2003.

[17] Hannes Heikinheimo, Eino Hinkkanen, Heikki Mannila, Taneli Mielikäinen, and Jouni K. Seppänen. Finding low-entropy sets and trees from binary data. In *Knowledge Discovery and Data Mining*, 2007.

[18] R. Jiroušek and S. Přeušil. On the effective implementation of the iterative proportional fitting procedure. *Computational Statistics and Data Analysis*, 19:177–189, 1995.

[19] R. Kohavi, C. Brodley, B. Frasca, L. Mason, and Z. Zheng. KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explorations*, 2(2):86–98, 2000.

[20] S. Kullback. *Information Theory and Statistics*. Dover Publications, Inc., 1968.

[21] H. Mannila and T. Mielikäinen. The pattern ordering problem. In *Principles of Data Mining and Knowledge Discovery*, pages 327–338, 2003.

[22] G. N. Norén, A. Bate, and I. R. Edwards. Extending the methods used to screen the who drug safety database towards analysis of complex associations and improved accuracy for rare events. *Statistics in Medicine*, 25:3740–3757, 2007.

[23] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. *Lecture Notes in Computer Science*, 1540:398–416, 1999.

[24] D. Pavlov, H. Mannila, and P. Smyth. Beyond independence: Probabilistic models for query approximation on binary transaction data. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1409–1421, 2003.

[25] N. Tatti. Computational complexity of queries based on itemsets. *Information Processing Letters*, pages 183–187, June 2006.

[26] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[27] G. I. Webb. Discovering significant rules. In *Knowledge discovery and data mining*, pages 434–443, 2006.

A Asymptotic Behaviour of the Divergence

By asymptotic behaviour we mean the following: We assume that we have an ensemble of datasets D_i such that $|D_i| \to \infty$. We assume that G is non-derivable in each D_i and that the frequencies of $\mathcal{F}_G$ are all equal.

Define $N = |D|$ and $M = |\mathcal{H}|$. Let $\mathbb{P}$ be the set of distributions satisfying the itemsets $\mathcal{F}_G$. It is easy to see that we can parameterize $\mathbb{P}$ with frequencies of $\mathcal{H}$. In other words, let $\mathcal{H} = \{H_1, \ldots, H_M\}$. Then for each $p \in \mathbb{P}$, there is a unique frequency vector $\theta \in \mathbb{R}^M$ such that $\theta_i = p(H_i = 1)$. Let Θ be the set of all possible frequency vectors. The set Θ is a closed polytope — the vectors located on the boundary of Θ corresponds to the distributions in which at least one entry is 0.

Let θ^{ME} be a frequency vector corresponding to the Maximum Entropy distribution p^{ME}. We need to show that θ^{ME} is not a boundary vector. Assume the converse, then p^{ME} must have $p^{ME}(\omega) = 0$ for some ω. We know that this implies that $p(\omega) = 0$ for all $p \in \mathbb{P}$ [9, Theorem 3.1]. Let Y be the itemset containing the elements for which ω has positive entries. This in turns (see [7]) implies that for each $p \in \mathbb{P}$

$$p(G = 1) = \sum_{Y \subseteq Z \subsetneq G} (-1)^{|G|-|Z|} p(Z = 1),$$

making G derivable and contradicting the statement.

Since θ^{ME} is an inner point of Θ, let $B \subset \Theta$ be an open ball around θ^{ME}. Assume that $\theta \in B$. By taking the expectation of the second-degree Taylor expansion of $\log \frac{p(\omega;\theta^{ME})}{p(\omega;\theta)}$ around θ we arrive to

$$-\mathrm{KL}\left(\theta \| \theta^{ME}\right) = \frac{1}{2} \Delta\theta^T \mathrm{E}_\theta \left[H(\omega; \eta)\right] \Delta\theta,$$

where $\Delta\theta = \theta^{ME} - \theta$ and η is a vector lying between θ and θ^{ME}, and H is the Hessian matrix of $\log p(\omega; \eta)$.

Let θ_N be the frequencies of $\mathcal{H}$ obtained from a dataset containing N points. According to 0-hypothesis we have $\theta_N \rightsquigarrow \theta^{ME}$ and $\sqrt{N}\left(\theta_N - \theta^{ME}\right) \rightsquigarrow N(0, \Sigma)$, where Σ is a covariance matrix,

$$\begin{aligned} \Sigma_{ij} = & p^{ME}(H_i = 1, H_j = 1) \\ & - p^{ME}(H_i = 1) p^{ME}(H_j = 1). \end{aligned}$$

If $\theta_N \in B$, we let η_N correspond to η in the Taylor expansion, otherwise we set $\eta_N = 0$. We can show that $\eta_N \rightsquigarrow \theta^{ME}$ [26, Theorem 2.7]. Consider a function

$$g(a, b, c, d) = \left\{ \begin{array}{ll} -a^T \mathrm{E}_c \left[H(\omega; b)\right] a, & c \in B \\ (2/d)\, \mathrm{KL}\left(c \| \theta^{ME}\right), & c \notin B \end{array} \right. .$$

This function is continuous in $\left(\mathbb{R}^M, \theta^{ME}, \theta^{ME}, 0\right)$. Hence, we can apply continuous map theory [26, Theorem 2.3] to obtain that

$$\begin{aligned} 2N\mathrm{KL}\left(\theta_N \| \theta^{ME}\right) &= g\left(\sqrt{N}\left(\theta_N - \theta^{ME}\right), \eta_N, \theta_N, \frac{1}{N}\right) \\ &\rightsquigarrow -X^T \mathrm{E}_{\theta^{ME}}\left[H\left(\omega; \theta^{ME}\right)\right] X, \end{aligned}$$

where X is a random variable distributed as $N(0, \Sigma)$. We know that $\mathrm{E}_{\theta^{ME}}\left[H\left(\omega; \theta^{ME}\right)\right] = -\Sigma^{-1}$ [20, Lemma 4.11]. Theorem follows since $X^T \Sigma^{-1} X$ is distributed as χ^2 with M degrees of freedom [26, Lemma 17.1].