Chapter 2
Quality Measures in Pattern Mining

Abstract In this chapter different quality measures to evaluate the interest of the patterns discovered in the mining process are described. Patterns represent major features of data so their interestingness should be accordingly quantified by considering metrics that determine how representative a specific pattern is for the dataset. Nevertheless, a pattern can also be of interest for a user despite the fact that this pattern does not describe useful and intrinsic properties of data. Thus, any quality measure can be divided into two main groups: objective and subjective quality measures. Whereas objective measures describe statistical properties of data, subjective quality measures take into account both the data properties and external knowledge provided by the expert in the application domain.

2.1 Introduction

Pattern mining is defined as the process of extracting patterns of special interest from raw data [1]. Generally, the user does not have any information about data so any knowledge extracted from that data is completely new and the interest of the mined patterns is hardly quantifiable. Sometimes, though, the user previously knows what type of knowledge is useful to be obtained from data, and it makes possible to quantify the level of interest of the patterns [15] discovered by different pattern mining algorithms.

In general, patterns represent major features of data so their interest should be quantified by considering metrics that determine how representative a specific pattern is for the dataset [5, 11]. Good metrics should choose and rank patterns based on their potential interest to the user. Nevertheless, many application domains require specific knowledge to be discovered and any expert in the domain needs to quantify how promising a pattern is [26]. All of this give rise to the division of pattern mining metrics into two different groups: objective and subjective metrics.

Objective metrics are usually defined from a statistical point of view and they provide structural properties of data [27]. Some of these metrics have been widely used to evaluate the interest of association patterns [21], determining that the stronger is the dependence relationship, the more interesting is the pattern. Support, confidence, lift, and leverage are some of these objective metrics, which are defined in terms of the frequency of occurrence of the patterns.

S. Ventura, J.M. Luna, *Pattern Mining with Evolutionary Algorithms*,
DOI 10.1007/978-3-319-33858-3_2

As for the subjective metrics, they usually incorporate some users' knowledge in the application field. Blöttcher et al. [7] described that any association pattern is considered interesting if it is either actionable or unexpected. Actionability of a relationship means that the user might act upon it to his own advantage. Additionally, the unexpectedness refers to how far the discovered pattern contradicts the user's knowledge about the domain. Thus, most of the approaches for quantifying the subjective interest of a pattern require comparisons of the discovered knowledge with regard to the user's knowledge.

In general, most of the existing proposals for mining patterns and associations between patterns follow an optimization based on objective quality measures. This huge attraction for this type of metrics lies in the fact that pattern mining mainly aims at discovering hidden and previously unknown knowledge from datasets [14], so there is no possibility to compare the extracted knowledge with the subjective knowledge provided by the expert. Besides, the knowledge of two different users into a specific field can differ greatly, which causes inaccuracy in the metrics. Thus, the values obtained for different subjective metrics cannot be properly compared.

Finally, it should be noted that both objective and subjective measures can be used to select interesting rules. First, objective metrics serves as a kind of filter to select a subgroup of potentially interesting patterns and association between patterns. Second, subjective metrics might be used as a second filter to keep only those patterns that are truly interesting for both the user and the application field.

2.2 Objective Interestingness Measures

As described in the previous chapter, among the objectives of KDD (Knowledge Discovery in Databases) the production of patterns of interest plays one of the most important roles. In general terms, a pattern is an entity that should be valid, new and comprehensive [9]. Nevertheless, the mining of unknown patterns in a database might produce a large amount of different patterns, which causes a hardly post-process for the end user who needs to analyse and study each pattern individually. Besides, a large percentage of this set of patterns may be uninteresting and useless, so the end user has to face two different problems: the quantity and the quality of the rules [13]. To solve this issue, different quality measures based on the analysis of the statistical properties of data have been proposed by different authors [5, 26].

Let us consider an association rule $X \rightarrow Y$ defined from a pattern $P \subseteq I = \{i_1, i_2, \ldots, i_k\}$ obtained from a dataset, where X and Y are subset of P with no item in common, i.e. $\{X \subset P \wedge Y \subset P : X \cap Y = \emptyset, P \setminus X = Y, P \setminus Y = X\}$. The absolute frequencies for any association rule $X \rightarrow Y$ comprising an antecedent X and a consequent Y can be tabulated as shown in Table 2.1. We note n as the total number of transactions in the dataset. We also define n_x and n_y as the number of transactions that satisfies X and Y, respectively. n_{xy} is defined as the number of transactions that satisfies both X and Y, i.e. the number of transactions that satisfies the pattern P in which the association rule was defined. Additionally, $n_{x\bar{y}}$ is defined as the number

Table 2.1 Absolute frequencies for the antecedent X and consequent Y of any association rule of the form $X \rightarrow Y$

	Y	$\overline{Y}$	Σ
X	n_{xy}	$n_{x\overline{y}}$	n_x
$\overline{X}$	$n_{\overline{x}y}$	$n_{\overline{xy}}$	$n_{\overline{x}}$
Σ	n_y	$n_{\overline{y}}$	n

Table 2.2 Relative frequencies for the antecedent X and consequent Y of any association rule of the form $X \rightarrow Y$

	Y	$\overline{Y}$	Σ
X	p_{xy}	$p_{x\overline{y}}$	p_x
$\overline{X}$	$p_{\overline{x}y}$	$p_{\overline{xy}}$	$p_{\overline{x}}$
Σ	p_y	$p_{\overline{y}}$	1

of transactions that satisfies X but not Y, i.e. $n_{x\overline{y}} = n_x - n_{xy}$. Finally, $n_{\overline{x}y}$ is defined as the number of transactions that satisfies Y but not X, i.e. $n_{\overline{x}y} = n_y - n_{xy}$.

All these values can also be represented by considering the relative frequencies rather than the absolute ones. Thus, given the antecedent X and the number of transactions that it satisfies n_x, it is possible to calculate its relative frequency as $p_x = n_x/n$. The relative frequency of X describes, in per unit basis, the percentage of transactions satisfied by X. Table 2.2 illustrates the relative frequencies for a sample association rule $X \rightarrow Y$ comprising an antecedent X and a consequent Y. Analysing the relative frequencies, it should be noted that a rule is useless or misleading if $p_{xy} = 0$ since it does not represent any transaction. It could be caused by two different situations: (1) the antecedent X (or the consequent Y) does not satisfies any transaction within the dataset, so both X and Y does not have any transaction satisfied in common, i.e. $P_{xy} = 0$. (2) either the antecedent X and the consequent Y satisfy up to 50 % of the transactions within the dataset, i.e. $p_x \leq 0.5$ and $p_y \leq 0.5$, but they do not satisfy any transaction in common. In case the sum of the probabilities is greater than 1, then P_{xy} is always greater than 0, so it is impossible to have a probability of 0 if $p_x + p_y > 1$. In fact, it should be noted that the maximum value of P_{xy} is equal to the minimum value among P_x and P_y, i.e. $P_{xy} \leq Min\{P_x, P_y\}$. Figure 2.1 illustrates this behaviour, describing that in cases where $P_x + P_y \leq 1$ the value of P_{xy} is defined in the range $[0, Min\{P_x, P_y\}]$. On the contrary, in those situations where $P_x + P_y > 1$ the value of P_{xy} is defined in the range $[P_x + P_y - 1, Min\{P_x, P_y\}]$. All of this led us to the conclusion that when both P_x and P_y are maximum, i.e. $P_x = P_y = 1$, then P_{xy} is equal to 1.

The behaviour of both p_x and p_y with regard to p_{xy} is described by the equation $2 \times p_{xy} \leq p_x + p_y \leq 1 + p_{xy}$, which is graphically illustrated in Fig. 2.2. As shown, $P_{xy} > 0.5$ if and only if $P_x + Py > 1$. Considering the aforementioned equation, i.e. $P_{xy} \leq Min\{P_x, P_y\}$, P_{xy} will have a value greater or equal to 0.5 when $P_x + P_y \geq 1$.

As previously stated, P_{xy} is defined as the probability of occurrence of a pattern $P = \{X, Y\}$ that represent an association rule $X \rightarrow Y$ to be satisfied in a dataset. Here, P_x stands for the probability or relative frequency of the antecedent X of the rule, whereas P_y describes the relative frequency of the consequent Y of the rule. In pattern mining, this frequency of occurrence is widely known as the support [3] of

Fig. 2.1 P_{xy} may be 0 in situations where $P_x + P_y \leq 1$. On the contrary (example on the right), $P_{xy} > 0$ in situations where $P_x + P_y > 1$, taking a minimum value of $P_x + P_y - 1$

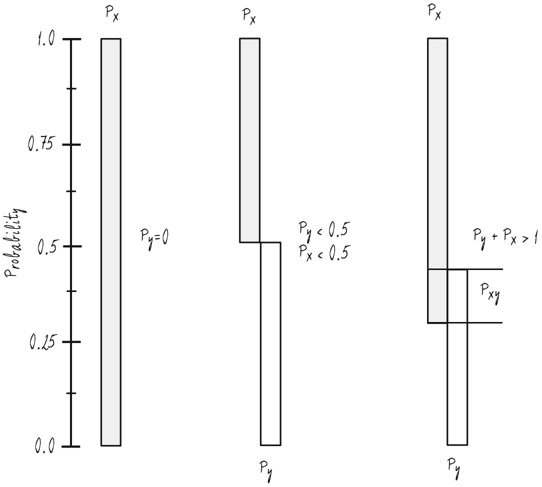
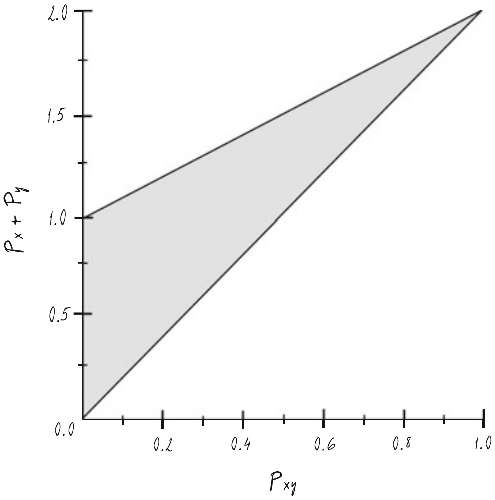


Fig. 2.2 Relationship between the probability of a rule P_{xy} and the sum of the probabilities of both the antecedent P_x and consequent P_y



the pattern, being one of the major quality measures used in this field. Hence, given an association rule of the form $X \rightarrow Y$, we define its frequency of occurrence as $support(X \rightarrow Y) \equiv P_{xy}$.

2.2.1 Quality Properties of a Measure

In 1991, Piatetsky-Shapiro [20] suggested that any quality measure $\mathscr{M}$ defined to quantify the interest of an association within a pattern should verify three specific properties in order to separate strong and weak rules so high and low values can be assigned, respectively. These properties can be described as follows:

Table 2.3 Properties satisfied by a set of different objective quality measures

Measure	Equation	Range	Property 1	Property 2	Property 3
Support	P_{xy}	[0, 1]	No	Yes	No
Coverage	P_x	[0, 1]	No	No	No
Confidence	P_{xy}/P_x	[0, 1]	No	Yes	No
Lift	$P_{xy}/(P_x \times P_y)$	[0, n]	Yes[a]	Yes	Yes
Leverage	$P_{xy} - (P_x \times P_y)$	[-0.25, 0.25]	Yes	Yes	Yes
Cosine	$P_{xy}/\sqrt{(P_x \times P_y)}$	[0, 1]	No	Yes	Yes
Conviction	$(P_x \times P_{\bar{y}})/P_{x\bar{y}}$	[0, ∞)	No	Yes	No
Gain	$(P_{xy}/P_x) - P_y$	[-1, 1)	Yes	Yes	Yes
Certainty factor	$((P_{xy}/P_x) - P_y)/(1 - P_y)$	[-1, 1]	Yes	Yes	Yes

[a]This property is satisfied just in case that the measure will be normalized

- Property 1: $\mathscr{M}(X \to Y) = 0$ when $P_{xy} = P_x \times P_y$. This property claims that any quality measure $\mathscr{M}$ should test whether X and Y are statistically independent.
- Property 2: $\mathscr{M}(X \to Y)$ monotonically increases with P_{xy} when P_x and P_y remain the same.
- Property 3: $\mathscr{M}(X \to Y)$ monotonically decreases with P_x or with P_y when other parameters remain the same, i.e. P_{xy} and P_x or P_y remain unchanged.

Following with the analysis of objective quality measures, all of them will be analysed by considering the properties defined by *Piatetsky-Shapiro*. Table 2.3 summarizes all the quality measures describes in this section, which will be described in depth. Beginning with the support quality measure, it should be noted that this measure does not satisfy first property since support$(X \to Y) \neq 0$ when $P_{xy} = P_x \times P_y$. For example, if $P_{xy} = P_x = P_y = 1$ then $P_{xy} = P_x \times P_y$, and support$(X \to Y) = 1$, so the first property defined by *Piatetsky-Shapiro* is not satisfied. Similarly, this quality measure does not satisfy the third property since support$(X \to Y) = P_{xy}$ so support measure cannot monotonically decrease when P_{xy} remains unchanged. Finally, it is interesting to note that a general belief in pattern mining is that the greater the support, the better the pattern discovered in the mining process. Nevertheless, this assertion should be taken with a grain of salt as it is only true to some extent. Indeed, patterns that appear in all the transactions are misleading since they do not provide any new knowledge about the data properties.

Similarly to support, there is a metric that calculates the generality [2] of a rule based on the percentage of transactions satisfied by the antecedent P_x of a rule, also known as body of a rule. This measure, known as coverage, determines the comprehensiveness of a rule. If a pattern characterizes more information in the dataset, it tends to be much more interesting. Nevertheless, this quality measure does not satisfy any of the properties described by *Piatetsky-Shapiro* as shown in Table 2.3. Besides, it is highly related to the support metric so most of the proposals have considered support instead of coverage.

As well as support, confidence is a quality measure that appears in most of the problems where the mining of association between patterns is a dare [29].

This quality measure determines the reliability or strength of implication of the rule, so the higher its value, the more accurate the rule is. In a formal way, the confidence measure is defined as the proportion of transactions that satisfy both the antecedent X and consequent Y among those transactions that contain only the antecedent X. This quality measure can be formally expressed as $\mathit{confidence}(X \rightarrow Y) = P_{xy}/P_x$, or as an estimate of the conditional probability $P(Y|X)$.

Support and confidence are broadly conceived as the finest quality measures in quantifying the quality of association rules and, consequently, a great variety of proposals make use of them [23]. These proposals attempt to discover rules whose support and confidence values are greater than certain thresholds. Nevertheless, many authors have considered that the mere fact of exceeding these quality thresholds does not guarantee that the rules are interesting at all [5]. For instance, the support-confidence framework does not provide a test for capturing the correlation of two patterns. Let us consider a real example obtained in [8], which described the rule IF *past active duty in military* THEN *no service in Vietnam*. This rule is calculated with a very high accuracy, which is quantified by the confidence value of 0.90. Thus, the rule suggests that knowing that a person served in military we should believe that he or she did not serve in Vietnam with a probability of 90 %. However, the item *no service in Vietnam* has a support of 0.95, so the probability that a person did not serve in Vietnam decreases (from 0.95 to 0.90) when we know he or she served in military. All of this led us to determine that the rule is misleading since the previously known information is more accurate than the probability obtained when more descriptive information is added.

Analysing whether the confidence quality measure satisfies or not the three properties provided by Piatetsky-Shapiro [20], we obtain that it only satisfies the second property (see Table 2.3). Let us consider the following probabilities: $P_{xy} = 1/3$, $P_x = 1/2$, and $P_y = 2/3$. The first property proposed by *Piatetsky-Shapiro* determines that if $P_{xy} = P_x \times P_y$ is satisfied by a measure $\mathscr{M}$, then this measure should satisfies that $\mathscr{M}(X \rightarrow Y) = 0$. Considering the aforementioned probabilities, $P_{xy} = P_x \times P_y = 1/3$ so the confidence will satisfy the first property if and only if $\mathit{confidence}(X \rightarrow Y) = 0$. Nevertheless, computing this value, we obtain that $\mathit{confidence}(X \rightarrow Y) = P_{xy}/P_x = 2/3 \neq 0$ so the confidence metric does not satisfy the property number one. Continuing with the second property, it is trivial to demonstrate that $\mathit{confidence}(X \rightarrow Y) = P_{xy}/P_x$ monotonically increases with P_{xy} when P_x remains the same, so this second property is satisfied by the confidence metric. Finally, the third property is partially satisfied by confidence measure since this measure does not include P_y, so we cannot state that third property is satisfied by confidence measure.

Support and confidence have been widely used in the mining of associations between patterns [29], and they are still considered by numerous authors that apply association to specific application fields [18, 22]. These two metrics are related as shown in Fig. 2.3, the shaded area illustrates the feasible area in which any association rule can obtain the values for the support and confidence measures. In order to understand the existing relation between these two quality measures,

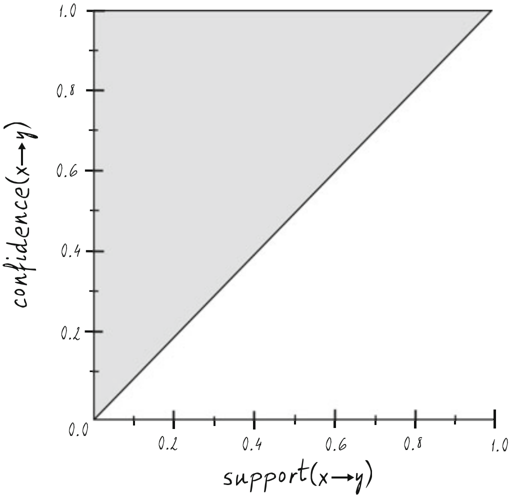


Fig. 2.3 Relationship between the support and the confidence measures

it should be noted that $P_{xy} \leq Min\{P_x, P_y\}$, and $P_{xy} \leq P_x$. Thus, given a value P_{xy}, then P_x will have a value in the range $P_x \in [P_{xy}, 1]$, so *confidence*$(X \to Y)$ is always greater or equal to *support*$(X \to Y)$.

Brin et al. [8] proposed a different metric to quantify the interest of the associations extracted in the mining process. This quality measure, which is known as lift, calculates the relationship between the confidence of the rule and the expected confidence of the consequent. Lift quality measure is described as *lift*$(X \to Y)$= $P_{xy}/(P_x \times P_Y) = $ *confidence*$(X \to Y)/P_y$. This measure calculates the degree of dependence between the antecedent X and the consequent Y of an association rule, obtaining a value < 1 if they are negative dependent; a value > 1 if they are positive dependent; and 1 in case of independence. As shown in Table 2.3, if *lift*$(X \to Y) = 1$, then $P_{xy} = P_x \times P_y$ so X and Y are independent and this measure satisfies the first property of Piatetsky-Shapiro [20] just in case that the measure will be normalized; if *lift*$(X \to Y) > 1$, then this measure monotonically increases with P_{xy} when other parameters remain the same (Property 2 of *Piatetsky-Shapiro*); finally, if *lift*$(X \to Y) < 1$, then lift monotonically decreases with P_x (or P_y) when other parameters remain the same (Property 3 of *Piatetsky-Shapiro*).

Let us consider the same real example described for the confidence measure and proposed by Brin et al. [8]. This example determined that the rule IF *past active duty in military* THEN *no service in Vietnam* has a confidence value of 0.90, and the support of the pattern *no service in Vietnam* is 0.95. In this regard, considering the same rule, the lift measure obtains a value of *lift* $= 0.90/0.95 = 0.947$, describing a negative dependence between the antecedent and consequent. Thus, this quality measure describes much more information than the one obtained by the confidence metric. Finally, it should be noted that most authors look for a positive correlation among the antecedent X and the consequent Y so only values greater than 1 are desired, that is, the confidence of the rule is greater than the support of the consequent.

Similarly to the lift measure, *Piatetsky-Shapiro* proposed a metric that calculates how different is the co-occurrence of the antecedent and the consequent from independence [16]. This quality measure is known as novelty or leverage [4], and is defined as $leverage(X \rightarrow Y) = P_{xy} - (P_x \times P_y)$. Leverage takes values in the range $[-0.25, 0.25]$, and its value is zero in those cases where the antecedent X and consequent Y are statistically independent, so values close to zero imply uninteresting rules. A important feature of this quality measure is that it satisfies the three properties proposed by *Piatetsky-Shapiro*. First, $leverage(X \rightarrow Y) = 0$ if $P_{xy} = P_x \times P_y$, so it satisfies the first property. Additionally, $leverage(X \rightarrow Y)$ monotonically increases with P_{xy} (property 2), and monotonically decreases with P_x or with P_y (property 3).

Another quality metric derived from lift is the IS measure, also known as cosine, which is formally defined as $IS(X \rightarrow Y) = \sqrt{Lift \times P_{xy}}$. As described by the authors [26], the IS measure presents many desiderable properties despite it does not satisfies the first property described by *Piatetsky-Shapiro*. First, it takes into account both the interestingness and the significance of an association rule since it contains two important quality measures in this sense, i.e. support and lift. Second, IS is equivalent to the geometric mean of confidence, i.e. $IS = \sqrt{Lift \times P_{xy}} = \sqrt{P_{xy}^2/(P_x \times P_y)}$, which can also be described as $IS(X \rightarrow Y) = \sqrt{confidence(X \rightarrow Y) \times confidence(Y \rightarrow X)}$. Finally, this quality measure can be described as the cosine angle, i.e. $IS(X \rightarrow Y) = P_{xy}\sqrt{P_x \times P_y}$.

A major drawback of the lift quality measure is its bi-directional meaning, which determines that $lift(X \rightarrow Y) = lift(Y \rightarrow X)$ so it measures co-occurrence, not implication. It should be noted that the final aim of any association rule is the discovery and description of implications between the antecedent and consequent of the rule, so the direction of the implication is quite important and does not always reflect the same meaning and, therefore, cannot be measured with the same values. In this regard, the conviction quality measure (see Table 2.3) was proposed [8] as $conviction(X \rightarrow Y) = (P_x \times P_{\bar{y}})/P_{x\bar{y}}$. Conviction represents the degree of implication of a rule, and values far from the unity indicate interesting rules. According to the authors, this quality measure is useful for the following reasons:

- Conviction is related to both P_x and P_y, which is a great advantage with regard to confidence. As previously described, confidence only considers P_x so it may give rise to a confidence value that is smaller than the support of the consequent, i.e. a negative dependence between the antecedent and consequent.
- The value of this measure is always equal to 1 when the antecedent X and the consequent Y are completely unrelated. For example, a P_y value equal to P_{xy} produces a value of 1, which means independence between both X and Y.
- Unlike lift, rules which hold 100 % of the time have the highest possible conviction value. The confidence measure also has this property, providing a value of 1 (the maximum value for confidence) to these rules.
- Conviction is a measure of implication since it is directional, so it behaves differently to lift.

Considering the conviction quality measure, it should be noted that its main drawback lies in the fact that it is not a bounded measure, i.e. its range is [0, ∞]. This quality makes impossible to determine an optimum quality threshold, which is a really big handicap for the use of this metric.

A different quality measure that is based on both the confidence and the support of the consequent is defined as gain of a rule or relative accuracy [16]. This quality measure, which is formally defined as $gain(X \to Y) = confidence(X \to Y) - P_y$, satisfies the three properties described by *Piatetsky-Shapiro* as shown in Table 2.3. The gain measure can also be defined as $gain(X \to Y) = (P_{xy} - (P_x \times P_y))/P_x$ so when $P_{xy} = P_x \times P_y$, then $gain(X \to Y) = 0$. Thus, this metric satisfies the first property described provided by *Piatetsky-Shapiro*. Additionally, $gain(X \to Y)$ monotonically increases with P_{xy} (property 2), and monotonically decreases with P_x or P_y (property number 3).

Based on the same features provided by the gain measure, a different quality measure was defined in [24] as the gain normalized into the interval [-1,1]. This quality measure, known as certainty factor (CF), calculates the variation of the probability P_Y of the consequent of a rule when consider only those transactions satisfied by X. Formally, CF is defined as $CF(X \to Y) = gain(X \to Y)/(1 - P_y)$ if $gain(X \to Y) \geq 0$), and $CF(X \to Y) = gain(X \to Y)/P_y$ if $gain(X \to Y) < 0$).

2.2.2 Relationship Between Quality Measures

The number of objective quality measures is enormous [13], and there is no clear set of metrics appropriate for the pattern mining task. The aforementioned subset of measures (see Table 2.3) comprises different metrics widely used by many experts in the field [13, 27, 29], but we cannot define this group as the best or the optimum one. Considering the previously described group of measures, Fig. 2.4 illustrates how the different pairs of measures are related, which is really interesting to know the behaviour of the metrics. Here, each dot on the scatter plot represents one association rule from different datasets. The position of the dot on the scatterplot represents its A (measure in the x-axis) and B (measure in the y-axis) value. As a matter of example, it is interesting to note that the existing relationship between both support and confidence (see Fig. 2.4) for a varied set of different rules is the same as the one described in Fig. 2.3, which was demonstrated mathematically.

Analysing the properties described by Piatetsky-Shapiro [20], and considering the set of quality measures shown in Table 2.3, the three properties are satisfied by the following metrics: lift, leverage, gain and certainty factor. In this regard, it is interesting to analyse in depth these four quality measures, so the existing relationship between them and the support measure is described below. Despite the fact that support is not considered as a quality measure according to the properties of *Piatetsky-Shapiro*, it is the most well-known metric used in pattern mining [3] and any existing proposal in this field includes this metric. First, we analyse the existing relationship between support ($support(X \to Y) = P_{xy}$) and lift

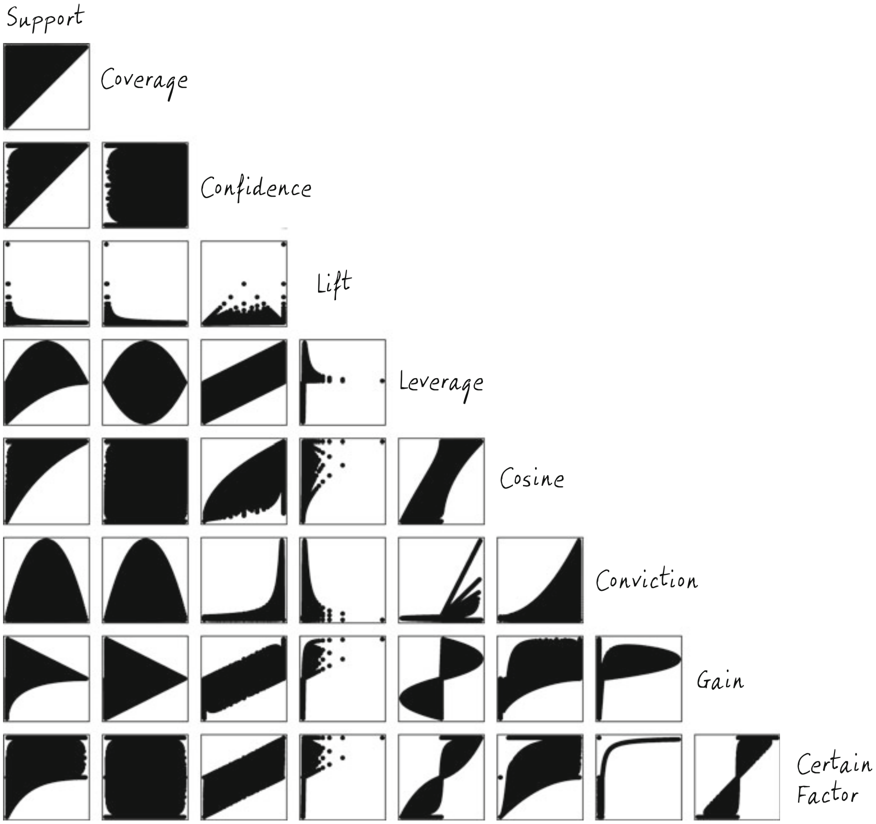


Fig. 2.4 Relationship between pairs of measures

($lift(X \rightarrow Y) = P_{xy}/(P_x \times P_y)$) as shown Fig. 2.5. As it is illustrated, the smaller the support value the higher the lift value, denoting a high positive dependence between the antecedent X and the consequent Y of an association rule. Nevertheless, if we just analyse the lift values in the range [0, 1], we discover that small support values also imply a high negative dependence between the antecedent X and the consequent Y. Finally, it should be noted that the higher the support the more independent are both X and Y, and a support value of 1 implies that $P_{xy} = P_x \times P_y$ so X and Y are statistically independent.

In a second analysis, we study the relationship between support and leverage, which is defined as $leverage(X \rightarrow Y) = P_{xy} - (P_x \times P_y)$ in the range $[-0.25, 0.25]$. It is quite interesting to note that, similarly to the lift quality measure, leverage denotes independence between X and Y for maximum support values (see Fig. 2.6). For negative leverage values, the behaviour is quite similar to the one obtained for lift. The main different lies on positive leverage values, where the maximum value is bounded. Noted that the upper bound in the lift quality measure is the total number n of transactions in the dataset, whereas in the leverage metric is 0.25 regardless the dataset used.

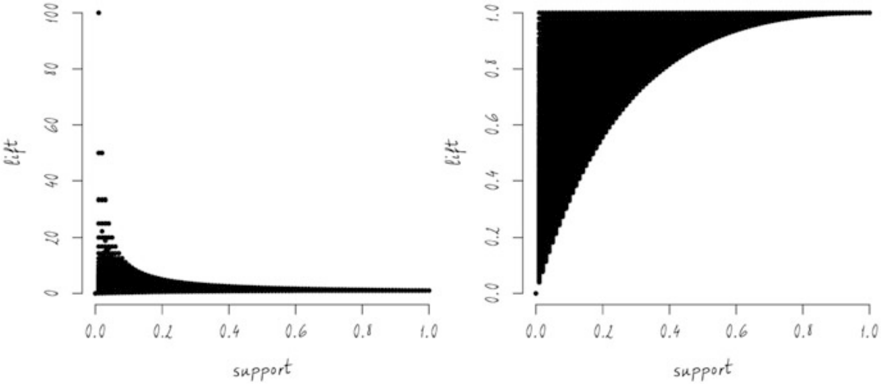
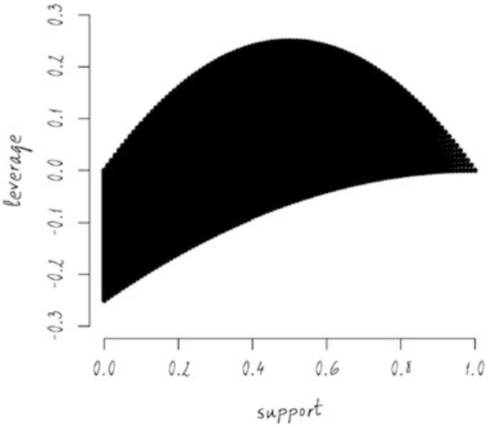


Fig. 2.5 Relationship between support and lift quality measures. Scatter plot on the right illustrates lift values lower than 1, whose distribution is completely different

Fig. 2.6 Relationship between support and leverage quality measures



Following with the same analysis and considering now support against gain, which is defined as $gain(X \rightarrow Y) = (P_{xy} - (P_x \times P_y))/P_x$, the existing relationship is shown in Fig. 2.7. Similarly to the leverage metric, the gain measure can obtain any value in a bounded range of values, i.e. $[-1, 1]$, which is a great advantage with regard to the lift metric. Furthermore, negative gain values describe a behaviour quite similar to the aforementioned metrics (lift and leverage). In fact, the behaviour described by this metric is quite similar to the one described by the lift. Two are the main differences: (1) the upper bound is delimited, which is a great advantage since it does not depend on the dataset under study and the rules can be properly quantified; (2) the value 0 implies independence between the antecedent X and the consequent Y.

Finally, we analyse the existing relationship between support and certainty factor (CF), which is illustrated in Fig. 2.8. This metric, which is defined as $CF(X \rightarrow Y) = ((P_{xy}/P_x) - P_y)/(1 - P_y)$, obtains negative values similarly to the other analysed

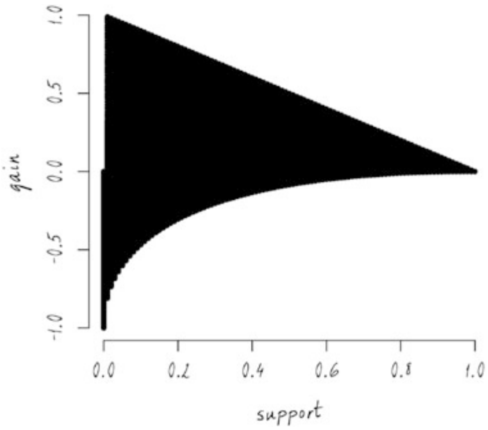


Fig. 2.7 Relationship between support and gain quality measures

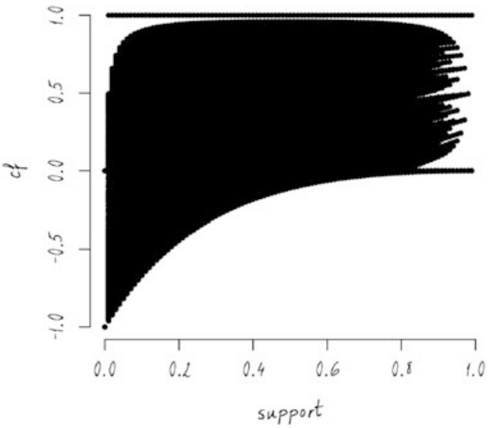


Fig. 2.8 Relationship between support and certainty factor (CF) quality measures

measures, i.e. lift, leverage and gain. Furthermore, a value of 0 for the CF measure implies an independence between X and Y. Another interesting feature of CF is its bounded range of values, which is the same as the one of gain, i.e. $[-1, 1]$. The main difference between CF and the other metrics lies on the fact that high support values can imply high CF values. Finally, it should be noted that maximum support values, i.e. $support(X \rightarrow Y) = 1$, produce a value $CF(X \rightarrow Y) = 0$, whereas maximum CF values can be obtained for any support value lower than 1 and greater than 0.

2.2.3 Other Quality Properties

Since the proposal described by Piatetsky-Shapiro [20], which determined three different properties to quantify the interest of any metric, many authors have studied other properties in this regard [27]. The first property (O1) is related to the symmetry

Table 2.4 Summary of properties satisfied by a set of different objective quality measures

Measure	Property 1	Property 2	Property 3	O1	O2	O3	O4
Support	No	Yes	No	Yes	No	No	No
Coverage	No	No	No	No	No	No	No
Confidence	No	Yes	No	No	No	No	Yes
Lift	Yes[a]	Yes	Yes	Yes	No	No	No
Leverage	Yes	Yes	Yes	Yes	Yes	Yes	No
Cosine	No	Yes	Yes	Yes	No	No	Yes
Conviction	No	Yes	No	No	No	Yes	No
Gain	Yes	Yes	Yes	No	No	No	No
Certainty factor	Yes	Yes	Yes	No	No	Yes	No

[a] This property is satisfied just in case that the measure will be normalized

Table 2.5 Sample relative frequencies for an association rule of the form $X \to Y$

	Y	$\overline{Y}$	Σ
X	0.55	0.05	0.60
$\overline{X}$	0.15	0.25	0.40
Σ	0.70	0.30	1.00

under variable permutation. In this sense, a measure $\mathscr{M}$ satisfies this property if and only if $\mathscr{M}(X \to Y) = \mathscr{M}(Y \to X)$. Otherwise, $\mathscr{M}$ is known as an asymmetric measure. As shown in Table 2.4, the symmetric measures considered in this study include support, lift, leverage and cosine. As for asymmetric measures, the following metrics are considered: coverage, confidence, conviction, gain and certainty factor. This type of measures are used for situations in which there is a need to distinguish between the rule $X \to Y$ and the rule $Y \to X$.

Let us consider the confidence quality measure, which was defined as P_{xy}/P_x, and the relative frequencies shown in Table 2.5. The confidence quality measure is computed as $confidence(X \to Y) = 0.55/0.60 = 0.9166$, so this quality measure does not satisfy the property O1, since $confidence(Y \to X) = 0.55/0.70 = 0.7857$ and, therefore, $confidence(X \to Y) \neq confidence(Y \to X)$. On the contrary, the lift quality measure $(P_{xy}/(P_x \times P_y))$ is considered as a symmetric measure since $lift(X \to Y) = lift(Y \to X)$. Considering again the relative frequencies shown in Table 2.5, it is obtained that $lift(X \to Y) = 0.55/(0.60 \times 0.70) = 1.3095 = lift(Y \to X)$.

The second property (O2) describes the antisymmetry under row/column permutation. A normalized measure $\mathscr{M}$ is antisymmetric under the row permutation operation if $\mathscr{M}(T')$=-$\mathscr{M}(T)$, considering T as the table of frequencies (see Table 2.2) and T' as the table of frequencies with a permutation on the rows; whereas the measure $\mathscr{M}$ is antisymmetric under the column permutation operation if $\mathscr{M}(T'') = -\mathscr{M}(T)$, considering T'' as the table of frequencies with a permutation on the columns. From the set of metrics used in this study, only the leverage or novelty satisfies the antisymmetry property under row/column permutation (see Table 2.4). According to the authors [27] measures that are symmetric under the row and column permutation

Table 2.6 Row permutation of the sample relative frequencies shown in Table 2.5

	Y	$\overline{Y}$	Σ
X	0.15	0.25	0.40
$\overline{X}$	0.55	0.05	0.60
Σ	0.70	0.30	1.00

Table 2.7 Row and column permutations of the sample relative frequencies shown in Table 2.5

	Y	$\overline{Y}$	Σ
X	0.25	0.15	0.40
$\overline{X}$	0.05	0.55	0.60
Σ	0.30	0.70	1.00

operations do not distinguish well between positive and negative correlations so it should be careful when using them to evaluate the interestingness of a pattern.

Considering the leverage quality measure, which was defined as $P_{xy} - (P_x \times P_y)$, and the relative frequencies shown in Table 2.5, it is possible to assert that this quality measure is antisymmetric under the row permutation (see Table 2.6). Noted that $leverage(X \to Y) = 0.55 - (0.60 \times 0.70) = 0.13$ on the original contingency table (Table 2.5), and $leverage(X \to Y) = 0.15 - (0.40 \times 0.70) = -0.13$ when using the row permutation (Table 2.6). Hence, it is possible to assert that leverage satisfies the second property under the row permutation.

A third property (O3) is related to inversion as a special case of the row/column permutation where both rows and columns are swapped simultaneously. This third property describes symmetric binary measures, which are invariant under the inversion operation. A measure $\mathscr{M}$ is a symmetric binary measure if $\mathscr{M}(T) = \mathscr{M}(T''')$, considering T as the table of frequencies (see Table 2.5) and T''' as the table of frequencies with a permutation on both rows and columns (see Table 2.7).

Taking again leverage as a quality measure that satisfies the third property (O3), it is possible to calculate $leverage(X \to Y) = 0.55 - (0.60 \times 0.70) = 0.13$ on the original contingency table (Table 2.5). On the contrary, when using the table of frequencies with a permutation on both rows and columns (see Table 2.7), the leverage value is calculated as $leverage(X \to Y) = 0.25 - (0.40 \times 0.30) = 0.13$, so it is demonstrated that the O3 property is satisfied. Finally, let us consider now the confidence measure as a quality measure that does not satisfy this property. Taking the original contingency table (Table 2.5), the confidence value is obtained as $confidence(X \to Y) = 0.55/0.60 = 0.9166$, whereas for the permutated table (see Table 2.7), the value obtained is $confidence(X \to Y) = 0.25/0.40 = 0.6250$, so this property is not satisfied by the confidence quality measure.

To complete this analysis, the null-invariant property (O4) is also included, which is satisfied by those measures that do not vary when adding more records that do not contain the two variables X and Y. This property may be of interesting in domains where co-presence of items is more important than co-absence. Table 2.8 shows the relative frequencies when more records that do not contain the two variables X and Y have been added to the original relative frequencies (see Table 2.5). Here, it is possible to demonstrate that confidence satisfies this property since

Table 2.8 Sample relative frequencies when adding more records that do not contain the two variables X and Y to relative frequencies shown in Table 2.5

	Y	$\overline{Y}$	Σ
X	0.275	0.025	0.300
$\overline{X}$	0.075	0.625	0.700
Σ	0.350	0.650	1.000

$confidence(X \to Y) = 0.55/0.60 = 0.9166$ on the original table of frequencies and this value remains the same for the new table of frequencies, i.e. $confidence(X \to Y) = 0.275/0.300 = 0.9166$.

2.3 Subjective Interestingness Measures

In previous section, we have described a set of quality measures that provide statistical knowledge about the data distribution. Many of these measures state for the interest and novelty of the knowledge discovered, but none of them describe the quality based on the background knowledge of the user [12]. Sometimes, this background knowledge is highly related to the application domain, and the use of objective measures does not provide useful knowledge.

Many authors have considered that the use of any subjective measure is appropriate only if one of the following conditions are satisfied [11]. First, the background knowledge of users varies due to the application field is highly prone to changes. Second, the interest of the users vary, so a pattern that can be of high interest in a quantum of time t_1 may be useless in a quantum of time t_2. Thus, subjective quality measures cannot be mathematically formulated as objective interestingness measures do.

Knowledge comprehensibility can be described as a subjective measure, in the sense that any pattern can be little comprehensible for a specific user and, at the same time, very comprehensible for a different one [10]. Nevertheless, many authors have considered this metric as an objective measure with a fixed formula: the fewer the number of items, the more comprehensible a pattern is. This concept of comprehensibility can also be applied to the number of patterns discovered [17]. It should be noted that, when an expert want to extract knowledge from a database, it is expected that the knowledge is comprehensible so to provide the user with tons of interesting patterns may be counter-productive. Thus, the fewer the number of patterns provided to the user, the more comprehensible the extracted knowledge is. Nevertheless, the comprehensibility problem is not an easy issue, and it is not only related to the length of the patterns or the number of patterns discovered in the mining process. Comprehensibility can be associated with subjective human preferences and the level of abstraction. As a matter of example, let us consider the representation of the invention of the first light bulb by Thomas Edison. At a low level of abstraction, it is possible to assert that Edison filed his first patent application on October 14th, 1878. Depending on the application domain, this date may be extremely accurate, so for some users it might be enough by describing that the invention of the first light bulb is dated in 1878. Even more, for a young students, it is perfectly fine to know that this invention was carried out in the nineteenth century. The three aforementioned representations are perfectly accurate, but depending on the human preferences, the last representation is more comprehensible than the first one.

The term unexpectedness has also been used in describing subjective quality measures [19]. A pattern is defined as unexpected and of high interest if it contradicts the user's expectations. Let us consider a dataset comprising information about students of a specific course, where the instructor previously knows that 80 % of the students passed the first test. It is expected that a huge percentage of the students also passed the second test according to the previous knowledge. Nevertheless, it is discovered that only 5 % of the students passed this second test. This pattern is completely unexpected to the instructor, denoting that something abnormal has occurred during the second lesson.

In subjective quality measures, actionability plays an important role [25]. A pattern is interesting if the user can do something with it or react to ti to his advantage. Considering the same example described before, which described that only 5 % of the students passed the second test whereas 80 % of these students passed the first test, it is possible to consider this pattern as actionable if it serves to improve both the teaching skills and the resources provided to students. This is a clear example of a pattern that is unexpected and actionable at the same time. Nevertheless, these two concepts are not always associated since an unexpected pattern might not be actionable if the results obtained do not depend on the user but on external systems.

Interactive data exploration is another subjective way of evaluating the quality of the extracted knowledge [28]. This novel task defines methods that allow the user to be directly involved in the discovery process. In interactive data exploration, patterns result to be much more relevant and interesting to the user [6] since he or she explores the data and identifies interesting structure through the visualization and interaction of different models. Nevertheless, this task has an enormous drawback since it can only be applied to relatively small datasets since the mining process cannot be tedious for the end user.

References

1. C. C. Aggarwal and J. Han. *Frequent Pattern Mining*. Springer International Publishing, 2014.
2. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.
3. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD Conference '93, pages 207–216, Washington, DC, USA, 1993.

4. J. L. Balcázar and F. Dogbey. Evaluation of association rule quality measures through feature extraction. In *Proceedings of the 12th International Symposium on Advances in Intelligent Data Analysis*, IDA 2013, pages 68–79, London, UK, October 2013.
5. F. Berzal, I. Blanco, D. Sánchez, and M. A. Vila. Measuring the Accuracy and Interest of Association Rules: A new Framework. *Intelligent Data Analysis*, 6(3):221–235, 2002.
6. M. Bhuiyan, S. Mukhopadhyay, and M. A. Hasan. Interactive pattern mining on hidden data: A sampling-based solution. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 95–104, New York, NY, USA, 2012. ACM.
7. M. Böttcher, G. Ruß, D. Nauck, and R. Kruse. From Change Mining to Relevance Feedback: A Unified View on Assessing Rule Interestingness Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction. In Y. Zhao, L. Cao, and C. Zhang, editors, *Information Science Reference*, pages 12–37. IGI Global, Hershey, New York, 2009.
8. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, SIGMOD '97, pages 255–264, Tucson, Arizona, USA, 1997. ACM.
9. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.
10. A. A. Freitas. *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag Berlin Heidelberg, 2002.
11. L. Geng and H. J. Hamilton. Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys*, 38, 2006.
12. B. Goethals, S. Moens, and J. Vreeken. MIME: A Framework for Interactive Visual Pattern Mining. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6913 of *Lecture Notes in Computer Science*, pages 634–637. Springer Berlin Heidelberg, 2011.
13. F. Guillet and H. Hamilton. *Quality Measures in Data Mining*. Springer Berlin / Heidelberg, 2007.
14. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.
15. A. Jiménez, F. Berzal, and J. C. Cubero. Interestingness measures for association rules within groups. In *Proceedings of the 13th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, IPMU 2010, pages 298–307. Springer, 2010.
16. N. Lavrač, P. A. Flach, and B. Zupan. Rule Evaluation Measures: A Unifying View. In *Proceedings of the 9th International Workshop on Inductive Logic Programming*, ILP '99, pages 174–185, London, UK, 1999. Springer-Verlag.
17. J. M. Luna, J. R. Romero, C. Romero, and S. Ventura. On the use of genetic programming for mining comprehensible rules in subgroup discovery. *IEEE Transactions on Cybernetics*, 44(12):2329–2341, 2014.
18. A. Merceron and K. Yacef. Interestingness measures for association rules in educational data. In *Proceedings of the 1st International Conference on Educational Data Mining*, EDM 2008, pages 57–66, Montreal, Canada, 2008.
19. B. Padmanabhan and A. Tuzhilin. Unexpectedness as a Measure of Interestingness in Knowledge Discovery. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 275–281, New York, NY, USA, 1999. AAAI Press.
20. G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI Press, 1991.
21. C. Romero, J. M. Luna, J. R. Romero, and S. Ventura. RM-Tool: A framework for discovering and evaluating association rules. *Advances in Engineering Software*, 42(8):566–576, 2011.
22. D. Sánchez, J. M. Serrano, L. Cerda, and M. A. Vila. Association Rules Applied to Credit Card Fraud Detection. *Expert systems with applications*, (36):3630–3640, 2008.

23. T. Scheffer. Finding association rules that trade support optimally against confidence. In *Proceedings of the 5th European Conference of Principles and Practice of Knowledge Discovery in Databases*, PKDD 2001, pages 424–435, Freiburg, Germany, 2001.
24. E. H. Shortliffe and B. G. Buchanan. A model of inexact reasoning in medicine. *Mathematical biosciences*, 23:351–379, 1975.
25. A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In *Proceedings of the 1st international conference on Knowledge Discovery and Data mining*, pages 275–281, Montreal, Quebec, Canada, 1995. AAAI Press.
26. P. Tan and V. Kumar. Interestingness Measures for Association Patterns: A Perspective. In *Proceedings of the Workshop on Postprocessing in Machine Learning and Data Mining*, KDD '00, New York, USA, 2000.
27. P. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.
28. M. van Leeuwen. Interactive data exploration using pattern mining. In A. H. Gandomi and Conor Alavi, A. H. Ryan, editors, *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, volume 8401 of *Lecture Notes in Computer Science*, pages 169–182. Springer Berlin Heidelberg, 2015.
29. C. Zhang and S. Zhang. *Association rule mining: models and algorithms*. Springer Berlin / Heidelberg, 2002.

Springer

http://www.springer.com/978-3-319-33857-6

Pattern Mining with Evolutionary Algorithms
Ventura, S.; Luna, J.M.
2016, XIII, 190 p. 126 illus., 4 illus. in color., Hardcover
ISBN: 978-3-319-33857-6