Data Min Knowl Disc (2010) 21:371–397
DOI 10.1007/s10618-009-0161-2

Re-examination of interestingness measures in pattern mining: a unified framework

Tianyi Wu · Yuguo Chen · Jiawei Han

Received: 12 February 2009 / Accepted: 10 December 2009 / Published online: 6 January 2010
© The Author(s) 2009

Abstract Numerous interestingness measures have been proposed in statistics and data mining to assess object relationships. This is especially important in recent studies of association or correlation pattern mining. However, it is still not clear whether there is any intrinsic relationship among many proposed measures, and which one is truly effective at gauging object relationships in *large data sets*. Recent studies have identified a critical property, *null-*(*transaction*) *invariance*, for measuring associations among events in large data sets, but many measures do not have this property. In this study, we re-examine a set of null-invariant interestingness measures and find that they can be expressed as the generalized mathematical mean, leading to a total ordering of them. Such a unified framework provides insights into the underlying philosophy

Responsible editor: M.J. Zaki.

The work was supported in part by the U.S. National Science Foundation NSF IIS-08-42769, NSF IIS-09-05215, NSF DMS-05-03981, and NSF DMS-08-06175. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. This paper is a major-value added extension of the paper: Tianyi Wu, Yuguo Chen and Jiawei Han, Association Mining in Large Databases: A Re-Examination of Its Measures", Proc. 2007 Int. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Warsaw, Poland, Sept. 2007, pp. 621–628.

T. Wu (✉) · J. Han
Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA
e-mail: twu5@uiuc.edu

J. Han
e-mail: hanj@cs.uiuc.edu

Y. Chen
Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA
e-mail: yuguo@uiuc.edu

of the measures and helps us understand and select the proper measure for different applications. Moreover, we propose a new measure called *Imbalance Ratio* to gauge the degree of skewness of a data set. We also discuss the efficient computation of interesting patterns of different null-invariant interestingness measures by proposing an algorithm, GAMiner, which complements previous studies. Experimental evaluation verifies the effectiveness of the unified framework and shows that GAMiner speeds up the state-of-the-art algorithm by an order of magnitude.

Keywords Association rules · Frequent pattern · Interestingness measure · Null-invariant measure · Generalized mean

1 Introduction

Association pattern mining has been studied for more than a decade (Agrawal and Srikant 1994; Han et al. 2007). However, it has been well recognized that the association rules typically defined in a support-confidence framework (Agrawal and Srikant 1994) may not disclose truly interesting event relationships (Brin et al. 1997). For example, mining in a typical market basket data set may result in a rule, "*coffee* → *milk*", with nontrivial support (*e.g.*, 10%) and high confidence (*e.g.*, 60%; that is, 60% of transactions containing *coffee* contain *milk* also.), but this rules does not necessarily imply that the event *buying coffee* and the event *buying milk* are strongly correlated, because *milk* itself could be a very popular item that 70% of all transactions contain it already, which means that *coffee* actually discounts the chance of *milk*. Thus, researchers have proposed various kinds of measures in pattern mining to assess the interestingness of the relationships among various events (Brin et al. 1997; Savasere et al. 1998; Omiecinski 2003; Grahne et al. 2000).

Many association, correlation, and similarity measures have been proposed in the fields of statistics, information retrieval, machine learning and data mining, for analyzing the relationships among discretized events (Hilderman and Hamilton 2001; Tan et al. 2002). For example, χ^2 is a typical correlation measure in analyzing correlations among discretized events in classical statistics (Kachigan 1991), which is also used for mining association rules (Brin et al. 1997). However, it may not be an appropriate measure for analyzing event associations in large transaction databases (as shown in Ex. 1 below). Notice that in a typical transaction database, a particular item i (such as *coffee*) appearing in a transaction T (*i.e.*, $i \in T$) is often a *small probability event*; that is, i has a small number of occurrences compared to the total number of transactions. Small probability event is ubiquitous in real-world applications. For example, in a market basket data set, the event that a particular product is contained in a transaction has a very small probability. Also, in a publication database like DBLP,[1] the event that a particular name is on the author list of a paper also has a very small probability. A transaction that does not contain an item i is called a *null transaction with respect to (w.r.t.) item* i. If a measure that assesses the association among a set of events being analyzed is affected by the number of transactions that contain none of them

[1] http://www.informatik.uni-trier.de/~ley/db/.

Table 1 Two-event contingency table

	$milk$	$\overline{milk}$	Σ_{row}
$coffee$	mc	$\overline{m}c$	c
$\overline{coffee}$	$m\overline{c}$	$\overline{mc}$	$\overline{c}$
Σ_{col}	m	$\overline{m}$	Total

(*i.e.*, null-transactions), such a measure is unlikely to be so interesting. Thus it is desirable to select a measure that is not influenced by the number of null transactions, *i.e.*, *null-invariant*. Recent studies (Tan et al. 2002; Wu et al. 2007; Lee et al. 2003; Omiecinski 2003) have shown that null-(transaction) invariance is critically important for an interestingness measure. We use the following example to illustrate this.

Example 1 In a typical shopping transaction database, a product appearing in a transaction is called an **event**, and a set of products appearing in a transaction is called an **event-set**. Association analysis is to identify interesting associations among a set of events. It is expected that a particular event happens with a very low probability.

In Table 1, the purchase history of two events *milk* and *coffee* is summarized by their support counts, where, for instance, mc denotes the support of the event-set "*coffee and milk*", *i.e.*, the number of occurrences of transactions containing both of them. Each such contingency table can be uniquely characterized by four variables, mc, $\overline{m}c$, $m\overline{c}$, and $\overline{mc}$, where $\overline{mc}$ is the null-transactions w.r.t. m and c. Table 2 lists six data sets in terms of a "flattened" contingency table. We then select seven interestingness measures: χ^2, *Lift*, *AllConf*, *Coherence*, *Cosine*, *Kulczynski* (denoted as *Kulc* for simplicity hereafter), and *MaxConf*, and show their resulting measure values for each data set.

The definitions of the seven measures on two events a and b are given in Table 3. Among them, *AllConf*, *Coherence*, *Cosine*, and *MaxConf* are the only ones that do not depend on the number of null-transactions w.r.t. a and b (hence called *null-invariant measures*) among over 20 existing interestingness measures studied in Tan et al. (2002). *Kulc* is another null-invariant measure proposed in Bradshaw (2001). Two popular but not null-invariant measures, χ^2 and *Lift* (Brin et al. 1997; Hilderman and Hamilton 2001), are listed here for comparison. Notice that $Coherence$, though introduced lately (Omiecinski 2003) and defined differently, is essentially the commonly used *Jaccard coefficient* (Tan et al. 2002).[2] In addition to that, we use the name *MaxConf* instead of *Confidence* in Tan et al. (2002) in order to avoid any confusion with the directional "confidence" measure in traditional association rule mining (Agrawal and Srikant 1994) (here "directional" refers to the fact that the confidence of a and b may not be equal to the confidence of b and a). To the best of our knowledge, *AllConf*, *Coherence*, *Cosine*, *Kulc*, and *MaxConf* are the only null-invariant measures in the literature.

Let's first examine data sets D_1 and D_2, where *milk* and *coffee* are positively associated because mc (10,000) is considerably greater than $\overline{m}c$ (1,000) and $m\overline{c}$ (1,000). Intuitively, for people who bought *milk* ($m = 10{,}000 + 1{,}000 = 11{,}000$), it is very

[2] $Coherence(a, b)$, though introduced lately (Omiecinski 2003) and defined differently, is essentially the popularly used *Jaccard coefficient* (Tan et al. 2002) because $Jaccard(a, b) = sup(ab)/(sup(ab) + sup(\bar{a}b) + sup(a\bar{b})) = sup(ab)/(sup(b) + sup(a) - sup(ab)) = Coherence(a, b)$.

Table 2 Example data sets

Data set	mc	$\overline{m}c$	$m\bar{c}$	$\overline{mc}$	χ^2	*Lift*	*AllConf*	*Coherence*	*Cosine*	*Kulc*	*MaxConf*
D_1	10,000	1,000	1,000	100,000	90557	9.26	0.91	0.83	0.91	0.91	0.91
D_2	10,000	1,000	1,000	100	0	1	0.91	0.83	0.91	0.91	0.91
D_3	100	1,000	1,000	100,000	670	8.44	0.09	0.05	0.09	0.09	0.09
D_4	1,000	1,000	1,000	100,000	24740	25.75	0.5	0.33	0.5	0.5	0.5
D_5	1,000	100	10,000	100,000	8173	9.18	0.09	0.09	0.29	0.5	0.91
D_6	1,000	10	100,000	100,000	965	1.97	0.01	0.01	0.10	0.5	0.99

Table 3 Interestingness measure definitions

Measure	Definition	Range	Null-invariant
$\chi^2(a, b)$	$\sum_{i,j=0,1} \frac{\left(e(a_i,b_j)-o(a_i,b_j)\right)^2}{e(a_i,b_j)}$	$[0, \infty]$	No
$Lift(a, b)$	$\frac{P(ab)}{P(a)P(b)}$	$[0, \infty]$	No
$AllConf(a, b)$	$\frac{sup(ab)}{max\{sup(a),sup(b)\}}$	[0, 1]	Yes
$Coherence(a, b)$	$\frac{sup(ab)}{sup(a)+sup(b)-sup(ab)}$	[0, 1]	Yes
$Cosine(a, b)$	$\frac{sup(ab)}{\sqrt{sup(a)sup(b)}}$	[0, 1]	Yes
$Kulc(a, b)$	$\frac{sup(ab)}{2}\left(\frac{1}{sup(a)}+\frac{1}{sup(b)}\right)$	[0, 1]	Yes
$MaxConf$(a,b)	$max\left\{\frac{sup(ab)}{sup(a)}, \frac{sup(ab)}{sup(b)}\right\}$	[0, 1]	Yes

likely that they bought *coffee* also ($mc/m = 10/11 = 91\%$), and vice versa. The results of the five null-invariant measures show that m and c are strongly positively associated in both data sets by producing a measure value of at least 0.83. However, *Lift* and χ^2 generate dramatically different measure values for D_1 and D_2, due to their sensitivity to $\overline{mc}$. In fact, in many real-world scenarios $\overline{mc}$ is usually huge and unstable. For example, in a market basket database, the total number of event-sets could fluctuate on a daily basis and can overwhelmingly exceed the number of any particular event-set. In the DBLP database, the total number of publications is also growing rapidly day-to-day, while the total number of publications of a particular author may remain the same in months. Therefore, a good interestingness measure should not be affected by null-transactions; otherwise, it would generate unstable results as illustrated in D_1 and D_2. Similarly, in D_3, the five null-invariant measures correctly show that m and c are strongly negatively associated, because $mc : c = mc : m = 100 : 1{,}100 = 9.1\%$; whereas *Lift* and χ^2 judge it in an incorrect or controversial way: D_2's association is between that of D_1 and D_3.

For data set D_4, both *Lift* and χ^2 indicate a highly positive association between m and c, whereas the others a "neutral" association, because $mc : \overline{m}c = mc : m\overline{c} = 1 : 1$. This means that given the event *coffee* (or *milk*), the probability that the event *milk* (or *coffee*) happens is exactly 50%. (The neutral point of *Coherence* is at 0.33 instead of 0.5 (Lee et al. 2003).)

Based on the null-invariance property for small probability events, our subsequent discussion will be focused only on the null-invariant interestingness measures, *i.e.*, *AllConf*, *Cosine*, *Coherence*, *Kulc*, and *MaxConf*. Although a comparative study of various interestingness measures has been done in Tan et al. (2002), there are still very important problems left open on those null-invariant measures, such as:

- Are there inherent relationships among those null-invariant measures?
- What are their intrinsic differences and how would they influence pattern mining results?
- Which measure would be more effective for evaluating interesting associations among small probability events? And in what context?

These questions motivate us to conduct an in-depth study of these measures, which weaves a well-organized picture over the null-invariant measures. Specifically, our study makes the following contributions.

1. (Section 2) Show that there exists a total ordering among these measures based on a mathematical analysis. This not only explains the inherent relationships and underlying philosophy of them, but also provides a unified view of association analysis in large transaction data sets.
2. (Section 2) Propose a generalized null-invariant measure to extend the existing association measures to handle multiple events under the unified framework. Moreover, we propose an *Imbalance Ratio* measure to quantify the degree of imbalance of a data set based on a set of desirable properties.
3. (Section 3) Systematically study the efficient computation of interesting patterns with respect to different null-invariant measures. Specifically, we propose an algorithm, GAMiner, to efficiently mine frequent and highly associated *Kulc* and *Cosine* patterns.
4. (Section 4) Present comprehensive experimental evaluation on two real data sets to illustrate the difference of the measures being examined here; we also evaluate the efficiency of our proposed algorithm by comparing it to the state-of-the-art method on a standard synthetic data set.
5. (Section 5) Discuss the selection of an appropriate null-invariant measure.

In comparison to Wu et al. (2007), this paper contains the following main technical novelties:

- We extend the intuition behind the null-invariant measures through examples and give details for the mathematical proof of the ordering of the measures, where the boundary cases are thoroughly explained;
- We propose a novel *Imbalance Ratio* measure to complement the re-examined measures;
- We propose a novel pattern mining algorithm to address the efficient computation problem as an integral part of the overall association pattern mining framework;
- We present additional experimental studies to (i) evaluate *Imbalance Ratio* on a real data set, (ii) evaluate the null-invariant measures on a Web search log data and present quantitative analysis of the their differences, and (iii) verify the efficiency gain of our proposed pattern mining algorithm;
- We discuss techniques for the measure selection problem.

The rest of the paper is organized as follows. Section 2 presents the re-examination and the unified framework of the null-invariant measures, and we also propose a generalized measure and a complementary *Imbalance Ratio* measure. In Section 3, we systematically discuss the problem of efficiently mining interesting patterns of different measures and propose the GAMiner algorithm. This is followed by an empirical evaluation of the unified framework on real data sets as well as a performance evaluation of GAMiner in Section 4. Section 5 discusses the selection of an appropriate null-invariant measure and reviews the related work, and finally, Section 6 concludes the study.

2 A re-examination of null-invariant measures

In this section, the five existing null-invariant measures for association mining in large databases are re-examined. We first re-examine their *similarities* using a set of fundamental, desired properties. Next, we generalize the measures using the mathematical generalized mean, which provides insights into the *differences* and underlying philosophy of the measures. The generalized mean approach is subsequently extended to support multiple events.

2.1 Inherent ordering among the measures

Given two arbitrary events a and b, we denote the support of a, b, and ab as $sup(a), sup(b)$, and $sup(ab)$, respectively, and use a and $sup(a)$ interchangeably when there is no ambiguity. Let $\mathcal{M}$ be any of the five null-invariant measures. From the definitions in Table 3, we immediately have the following fundamental properties.

- **P1** $\mathcal{M} \in [0, 1]$;
- **P2** $\mathcal{M}$ monotonically increases with $sup(ab)$ when $sup(a)$ and $sup(b)$ remain unchanged; and it monotonically decreases with $sup(a)$ (or $sup(b)$) when $sup(ab)$ and $sup(b)$ (or $sup(a)$) stay the same;
- **P3** $\mathcal{M}$ is symmetric under item permutations; and
- **P4** $\mathcal{M}$ is invariant to scaling, *i.e.*, multiplying a scaling factor to $sup(ab)$, $sup(a)$, and $sup(b)$, will not affect the measure.

These four properties justify our "conventional wisdom" about association analysis in large databases, and therefore are desired for the measures. Specifically speaking, property P1 states that the value domain of $\mathcal{M}$ is normalized so that 0 indicates no co-occurrence of the events and 1 indicates that they always appear together. The only exception is that $MaxConf(a, b) = 1$ may not indicate that a and b always co-occur. It reflects a slightly weaker association condition where one event always co-occurs with the other but the converse may not necessarily be true. Property P2 is consistent with the basic intuition that the more the co-occurrences of the two events, the greater the value of the measure, and vice versa. Also, if the individual occurrences of a single event increase, its association with the other one must not become larger. Property P3 and P4 show the robustness of $\mathcal{M}$ in terms of the event ordering and the measurement scale. Such symmetric association relationship is opposed to the asymmetric nature of the traditional confidence measure for association rule mining (Agrawal and Srikant 1994).

Despite such "conventional wisdom", however, there are subtle cases that cannot be resolved straightforwardly by our common sense.

Example 2 Turn to data sets D_5 and D_6 in our running example in Table 2, where the two events m and c have unbalanced conditional probabilities. That is, $mc : c > 0.9$, meaning that knowing that c happens would strongly suggest that m happens also, whereas $mc : m < 0.1$, indicating that m implies that c is quite unlikely to happen. *AllConf*, *Coherence*, and *Cosine* view both cases as negatively associated and *Kulc* takes both as neutral, but *MaxConf* claims strongly positive associations for these

cases. Having observed such divergent results, one may ask, "*which measure intuitively reflects the true relationship between milk and coffee?*" Unfortunately, there is no commonly agreed judgment for such cases due to the "*balanced*" skewness of the data. It is difficult to argue whether the two data sets have positive or negative association. On the one hand, only $mc/(mc+m\bar{c}) = 1{,}000/(1{,}000+10{,}000) = 9.1\%$ of *milk*-related event-sets contain *coffee* in D_5 and this percentage is $1{,}000/(1{,}000+100{,}000) = 1\%$ in D_6, both indicating a negative association. On the other hand, $mc/(mc+\overline{m}c) = 1{,}000/(1{,}000+100) = 91\%$ of transactions in D_5 and $1{,}000/(1{,}000+10) = 99\%$ in D_6 containing *coffee* contain *milk* as well, which indicates a positive association between *milk* and *coffee*. We draw totally different conclusions from different perspectives. Nor are we able to seek a straightforward answer to another question—"*Which one of D_5 and D_6 has stronger (or more positive) association?*"—because for D_5, the two "one-way" associations are 9.1% (in terms of *milk*) and 91% (in terms of *coffee*), respectively, while D_6 demonstrates even more extreme, unbalanced associations. That is, for *milk*-related event-sets in D_6, *coffee* shows a more negative relationship than D_5 does (*i.e.*, $1 < 9.1\%$) and for *coffee*-related event-sets, *milk* has a more positive relationship (*i.e.*, $99 > 91\%$). It is therefore difficult to judge whether it is D_5 or D_6 that has a more positive relationship.

Interestingly, we show that there exists a unified framework that can seamlessly explain the above controversial cases. This framework discloses the underlying philosophies of the measures and explains their inherent relationships, which in turn may help a user's decision-making in selecting the right measure for different application domains.

To begin with, we rewrite the definitions in Table 3 into the form of conditional probabilities as shown in Table 4. We convert the support counts into conditional probabilities using the following equations:

$$P(a|b) = \frac{sup(ab)}{sup(b)} = \frac{sup(ab)}{sup(ab) + sup(\overline{a}b)}$$
$$P(b|a) = \frac{sup(ab)}{sup(a)} = \frac{sup(ab)}{sup(ab) + sup(a\overline{b})}.$$

After rewriting, the variables in the contingency table, $sup(ab)$, $sup(a\overline{b})$, and $sup(\overline{a}b)$ can be removed, and hence all five null-invariant measures can be expressed as a function of only two variables $P(a|b)$ and $P(b|a)$. Note that, among the original defini-

Table 4 Null-invariant measures defined by conditional probabilities

Measure	Definition	Exponent
$AllConf(a,b)$	$\min\{P(a\|b), P(b\|a)\}$	$k \to -\infty$
$Coherence(a,b)$	$(P(a\|b)^{-1} + P(b\|a)^{-1} - 1)^{-1}$	$k = -1$
$Cosine(a,b)$	$\sqrt{P(a\|b)P(b\|a)}$	$k \to 0$
$Kulc(a,b)$	$(P(a\|b) + P(b\|a))/2$	$k = 1$
$MaxConf(a,b)$	$\max\{P(a\|b), P(b\|a)\}$	$k \to +\infty$

tions of the five null-invariant measures in Table 3, *AllConf* and *Coherence* require $sup(a) \neq 0$ or $sup(b) \neq 0$, while *Cosine*, *Kulc*, and *MaxConf* require $sup(a) \neq 0$ and $sup(b) \neq 0$. For the definitions in Table 4, we require $sup(a) \neq 0$ and $sup(b) \neq 0$, because otherwise $P(a|b)$ and/or $P(b|a)$ would be undefined. The rewritten definition of *Coherence* further requires the assumption that $sup(ab) \neq 0$. For simplicity, we hereafter assume that all five measures will be equal to 0 if $sup(ab) = 0$.

Following the rewritten definitions, we can generalize all five measures using the mathematical generalized mean (Kachigan 1991). Specifically, each of the null-invariant measures can be represented by the generalized mean of the two conditional probabilities $P(a|b)$ and $P(b|a)$ as

$$\mathbb{M}^k(P(a|b), P(b|a)) = \left(\frac{P(a|b)^k + P(b|a)^k}{2}\right)^{\frac{1}{k}}, \tag{1}$$

where $\mathbb{M}$ denotes the mathematical generalized mean and $k \in (-\infty, +\infty)$ is the exponent (k is a real number). We have the following lemma.

Lemma 1 *Each null-invariant measure in Table 4 can be expressed using Eq. (1) with its corresponding exponent.*

Proof We now prove the correctness of the general representation. We first prove that

$$AllConf(a, b) = \lim_{k \to -\infty} \mathbb{M}^k(P(a|b), P(b|a)). \tag{2}$$

Without loss of generality, let's assume that $0 \leq P(a|b) \leq P(b|a) \leq 1$. The proof follows from

$$\begin{aligned}
\lim_{k \to -\infty} \mathbb{M}^k(P(a|b), P(b|a)) &= \lim_{k \to -\infty} \left(\frac{1 + (P(b|a)/P(a|b))^k}{2}\right)^{1/k} P(a|b) \\
&= P(a|b) \\
&= min\{P(a|b), P(b|a)\} \\
&= AllConf(a, b).
\end{aligned}$$

For *Coherence*, the equation $Coherence(a, b) = \mathbb{M}^{-1}(P(a|b), P(b|a))$ does not hold. In fact, we have

$$\begin{aligned}
Coherence(a, b) &= \left(P(a|b)^{-1} + P(b|a)^{-1} - 1\right)^{-1} \\
&= \left(\frac{2}{\mathbb{M}^{-1}(P(a|b), P(b|a))} - 1\right)^{-1}.
\end{aligned}$$

For simplicity, we define a new measure $Coherence'(a, b) = \mathbb{M}^{-1}(P(a|b), P(b|a))$ as a replacement measure of *Coherence* in our subsequent discussions. This is a reasonable replacement because *Coherence*$'$ preserves the ordering of *Coherence*; that is, given any events a_1, b_1, a_2, and b_2,

$$Coherence'(a_1, b_1) \leq Coherence'(a_2, b_2) \Leftrightarrow Coherence(a_1, b_1) \leq Coherence(a_2, b_2).$$

For *Cosine*, let $x = P(a|b)/P(b|a)$ (notice that we are assuming $P(b|a) \neq 0$ because $sup(ab) \neq 0$). To prove

$$Cosine(a, b) = \lim_{k \to 0} \mathbb{M}^k(P(a|b), P(b|a)), \tag{3}$$

is equal to proving that

$$\lim_{k \to 0} \ln \left(\frac{x^k + 1}{2} \right)^{1/k} = \ln(x^{1/2}), \tag{4}$$

because

$$\lim_{k \to 0} \left(\frac{x^k + 1}{2} \right)^{1/k} = \frac{1}{P(b|a)} \lim_{k \to 0} \left(\frac{P(a|b)^k + P(b|a)^k}{2} \right)^{1/k}, \text{ and}$$
$$x^{1/2} = \frac{1}{P(b|a)} \sqrt{P(a|b)P(b|a)}.$$

In fact, we have

$$\begin{aligned} \lim_{k \to 0} \ln \left(\frac{x^k + 1}{2} \right)^{1/k} &= \frac{\lim_{k \to 0} \partial \left(\ln \frac{x^k + 1}{2} \right) / \partial k}{\lim_{k \to 0} \partial k / \partial k} \\ &= \lim_{k \to 0} \frac{\frac{1}{2} x^k \ln x}{(x^k + 1)/2} \\ &= \frac{1}{2} \ln x, \end{aligned}$$

which means that Eqs. (4) and (3) hold.

The proof for $Kulc(a, b) = \mathbb{M}^1(P(a|b), P(b|a))$ is omitted here because it is straightforward.

The proof for

$$MaxConf(a, b) = \lim_{k \to +\infty} \mathbb{M}^k(P(a|b), P(b|a)). \tag{5}$$

follows a similar argument as the one for *AllConf*. □

Notice that all five null-invariant measures except *Coherence* (or *Jaccard coefficient*) can be expressed nicely as the generalized mean of $P(a|b)$ and $P(b|a)$ with different exponents. *Coherence'*, transformed from the original *Coherence* measure, can be also expressed using the generalized mean. The generalization of the association measures to $\mathbb{M}^k(P(a|b), P(b|a))$ (note that these association measures only differ in terms of the exponent k) has two implications, which we summarize into the following lemmas.

Lemma 2 *For any $k \in (-\infty, +\infty)$, $\mathbb{M}^k(P(a|b), P(b|a))$ always satisfies the fundamental properties P1 through P4 and the null-invariance property as well.*

Proof Both $P(a|b)$ and $P(b|a)$ have range $[0, 1]$. Thus the generalized mean of these two conditional probabilities must have a range of $[0, 1]$ based on Eq. (1). Also, $\mathbb{M}^k(P(a|b), P(b|a))$ is a monotone function with respect to $sup(a)$, $sup(b)$, and $sup(ab)$. When $sup(ab)$ increases while $sup(a)$ and $sup(b)$ are kept fixed, both $P(a|b)$ and $P(b|a)$ will also increase and result in a larger measure value. This corresponds to the case where the association relationship between event a and event b becomes positively stronger because they tend to have more co-occurrences. When $sup(a)$ and/or $sup(b)$ increase and $sup(ab)$ remain fixed, the association between a and b would become smaller in that $P(b|a)$ and/or $P(a|b)$ will decrease. This corresponds to the case where the co-occurrences of a and b become relatively fewer when the support count of a single event increases. Furthermore, $\mathbb{M}^k(P(a|b), P(b|a))$ is invariant to event permutation because the generalized mean has a natural symmetric expression in terms of input a and b, *i.e.*, $\mathbb{M}^k(P(a|b), P(b|a)) = \mathbb{M}^k(P(b|a), P(a|b))$. The generalized expression is also invariant to scaling and null-transactions, which can be derived from the definition. □

Lemma 3 *Given any two events a and b, we have*

$$AllConf(a, b) \leq Coherence'(a, b) \leq Cosine(a, b) \leq Kulc(a, b) \leq MaxConf(a, b). \tag{6}$$

Proof Given any exponents $k \in (-\infty, +\infty)$ and $k' \in (-\infty, +\infty)$ $(k < k')$, we have $\mathbb{M}^k(P(a|b), P(b|a)) \leq \mathbb{M}^{k'}(P(a|b), P(b|a))$, where the equality holds if and only if $P(a|b) = P(b|a)$. The detailed proof of this statement can be found in Kachigan (1991). □

These two lemmas provide insights into both sides of the coin. The first lemma provides a general justification to the *common*, desired properties of the null-invariant association measures for mining small probability events, whereas the second lemma presents an organized picture of the *differences* between them.

The total ordering of the null-invariant measures clearly exhibits their relationships. First, higher-order (*i.e.*, with larger k) measures provide an upper-bound to lower-order (*i.e.*, with smaller k) measures. Therefore, given a fixed association threshold (*e.g.*, 0.9), the patterns output by a higher-order measure must be a superset of the patterns output by a lower-order measure. This is helpful to interesting null-invariant pattern mining, because computationally expensive association measures such as *Cosine* that involves square root computation, is bounded by computationally cheaper measures like *Kulc*, which can be pushed deep into the mining process. We will discuss extensively the pattern mining problem in the next section. Intuitively, a lower-order measure is more strict (*i.e.*, prune more patterns), because a small k tends to mitigate the impact of the larger one of the two conditional probabilities, whereas a large k tends to aggravate it.

While the generalized mean can be used to represent a family of null-invariant association measures, there is no universally accepted measure for association analysis in large databases, because no particular value of k is generally "better" than the others.

Thus the selection of association measure should be considered on a case-by-case basis, where an appropriate value of k should be determined. It is worth mentioning that each of the five null-invariant measures being examined corresponds to a special case of the whole spectrum of exponent k. In particular, *AllConf* ($k \to -\infty$) and *MaxConf* ($k \to +\infty$) correspond to the minimum and maximum of the conditional probabilities, whereas *Coherence*$'$ ($k = -1$), *Cosine* ($k \to 0$), and *Kulc* ($k = 1$) correspond to the *harmonic mean*, *geometric mean*, and *arithmetic mean* of the conditional probabilities.

2.2 Multiple events

Our extension of the measure to multiple events is based on the same philosophy for two events. In order to preserve the fundamental properties and take the "generalized mean" approach for balancing conditional probabilities, we have the following definition.

Definition 1 (*Generalized null-invariant measure*) Let X be an event-set containing n ($n \geq 2$) events $\{a_1, a_2, \ldots, a_n\}$, and then we have

$$\begin{aligned}\mathbb{M}^k(X) &= \mathbb{M}^k\left(P(a_2, \ldots, a_n|a_1), \ldots, P(a_1, \ldots, a_{n-1}|a_n)\right) \\ &= \sqrt[k]{\frac{sup(X)^k}{n}\left(\frac{1}{sup(a_1)^k} + \cdots + \frac{1}{sup(a_n)^k}\right)},\end{aligned}$$

where $P(a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n|a_i) = sup(X)/sup(a_i)(1 \leq i \leq n)$.

The generalized null-invariant measure is the generalized mean of the conditional probabilities of all events. For a fixed pattern X, the total ordering among measures with different exponents still applies to the extension in that the smaller k is, the smaller value the measure will produce. It is worth mentioning that *AllConf* has been previously defined on more than two events (Omiecinski 2003; Lee et al. 2003), which is $AllConf(X) = sup(X)/max\{sup(a_1), \ldots, sup(a_n)\}$, and $\mathbb{M}^k(X)$ is able to be instantiated to accommodate the previous definition of *AllConf* based on the following equation.

$$\lim_{k \to -\infty} \mathbb{M}^k(X) = AllConf(X) = \min_{1 \leq i \leq n}\left\{\frac{sup(X)}{sup(a_i)}\right\}. \tag{7}$$

Similar to pairwise association, the total ordering still holds with respect to multiple events. Given an event-set of arbitrary length, its generalized association measure value, $\mathbb{M}^k(X)$, would monotonically increase when the exponent k is increased. Thus, we generalize Lemma 1 to the following lemma.

Lemma 4 *Given any event-set X with an arbitrary length $n \geq 2$, we have that $\mathbb{M}^{k_1}(X) \leq \mathbb{M}^{k_2}(X)$ holds if and only if $k_1 \leq k_2$.*

Proof The proof on event-sets with arbitrary length is similar to that for Lemma 1. We refer to Polya et al. (1998) for its general idea. □

2.3 The imbalance ratio measure

In many applications it is important to quantify to what extent a data set is "controversial" in order for a data analyst to correctly understand the data. Therefore, as a complement of the null-invariant association measures, we propose a new measure called *Imbalance Ratio* to gauge the degree of imbalance between two events. Denote by $IR(a, b)$ the imbalance ratio of events a and b. Mathematically, we would like $IR(a, b)$ to satisfy the following set of fundamental, desirable properties:

- **Q1** $IR \in [0, 1]$; in other words, $IR = 0$ when the two conditional probabilities $P(a|b)$ and $P(b|a)$ are the same so that the case is perfectly balanced; on the other hand, $IR = 1$ when $|P(a|b) - P(b|a)| = 1$;
- **Q2** IR should be symmetric: $IR(a, b) = IR(b, a)$;
- **Q3** IR should be monotonically decreasing when $sup(ab)$ becomes larger, and $sup(a\bar{b})$ and $sup(\bar{a}b)$ are fixed; and
- **Q4** Let us keep $sup(ab)$ and $sup(a\bar{b})$ fixed. If $sup(\bar{a}b) \geq sup(a\bar{b})$, IR should be monotonically increasing when $sup(\bar{a}b)$ becomes larger. If $sup(\bar{a}b) \leq sup(a\bar{b})$, IR should be monotonically increasing when $sup(\bar{a}b)$ becomes smaller. Symmetrically, when keeping $sup(ab)$ and $sup(\bar{a}b)$ fixed, IR should have similar monotonicity with respect to $sup(a\bar{b})$.

We elaborate on these properties as follows. For $Q1$, when the conditional probabilities are identical and totally different, the *Imbalance Ratio* should be 0 and 1, respectively. $Q2$ is self-explanatory. $Q3$ follows from the fact that the degree of imbalance should decrease when a and b share more common event-sets. Finally, the intuition behind $Q4$ is that, when the gap between $sup(\bar{a}b)$ and $sup(a\bar{b})$ becomes larger due to an increase or decrease of either $sup(\bar{a}b)$ or $sup(a\bar{b})$, the *Imbalance Ratio* should be larger. As can be seen, these properties model the underlying principles an imbalance measure should have. They lead to the following definition of IR.

Definition 2 (*The imbalance ratio measure*) Given any two events a and b, define the imbalance ratio, $IR(a, b)$, to be

$$IR(a, b) = \frac{|sup(a\bar{b}) - sup(\bar{a}b)|}{sup(ab) + sup(a\bar{b}) + sup(\bar{a}b)} \tag{8}$$

$$= \frac{|P(a|b) - P(b|a)|}{P(a|b) + P(b|a) - P(a|b) \times P(b|a)}. \tag{9}$$

Lemma 5 *The imbalance ratio measure in Definition 2 satisfies properties $Q1$ through $Q4$.*

Proof For $Q1$, we can derive $P(a|b) = P(b|a) \Rightarrow IR(a, b) = 0$ directly from Eq. 9. Without loss of generality, assume $P(a|b) \leq P(b|a)$. Thus $|P(a|b) - P(b|a)| = 1 \Rightarrow P(a|b) = 0 \wedge P(b|a) = 1 \Rightarrow IR(a, b) = 1$. $Q2$ can be proved straightforwardly and $Q3$ can be derived from Eq. 8. $Q4$ can be proved by observing that, when fixing $sup(ab)$ and $sup(a\bar{b})$ and varying $sup(\bar{a}b)$ such that $|sup(a\bar{b}) - sup(\bar{a}b)|$ increases by Δ (> 0), the denominator of Eq. 8 would either increase by Δ or decrease by Δ, so

IR would become larger. Similar argument holds when fixing $sup(ab)$ and $sup(\overline{a}b)$ and varying $sup(a\overline{b})$. □

Indeed, there exists an infinite number of measures which are able to gauge the degree of imbalance between a and b while satisfying properties $Q1$ through $Q4$. However, our proposed *IR* measure has an intuitive expression: in Eq. 8, the numerator is the *difference* between the cardinalities of a and b's event-sets, whereas the denominator is the cardinality of the *union* of the event-sets of a and b.

It is worth mentioning, however, that a simple measure $|P(a|b) - P(b|a)|$, does not satisfy all the above properties. In particular, it violates property $Q3$ as illustrated using the following counterexample. We fix $sup(a\overline{b}) = 6$ and $sup(\overline{a}b) = 10$. When $sup(ab) = 3$, we have $|P(a|b) - P(b|a)| = |3/(3+6) - 3/(3+10)| = 0.10$; when $sup(ab) = 8$, we have $|P(a|b) - P(b|a)| = |8/(8+6) - 8/(8+10)| = 0.13$. Thus, the measure $|P(a|b) - P(b|a)|$ would consider the latter case more imbalanced. This may not be desirable since a and b share more common event-sets and tend to be more similar with each other in the latter case.

3 Efficient computation of interesting null-invariant patterns

The efficient computation of interesting patterns in large databases is an important problem in data mining. In many applications, users are interested in strongly positively associated event-sets. We formulate such computation problem as follows. Given a small probability event data set DB, an interestingness measure $\mathcal{M}$, a minimum support threshold θ ($0 \leq \theta \leq 1$), and a minimum association threshold γ ($0 \leq \gamma \leq 1$), our goal is to find $\mathbb{S}_{\mathcal{M}}$, the complete set of event-sets, so that for each event-set $X \in \mathbb{S}_{\mathcal{M}}$, $sup(X) \geq \theta$ and $\mathcal{M}(X) \geq \gamma$. Notice that the purpose of the support threshold θ is to help filter a large number insignificant or noisy patterns, which are often not interesting. If one need to ignore the support threshold, θ can be simply set to 0.

It has been shown in Lee et al. (2003) that *AllConf* is anti-monotonic, *i.e.*, if an event-set cannot pass the association threshold γ, none of its supersets will pass it. Generalized *MaxConf*, on the other hand, is monotonic, *i.e.*, if an event-set's association is no less than γ, all of its supersets will be no less than it. The (anti-)monotonicity properties can facilitate effective pruning and therefore the measures can be pushed into the existing frequent pattern mining methods (Agrawal and Srikant 1994; Han et al. 2000). In Lee et al. (2003), an efficient algorithm, CoMine, is developed for mining interesting *AllConf* or *Coherence* patterns (*i.e.*, $\mathcal{M}$ is *AllConf* or *Coherence*). Generalized *Cosine* and *Kulc*, on the other hand, do not have such nice property of monotonicity, and thus cannot be pushed into the existing pattern mining methods straightforwardly. To the best of our knowledge, the efficient computation of interesting patterns of *Cosine* and *Kulc* has not been studied before. So we focus on the two computationally challenging null-invariant measures, *Cosine* and *Kulc*, and propose an efficient algorithm, GAMiner, to discover interesting patterns with respect to the two measures, *i.e.*, $\mathcal{M}$ is *Cosine* or *Kulc*.

Our key idea here is to push the association computation process by exploiting the upper bound for both *Cosine* and *Kulc*. Given a seen pattern (*i.e.*, both its support and

association measure values have already been computed), we compute the maximum association value that can possibly be achieved by appending arbitrary events to this seen pattern. Then, any pattern with upper bound association value no more than the given threshold can be safely pruned from the mining process.

Before we describe the detailed upper bound computation process, observe that the *Cosine* measure, involving the n-th root computation, can be effectively bounded by leveraging the upper bound for *Kulc*, as shown in the following theorem.

Theorem 1 (Cosine bounding) *Given a fixed support threshold θ and a fixed association threshold γ, the complete set of patterns found by Kulc (denoted as $\mathbb{S}_{kulc}$) is a superset of those found by Cosine (denoted as $\mathbb{S}_{cosine}$). That is, $\mathbb{S}_{cosine} \subseteq \mathbb{S}_{kulc}$.*

Proof Following from the total ordering in Eq. (6), given any event-set $X \in \mathbb{S}_{cosine}$ which contains two events, we have $Kulc(X) \geq Cosine(X) \geq \gamma$, which means that $X \in \mathbb{S}_{kulc}$ must hold. When an event-set $X \in \mathbb{S}_{cosine}$ contains more than two events, the inequality also holds because of the definition in Eq. (7). □

The theorem implies that we can reduce the *Cosine* pattern mining problem to a *Kulc* pattern mining problem. Specifically, if we can efficiently obtain $\mathbb{S}_{kulc}$, the results for *Kulc*, we can check each event-set $X \in \mathbb{S}_{kulc}$ to see if $Cosine(X) \geq \gamma$ holds or not. The patterns which pass the *Cosine* threshold can be added into $\mathbb{S}_{cosine}$. This would guarantee that the complete set of *Cosine* patterns is found. As a result, the computation of $\mathbb{S}_{cosine}$ will be at least as efficient as $\mathbb{S}_{kulc}$.

Now we are ready to present the algorithm, GAMiner (**G**eneralized **A**ssociation Measure-based Pattern **Miner**), to efficiently mine the *Kulc* patterns (*i.e.*, $\mathcal{M}$ is *Kulc*). The algorithm extends the FP-Growth algorithm to support association pattern generation.

In a nutshell, FP-Growth can efficiently mine frequent patterns in the following way. First, FP-Growth builds an FP-tree, which serves as a concise representation of the input event-set database. All information regarding the input event-sets are completely stored inside the FP-tree. It then adopts a depth-first style strategy that recursively grows the current pattern by attaching one of the children items to it. Next, it projects the FP-tree on the current pattern (that is, the projected FP-tree represents all input event-sets conditioned on the current pattern) and recursively compute all frequent patterns of the projected tree.

Intuitively, our GAMiner algorithm grows patterns in a way similar to FP-Growth, and it also takes into consideration the association threshold γ. GAMiner computes the maximum *Kulc* measure value that the current pattern as well as all the children patterns can possibly achieve (*i.e.*, the maximum possible association along the current condition pattern). If the maximum possible *Kulc* measure value cannot pass the threshold γ, then it would be simply impossible to find any patterns by growing the current pattern. Therefore, the algorithm can stop growing the current pattern and terminate depth-first searching the current conditional FP-tree.

Specifically, during the pattern growth process, we identify the association upper bound of the current pattern and all its children patterns as follows. Suppose the current pattern is (*i.e.*, conditional pattern (Han et al. 2000)) $A = a_1 a_2 \ldots a_k (k \geq 1)$ which has just been grown. In A's conditional database (*i.e.*, the projection of the FP-tree

on A), assume that there exists at least one tree path $B = b_1 b_2 \dots b_m (m \geq 1)$. If no such path is left, then the projected FP-tree would be empty and nothing need to be pruned. Notice that although A must be a frequent pattern (otherwise the growth would not have happened), A might not be able to produce any pattern having a strongly positive association. However, by computing an association upper bound we are able to determine whether or not continue growing on A would yield interesting results. The rationale for the association upper bound is described in the theorem below.

Theorem 2 (Association upper bound for Kulc) *Suppose we have found in the current A's conditional FP-tree a both frequent and strongly positively associated n-pattern (an event-set containing n events)* $S = a_1 \cdots a_k b_{i_1} \cdots b_{i_{n-k}}$, *where* $k+1 \leq n \leq k+m$ *and* $1 \leq i_j \leq m (1 \leq j \leq n-k)$, *then the following inequality always holds:*

$$n \geq \frac{c}{1-\gamma},$$

where γ *is the minimum association threshold and c is a scaling factor.*

Proof By definition,

$$Kulc(S) = \frac{sup(S)}{n} \left(\sum_{l=1}^{k} \frac{1}{sup(a_l)} + \sum_{j=1}^{n-k} \frac{1}{sup(b_{i_j})} \right).$$

Furthermore, since any event-pattern must have support no larger than any of its subsets and the support of any event in the conditional FP-tree must be no less than that of the conditional pattern, we have

$$\begin{aligned} sup(S) &\leq sup(A) \\ &\leq \min\{sup(b_i) | b_i \text{ is in conditional FP-tree}\} \\ &= min\_sup\_b. \end{aligned}$$

Therefore,

$$\begin{aligned} &Kulc(S) \\ &\quad \leq \frac{min\_sup\_b}{n} \sum_{l=1}^{k} \frac{1}{sup(a_l)} + \frac{1}{n} \sum_{j=1}^{n-k} \frac{sup(S)}{sup(b_{i_j})} \\ &\quad \leq \frac{1}{n} \sum_{l=1}^{k} \frac{min\_sup\_b}{sup(a_l)} + \frac{n-k}{n} \\ &\quad \leq \frac{k-c}{n} + \frac{n-k}{n} \\ &\quad = 1 - \frac{c}{n}, \end{aligned}$$

where $c = \sum_{l=1}^{k} \left(1 - \frac{min\_sup\_b}{sup(a_l)}\right)$. Because $Kulc(S) \geq \gamma$, the lemma is proved. □

Note that this association upper bound is determined by pattern A and the length of the path B. Therefore, we could use the lower bound of pattern length to prune search space. If the sum of the length of pattern A and B is less than the lower bound, *i.e.*, $k + m < \lceil \frac{c}{1-\gamma} \rceil$, B can be safely pruned. More generally, we can keep track of the maximum depth of A's conditional subtree denoted as $tree\_depth$. If $k + tree\_depth < \lceil \frac{c}{1-\gamma} \rceil$, then the lower bound of pattern length is violated and thus we can immediately terminate the pattern growing process; that is, the whole conditional FP-tree of A can be be pruned.

We illustrate the generalized association measure-based pattern mining algorithm (GAMiner) in Fig. 1. Initially, the algorithm takes the global FP-tree constructed from the original transaction database as the input conditional FP-tree. The conditional pattern is initialized to be an empty pattern. Starting at the root, GAMiner enumerates each of the children events (Line 1) of the root and add it to the current pattern (Line 2). Since the structure of the FP-tree is able to guarantee the support of the current pattern Q passes the threshold θ, we only need to examine the true *Kulc* association value of Q to see if it qualifies to be an output pattern (Line 3). Next, the current pattern Q's conditional database is computed using the input FP-tree $Tree$ (Line 4). Given the conditional database, the variables, $min\_sup\_b, c, tree\_depth$, which will be used to derive the association upper bound of the current pattern Q, are computed (Line 5). GAMiner then checks if growing the current pattern by adding an event b_j would yield any interesting result (Lines 6–10). Specifically, a future pattern Qb_j must pass the support threshold θ (Lines 7–8) (otherwise none of its children pattern would pass it according to the monotonicity of the support measure), and it must also pass the association upper bound discussed earlier (Lines 9–10) (otherwise none of its children pattern would pass the association threshold γ). After all potentially uninteresting

Algorithm: GAMiner: Mining interesting association patterns through measure upper bounding.
Input: A database DB represented by a FP-Tree $Tree$, support threshold θ, association threshold γ, and an initially empty pattern P.
Output: The complete set of frequent and strongly associated patterns $\mathbb{S}_{kulc}$.

Procedure: GAMiner ($Tree$, θ, γ, P)
1 FOR each event a_i in the header of $Tree$ DO
2 Generate pattern $Q = P \cup a_i$;
3 IF $Kulc(Q) \geq \gamma$ THEN $\mathbb{S}_{kulc} = \mathbb{S}_{kulc} \cup Q$;
4 Get Q-projected database including a set I_Q of events;
5 Calculate min_sup_b, c, and $tree\_depth$ from the Q-projected database;
6 FOR each b_j in I_Q DO
7 IF $sup(Qb_j) < \theta$ THEN
8 Remove b_j from I_Q; //support pruning
9 IF $length(Qb_j) + tree\_depth < \lceil \frac{c}{1-\gamma} \rceil$ THEN
10 Remove b_j from I_Q; //association pruning
11 Construct Q-conditional FP-Tree $Tree_Q$ with events in I_Q;
12 IF $Tree_Q \neq \emptyset$ THEN
13 GAMiner ($Tree_Q$, θ, γ, Q);

Fig. 1 The GAMiner Algorithm

events have been removed, we construct Q's conditional FP-tree $Tree_Q$ using the pruned set of events, I_Q (Line 11). Finally, we recursively call the GAMiner process by taking the current pattern Q as the conditional pattern and $Tree_Q$ as the input conditional FP-tree (Lines 12–13). The output of the algorithm would be the complete set of patterns which satisfy both θ and γ.

4 Experimental evaluation

In this section, we evaluate both the unified framework of the null-invariant measures and the proposed pattern mining algorithm GAMiner. We evaluate them using two real data sets as well as a synthesized data set. Our methodology can be described as follows.

- *Two real data sets (DBLP and Web Search Log):* We take the DBLP data set and a Web search log data set as examples to discuss the underlying philosophy of the generalized measure. Also, the proposed *Imbalance Ratio* measure is evaluated on the DBLP data set;
- *Standard synthetic data set:* We then use the generator in Han et al. (2000) to generate a synthetic data set and evaluate the efficiency of GAMiner by comparing it to the state-of-the-art frequent pattern mining algorithm, FP-Growth.

All the experiments were done on a Pentium machine with 3.0 GHz CPU, 1GB of RAM, and 160G hard disk. All source code was written in C++ and compiled using Microsoft Visual C++ in Windows XP.

4.1 Event association on the DBLP data set

We choose the DBLP data set for our empirical evaluation of the generalized null-invariant measure because it is clean and the results are easy to interpret and understand. A collection of papers from the top data mining and database conferences including *KDD*, *SIGMOD*, and *VLDB* in the recent 10 years were extracted. We generated a small-probability-event database where each event-set corresponds to a conference paper and each event to an author of that paper. Notice that "an author publishing a particular paper" can be regarded as a small probability event because there are hundreds of thousands of authors in the database. We extract a set of 10 typical pairs of productive authors with at least 10 papers, and rank them according to their number of joint papers (*i.e.*, $sup(ab)$), as shown in Table 5. The columns ab, a, and b refer to $sup(ab)$, $sup(a)$, and $sup(b)$, respectively.

Notice that we choose the pairs of authors with unbalanced number of papers (as shown in the a and b columns) in order to highlight the differences between different null-invariant measures. Recall that an extremely balanced (hence less interesting in our study) case is that when $sup(a)$ is equal to $sup(b)$, all the null-invariant measures would reach consensus, producing exactly the same measure values. Moreover, since *AllConf* and *MaxConf* have a straightforward philosophy (*i.e.*, minimum and maximum of the conditional probability $P(a|b)$ and $P(b|a)$), we list only the measure value of the other three measures, *Coherence*, *Cosine*, and *Kulc* for ease of exposition.

Table 5 Experiment on the DBLP data set

ID	Author *a*	Author *b*	*ab*	*a*	*b*	*Coherence*	*Cosine*	*Kulc*	*IR*
1	Hans-Peter Kriegel	Martin Ester	28	146	54	0.163 (2)	0.315 (7)	0.355 (9)	0.53 (9)
2	Michael Carey	Miron livny	26	104	58	0.191 (1)	0.335 (4)	0.349 (10)	0.34 (10)
3	Hans-Peter Kriegel	Jorg Sander	24	146	36	0.152 (3)	0.331 (5)	0.416 (8)	0.70 (8)
4	Christos Faloutsos	Spiros Papadimitriou	20	162	26	0.119 (7)	0.308 (10)	0.446 (7)	0.81 (6)
5	Hans-Peter Kriegel	Martin Pfeifle	18	146	18	0.123 (6)	0.351 (2)	0.562 (2)	0.88 (2)
6	Hector Garcia-Molina	Wilburt Labio	16	144	18	0.110 (9)	0.314 (8)	0.500 (4)	0.86 (4)
7	Divyakant Agrawal	Wang Hsiung	16	120	16	0.133 (5)	0.365 (1)	0.567 (1)	0.87 (3)
8	Elke Rundensteiner	Murali Mani	16	104	20	0.148 (4)	0.351 (3)	0.477 (6)	0.78 (7)
9	Divyakant Agrawal	Oliver Po	12	120	12	0.100 (10)	0.316 (6)	0.550 (3)	0.90 (1)
10	Gerhard Weikum	Martin Theobald	12	106	14	0.111 (8)	0.312 (9)	0.485 (5)	0.85 (5)

It can be seen from the support count in the table that at least 3 pairs of authors ($ID = 5, 7, 9$) demonstrate advisor-advisee style relationship because $sup(a) \gg sup(b)$ and b *always* coauthors with a, but conversely, a, as an advisor, only coauthors a small portion of his/her papers with b. While *Kulc* shows relative preferences for such skewed patterns by ranking them the top-3, *Cosine* and *Coherence* rank "balanced" data (*i.e.*, pairs which do not demonstrate clear advisor-advisee relationship) higher. The author pairs ranked top 3 ($ID = 1, 2, 3$) by *Coherence* are considered to be the bottom 3 by *Kulc*, because these 3 pairs have relatively large $sup(ab)$ but the conditional probabilities are more balanced. The *Cosine* measure, as expected, stands in the middle of the other two: the top *Cosine* patterns ($ID = 5, 7, 8$) are ranked by *Coherence* as 4th, 5th, and 6th, and by *Kulc* as 1st, 2nd, and 6th. The same observation can be made to the bottom 3 patterns of *Cosine*. Notice that the first group of authors ($ID = 5, 7, 9$) and the second group ($ID = 1, 2, 3$) are two controversial cases that cannot be resolved by our common sense. That is, there exists no universally accepted way to compare the association of the two groups and different users may reach opposite conclusions.

Thus, the selection of a right association measure should be tailored to the specific need of a user: *Kulc* tends to give more credits to skewed patterns (*e.g.*, advisor-advisee relationships), *Coherence* prefers balanced patterns (*e.g.*, two comparable collaborators), and *Cosine* is in-between. *AllConf* and *MaxConf* on the other hand, would yield more extreme results by only looking at the weakest or strongest link.

4.2 Evaluation of the imbalance ratio on the DBLP data set

We now evaluate the performance of the proposed *Imbalance Ratio* measure on the DBLP data set. The last column of Table 5 displays the *IR* measure value for each pair of authors. We can see that *IR* indeed characterizes the degree of imbalance well for each pair. For example, the top-3 pairs with the highest *IR* values are $ID = 9, 5$, and 7, respectively, which are the only 3 tuples showing a perfect advisor-advisee relationship (*i.e.*, $P(a|b) = 1.0$). In contrast, the bottom-3 pairs with the lowest *IR* values are $ID = 2$, 1, and 3, which are clearly less skewed due to that $sup(ab)$ makes a relatively small portion of $sup(b)$.

Moreover, because these 10 pairs of authors are more or less skewed, we can see that the results produced by *IR* is highly correlated with *Kulc*'s results and less correlated with *Cosine* and *Coherence*. This matches our intuition since *Kulc* also favors skewed patterns more than *Cosine* and *Coherence* do. On the other hand, the difference between *IR* and *Kulc* can be demonstrated in the following. For example, the author pair with $ID = 9$ is ranked as the 1st and 3rd by *IR* and *Kulc*, respectively, whereas the author pair with $ID = 7$ is ranked 3rd and 1st, respectively. For $ID = 9$, we have $sup(ab) = sup(b) = 12$ and $sup(a) = 120$, and for $ID = 7$ we have $sup(ab) = sup(b) = 16$ and $sup(a) = 120$. From *IR*'s perspective, the former case is clearly more skewed because $sup(a)$ is relatively larger than $sup(b)$ and $sup(ab)$ (*i.e.*, intuitively, author a is relatively more prominent in the former case, therefore more different from author b). From *Kulc*'s perspective, the latter case has higher association because author a is more "dedicated" to author b. Such results verify that

IR can serve as a complimentary measure to further evaluate the degree of skewness of the data.

4.3 Event association on the search log data set

We further examine the null-invariant measures using a Web search log data obtained from a search engine. This log data consists of about $300K$ small probability event-sets corresponding to user search sessions, and $1M$ distinct events corresponding to various query terms and URLs visited. Thus, analyzing associations over such data set can naturally reveal strong associations between various user search activities (note that frequent pattern mining may not be effective any more since two queries (events) "weather" and "stock price" may co-occur many times but have no intrinsic correlation.

Table 6 displays 8 association patterns mined from the log, ordered descendingly according to the pattern's support (shown in the third column). All of the 8 patterns contain three events, where each event could be a query term like *"movies"* or a URL

Table 6 Association patterns obtained from a Web search log data

ID	Event-set	*sup*	*Coherence*	*Cosine*	*Kulc*
1	*"camero" (11166)* http://www.chevrolet.com (12269) http://microsite.chevrolet.com (3763)	1984	0.22 (5)	0.24 (2)	0.29 (5)
2	*"murcielago" (2893)* *"lamborghini" (9555)* *"lp640" (733)*	539	0.12 (7)	0.20 (7)	0.33 (3)
3	*"watch" (1378)* *"movies" (2188)* http://www.watchmovieslinks.net (1448)	380	0.23 (1)	0.23 (4)	0.24 (7)
4	*"fiat" (3700)* *"500" (3187)* http://www.fiat500.com (295)	294	0.12 (6)	0.19 (8)	0.39 (1)
5	*"kawasaki" (1278)* *"motorcycles" (1825)* http://www.kawasaki.com (791)	289	0.22 (3)	0.24 (3)	0.25 (6)
6	http://www.google.ro (4990) http://cautare.mobile.ro (527) http://anunturi.automarket.ro (513)	226	0.11 (8)	0.20 (6)	0.30 (4)
7	*"jon" (505)* *"kate" (565)* *"plus" (602)*	126	0.23 (2)	0.23 (5)	0.23 (8)
8	http://www.ebay.co.uk (937) http://signin.ebay.co.uk (153) http://my.ebay.co.uk (503)	117	0.22 (4)	0.28 (1)	0.37 (2)

like http://www.ebay.co.uk. We can see that all these patterns present some meaningful correlations between the events. For example, query terms *"watch"* and *"movies"* are highly associated with the URL http://www.watchmovieslinks.net. Note that a pattern having high association may or may not have frequent co-occurrence (*e.g.*, the case for $ID = 2$).

The last three columns in the table list the generalized *Coherence*, *Cosine*, and *Kulc* measure values for multiple events, along with the corresponding rank of each pattern among the 8 patterns, respectively. Clearly, these association measures tend to disagree with each other. To quantify the difference between the orderings and hence the measures, we apply to the table the well-known *Kendall's τ rank correlation coefficient*, defined as follows:

$$\tau = \frac{n_c - n_d}{n(n-1)/2},$$

where given n patterns and two orderings, n_c and n_d denote the number of concordant (*i.e.*, ordered in the same way) and discordant (*i.e.*, ordered reversely) pairs of patterns, respectively. We have $\tau = 1$ when two orderings are identical and $\tau = -1$ when they are the reverse of each other. As a result, for Table 6, we have

$$\begin{aligned} \tau(Coherence, Cosine) &= 0.21, \\ \tau(Coherence, Kulc) &= -0.5, \\ \text{and } \tau(Cosine, Kulc) &= -0.14. \end{aligned}$$

Thus, on this particular data set, the two "closest" measures *Coherence* and *Cosine* only have a rank correlation of 0.21. Indeed, in real applications, *Coherence*, *Cosine*, and *Kulc* can yield dramatically different orderings of patterns in a rigorous, mathematical sense. In Sect. 5.1 we will discuss several methods for selecting an appropriate null-invariant measure for association mining.

4.4 Experiments on interesting pattern mining algorithms

In this section we evaluate the performance of the proposed pattern mining algorithm, GAMiner. We compare GAMiner with the state-of-the-art frequent pattern mining algorithm FP-Growth, which only considers the support threshold θ. To ensure that FP-Growth is aware of the association threshold, it is modified so that the complete pattern set with respect to the support threshold is first obtained and then all patterns not qualifying for the association threshold are removed. For performance test, we use a synthetic data set generated in Han et al. (2000), which contains $100K$ event-sets and $10K$ different events, each event-set having an average of 50 events. All the algorithms run in memory and the total execution time is our primary measure of evaluation.

In Fig. 2, we plot the relations of user parameters versus algorithm execution time as well as the number of interesting patterns generated. Specifically, as depicted in Fig. 2a, we draw the relation between the threshold θ and the execution time in four settings, which are FP-Growth and GAMiner with $\gamma = 0.8$, 0.9, and 0.95, respectively.

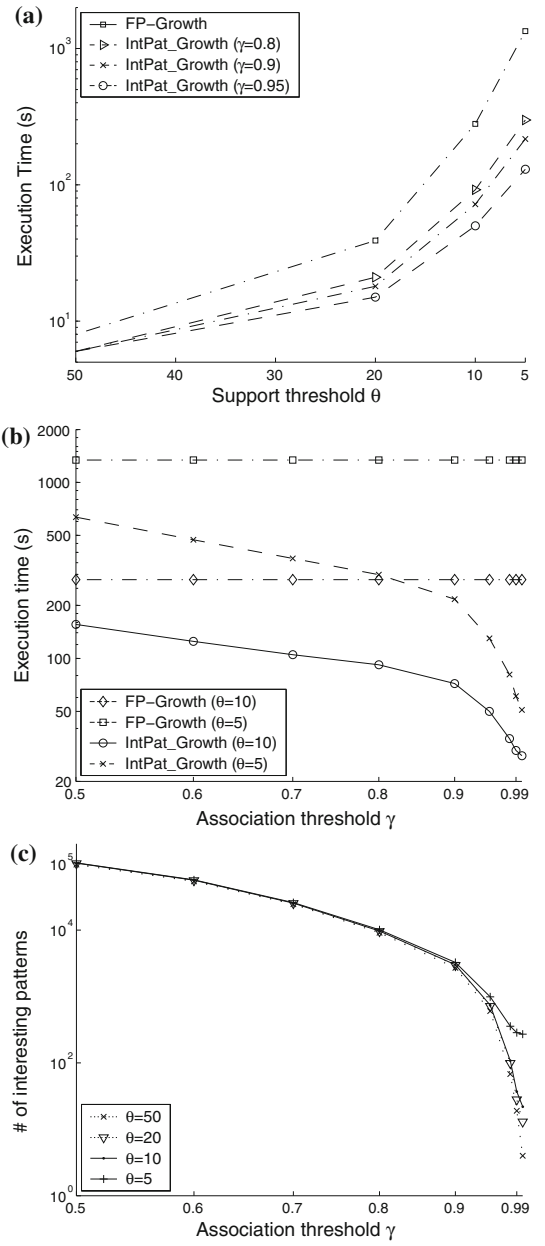

Fig. 2 Performance evaluation of GAMiner **a** support threshold θ vs. execution time **b** association threshold γ vs. execution time **c** association threshold γ and support threshold θ vs. the number of patterns

Notice that the execution time of FP-Growth is insensitive to the association threshold γ due to the way of modification discussed earlier. We can see that, as the support threshold decreases, the difference in computation time between GAMiner and different instances of FP-Growth becomes larger. Eventually GAMiner with $\gamma = 0.95$ outperforms FP-Growth by an order of magnitude. The reason that GAMiner runs much faster than FP-Growth lies in the fact that GAMiner is able to utilize the

association threshold and thereby achieving higher efficiency when the threshold is larger. Among the three instances of GAMiner (*i.e.*, $\gamma = 0.8, 0.9, 0.95$), GAMiner with $\gamma = 0.95$ consistently outperforms the other two. This justifies that the association upper bound derived in the previous section is indeed useful. It is worth mentioning that when the support threshold θ is large (*e.g.*, $\theta = 50$), the size of the pattern set tends to be dominated by the support threshold; in other words, most event-sets cannot pass it. As a result, the efficiency gain of GAMiner over FP-Growth may not be as high as in the case where the support threshold is relatively low.

The execution time with respect to the association threshold γ is shown in Fig. 2b. We plot four lines corresponding to FP-Growth with $\theta = 5$ and 10, and GAMiner with $\theta = 5$ and 10. For GAMiner, a greater γ value would lead to shorter execution time because the pruning power indicated by $\lceil \frac{c}{1-\gamma} \rceil$ (Fig. 1) is monotonically increasing with respect to γ. In contrast, FP-Growth remains unchanged for all γ values. Again, GAMiner consistently outperforms the baseline method. At $\theta = 10$ and $\gamma = 0.95$, the execution time of GAMiner is less than 1/26 of that of FP-Growth, while at $\theta = 5$ and $\gamma = 0.95$, the ratio of efficiency gain is 1/10. This shows the significant improvement of our proposed algorithm over FP-Growth.

To further explain the difference between the two algorithms, we show in Fig. 2c the relation between the association/support threshold with respect to the number of patterns in the final output set. The size of the output pattern set becomes exponentially smaller when the association threshold γ increases. This verifies the effectiveness of the association threshold that can prune a large number of uninteresting patterns. Also, Fig. 2b and c together demonstrate that the pruning power of the association upper bound is correlated with the output size. Moreover, when $\gamma \geq 0.5$ (*i.e.*, it requires a "positive" *Kulc* association of any output pattern), the number of output patterns is not that sensitive to the support threshold (here θ ranges from 5 to 50). This is in contrast to the steady shrinking size of the output pattern set with respect to γ, which indicates that the association threshold is a fundamentally different measure from the traditional support measure and cannot be replaced.

5 Discussion

5.1 Selecting an appropriate null-invariant measure

In this subsection we discuss two methods of selecting an appropriate null-invariant measure for some given application.

5.1.1 Rank-guided selection:

When a user is concerned about the ordering of the output association patterns, one can ask a domain expert to label the order of a few training patterns and then find the best exponent k so that the results produced can match the maximum number of the labeled patterns. Mathematically, this can be modeled as an optimization problem that aims at finding $k^* = \arg\max_k \tau(O_{train}, O_{\mathbb{M}^k})$, where O_{train} denotes the labeled ordering of patterns, $O_{\mathbb{M}^k}$ denotes the ordering of patterns produced using the

generalized measure with exponent k, and τ denotes the rank correlation coefficient discussed earlier. Because the search space for k can be infinite, it would be simply impossible to enumerate all k values. One may in turn narrow down k to the particular measures discussed in this paper (hence the search space would be finite), or conduct heuristic search to first find a good range for k and then further reduce this range by comparing samples k values within the range.

5.1.2 Threshold-guided selection:

A user may often intend to specify an absolute association threshold. This threshold would be useful when the total number of interesting patterns discovered is not known a priori, and a threshold such as $\gamma = 0.5$ can help filter out a large number of less associated patterns. In such cases, k can be heuristically selected based on the user's intention. If one would like to see only a small set of balanced, highly associated patterns, she may choose *AllConf* or *Coherence* as these patterns would have strong mutual associations, while skewed patterns would be pruned out. Conversely, if one wants to generate a larger set of association patterns to trade for a better "recall", *i.e.*, any patterns with a potential high association should be discovered, *Cosine* or *Kulc* could be a better choice.

5.2 Related work

Correlation and association mining is an important topic in data mining since it is emerging as an augment to frequent pattern mining methods in order to discover more interesting patterns in market basket analysis. Various existing metrics and newly proposed interestingness measures have been studied to facilitate association analysis (Savasere et al. 1998; Brin et al. 1997; Tan et al. 2002; Omiecinski 2003; Hilderman and Hamilton 2001).

There are several statistics-based association and correlation analysis methods. In Brin et al. (1997), a typical statistical measure χ^2 (Kachigan 1991) is introduced for correlation mining, which takes into consideration both the absence and presence of items. The measure in turn serves as a basis for discovering interesting association rules. In addition, TAPER (Xiong et al. 2004), an algorithm developed to efficiently answer all-strong-pairs correlation queries, is grounded on the *Pearson*'s correlation coefficient. Another line of work explores constraints in pattern mining (Grahne et al. 2000), where the interestingness measures like *Confidence* and *Lift* are used to assist in finding interesting association rules. In Omiecinski (2003), Lee et al. (2003) two new interestingness measures *AllConf* and *Coherence* are defined and studied based on the observation of a few desirable properties, such as the anti-monotonicity of the *AllConf* measure discussed earlier. There are also interestingness measures and metrics used in a variety of other fields and applications, such as information theory, Web search, chemistry applications, biological data analysis, and text mining. For instance, $H\text{-}Measure$ (He et al. 2004) is a correlation measure tailored specifically to Web applications for answering complex matching queries. Similarity metrics like *Cosine* distance function and *Jaccard coefficient* (as shown here *Coherence* is just

another form of *Jaccard coefficient*) have been very popularly used. *Kulc*, proposed in Kulczynski (1927), has been discussed in the chemistry literature (Bradshaw 2001).

An extensive investigation of the implications and connections between different measures has been done in Tan et al. (2002). The authors conduct a comparative and thorough study of a list of twenty-one existing interestingness measures. Specifically, the paper discusses three desirable properties and five other key properties to compare different measures. It is claimed that no measure is generally better than others because the intrinsic properties of the measures vary from one to another. Thus, one should match the desired properties of an application against the intrinsic properties of existing measures. The paper also discusses the problem of finding a proper measure using a small set of training data identified by an expert. In Wu et al. (2007), the authors further discuss the similarities and differences among a set of null-invariant measures. Our study can be viewed as a continued study of Tan et al. (2002), Wu et al. (2007) in the context of small probability events, which not only proposes the unified framework, but also addresses the problem of efficient pattern mining.

6 Conclusion

We have presented a comprehensive study of null-invariant interestingness measures for mining the associations among small probability events. We re-examined a set of five existing null-invariant measures and showed a generalization and the total ordering among them, which provides insights into the dichotomy between the measures' similarities and differences. Thus, a well-organized picture has been weaved to understand the underlying philosophy of the null-invariant measures. To the best of our knowledge, this is the first work to provide such a unified framework to the existing null-invariant measures. We also extended the definitions of measures to support multiple events based on the generalized mean approach, and we studied a complementary measure called *Imbalance Ratio* to quantify the degree of skewness of an event pair. Moreover, we proposed an efficient algorithm, GAMiner, to mine interesting *Kulc* and *Cosine* patterns. Experiments on both synthetic and real data sets verify the effectiveness and efficiency of our methods.

With this re-examination of null-invariant interesting measures, we believe it is crucial to choose the right interestingness measures at mining large data sets in which many events are essentially small probability events. Therefore, the re-examination of the interestingness measures in many advanced applications, including social network analysis, biomedical data mining, and pattern-based classification and clustering, is an important task for future research.

References

Agrawal R, Srikant R (1994, Sept) Fast algorithms for mining association rules. In: Proceedings of 1994 international conference on very large data bases (VLDB'94), Santiago, Chile, pp 487–499

Bradshaw J (2001) YAMS–Yet another measure of similarity. EuroMUG, http://www.daylight.com/meetings/emug01/bradshaw/similarity/YAMS.html

Brin S, Motwani R, Silverstein C (1997, May) Beyond market basket: generalizing association rules to correlations. In: Proceedings of 1997 ACM-SIGMOD international conference on management of data (SIGMOD'97), Tucson, AZ, pp 265–276

Brin S, Motwani R, Ullman JD, Tsur S (1997, May) Dynamic itemset counting and implication rules for market basket analysis. In: Proceedings of 1997 ACM-SIGMOD international conference on management of data (SIGMOD'97), Tucson, AZ, pp 255–264

Grahne G, Lakshmanan L, Wang X (2000, Feb) Efficient mining of constrained correlated sets. In: Proceedings of 2000 international conference on data engineering (ICDE'00), San Diego, CA, pp 512–521

Han J, Cheng H, Xin D, Yan X (2007) Frequent pattern mining: current status and future directions. Data Min Knowl Discov 15:55–86

Han J, Pei J, Yin Y (2000, May) Mining frequent patterns without candidate generation. In: Proceedings of 2000 ACM-SIGMOD international conference on management of data (SIGMOD'00), Dallas, TX, pp 1–12

He B, Chang KC-C, Han J (2004, Aug) Discovering complex matchings across web query interfaces: a correlation mining approach. In: Proceedings of 2004 ACM SIGKDD international conference on knowledge discovery in databases (KDD'04), Seattle, WA, pp 148–157

Hilderman RJ, Hamilton HJ (2001) Knowledge discovery and measures of interest. Kluwer Academic, Dordrecht

Kachigan S (1991) Multivariate statistical analysis: a conceptual introduction. Radius Press, New York

Kulczynski S (1927) Die pflanzenassoziationen der pieninen. In: Bulletin International de l'Académie Polonaise des Sciences et des Lettres, pp 57–203

Lee Y-K, Kim W-Y, Cai YD, Han J (2003, Nov) CoMine: efficient mining of correlated patterns. In: Proceedings of 2003 international conference on data mining (ICDM'03), Melbourne, FL, pp 581–584

Omiecinski E (2003) Alternative interest measures for mining associations. IEEE Trans Knowl Data Eng 15:57–69

Polya G, Harold HG, Littlewood JE (1988) Inequalities. Cambridge University Press, Cambridge

Savasere A, Omiecinski E, Navathe S (1998, Feb) Mining for strong negative associations in a large database of customer transactions. In: Proceedings of 1998 international conference on data engineering (ICDE'98), Orlando, FL, pp 494–502

Tan P-N, Kumar V, Srivastava J (2002, July) Selecting the right interestingness measure for association patterns. In: Proceedings of 2002 ACM SIGKDD international conference on knowledge discovery in databases (KDD'02), Edmonton, Canada, pp 32–41

Wu T, Chen Y, Han J (2007, Sept) Association mining in large databases: a re-examination of its measures. In: Proceedings of 2007 international conference on principles and practice of knowledge discovery in databases (PKDD'07), Warsaw, Poland, pp 621–628

Xiong H, Shekhar S, Tan P-N, Kumar V (2004, Aug) Exploiting a support-based upper bound of Pearson's correlation coefficient for efficiently identifying strongly correlated pairs. In: Proceedings of 2004 ACM SIGKDD international conference on knowledge discovery in databases (KDD'04), Seattle, WA, pp 334–343