Balancing the Analysis of Frequent Patterns

Arnaud Giacometti, Dominique H. Li, and Arnaud Soulet

Université François Rabelais Tours, LI EA 6300
3 Place Jean Jaurès, F-41029 Blois, France
firstname.lastname@univ-tours.fr

Abstract. A main challenge in pattern mining is to focus the discovery on high-quality patterns. One popular solution is to compute a numerical score on how well each discovered pattern describes the data. The best rating patterns are then the most analyzed by the data expert. In this paper, we evaluate the quality of discovered patterns by anticipating of how user analyzes them. We show that the examination of frequent patterns with the notion of support led to an unbalanced analysis of the dataset. Certain transactions are indeed completely ignored. Hence, we propose the notion of balanced support that weights the transactions to let each of them receive user specified attention. We also develop an algorithm ABSOLUTE for calculating these weights leading to evaluate the quality of patterns. Our experiments on frequent itemsets validate its effectiveness and show the relevance of the balanced support.

Keywords: Pattern mining, stochastic model, interestingness measure.

1 Introduction

For twenty years, the pattern mining algorithms have gained performance and now arrive to quickly extract patterns from large amounts of data. However, evaluate and ensure the quality of extracted patterns remains a very open issue. In general, a pattern is considered to be relevant if it deviates from what was expected from a knowledge model. In the literature, there are two broad categories of knowledge models [1,2,3]: user-driven and data-driven ones. *User-driven* approaches discover interesting patterns with subjective models based on user oriented information, such as domain knowledge, beliefs or preferences. *Data-driven* approaches discover interesting patterns with objective models based on the statistical properties applied to data, such as frequency of patterns. Most often these methods neglect how the user will analyze the collection of patterns. In this paper, we present a novel approach, named *analysis-driven*, to evaluate discovered patterns by foreseeing how the collection will be analyzed.

Before presenting in depth our motivations, we first recall the context of frequent itemset mining [4]. Let $\mathcal{I}$ be a set of distinct literals called *items*, an itemset corresponds to a non-null subset of $\mathcal{I}$. A transactional dataset is a multi-set of itemsets, named *transactions*. Table 1 (a) presents such a dataset $\mathcal{D}$ where 4 transactions $t_1, \ldots, t_4$ are described by 4 items $A, \ldots, D$. The *support* of an

V.S. Tseng et al. (Eds.): PAKDD 2014, Part I, LNAI 8443, pp. 53–64, 2014.

itemset X in a dataset $\mathcal{D}$ is the fraction of transactions containing X. An itemset is *frequent* if it appears in a database with support no less than a user specified threshold. For instance, Table 1 (b) shows the frequent itemsets of $\mathcal{D}$ (with 0.25 as minimal support) that are ranked according to the support.

Table 1. An unbalanced analysis of a dataset by frequent patterns

(a) Dataset $\mathcal{D}$

TID	Items
t_1	AB
t_2	AC
t_3	BC
t_4	D

(b) The mined pattern set P

PID	Pattern	Support
p_1	A	0.50
p_2	B	0.50
p_3	C	0.50
p_4	D	0.25
p_5	AB	0.25
p_6	AC	0.25
p_7	BC	0.25

(c) Analysis resulting from the scoring

		Reality	Expected
TID	Length	Proportion Π	Preference ρ
t_1	1.25	0.3125	> 0.25
t_2	1.25	0.3125	> 0.25
t_3	1.25	0.3125	> 0.25
t_4	0.25	0.0625	< 0.25
Sum:	4	1	1

In a post-analysis process based on the scoring of patterns, we assume that an analyst examines each pattern with a diligence proportional to an interestingness measure (here, the support) for analyzing the dataset (we will justify our proposal of analysis model in Section 3). Hence, it is not difficult to see that the attention paid to A is twice that for AB since their support is respectively 0.5 and 0.25. We also assume that the analysis proportion of each transaction is proportional to the sum of the time spent on each pattern covering it. The transaction t_1 covered by A, B and AB is then analyzed for a length of 1.25 ($= 0.5 + 0.5 + 0.25$). Therefore, according to this observation, the individual analysis of t_1, t_2, or t_3 in the analysis is 5 times greater than that of t_4 as shown by the third column of Table 1 (c). Assuming that the user considers all transactions equally interesting (see the preference vector ρ in the fourth colmun), we say that such an analysis is *unbalanced*. It means that some transactions are understudied while others are overstudied! For us, a good way to balance the analysis is to increase the score (here the support) of the pattern D. Its score increase will also increase the analysis proportion of t_4 and decrease that of the other three transactions. More interestingly, we think that D is peculiar and deserves a higher valuation because it describes a singular transaction that is described by no other pattern.

In this work, we seek to balance the analysis of the dataset by proposing a new interestingness measure to rate the patterns. More precisely, a preference vector $\rho : \mathcal{D} \to (0, 1]$ indicates the intensity of his/her interest for each transaction such that $\sum_{t \in \mathcal{D}} \rho(t) = 1$. Indeed, the user does not always devote the same interest in all transactions as in our example above. Sometimes he/she prefers to focus on rare data as it is often the case in fraud detection or with biomedical data. An analysis is balanced when each transaction t is studied with an acuity $\Pi(t, \mathcal{M})$ corresponding to that specified by a user-preference vector $\rho(t)$: $\Pi(t, \mathcal{M}) = \rho(t)$ where $\Pi(t, \mathcal{M})$ is the analysis proportion of the transaction t given an analysis model $\mathcal{M}$ that simulates analysis sessions conducted to understand the data. To the best of our knowledge, we propose the first model, called *scoring*

analysis model, to simulate the sessions of analyzing pattern sets according to an interestingness measure. Its strength is to rely on a stochastic model successfully used in Information Retrieval [5,6] while integrating the preferences given by the user. As main contribution, we then introduce the *balanced support* that induces balanced analysis under our model. This measure removes the equalized axiom of support saying that every transaction has the same weight in its calculation. It gives a higher weight to the most singular transactions in the calculation of the balanced support. For instance, in Table 1, the transaction t_4 will be weighted 5 times more than others so that it receives the same attention as other transactions. We also develop an algorithm, ABSOLUTE, to compute the balanced support. Our experiments show its effectiveness for balancing frequent itemsets and compare the balanced support with the traditional support.

The rest of this paper is organized as follows. Section 2 introduces the basic notations. In Section 3, we introduce the scoring analysis model allowing us to simulate the behavior of an analyst. Under this model, we propose in Section 4 the balanced support and the algorithm ABSOLUTE to compute it. Section 5 presents experiments demonstrating its efficiency and the interest of the balanced support. Section 6 reviews some related work. We conclude in Section 7.

2 Preliminaries

For the sake of clarity, we illustrate our definitions with the notion of itemsets but, our problem is not limited to a particular type of pattern. We consider a *language* $\mathcal{L}$ and a dataset $\mathcal{D}$ that is a multiset of $\mathcal{L}$ (or another language). A *specialization relation* $\preceq$ is a partial order relation on $\mathcal{L}$ [7]. Given a specialization relation $\preceq$ on $\mathcal{L}$, $l \preceq l'$ means that l is more general than l', and l' is more specific than l. For instance, A is more general than AB w.r.t $\subseteq$.

Given two posets $(\mathcal{L}_1, \preceq_1)$ and $(\mathcal{L}_2, \preceq_2)$, a binary relation $\lhd \subseteq \mathcal{L}_1 \times \mathcal{L}_2$ is a *cover relation* iff for any $l_1 \lhd l_2$, we have $l'_1 \lhd l_2$ (resp. $l_1 \lhd l'_2$) for any pattern $l'_1 \preceq_1 l_1$ (resp. $l_2 \preceq_2 l'_2$). The relation $l_1 \lhd l_2$ means that l_1 covers l_2, and l_2 is covered by l_1. The cover relation is useful to relate different languages together (e.g., for linking patterns to data). Note that a specialization relation on $\mathcal{L}$ is also a cover relation on $\mathcal{L} \times \mathcal{L}$. For instance, the set inclusion is used for determinating which patterns of P cover a transaction of $\mathcal{D}$. Given two pattern sets $L \subseteq \mathcal{L}$, $L' \subseteq \mathcal{L}'$ and a cover relation $\lhd \subseteq \mathcal{L} \times \mathcal{L}'$, the *covered patterns* of L' by $l \in L$ is the set of patterns of L' covered by the pattern l: $L'_{\rhd l} = \{l' \in L' | l \lhd l'\}$. Dually, the *covering patterns* of L for $l' \in L'$ is the set of patterns of L covering the pattern l': $L_{\lhd l'} = \{l \in L | l \lhd l'\}$. With Table 1, we obtain that $\mathcal{D}_{\supseteq A} = \{t_1, t_2\}$ and $P_{\subseteq t_1} = \{A, B, AB\}$.

Pattern discovery takes advantage of interestingness measures to evaluate the relevancy of a pattern. The *support* of a pattern φ in the dataset $\mathcal{D}$ can be considered as the proportion of transactions covered by φ [4]: $Supp(\varphi, \mathcal{D}) = |\mathcal{D}_{\rhd\varphi}|/|\mathcal{D}|$. A pattern is said to be *frequent* when its support exceeds a user-specified minimal threshold. For instance, with a minimal threshold 0.25, the pattern A is frequent because $Supp(A, \mathcal{D}) = |\{t_1, t_2\}|/4$ (≥ 0.25). Thereafter, any

function $f : \mathcal{L} \to \mathbb{R}$ is extended to any pattern set $P \subseteq \mathcal{L}$ by considering $\widetilde{f}(P) = \sum_{\varphi \in P} f(\varphi)$. For instance, $\widetilde{Supp}(P, \mathcal{D})$ corresponds to $\sum_{\varphi \in P} Supp(\varphi, \mathcal{D}) = 2.5$ with $P = \{A, B, C, D, AB, AC, BC\}$ in Table 1.

3 Simulating Analysis Using a Scoring of Patterns

3.1 Scoring Analysis Model

In this section, we propose the *scoring analysis model* to simulate an analyst faced with a set of scored patterns. This model generates sessions by randomly picking patterns taking into account the scoring of patterns. More precisely, the "simulated analyst" randomly draws a pattern by favoring those with the highest measure, and then studies each transaction covered by this pattern during a constant period weighted by its preference vector. Indeed, it is important to benefit from these preferences for better approximating the user behavior. After each pattern analysis, the session can be interrupted (if the analyst is satisfied, no longer has time to pursue, etc.) or continued (if the analyst is dissatisfied, wants more information, etc). This interruption of the session of analysis can be modeled by a halting probability. We now formalize this model:

Definition 1 (Scoring analysis model). *Let $\mathcal{D}$ be a dataset, $P \subseteq \mathcal{L}$ a pattern set, $m : \mathcal{L} \to [0, 1]$ an interestingness measure and ρ a preference vector.*

The scoring analysis model with a halting probability $\alpha \in (0, 1)$ and a unit length $\delta > 0$, denoted by $\mathcal{S}_{\rho,\alpha,\delta}$, generates sessions with the following process:

1. *Pick (with replacement) a pattern φ of P with probability distribution $p(\gamma) = m(\gamma)/\widetilde{m}(P)$ (where $\gamma \in P$).*
2. *Study each transaction $t \in \mathcal{D}$ covered by φ during a length $\delta \times \rho(t)$.*
3. *Stop the session with probability α or then, continue at Step 1.*

Basically, Step 1 favors the analysis of patterns having the highest measure (with replacement because the end-user can re-analyze a pattern in the light of another). Step 2 takes into account the user-preferences for the analysis of transactions. Simulating a data expert by randomly picking patterns may seem strange and unrealistic at first. However, this mechanism has been successfully used in other high-level tasks such as web browsing [5] and text analysis [6]. We think that the strength of our stochastic model is to describe the average behavior of users. By analogy with the random surfer model, each pattern would be a web page. The web pages would then be completely interconnected where each link is weighted by the support of the destination page. In this context, the probability α would correspond to the probability of interrupting navigation.

3.2 Analysis Proportion of a Transaction under $\mathcal{S}_{\rho,\alpha,\delta}$

Starting from the scoring analysis model, we desire to derive the analysis proportion of each transaction.

Theorem 1 (Analysis proportion). *The analysis proportion* $\Pi(t, \mathcal{S}_{\rho,\alpha,\delta})$ *of the transaction* t *is:*

$$\Pi(t, \mathcal{S}_{\rho,\alpha,\delta}) = \frac{\widetilde{m}(P_{\lhd t}) \times \rho(t)}{\sum_{t' \in \mathcal{D}} \widetilde{m}(P_{\lhd t'}) \times \rho(t')}$$

Theorem 1 (proofs are omitted due to lack of space) means that the analysis proportion of a pattern is independent of the parameters α and δ. Therefore, in the following, $\Pi(t, \mathcal{S}_{\rho,\alpha,\delta})$ is simply denoted $\Pi(t, \mathcal{S}_\rho)$. Let us consider the analysis proportion of each transaction of Table 1 in light of Theorem 1 using a uniform preference vector. As the itemsets A, B and AB cover t_1, we obtain that $\widetilde{Supp}(P_{\lhd t_1}, \mathcal{D}) = 0.5 + 0.5 + 0.25 = 1.25$. The same result is obtained for t_2 and t_3 and similarly, $\widetilde{Supp}(P_{\lhd t_4}, \mathcal{D}) = 0.25$. Finally, $\Pi(t_1, \mathcal{S}_\rho) = \widetilde{Supp}(P_{\lhd t_1}, \mathcal{D}) / \sum_{t' \in \mathcal{D}} \widetilde{Supp}(P_{\lhd t'}, \mathcal{D}) = 1.25/(3 \times 1.25 + 0.25) = 5/16 = 0.3125$ and $\Pi(t_4, \mathcal{S}_\rho) = 0.25/4 = 1/16 = 0.0625$. It means that under the scoring analysis model, the transaction t_4 will be less analyzed than the transactions t_1, t_2 or t_3 as indicated in Table 1 (c).

3.3 Balanced Analysis under $\mathcal{S}_{\rho,\alpha,\delta}$

We now deduce what a balanced analysis with respect to ρ is under the scoring analysis model:

Property 1 (Balanced analysis). The analysis of $\mathcal{D}$ by the pattern set P with m is balanced with respect to ρ under the scoring analysis model iff for any transaction $t \in \mathcal{D}$, the following relations holds:

$$\widetilde{m}(P_{\lhd t}) = \frac{1}{|\mathcal{D}|} \times \sum_{t' \in \mathcal{D}} \widetilde{m}(P_{\lhd t'})$$

The crucial observation highlighted by Property 1 is that the balance of an analysis is independent of the preference vector specified by the user, under the scoring analysis model. Indeed, the preference vector ρ involved in the right side of the equation $\Pi(t, \mathcal{M}) = \rho(t)$ is canceled by the one appearing in the analysis proportion (see Theorem 1). Consequently, if the analysis of a dataset $\mathcal{D}$ by a pattern set P with a measure m is balanced with respect to a given preference vector ρ, then it is also balanced with respect to any other preference vector. However, note that the analysis length of a transaction will take into account the considered preference vector.

Let us compute whether the analysis of the dataset $\mathcal{D}$ by the pattern set P with the measure *Supp* (see Table 1) is balanced under the model $\mathcal{S}_\rho$ using Property 1. First, the transaction t_1 is too much studied because $\widetilde{Supp}(P_{\lhd t_1}, \mathcal{D}) = \widetilde{Supp}(\{A, B, AB\}, \mathcal{D}) = 1.25$ and $1/|\mathcal{D}| \times \sum_{t' \in \mathcal{D}} \widetilde{Supp}(P_{\lhd t'}, \mathcal{D}) = 1/4 \times (3 \times 1.25 + 0.25) = 1$. Conversely, as $\widetilde{Supp}(P_{\lhd t_4}, \mathcal{D}) = \widetilde{Supp}(\{D\}, \mathcal{D}) = 0.25$ (< 1), the transaction t_4 is not studied enough. In Section 5, we observe that the use of frequent patterns with the support for the analysis of datasets coming from the UCI repository always leads to an unbalanced analysis.

4 Balancing the Analysis of Patterns

4.1 Axiomatization of Support

Under the scoring analysis model, we aim at balancing the analysis of the dataset by proposing a new interestingness measure that satisfies the equation of Property 1. At this stage, the right question is "what characteristics should satisfy this measure?" Unfortunately we found that the support does not lead to balanced analysis. However, this extremely popular measure is both intuitive for experts and useful in many applications. Moreover, it is an essential atomic element to build many other interestingness measures. For all these reasons, we desire a measure that leads to balanced analysis while maintaining the fundamental properties of the support. To achieve this goal, we first dissect the support by means of its axiomatization (we only focus on the support measure and not on the frequent itemset mining as proposed in [8]).

Property 2 (Support axioms). The support is the only interestingness measure m that simultaneously satisfies the three below axioms for any dataset $\mathcal{D}$:

1. **Normalized:** If a pattern φ covers no transaction (resp. all transactions), then its value $m(\varphi)$ is equal to 0 (resp. 1).
2. **Cumulative:** If patterns φ_1 and φ_2 cover respectively the set of transactions T_1 and T_2 such that $T_1 \cap T_2 = \emptyset$, then the value $m(\varphi)$ of a pattern φ covering exactly $T_1 \cup T_2$ is $m(\varphi_1) + m(\varphi_2)$.
3. **Equalized:** If two patterns cover the same number of transactions, then they have the same value for m.

Clearly the first axiom does not constitute the keystone of support, since similar normalizations are widely used by other measures (e.g., confidence or J-measure). Furthermore, it has no impact on the fact that an analysis is balanced or not, since Step 1 of scoring analysis model performs another normalization. Conversely, we believe that the other two axioms (not verified by other measures) are the main characteristics of the support. If we do not find reason to reconsider the cumulative axiom, we think the third is not fair. Ideally, an interestingness measure should favor the patterns covering the least covered transactions as explained in the introduction. Thus, the value of a measure should not only depend on the number of transactions covered but also on the *singularity* of these transactions. To this end, we propose to retain the first two axioms and to substitute the equalized axiom by the axiom of balance: *a measure of interest must lead to the balanced analysis of the dataset by the pattern set.*

4.2 Balanced Support

We first introduce a relaxation of the support by removing the constraint due to the equalized axiom:

Definition 2 (Weighted support). *Given a function* $w : \mathcal{D} \to \Re^+$*, the weighted support of a pattern* φ *in the dataset* $\mathcal{D}$ *is defined as:* $Supp_w(\varphi, \mathcal{D}) = \sum_{t \in \mathcal{D}_{\triangleright\varphi}} w(t) / \sum_{t \in \mathcal{D}} w(t)$.

It is not difficult to see that the weighted support satisfies the normalized axiom and the cumulative axiom. Now it only remains to choose the right vector w to get a balanced analysis. A naive idea would be to use the preference vector ρ to weight the support. That does not work in the general case: $Supp_{\rho_u}$ (where $\rho_u : t \mapsto 1/|\mathcal{D}|$) corresponds exactly to the support $Supp$ which lead to an unbalanced analysis as shown by our running example (see Section 3.3) or observed in experimental study (see Section 5). In fact, to find the right weights, it is necessary to solve the equation of Property 1 by using the definition of the weighted support. Then, the weighted support induced by these weights defines the *optimal balanced support*:

Definition 3 (Optimal balanced support). *If it exists, the optimal balanced support of a pattern $\varphi \in P$ in the dataset $\mathcal{D}$ with the pattern set P, denoted by* $\mathrm{BS}^*(\varphi, \mathcal{D}, P)$*, is the weighted support where the weight w satisfies the following equation for all transactions $t \in \mathcal{D}$:*

$$\widetilde{Supp_w}(P_{\lhd t}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \times \sum_{t' \in \mathcal{D}} \widetilde{Supp_w}(P_{\lhd t'}, \mathcal{D})$$

Interestingly, this definition underlines that the whole set of mined patterns P is necessary to compute the optimal balanced support of any individual pattern. Let us illustrate the above equation with the example given by Table 1. With the weight w_{bal} where $t_1, t_2, t_3 \mapsto 1/8$ et $t_4 \mapsto 5/8$, we obtain $Supp_{w_{bal}}(A, \mathcal{D}) = Supp_{w_{bal}}(B, \mathcal{D}) = 1/8 + 1/8 = 2/8$; $Supp_{w_{bal}}(AB, \mathcal{D}) = 1/8$ and $Supp_{w_{bal}}(D, \mathcal{D}) = 5/8$. Then, we can check that the equation of Definition 3 is satisfied $\widetilde{Supp_{w_{bal}}}(P_{\lhd t_1}, \mathcal{D}) = \widetilde{Supp_{w_{bal}}}(\{A, B, AB\}, \mathcal{D}) = 2/8+2/8+1/8 = 5/8$ (similar for t_2 and t_3) and $\widetilde{Supp_{w_{bal}}}(P_{\lhd t_4}, \mathcal{D}) = \widetilde{Supp_{w_{bal}}}(\{D\}, \mathcal{D}) = 5/8$. In other words, $Supp_{w_{bal}}$ corresponds exactly to the optimal balanced support $\mathrm{BS}^*(\varphi, \mathcal{D}, P)$.

Theorem 2. *The optimal balanced support (if it exists) is the single interestingness measure that satisfies the normalized and cumulative axioms, and that leads to a balanced analysis.*

Theorem 2 achieves our main goal as stated in introduction. However, the equation of Definition 3 can admit no solution and then the optimal balanced support is not defined. For instance, it is impossible to adjust the weighted support for balancing the analysis of $\mathcal{D} = \{A, B, AB\}$ by $P = \{A, B, AB\}$. Indeed, whatever the weighted support, the transaction AB is still more analyzed than the other two since it is covered by all patterns. So, the next section proposes an algorithm to approximate the optimal balanced support by minimizing the deviation between $\widetilde{Supp_w}(P_{\lhd t}, \mathcal{D})$ and $\sum_{t' \in \mathcal{D}} \widetilde{Supp_w}(P_{\lhd t'}, \mathcal{D})/|\mathcal{D}|$.

4.3 Approximating the Balanced Support

Absolute (for an anagram of balanced support) returns the weights w such that the analysis $\mathcal{S}_{\rho,\alpha,\delta}(\mathcal{D}, P, Supp_w)$ is balanced as better as possible. Its input parameters consist in a pattern set P, a dataset $\mathcal{D}$ and a threshold ϵ. The

latter is the maximal difference expected between two weight vectors stemming from consecutive iterations before terminating the algorithm. The weights outputted by ABSOLUTE enable us to define the *(approximated) balanced support* $\mathrm{BS}(\varphi, \mathcal{D}, P)$.

Algorithm 1. ABSOLUTE

Input: a dataset $\mathcal{D}$, a set of patterns P, a difference threshold ϵ
Output: a weight vector that balances the support

1: **for all** $t \in \mathcal{D}$ **do**
2: $w_0[t] \leftarrow 1/|\mathcal{D}|$
3: **end for**
4: $i \leftarrow 0$
5: **repeat**
6: $W \leftarrow 0$
7: **for all** $t \in \mathcal{D}$ **do**
8: $w_{i+1}[t] \leftarrow w_i[t] \times \frac{\frac{1}{|\mathcal{D}|} \times \sum_{t' \in \mathcal{D}} \widetilde{Supp}_w(P_{\lhd t'}, \mathcal{D})}{\widetilde{Supp}_{w_i}(P_{\lhd t}, \mathcal{D})}$ *// Correct the weight of t*
9: $W \leftarrow W + w_{i+1}[t]$
10: **end for**
11: $diff \leftarrow 0$
12: **for all** $t \in \mathcal{D}$ **do**
13: $w_{i+1}[t] \leftarrow w_{i+1}[t]/W$ *// Normalize the weight of t*
14: $diff \leftarrow diff + |w_{i+1}[t] - w_i[t]|$ *// Update diff*
15: **end for**
16: $i \leftarrow i + 1$
17: **until** $diff/|\mathcal{D}| < \epsilon$
18: **return** w_i

Note that in Algorithm 1, w_i are symbol tables where the keys are transactions. Lines 1-3 initialize all the weigths with $1/|\mathcal{D}|$. The main loop (Lines 5-17) adjusts the weigths until the sum of differences between w_{i+1} and w_i is less than ϵ. More precisely, Lines 7-10 correct the weight of each transaction. Using Definition 3, Line 8 computes the new weight $w_{i+1}[t]$ by multiplying the previous weight $w_i[t]$ by the ratio between the average coverage (i.e., a constant $1/|\mathcal{D}| \times \sum_{t' \in \mathcal{D}} \widetilde{Supp}_{w_i}(P_{\lhd t'}, \mathcal{D})$ shared by all transactions) and the coverage of t (i.e., $\widetilde{Supp}_{w_i}(P_{\lhd t}, \mathcal{D})$). For instance, if the coverage of t is below the average coverage, the ratio is above 1 and the new weight is stronger. Thus, it increases the support of all the patterns covering this transaction. This operation therefore operates a local balance for each transaction. Nevertheless, there is also a global modification since a normalization is performed on these weights at Line 13 (where W is computed Line 9). Line 14 updates *diff* (initialized Line 11) accumulating the difference between w_{i+1} and w_i for all the transactions. Finally, Line 18 returns the last weights that correspond to a balanced analysis.

5 Experimental Evaluation

This section evaluates the effectiveness of the algorithm for balancing the analysis and to compare the quality of the balanced support with respect to the usual one. All experiments reported below were conducted with a difference threshold $\epsilon = 10^{-5}$ on datasets coming from the UCI Machine Learning Repository

(`www.ics.uci.edu/~mlearn/MLRepository.html`). Given a minimal support threshold *minsupp*, we select all the frequent itemsets for P. Increasing the weight of singular transactions does not cause the extraction of random noise patterns because the final patterns are selected from the collection of frequent patterns. For simplicity, we use the uniform preference vector $\rho_u : t \mapsto 1/|\mathcal{D}|$ for ρ.

Table 2. ABSOLUTE on UCI benchmarks for frequent itemsets ($minsupp = 0.05$)

Dataset	$\|P\|$	# of iter.	$D_{KL}{}^{\rho_u}_{Supp}$	$D_{KL}{}^{\rho_u}_{\mathrm{BS}}$	Gain
abalone	2,527	17	0.180	0.021	8.72
anneal	25,766	24	0.339	0.013	**26.2**
austral	20,386	29	0.076	0.006	**12.6**
breast	2,226	41	0.145	0.006	**22.5**
cleve	11,661	35	0.172	0.012	**14.7**
cmc	2,789	23	0.091	0.002	**50.8**
crx	34,619	24	0.122	0.011	**10.9**
german	124,517	17	0.172	0.029	5.96
glass	3,146	52	0.084	0.005	**17.7**
heart	16,859	37	0.116	0.014	7.95
hepatic	511,071	13	0.568	0.040	**14.1**
horse	17,084	19	0.275	0.017	**15.9**
lymph	275,278	34	0.138	0.016	8.34
page	3,190	42	0.054	0.004	**14.8**
vehicle	187,449	24	0.636	0.020	**31.3**
wine	12,656	46	0.154	0.009	**17.8**
zoo	586,579	34	0.353	0.019	**18.2**

Efficiency of ABSOLUTE Table 2 (columns 2-3) presents the number of patterns and the number of iterations required by ABSOLUTE for balancing all the frequent patterns. Note that we do not provide running times because they are very low. Indeed, the worst case is the balancing time for all the frequent patterns on `zoo`, but it does not exceed 16 seconds performed on a 2.5 GHz Xeon processor with the Linux operating system and 2 GB of RAM memory (ABSOLUTE is implemented in C++). Table 2 shows that the number of iterations varies between 13 and 52. No simple relationship was found between the number of iterations and the features of datasets.

Table 2 also reports the Kullback-Leibler divergence for support and BS (columns 4-6). Let us recall that Kullback-Leibler divergence defined by $D_{KL}(P||Q) = \sum_i P(i) \times \log \frac{P(i)}{Q(i)}$ measures the difference between two probability distributions P and Q [9]. For any transaction t, we fix $P(t) = \rho_u(t)$ as reference and $Q(t) = \Pi(t, \mathcal{M}_{\rho_u})$ as model. Table 2 shows that ABSOLUTE reaches its goal since the Kullback-Leibler divergence is always significantly reduced by benefiting from the balanced support. This divergence is at least divided by 5 and it is even divided by more than 10 in 13 datasets. The average gain is 17.56 for frequent itemsets. Similar experiments conducted on collections of free and closed itemsets [10] gave respectively an average gain of 12.36 and 11.94.

Effectiveness of Balanced Support. We desire to quantify the number of non-correlated patterns (i.e., the number of extracted patterns that are spurious) with a usual/balanced support. Unfortunately, the pattern discovery process is unsupervised and the (ir)relevant patterns are unknown. We tackle this issue

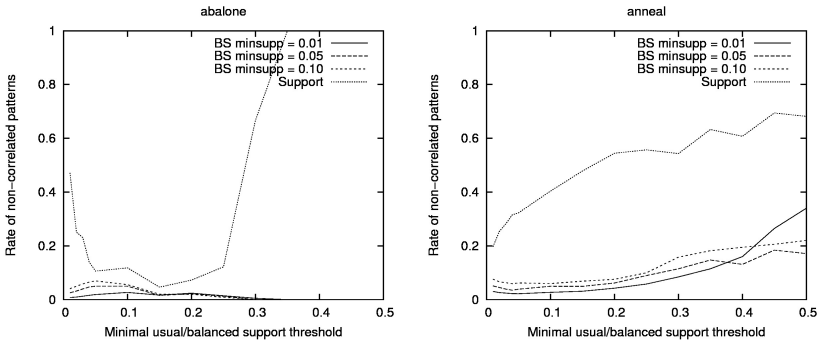

Fig. 1. Estimating the number of non-correlated patterns for *Supp* and BS

by using an experimental protocol inspired by [11]. The idea is to make the assumption that a pattern is non-correlated if this pattern is also extracted (by the same method) in a random dataset $\mathcal{D}^*$ having the same characteristics as $\mathcal{D}$ (i.e., the same dimensions and the same support for each item).

Figure 1 depicts the ratio of non-correlated patterns (averaged from 10 random datasets $\mathcal{D}^*$) for `abalone` and `anneal` for frequent itemsets with a minimal usual/balanced support varying between 0 and 0.5. This ratio is the number of non-correlated patterns divided by the total number of patterns. For the balanced support, we use three collections of frequent patterns P obtained with $minsupp = 0.01/0.05/0.10$ independently of the second threshold applied to balanced support. Given a minimal threshold (see horizontal axis), the ratio of non-correlated patterns for *Supp* is always higher than that of BS and most of times, with a significant difference. Interestingly, the change of $minsupp$ for the collection of patterns has a marginal impact on the ratio of non-correlated patterns. Recall that balanced support only differs from the traditional one by replacing the equalized axiom by the axiom of balance (see Section 4.1). So it is this axiom that enables our measure to keep out uncorrelated patterns. More generally, this experience justifies the interest of a balanced analysis and even the usefulness of the scoring analysis model for simulating an analysis.

6 Related Work

As mentioned in the introduction, many interestingness measures have been proposed for evaluating the pattern interest as alternative to the support [2,12,3]. They can be categorized into two sets [1]: user-driven measures and data-driven ones. Among the data-driven approaches, the statistical models are often based on the null hypothesis. A pattern is interesting if it covers more transactions than what was expected. Some models simply require the frequency of items forming the itemset [12], others rely on its subsets [13,14] or even, patterns already extracted [15]. These methods consider that all transactions have the same weight. However, in practice, the user tends to attach more importance

to information that describes the least common facts. Thus, the most singular transactions should have an important weight in the evaluation of patterns that describe such transactions. In this sense, this paper proposes another alternative resting on the integration of the analysis method into the metric. To the best of our knowledge, this way has not yet been explored in the literature. A major and original consequence of our approach lies in the fact that each transaction contributes with a different weight in the balanced support and this weight depends on the entire extracted collection.

However, the problem of unbalance induced by a pattern set is indirectly addressed by several approaches removing the patterns that describe transactions covered by other patterns. For instance, the condensed representations [10] which remove redundant patterns, often decrease the unbalanced of the analysis. But, empirical experiments have shown that the unbalance remains important (see Section 5). In the same way, global models based on patterns [16,17,18] favor balanced analyses of the dataset. Indeed, one goal of these approaches is to describe all the data by choosing the smallest set of patterns. The overlap between the coverings of the different patterns is very reduced (ideally each transaction should be described by a unique pattern as it is the case with a decision tree). Unfortunately, relevant patterns may be removed from such models. Our approach balances the analysis of the dataset by preserving the whole set of patterns to avoid losing information.

Rather than modifying the collection of mined patterns, it would be possible to modify the initial dataset in order to satisfy user preferences. Sampling methods [19,20] are widely used in machine learning and data mining in particular to correct a problem of unbalance between classes. There is no reason that the change of the dataset with a usual sampling method leads to a balanced analysis. We think that our approach is complementary to those of sampling.

7 Conclusion

In this paper, we introduce the scoring analysis model for simulating analysis sessions of a dataset by means of a pattern set. Under this model, we define the balanced support that induces a balanced analysis of the dataset for any user-specified preference vector. We propose the algorithm Absolute to iteratively calculate transaction weights leading to the balanced support. This new interestingness measure strongly balances the analysis and in parallel, it enables us to filter-out non-correlated patterns. The originality of our work is to show that the integration of the analysis method to drive the data mining is profitable.

In future work, we are interested in examining our approach on real-world data for better understanding the semantic of the balanced support: what are the patterns which balanced support is much higher than traditional support? What are domains and datasets where the balanced support is most appropriate? Dually, we must also study the properties of the weights resulting from Absolute that could be interesting to identify the outliers. Furthermore, the prospects of using the scoring analysis model are manifold. For instance, this model could be used to balance other measures of interest like the confidence.

References

1. Freitas, A.A.: Are we really discovering "interesting" knowledge from data. Expert Update (the BCS-SGAI Magazine) 9, 41–47 (2006)
2. McGarry, K.: A survey of interestingness measures for knowledge discovery. Knowledge Eng. Review 20(1), 39–61 (2005)
3. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Comput. Surv. 38(3) (2006)
4. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: VLDB, pp. 487–499. Morgan Kaufmann (1994)
5. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks 30(1-7), 107–117 (1998)
6. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: EMNLP, pp. 404–411. ACL (2004)
7. Mannila, H., Toivonen, H.: Levelwise search and borders of theories in knowledge discovery. Data Min. Knowl. Discov. 1(3), 241–258 (1997)
8. Calders, T., Paredaens, J.: Axiomatization of frequent itemsets. Theor. Comput. Sci. 290(1), 669–693 (2003)
9. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics 22, 49–86 (1951)
10. Calders, T., Rigotti, C., Boulicaut, J.F.: A survey on condensed representations for frequent sets. In: Boulicaut, J.-F., De Raedt, L., Mannila, H. (eds.) Constraint-Based Mining. LNCS (LNAI), vol. 3848, pp. 64–80. Springer, Heidelberg (2006)
11. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. TKDD 1(3) (2007)
12. Omiecinski, E.: Alternative interest measures for mining associations in databases. IEEE Trans. Knowl. Data Eng. 15(1), 57–69 (2003)
13. Tatti, N.: Probably the best itemsets. In: Rao, B., Krishnapuram, B., Tomkins, A., Yang, Q. (eds.) KDD, pp. 293–302. ACM (2010)
14. Webb, G.I.: Self-sufficient itemsets: An approach to screening potentially interesting associations between items. TKDD 4(1) (2010)
15. Mampaey, M., Vreeken, J., Tatti, N.: Summarizing data succinctly with the most informative itemsets. TKDD 6(4), 16 (2012)
16. Bringmann, B., Zimmermann, A.: The chosen few: On identifying valuable patterns. In: ICDM, pp. 63–72. IEEE Computer Society (2007)
17. Fürnkranz, J., Knobbe, A.: Guest editorial: Global modeling using local patterns. Data Min. Knowl. Discov. 21(1), 1–8 (2010)
18. Gamberger, D., Lavrac, N.: Expert-guided subgroup discovery: Methodology and application. J. Artif. Intell. Res. (JAIR) 17, 501–527 (2002)
19. Chawla, N.V.: Data mining for imbalanced datasets: An overview. In: Data Mining and Knowledge Discovery Handbook, pp. 875–886. Springer (2010)
20. Liu, H., Motoda, H.: On issues of instance selection. Data Min. Knowl. Discov. 6(2), 115–130 (2002)