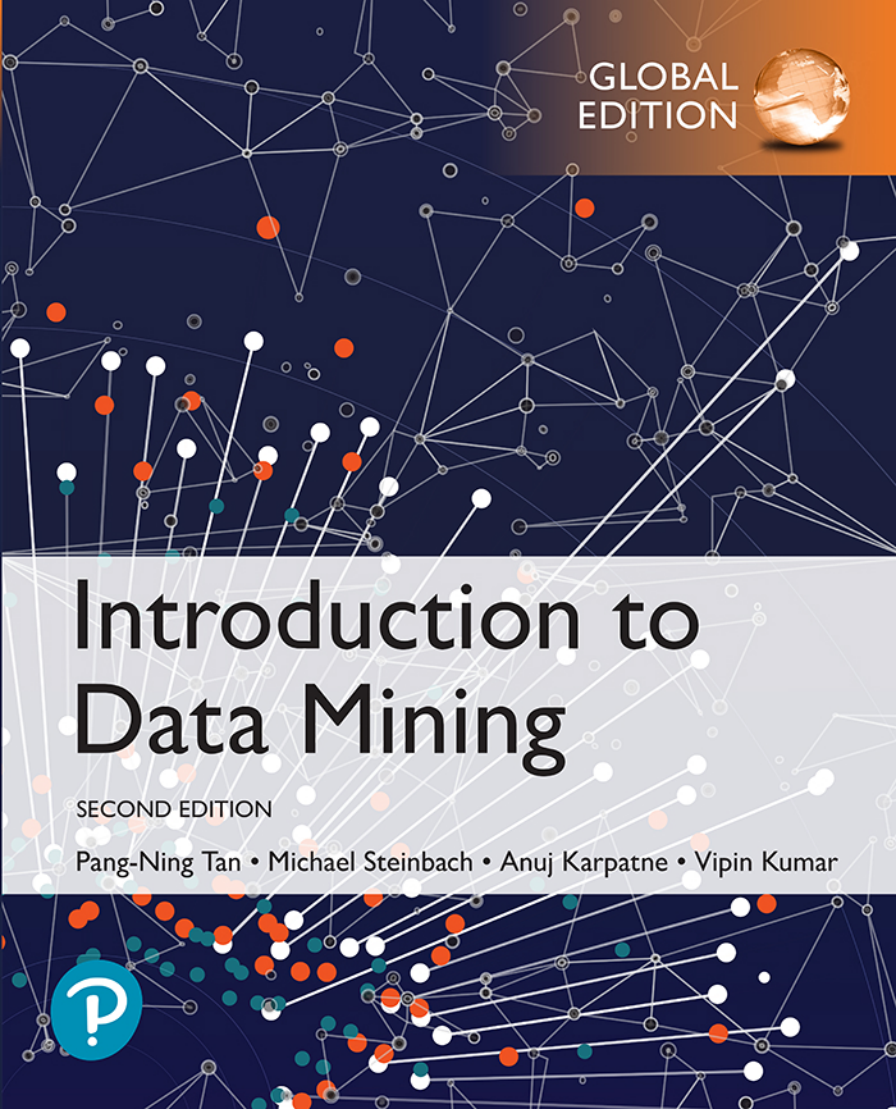

Introduction to Data Mining

SECOND EDITION

Pang-Ning Tan • Michael Steinbach • Anuj Karpatne • Vipin Kumar

INTRODUCTION TO DATA MINING

Table 4.7. Beverage preferences among a group of 1000 people.

	$Coffee$	$\overline{Coffee}$	
Tea	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

(75%) values are reasonably high. This argument would have been acceptable except that the fraction of people who drink coffee, regardless of whether they drink tea, is 80%, while the fraction of tea drinkers who drink coffee is only 75%. Thus knowing that a person is a tea drinker actually decreases her probability of being a coffee drinker from 80% to 75%! The rule $\{Tea\} \longrightarrow \{Coffee\}$ is therefore misleading despite its high confidence value.

Table 4.8. Information about people who drink tea and people who use honey in their beverage.

	$Honey$	$\overline{Honey}$	
Tea	100	100	200
$\overline{Tea}$	20	780	800
	120	880	1000

Now consider a similar problem where we are interested in analyzing the relationship between people who drink tea and people who use honey in their beverage. Table 4.8 summarizes the information gathered over the same group of people about their preferences for drinking tea and using honey. If we evaluate the association rule $\{Tea\} \longrightarrow \{Honey\}$ using this information, we will find that the confidence value of this rule is merely 50%, which might be easily rejected using a reasonable threshold on the confidence value, say 70%. One thus might consider that the preference of a person for drinking tea has no influence on her preference for using honey. However, the fraction of people who use honey, regardless of whether they drink tea, is only 12%. Hence, knowing that a person drinks tea significantly increases her probability of using honey from 12% to 50%. Further, the fraction of people who do not drink tea but use honey is only 2.5%! This suggests that there is definitely some information in the preference of a person of using honey given that she

drinks tea. The rule $\{Tea\} \longrightarrow \{Honey\}$ may therefore be falsely rejected if confidence is used as the evaluation measure.

■

Note that if we take the support of coffee drinkers into account, we would not be surprised to find that many of the people who drink tea also drink coffee, since the overall number of coffee drinkers is quite large by itself. What is more surprising is that the fraction of tea drinkers who drink coffee is actually less than the overall fraction of people who drink coffee, which points to an inverse relationship between tea drinkers and coffee drinkers. Similarly, if we account for the fact that the support of using honey is inherently small, it is easy to understand that the fraction of tea drinkers who use honey will naturally be small. Instead, what is important to measure is the change in the fraction of honey users, given the information that they drink tea.

The limitations of the confidence measure are well-known and can be understood from a statistical perspective as follows. The support of a variable measures the probability of its occurrence, while the support $s(A, B)$ of a pair of a variables A and B measures the probability of the two variables occurring together. Hence, the joint probability $P(A, B)$ can be written as

$$P(A, B) = s(A, B) = \frac{f_{11}}{N}.$$

If we assume A and B are statistically independent, i.e. there is no relationship between the occurrences of A and B, then $P(A, B) = P(A) \times P(B)$. Hence, under the assumption of statistical independence between A and B, the support $s_{\text{indep}}(A, B)$ of A and B can be written as

$$s_{\text{indep}}(A, B) = s(A) \times s(B) \quad \text{or equivalently,} \quad s_{\text{indep}}(A, B) = \frac{f_{1+}}{N} \times \frac{f_{+1}}{N}. \tag{4.4}$$

If the support between two variables, $s(A, B)$ is equal to $s_{\text{indep}}(A, B)$, then A and B can be considered to be unrelated to each other. However, if $s(A, B)$ is widely different from $s_{\text{indep}}(A, B)$, then A and B are most likely dependent. Hence, any deviation of $s(A, B)$ from $s(A) \times s(B)$ can be seen as an indication of a statistical relationship between A and B. Since the confidence measure only considers the deviance of $s(A, B)$ from $s(A)$ and not from $s(A) \times s(B)$, it fails to account for the support of the consequent, namely $s(B)$. This results in the detection of spurious patterns (e.g., $\{Tea\} \longrightarrow \{Coffee\}$) and the rejection of truly interesting patterns (e.g., $\{Tea\} \longrightarrow \{Honey\}$), as illustrated in the previous example.

Various objective measures have been used to capture the deviance of $s(A, B)$ from $s_{\text{indep}}(A, B)$, that are not susceptible to the limitations of the confidence measure. Below, we provide a brief description of some of these measures and discuss some of their properties.

Interest Factor The interest factor, which is also called as the "lift," can be defined as follows:

$$I(A, B) = \frac{s(A, B)}{s(A) \times s(B)} = \frac{N f_{11}}{f_{1+} f_{+1}}. \tag{4.5}$$

Notice that $s(A) \times s(B) = s_{\text{indep}}(A, B)$. Hence, the interest factor measures the ratio of the support of a pattern $s(A, B)$ against its baseline support $s_{\text{indep}}(A, B)$ computed under the statistical independence assumption. Using Equations 4.5 and 4.4, we can interpret the measure as follows:

$$I(A, B) \begin{cases} = 1, & \text{if } A \text{ and } B \text{ are independent;} \\ > 1, & \text{if } A \text{ and } B \text{ are positively related;} \\ < 1, & \text{if } A \text{ and } B \text{ are negatively related.} \end{cases} \tag{4.6}$$

For the tea-coffee example shown in Table 4.7, $I = \frac{0.15}{0.2 \times 0.8} = 0.9375$, thus suggesting a slight negative relationship between tea drinkers and coffee drinkers. Also, for the tea-honey example shown in Table 4.8, $I = \frac{0.1}{0.12 \times 0.2} = 4.1667$, suggesting a strong positive relationship between people who drink tea and people who use honey in their beverage. We can thus see that the interest factor is able to detect meaningful patterns in the tea-coffee and tea-honey examples. Indeed, the interest factor has a number of statistical advantages over the confidence measure that make it a suitable measure for analyzing statistical independence between variables.

Piatesky-Shapiro (PS) Measure Instead of computing the ratio between $s(A, B)$ and $s_{\text{indep}}(A, B) = s(A) \times s(B)$, the PS measure considers the difference between $s(A, B)$ and $s(A) \times s(B)$ as follows.

$$PS = s(A, B) - s(A) \times s(B) = \frac{f_{11}}{N} - \frac{f_{1+} f_{+1}}{N^2} \tag{4.7}$$

The PS value is 0 when A and B are mutually independent of each other. Otherwise, $PS > 0$ when there is a positive relationship between the two variables, and $PS < 0$ when there is a negative relationship.

Correlation Analysis Correlation analysis is one of the most popular techniques for analyzing relationships between a pair of variables. For continuous variables, correlation is defined using Pearson's correlation coefficient (see Equation 2.10 on page 103). For binary variables, correlation can be measured using the ϕ-coefficient, which is defined as

$$\phi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}. \tag{4.8}$$

If we rearrange the terms in 4.8, we can show that the ϕ-coefficient can be rewritten in terms of the support measures of A, B, and $\{A, B\}$ as follows:

$$\phi = \frac{s(A,B) - s(A) \times s(B)}{\sqrt{s(A) \times (1 - s(A)) \times s(B) \times (1 - s(B))}}. \tag{4.9}$$

Note that the numerator in the above equation is identical to the PS measure. Hence, the ϕ-coefficient can be understood as a normalized version of the PS measure, where that the value of the ϕ-coefficient ranges from -1 to $+1$. From a statistical viewpoint, the correlation captures the normalized difference between $s(A, B)$ and $s_{\text{indep}}(A, B)$. A correlation value of 0 means no relationship, while a value of $+1$ suggests a perfect positive relationship and a value of -1 suggests a perfect negative relationship. The correlation measure has a statistical meaning and hence is widely used to evaluate the strength of statistical independence among variables. For instance, the correlation between tea and coffee drinkers in Table 4.7 is -0.0625 which is slightly less than 0. On the other hand, the correlation between people who drink tea and people who use honey in Table 4.8 is 0.5847, suggesting a positive relationship.

IS Measure IS is an alternative measure for capturing the relationship between $s(A, B)$ and $s(A) \times s(B)$. The IS measure is defined as follows:

$$IS(A,B) = \sqrt{I(A,B) \times s(A,B)} = \frac{s(A,B)}{\sqrt{s(A)s(B)}} = \frac{f_{11}}{\sqrt{f_{1+}f_{+1}}}. \tag{4.10}$$

Although the definition of IS looks quite similar to the interest factor, they share some interesting differences. Since IS is the geometric mean between the interest factor and the support of a pattern, IS is large when both the interest factor and support are large. Hence, if the interest factor of two patterns are identical, the IS has a preference of selecting the pattern with higher support. It is also possible to show that IS is mathematically equivalent to the cosine

measure for binary variables (see Equation 2.6 on page 101). The value of IS thus varies from 0 to 1, where an IS value of 0 corresponds to no co-occurrence of the two variables, while an IS value of 1 denotes perfect relationship, since they occur in exactly the same transactions. For the tea-coffee example shown in Table 4.7, the value of IS is equal to 0.375, while the value of IS for the tea-honey example in Table 4.8 is 0.6455. The IS measure thus gives a higher value for the $\{Tea\} \longrightarrow \{Honey\}$ rule than the $\{Tea\} \longrightarrow \{Coffee\}$ rule, which is consistent with our understanding of the two rules.

Alternative Objective Interestingness Measures

Note that all of the measures defined in the previous section use different techniques to capture the deviance between $s(A, B)$ and $s_{\text{indep}}(A, B) = s(A) \times s(B)$. Some measures use the ratio between $s(A, B)$ and $s_{\text{indep}}(A, B)$, e.g., the interest factor and IS, while some other measures consider the difference between the two, e.g., the PS and the ϕ-coefficient. Some measures are bounded in a particular range, e.g., the IS and the ϕ-coefficient, while others are unbounded and do not have a defined maximum or minimum value, e.g., the Interest Factor. Because of such differences, these measures behave differently when applied to different types of patterns. Indeed, the measures defined above are not exhaustive and there exist many alternative measures for capturing different properties of relationships between pairs of binary variables. Table 4.9

Table 4.9. Examples of objective measures for the itemset $\{A, B\}$.

Measure (Symbol)	Definition
Correlation (ϕ)	$\frac{Nf_{11}-f_{1+}f_{+1}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$
Odds ratio (α)	$\left(f_{11}f_{00}\right)/\left(f_{10}f_{01}\right)$
Kappa (κ)	$\frac{Nf_{11}+Nf_{00}-f_{1+}f_{+1}-f_{0+}f_{+0}}{N^2-f_{1+}f_{+1}-f_{0+}f_{+0}}$
Interest (I)	$\left(Nf_{11}\right)/\left(f_{1+}f_{+1}\right)$
Cosine (IS)	$\left(f_{11}\right)/\left(\sqrt{f_{1+}f_{+1}}\right)$
Piatetsky-Shapiro (PS)	$\frac{f_{11}}{N} - \frac{f_{1+}f_{+1}}{N^2}$
Collective strength (S)	$\frac{f_{11}+f_{00}}{f_{1+}f_{+1}+f_{0+}f_{+0}} \times \frac{N-f_{1+}f_{+1}-f_{0+}f_{+0}}{N-f_{11}-f_{00}}$
Jaccard (ζ)	$f_{11}/\left(f_{1+} + f_{+1} - f_{11}\right)$
All-confidence (h)	$\min\left[\frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}}\right]$

Table 4.10. Example of contingency tables.

Example	f_{11}	f_{10}	f_{01}	f_{00}
E_1	8123	83	424	1370
E_2	8330	2	622	1046
E_3	3954	3080	5	2961
E_4	2886	1363	1320	4431
E_5	1500	2000	500	6000
E_6	4000	2000	1000	3000
E_7	9481	298	127	94
E_8	4000	2000	2000	2000
E_9	7450	2483	4	63
E_{10}	61	2483	4	7452

provides the definitions for some of these measures in terms of the frequency counts of a 2×2 contingency table.

Consistency among Objective Measures

Given the wide variety of measures available, it is reasonable to question whether the measures can produce similar ordering results when applied to a set of association patterns. If the measures are consistent, then we can choose any one of them as our evaluation metric. Otherwise, it is important to understand what their differences are in order to determine which measure is more suitable for analyzing certain types of patterns.

Suppose the measures defined in Table 4.9 are applied to rank the ten contingency tables shown in Table 4.10. These contingency tables are chosen to illustrate the differences among the existing measures. The ordering produced by these measures is shown in Table 4.11 (with 1 as the most interesting and 10 as the least interesting table). Although some of the measures appear to be consistent with each other, others produce quite different ordering results. For example, the rankings given by the ϕ-coefficient agrees mostly with those provided by κ and collective strength, but are quite different than the rankings produced by interest factor. Furthermore, a contingency table such as E_{10} is ranked lowest according to the ϕ-coefficient, but highest according to interest factor.