Properties of statistical association rules explained with the Mega Party example

Wilhelmiina Hämäläinen

Oct 2021

1 Origin and reference

The Mega Party association rules are not based on any real data, they are just the author's imagination. Inventing coherent examples of association rules that demonstrate important properties of the theory and are still funny is difficult. So, if you want to utilize this material or examples at your own work, please give the following reference:

W. Hämäläinen: Properties of statistical association rules explained with the Mega Party example. Learning material for CS-E4650 Methods of Data Mining. Aalto University, 2021.

2 Association rules and their properties

2.1 Summary of all candidate rules

The data contains information from 1000 students who participated Mega Party. The features describe what each of student ate or drank in the party and what happened to them next day (their exam success and health consequences).

The candidate rules to be analyzed are described in Table 1. All rules are presented in the form $\mathbf{X} \rightarrow C$, where the condition $\mathbf{X}$ is a conjunction of attributes and the consequent C is a single attribute, whose value is 1 (thus notated simply C).

All listed association rules express statistical dependence, which is a necessary condition of statistical association rules. The strength of association is measured by lift γ and leverage δ. Recall that for positive dependence, $\gamma > 1$ and $\delta > 0$ and for negative dependence, $\gamma < 1$ and $\delta < 0$. Precision (confidence), ϕ, of rules is also given, but we have already seen that it can be misleading concerning statistical association.

The following examples will first demonstrate properties of lift and leverage and potential problems in their interpretation and then problems of overfitted and other specious rules.

2.2 High lift can be misleading

Rules 1 and 2 demonstrate problems of looking at high lift alone.

Rule 1: One student got corona (=C)! The student was the only one who had combination $\mathbf{X}$ = peppermint tea, sushi, chili sauce, sour cream. $\gamma(\mathbf{X}, C) = \frac{P(\mathbf{X}C)}{P(\mathbf{X})P(C)} = \frac{1}{P(C)} = n$.

Table 1: Candidate rules of the form $\mathbf{X} \rightarrow C$, where $fr_X = fr(\mathbf{X})$, $fr_C = fr(C)$, $fr_{XC} = fr(\mathbf{X}C)$, $\phi = P(C|\mathbf{X})$, $\gamma = \gamma(\mathbf{X}, C)$, $\delta = \delta(\mathbf{X}, C)$.

num	rule	fr_X	fr_C	fr_{XC}	ϕ	γ	δ
1	*peppermint tea, sushi, chili sauce, sour cream* → *corona*	1	1	1	1.00	1000	0.0010
2	*vodka, sauerkraut, salmon* → *headache*	1	100	1	1.00	10	0.0009
3	*cake* → *exam failure*	500	500	270	0.54	1.1	0.0200
4	*magic mushrooms* → *intoxication*	20	20	20	1.00	50.0	0.0196
5	*vodka* → *headache*	100	100	80	0.80	8.0	0.0700
6	*vodka, salmon* → *headache*	40	100	30	0.75	7.5	0.0260
7	*alcohol* → *exam failure*	333	500	300	0.90	1.8	0.1335
8	*alcohol* → *cake*	333	500	200	0.60	1.2	0.0335

Rule 2: 100 students had headache (=C), including one with unique combination $\mathbf{X} =$ vodka, sauerkraut, salmon. $\gamma(\mathbf{X}, C) = \frac{1}{P(C)} = \frac{1}{0.1} = 10$.

Rule 1 has maximal possible lift (and maximal precision) and the dependence is also "perfect" in the sense that $P(\mathbf{X}C) = P(\mathbf{X}) = P(C)$ ($\mathbf{X}$ and C cover the same set of transactions). Rule 2 has also maximal precision but the consequent is much more frequent, which decreases lift, although lift is still pretty high. An important observation is that both patterns occur only on one row of data, so they are **very likely just due to chance**. This example warns that high lift alone is not a reliable indicator of good rules. It also shows how frequency of the consequent affects lift (rule with a very rare consequent gets better lift, when ϕ is the same). In general, lift tends to favor rarer rules.

2.3 Leverage does not tell significance

Rules 3 and 4 both have approximately the same leverage, but otherwise they are very different. **Question**: Which one (3 or 4) would you consider more significant association?

Rule 3: 500 students had chocolate cake ($\mathbf{X}$) and 500 failed the next day exam (C), including 270 cake eaters. $\delta(\mathbf{X}, C) = \frac{1}{n^2}(n \cdot 270 - 500 \cdot 500) = 0.02$.

Rule 4: 20 students tried magic mushrooms ($\mathbf{X}$) and only them got serious intoxication (C). $\delta(\mathbf{X}, C) = \frac{1}{n^2}(n \cdot 20 - 20 \cdot 20) = 0.0196$.

Rule 3 does not have particularly high lift. Instead, it has very high frequency, but this is only expected, since both condition and consequent are very frequent (expected frequency under independence is $n \cdot P(\mathbf{X}) \cdot P(C) = 1000 \cdot 0.5 \cdot 0.5 = 250$ – also very high!). If we present leverage as $\delta(\mathbf{X}, C) = P(\mathbf{X}C) - P(\mathbf{X})P(C) = P(\mathbf{X})(P(C|\mathbf{X}) - P(C))$, we can see that leverage increases with $P(\mathbf{X})$. Therefore, one can get relatively good leverage, if the rule is very frequent, even if precision ϕ wouldn't deviate much from $P(C)$.

Rule 4 has very high lift, since it expresses perfect dependence: $P(\mathbf{X}C) = P(\mathbf{X}) = P(C)$. However, the frequency is relatively low (still sufficient for statistical significance). Perfect dependence results best possible leverage, $\delta(\mathbf{X}, C) = P(\mathbf{X}, C) - P(\mathbf{X})P(C) = P(\mathbf{X})(1 - P(C)) = P(\mathbf{X})P(\neg\mathbf{X}) = P(C)P(\neg C)$. This expressions gets highest value when $P(C) = P(\mathbf{X}) = 0.5$. So, given two rules expressing perfect dependence, leverage favors the one with

$P(\mathbf{X})$ closer to 0.5. In general, leverage tends to favor more frequent rules than lift (as long as frequency is not too high, i.e., $P >> 0.5$).

This example shows that **leverage alone is not sufficient for evaluating significance of rules**. In addition, we should calculate probability of observing at least as strong association by chance, if $\mathbf{X}$ and C were actually independent. When we evaluate Fisher's p-value (from Fisher's exact test of independence) for rules 3 and 4, we obtain $ln(p) = -5.0$ ($p \approx 6.8e{-}3$) for rule 3 and $ln(p) = -95.6$ ($p \approx 3.0e{-}42$) for rule 4. So, rule 4 is much more significant, which fits our intuition.

2.4 Overfitted associations can be misleading

Rules 5 and 6 are an example of a more general rule and its specialization. These two rules are also related to headache (C), which had $fr(C) = 100$.

Rule 5: 100 students had vodka ($\mathbf{X}$) and 80 of them got headache next day.

Rule 6: 40 students had vodka and salmon ($\mathbf{X}$) and 30 of them got headache next day.

Rule 5 is a really good rule, very strong and extremely significant ($ln(p) = -180.7$). It is very likely that vodka explains most of headaches. However, if we consider rule 6 in isolation, it also looks very good (good lift, quite good leverage, and significant with $ln(p) = -53.7$). We could assume that combination of vodka and salmon is causing headaches and (if we didn't know rule 5) that vodka is safe as long as we avoid salmon with it. However, it turns out that salmon does not increase the likelihood of headache, given vodka, rather it can protect from the headache!

Let us notate V = vodka, S = salmon and H = headache. Then we will evaluate the expected absolute frequency of VSH, notated $E[fr(VSH)]$, assuming that H is conditionally independent of S given H. The conditional independence assumption means that $P(VSH) = P(VS)P(H|V)$. The expected absolute frequency is $E[fr(VSH)] = fr(VS)P(H|V) = 40 \cdot 0.80 = 32$. This is larger than the observed frequency $fr(VSH) = 30$, which means that there is negative conditional dependence between S and H given V. Salmon was slightly preventing vodka-headaches!

In this case rule 6 was an *overfitted specialization* of rule 5. This is an example of easily detected overfitted rules, because rule $\mathbf{XQ} \rightarrow C$ cannot improve rule $\mathbf{X} \rightarrow C$ [1], unless $P(C|\mathbf{XQ}) > P(C|\mathbf{X})$. So, we can simply compare the precisions. If we observe that $P(C|\mathbf{XQ}) \leq P(C|\mathbf{X})$, then there must be conditional independence or negative dependence between $\mathbf{Q}$ and C given $\mathbf{X}$ and rule $\mathbf{XQ} \rightarrow C$ is overfitted. However, this test does not catch all overfitted rules. The more specific rule could improve a more general a little bit just by chance (the more it improves, the less likely it is chance). To detect these overfitted rules, one must perform conditional significance testing.

Note that for testing overfitting of any rule $\mathbf{X} \rightarrow C$, one should consider all generalizations of the form $\mathbf{Y} \rightarrow C$, where $\mathbf{Y} \subsetneq \mathbf{X}$ (and $\mathbf{Y} \neq \emptyset$ – if $\mathbf{Y} = \emptyset$, one is actually testing unconditional dependence, if $P(C|\mathbf{X}) > P(C)$).

2.5 Specious associations are causally misleading

Rule 3, cake → failure, had quite good leverage, it is not overfitted, but it turns out that it is still misleading – eating chocolate cake does not cause exam failures! Here we show that rule 3 is *specious*, an association that is a side-product of other associations and disappears

[1] neither in value-based nor in variable-based interpretation, but this would require a proof

when conditioned on the actual cause factors (here alcohol). We know for certain that specious associations do not express causal relationships (the other associations may not present causal relations either, but we have no means to verify it).

Let us notate C = cake, F = failure, A = alcohol. We show that rule $C \rightarrow F$ is a side product of associations $A \rightarrow F$ (rule 7) and $A \rightarrow C$ (rule 8).

Rule $A \rightarrow F$ is very strong, with $P(F|A) = 0.9$ (vs. $P(F|\neg A) = 0.3$) and very high leverage. Rule $A \rightarrow C$ is not as strong, but still stronger (higher γ and δ) than $C \rightarrow F$.

Let us now evaluate what would be the expected frequency of event CF if C and F were conditionally independent given variable A. For this purpose, event CF is divided into two sub-events CAF and $C\neg AF$:

- Assuming that C is independent of F given A, $P(F|CA) = P(F|A) = 0.9$. So $CA \rightarrow F$ is also strong (overfitted or redundant).

- Similarly, assuming that C is independent of F given $\neg A$, $P(F|C\neg A) = P(F|\neg A)$. Here $P(F|\neg A) = \frac{fr(\neg AF)}{fr(\neg A)} = \frac{fr(F)-fr(AF)}{fr(\neg A)} = \frac{500-300}{1000-333} = \frac{200}{667} = 0.30$.

Now the expectation for $fr(CF) = fr(CAF) + fr(C\neg AF)$ is
$E[fr(CAF)] + E[fr(C\neg AF)] = fr(CA)P(F|A) + fr(C\neg A)P(F|\neg A) = 200 \cdot 0.9 + 300 \cdot 0.3 = 180 + 90 = 270$. This is the same frequency than we have observed in data! This confirms that C and F are indeed conditionally independent given variable A. (Here A is called a confounding variable.)

Note: Alternatively one could calculate *conditional leverage* of $C \rightarrow F$ given A and given $\neg A$ and evaluate their signs: $\delta_1 = P(CAF) - P(CA)P(F|A)$ and $\delta_2 = P(C\neg AF) - P(C\neg A)P(F|\neg A)$. If both δ_1 and δ_2 are ≤ 0.0, we know for certain that factor A cancels the observed positive association between C and F. More complicated cases (where at least one of conditional leverages is positive) require statistical significance testing of conditional independence in both cases.