Оценка паттернов и правил

Меры для оценки паттернов и правил

Пусть $\mathcal{I}$ - набор элементов, $\mathcal{T}$ - набор тидов и пусть $\boldsymbol{D} \subseteq \mathcal{T} \times \mathcal{I}$- двоичная база данных. Напомним, что правило ассоциации - это выражение $X \longrightarrow Y$, где X и Y - это наборы элементов, т.е. $X, Y \subseteq \mathcal{I}$ и $X \cap Y = \emptyset$. Мы называем X антецедентом правила, а Y - следствием.

Набор тидов для набора элементов X - это набор всех тидов, содержащих X, заданный как

$$t(X) = \{t \in \mathcal{T} \mid X \text{ содержится в } t\}$$

Таким образом, поддержка X равна $\sup(X) = |t(X)|$. В последующем обсуждении мы используем краткую форму XY для обозначения объединения XY наборов элементов X и Y.

Учитывая часто встречающийся набор элементов $Z \in F$, где F - множество всех часто встречающихся наборов элементов, мы можем вывести различные правила ассоциации, рассматривая каждое собственное подмножество Z как антецедент, а остальные элементы как консеквент, то есть для каждого $Z \in F$, мы можем вывести набор правил вида $X \longrightarrow Y$, где $X \subset Z$ и $Y = Z \setminus X$.

Меры по оценке правил
Поддержка (Support)

Поддержка правила определяется как количество транзакций, содержащих как X, так и Y, то есть

$$\sup(X \longrightarrow Y) = sup(XY) = |t(XY)| \qquad 12.1$$

Относительная поддержка — это доля транзакций, содержащих как X, так и Y, то есть эмпирическая совместная вероятность элементов, входящих в правило.

$$r\sup(X \longrightarrow Y) = P(XY) = r\sup(XY) = \frac{\sup(XY)}{|D|}$$

Обычно нас интересуют частые правила с $sup(X \longrightarrow Y) \geq minsup$, где $minsup$ - минимальный порог поддержки, определяемый пользователем. Когда минимальная поддержка указывается в виде дроби, подразумевается относительная поддержка. Обратите внимание, что (относительная) поддержка - это симметричная мера, потому что $\sup(X \longrightarrow Y) = \sup(Y \longrightarrow X)$.

Достоверность (Confidence)

Достоверность правила — это условная вероятность того, что транзакция содержит консеквент Y, при условии, что он содержит антецедент X:

$$conf(X \longrightarrow Y) = P(\,Y \mid X\,) = \frac{P(XY)}{P(X)} = \frac{rsup(XY)}{rsup(X)} = \frac{\sup(XY)}{\sup(X)}$$

Обычно нас интересуют правила с высокой степенью достоверности с $\mathrm{conf}(\mathrm{X} \longrightarrow \mathrm{Y}) \geq minconf$, где $minconf$ — это минимальное значение достоверности, определяемое пользователем. Достоверность не является симметричной мерой, потому что по определению она зависит от антецедента.

Лифт (Lift)

Лифт определяется как отношение наблюдаемой совместной вероятности X и Y к ожидаемой совместной вероятности, если бы они были статистически независимыми, то есть

$$lift(X \longrightarrow Y) = \frac{P(XY)}{P(X) \cdot P(Y)} = \frac{rsup(XY)}{rsup(X) \cdot rsup(Y)} = \frac{conf(X \longrightarrow Y)}{rsup(Y)}$$

Один из распространенных способов использования лифта - измерить неожиданность правила. Значение лифта, близкое к 1, означает, что ожидается поддержка правила с учетом поддержки его компонентов. Обычно мы ищем значения, которые намного больше (т. е. выше ожидаемого) или меньше 1 (т. е. ниже ожидаемого).

Обратите внимание, что лифт — это симметричная мера, и она всегда больше или равна достоверности, потому что это достоверность, деленная на вероятность консеквента. Лифт также не является нижним набором, т.е. если предположить, что $X' \subset X$ и $Y' \subset Y$, может случиться так, что $lift(X' \longrightarrow Y')$ может быть выше $lift(X \longrightarrow Y)$. Лифт может быть подвержен шуму в небольших наборах данных, поскольку редкие или нечастые наборы элементов, которые встречаются всего несколько раз, могут иметь очень высокие значения лифта.

Усиление (Leverage, Рычаг)

Усиление измеряет разницу между наблюдаемой и ожидаемой совместной вероятностью XY при условии, что X и Y независимы.

$$\text{leverage}\ (X \longrightarrow Y) = P(XY)\ -\ P(X) \cdot P(Y) = \text{rsup}(XY)\ -\ \text{rsup}(X) \cdot \text{rsup}(Y)$$

Усиление дает «абсолютную» меру того, насколько неожиданно правило, и его следует использовать вместе с лифтом. Как и лифт — симметричный.

Коэффициент Жаккара (Jaccard)

Коэффициент Жаккара измеряет сходство между двумя наборами. При применении в качестве меры оценки правил он вычисляет сходство между наборами значений X и Y:

$$jaccard(X \longrightarrow Y) = \frac{|t(X) \cap t(Y)|}{|t(X) \cup t(Y)|}$$

$$= \frac{sup(XY)}{sup(X) + sup(Y) - sup(XY)}$$

$$= \frac{P(XY)}{P(X) + P(Y) - P(XY)}$$

Жаккар - симметричная мера.

Убежденность (Conviction)

Все меры оценки правил, которые мы рассмотрели выше, используют только совместную вероятность X и Y. Определим $\neg X$ как событие, в котором X не содержится в транзакции, то есть $X \not\subseteq t \in \mathcal{T}$, и аналогично для $\neg Y$. Как правило, существует четыре возможных события в зависимости от возникновения или отсутствия наборов элементов X и Y, как показано в таблице непредвиденных обстоятельств, представленной в табл.12.8.

Таблица 12.8 — Таблица непредвиденных обстоятельств для *X* и *Y*

	Y	$\neg Y$	
X	$sup(XY)$	$sup(X\neg Y)$	$sup(X)$
$\neg X$	$sup(\neg XY)$	$sup(\neg X\neg Y)$	$sup(\neg X)$
	$sup(Y)$	$sup(\neg Y)$	$\|D\|$

Убежденность измеряет ожидаемую ошибку правила, то есть, как часто X встречается в транзакции, а Y - нет. Таким образом, это мера силы правила по отношению к дополнению консеквента, определяемого как

$$conv(X \longrightarrow Y) = \frac{P(X) \cdot P(\neg Y)}{P(X\neg Y)} = \frac{1}{lift(X \longrightarrow \neg Y)}$$

Если совместная вероятность $X\neg Y$ меньше ожидаемой при независимости от X и $\neg Y$, тогда убежденность высока, и наоборот. Это асимметричная мера.

Из таблицы 12.8 видно, что $P(X) = P(XY) + P(X\neg Y)$, откуда следует, что $P(X\neg Y) = P(X) - P(XY)$. Далее,$P(\neg Y) = 1 - P(Y)$. Таким образом, мы имеем

$$conv(X \longrightarrow Y) = \frac{P(X) \cdot P(\neg Y)}{P(X) - P(XY)} = \frac{P(\neg Y)}{1 - P(XY)/P(X)} = \frac{1 - r\sup(Y)}{1 - conf(X \longrightarrow Y)}$$

Мы заключаем, что убежденность бесконечна, если достоверность едина. Если X и Y независимы, то убежденность равна 1.

Соотношение шансов (Odds Ratio)

В соотношении шансов используются все четыре записи из таблицы непредвиденных обстоятельств, представленной в табл. 12.8. Давайте разделим набор данных на две группы транзакций – т.е., которые содержат X, и те, которые не содержат X. Определим шансы Y в этих двух группах следующим образом:

$$odds(Y \mid X) = \frac{P(XY)/P(X)}{P(X\neg Y)/P(X)} = \frac{P(XY)}{P(X\neg Y)}$$

$$odds(Y \mid \neg X) = \frac{P(\neg XY)/P(\neg X)}{P(\neg X\neg Y)/P(\neg X)} = \frac{P(\neg XY)}{P(\neg X\neg Y)}$$

Тогда соотношение шансов определяется как отношение этих двух шансов:

$$\text{oddsratio}(X \longrightarrow Y) = \frac{\text{odds}(Y|X)}{\text{odds}(Y|\neg X)} = \frac{P(XY) \cdot P(\neg X \neg Y)}{P(X\neg Y) \cdot P(\neg XY)}$$

$$= \frac{sup(XY) \cdot sup(\neg X \neg Y)}{sup(X\neg Y) \cdot sup(\neg XY)}$$

Соотношение шансов — это симметричная мера, и если X и Y независимы, тогда оно имеет значение 1. Таким образом, значения, близкие к 1, могут указывать на небольшую зависимость между X и Y. Отношение шансов больше 1 означает более высокие шансы Y встречается в присутствии X в отличие от его дополнения $\neg X$, тогда как шансы меньше единицы подразумевают более высокие шансы того, что Y встречается с $\neg X$.

Меры оценки модели
Поддержка (Support)

Самыми основными мерами являются поддержка и относительная поддержка, указывающие количество и долю транзакций в D, которые содержат набор элементов X:

$$\sup(X) = |t(X)| \quad rsup(X) = \frac{\sup(X)}{|D|}$$

Лифт (Lift)

Подъем k-элементного множества $X = \{x_1, x_2, \ldots, x_k\}$ в наборе данных D определяется как

$$lift(X, D) = \frac{P(X)}{\prod_{i=1}^{k} P(x_i)} = \frac{r\sup(X)}{\prod_{i=1}^{k} r\sup(x_i)} \quad (12.2)$$

то есть отношение наблюдаемой совместной вероятности элементов в X к ожидаемой совместной вероятности, если бы все элементы $\mathrm{x}_i \in X$ были независимыми.

Мы можем далее обобщить понятие лифта набора X, рассмотрев все различные способы его разбиения на непустые и непересекающиеся подмножества. Например, предположим, что множество $\{X_1, X_2, \ldots, X_q\}$ является q-разбиением X, т. е. разбиением X на q непустых и непересекающихся наборов элементов X_i, таких что $X_i \cap X_j = \emptyset$ и $\cup_i X_i = X$. Определим обобщенный лифт X над разбиениями размера q как:

$$lift_q(X) = \min_{X_1 \ldots, X_q} \left\{ \frac{P(X)}{\prod_{i=1}^{q} P(X_i)} \right\}$$

Это наименьшее значение лифта по всем q-разбиениям X. В этом свете $lift(X) = lift_k(X)$, то есть лифт - это значение, полученное из уникального k-разбиения X.

Меры, основанные на правилах

Учитывая набор элементов X, мы можем оценить его, используя меры оценки правил, рассматривая все возможные правила, которые могут быть сгенерированы из X. Пусть Θ - некоторая мера оценки правил. Мы генерируем все возможные правила из X в форме $X_1 \longrightarrow X_2$ и $X_2 \longrightarrow X_1$, где множество $\{X_1, X_2\}$ является 2-разбиением или двудольным разбиением X. Затем мы вычисляем меру Θ для каждого такого правила и используем суммарную статистику, такую как среднее, максимальное и минимальное значение для характеристики X.

Если Θ - симметричная мера, то $\theta(X_1 \longrightarrow X_2) = \theta(X_2 \longrightarrow X_1)$, и мы должны учитывать только половину правил. Например, если Θ — это повышение по правилу, то мы можем определить среднее, максимальное и минимальное значения подъема для X следующим образом:

$$AvgLift(X) = \underset{X_1,X_2}{avg}\{lift(X_1 \longrightarrow X_2)\}$$

$$MaxLift(X) = \underset{X_1,X_2}{max}\{lift(X_1 \longrightarrow X_2)\}$$

$$MinLift(X) = \underset{X_1,X_2}{min}\{lift(X_1 \longrightarrow X_2)\}$$

Мы также можем сделать то же самое для других мер правил, таких как усиление, достоверность и так далее. В частности, когда мы используем правило лифта, то $MinLift(X)$ идентичен обобщенному лифту $lift_2(X)$ по всем 2-разбиениям X.

Сравнение множественных правил и шаблонов
Сравнение наборов элементов

При сравнении нескольких наборов элементов мы можем сосредоточиться на максимальных наборах элементов, удовлетворяющих определенному свойству, или мы можем рассмотреть закрытые наборы элементов, которые захватывают всю информацию о поддержке. Мы рассматриваем эти и другие меры в следующих параграфах.

Максимальные наборы предметов.

Часто встречающийся набор элементов X является максимальным, если все его надмножества не являются частыми, то есть X является максимальным, если и только если

$$\sup(\mathrm{X}) \geq \text{ minsup }, \text{ и для всех } \mathrm{Y} \supset \mathrm{X}, \sup(\mathrm{Y}) < \text{ minsup}$$

Учитывая набор часто встречающихся наборов элементов, мы можем выбрать только максимальные из них, особенно среди тех, которые уже удовлетворяют некоторым другим ограничениям на меры оценки модели, такие как лифт или усиление.

Закрытые наборы предметов и минимальные генераторы

Набор элементов X *закрытый*, если все его надмножества имеют строго меньшую поддержку, т. е.

$$sup(X) > \ sup(Y) \ \text{для всех} \ Y \supset X$$

Набор элементов X является минимальным генератором, если все его подмножества имеют строго более высокий уровень поддержки, т. е.

$$sup(X) < \ sup(Y) \ \text{для всех} \ Y \subset X$$

Если набор X не является минимальным генератором, то это означает, что у него есть некоторые избыточные элементы, то есть мы можем найти некоторое подмножество $Y \subset X$, которое можно заменить еще меньшим подмножеством $W \subset Y$, не меняя носителя X , т. е. существует такое $W \subset Y$, что

$$\sup(X) = \sup\big(Y \cup (X \backslash Y)\big) = \sup\big(W \cup (X \backslash Y)\big)$$

Можно показать, что все подмножества минимального генератора сами должны быть минимальными генераторами.

Продуктивные наборы предметов

Набор элементов X является *продуктивным*, если его относительная поддержка выше, чем ожидаемая относительная поддержка по всем его разделам, при условии, что они независимы. Более формально, пусть $|X| \geq 2$, и пусть $\{X_1, X_2\}$ является двудольным X. Мы говорим, что X продуктивно, если

$rsup(X) > rsup(X_1) \times rsup(X_2)$, для всех двудольных $\{X_1, X_2\}$ из X (12.3)

Это сразу означает, что X является продуктивным, если его минимальный лифт больше единицы, поскольку

$$MinLift(X) = \min_{X_1, X_2} \left\{ \frac{rsup(X)}{rsup(X_1) \cdot rsup(X_2)} \right\} > 1$$

Что касается усиления, X является продуктивным, если его минимальное усиление нуля, потому что

$$\text{MinLeverage } (X) = \min_{X_1, X_2} \{rsup(X) - rsup(X_1) \times rsup(X_2)\} > 0$$

Сравнение правил
Неизбыточные правила

Учитывая два правила R: $X \longrightarrow Y$ и R': $W \longrightarrow Y$, которые имеют один и тот же консеквент, мы говорим, что R *более конкретен*, чем R', или, что эквивалентно, что R' *более общий*, чем R, при условии $W \subset X$.

Мы говорим, что правило R: $X \longrightarrow Y$ является избыточным, если существует более общее правило R': $W \longrightarrow Y$, которое имеет тот же носитель, то есть $W \subset X$ и $sup(R) = sup(R')$. С другой стороны, если $sup(R) < sup(R')$ по всем его обобщениям R', то R неизбыточно.

Правила улучшения и продуктивности

Определим *улучшение* правила $X \longrightarrow Y$ следующим образом

$$imp(X \longrightarrow Y) = conf(X \longrightarrow Y) - \max_{W \subset X}\{conf(W \longrightarrow Y)\}$$

Улучшение определяет минимальную разницу между достоверностью правила и любого из его обобщений. Правило R: $X \longrightarrow Y$ является продуктивным, если его улучшение больше нуля, что означает, что для всех более общих правил R': $W \longrightarrow Y$ имеем $conf(R) > conf(R')$. С другой стороны, если существует более общее правило R' с $conf(RR') > conf(R)$, то R *непродуктивно*. Если правило является избыточным, оно также непродуктивно, поскольку его улучшение равно нулю.

Чем меньше улучшение правила R: $X \longrightarrow Y$, тем больше вероятность, что оно будет непродуктивным. Мы можем обобщить это понятие, чтобы рассмотреть правила, которые имеют хотя бы некоторый минимальный уровень улучшения, то есть мы можем потребовать, чтобы $imp(X \longrightarrow Y) \geq$ t, где t - определяемый пользователем минимальный порог улучшения.

Точный тест Фишера для продуктивных правил

Пусть $R: X \to Y$ – ассоциативное правило. Рассмотрим его обобщение $R': W \to Y$, где $W = X \backslash Z$ – новый антецедент, полученный удалением из X подмножества $Z \subseteq X$. Имея исходный датасет $\mathbf{D}$, при условии наличия W, мы можем сформировать таблицу сопряжённости размера 2×2 между Z и консеквентом Y, как показано в таблице 12.17. Значения в ячейках таблицы формируются следующим образом:

$a = sup(WZY) = \sup(XY)$	$b = sup(WZ\neg Y) = sup(X\neg Y)$
$c = sup(W\neg XY)$	$d = sup(W\neg Z\neg Y)$

Таблица 12.17. Таблица сопряжённости для Z и Y при условии $W = X \backslash Z$.

W	Y	$\neg Y$	
Z	a	b	$a + b$
$\neg Z$	c	d	$c + d$
	$a + c$	$b + d$	$n = sup(W)$

Здесь a обозначает количество транзакций, которые содержат и X, и Y, b обозначает количество транзакций, которые содержат X, но не содержат Y, c обозначает количество транзакций с W и Y, но без Z, и, наконец, d обозначает количество транзакций, содержащих W, но без Z и Y. Предельные значения заданы следующим образом:

пределы рядов:	$a + b = sup(WZ) = sup(X), \qquad c + d = sup(W\neg Z)$
пределы столбцов:	$a + c = sup(WY), \qquad b + d = sup(W\neg Y)$

где пределы рядов задают частоту появления W с и без Z, а пределы рядов определяют количество вхождений W с и без Y. Наконец, мы можем видеть, что сумма всех ячеек – просто $n = a + b + c + d = sup(W)$. Заметим, что при $Z = X$ мы получим $W = \emptyset$, а таблица сопряжённости выродится в 12.8.

Имея таблицу сопряжённости при условии W, нам интересно получить отношение шансов, полученное сравнением наличия и отсутствия Z, то есть:

$$oddratio = \frac{a/(a+b)}{b/(a+b)} / \frac{c/(c+d)}{d/(c+d)} = \frac{ad}{bc}. \quad (12.4)$$

Напомним, что отношение шансов измеряет шансы X, то есть, W и Z, встречающиеся с Y, против подмножества W без Z, встречающиеся с Y. При нулевой гипотезе H_0 (в предположении, что Z и Y независимы при условии W) отношение шансов равно единице. Это легко видеть, так как в предположении независимости счётчик в ячейке таблицы сопряжённости равен произведению предельных значений соответствующего ряда и столбца, делённых на n, то есть, при H_0:

$a = (a+b)(a+c)/n$	$b = (a+b)(b+d)/n$
$c = (c+d)(a+c)/n$	$d = (c+d)(b+d)/n$

Подставляя эти значения в ур-е (12.4), получим:

$$oddsratio = \frac{ad}{bc} = \frac{(a+b)(c+d)(b+d)(a+c)}{(a+c)(b+d)(a+b)(c+d)}.$$

Таким образом, нулевая гипотеза соответствует $H_0: oddsratio = 1$, а альтернативная – $H_a: oddsratio > 1$. Если при нулевой гипотезе мы сделаем дополнительное предположение, зафиксировав пределы рядов и столбцов, тогда a однозначно определяет значения b, c и d, а функция вероятности для наблюдения a в таблице сопряжённости задаётся гипергеометрическим распределением. Напомним, что гипергеометрическое распределение задаёт вероятность выбрать s успехов в t испытаниях, если мы делаем выборку *без замены* из конечного набора размера T, в котором всего S успехов:

$$P(s|t,S,T) = \binom{S}{s} \cdot \binom{T-S}{t-s} / \binom{T}{t}.$$

В нашем случае мы трактуем появление Z как успех. Размер выборки – $T = sup(W) = n$, поскольку мы предполагаем, что W встречается всегда, а общее количество успехов – это поддержка Z при условии W, то есть, $S = a + b$. В $t = a + c$ испытаниях гипергеометрическое распределение даёт вероятность $s = a$ успехов:

$$P(a|(a+c),(a+b),n) = \frac{\binom{a+b}{a} \cdot \binom{n-(a+b)}{(a+c)-a}}{\binom{n}{a+c}} = \frac{\binom{a+b}{a} \cdot \binom{c+d}{c}}{\binom{n}{a+c}}$$

$$= \frac{(a+b)!\,(c+d)!}{a!\,b!\,c!\,d!} \Big/ \frac{n!}{(a+c)!\,\big(n-(a+c)\big)!}$$

$$= \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}. \quad (12.5)$$

Наша цель – сопоставить нулевую гипотезу H_0 о том, что $oddsratio = 1$, с альтернативной гипотезой H_a: $oddsratio > 1$. Поскольку a однозначно определяет остальные ячейки при фиксированных предельных значениях рядов и столбцов, из ур-я (12.4) мы видим, что чем больше a, тем больше отношение шансов, а следовательно, и больше доказательств для H_a. Мы можем получить p-значение для таблицы сопряжённости так же, как в таблице 12.17, суммируя все члены ур-я (12.5) с возможными значениями a и больше:

$$p - value(a) = \sum_{i=0}^{\min(b,c)} P(a+i|(a+c),(a+b),n)$$

$$= \sum_{i=0}^{\min(b,c)} \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,(a+i)!\,(b-i)!\,(c-i)!\,(d+i)!},$$

что следует из того факта, что, увеличивая количество a на i, мы, поскольку предельные значения рядов и столбцов фиксированы, должны уменьшить b и c на i, а d – увеличить на i, как показано в таблице 12.18. Чем меньше p-значение, тем выше уверенность в том, что отношение шансов выше единицы, и, таким образом, мы можем отклонить нулевую гипотезу H_0, если $p - value \leq \alpha$, где α – порог значимости (например, $\alpha = 0.01$). Этот тест известен как *точный тест Фишера*.

Таким образом, чтобы проверить, продуктивно ли правило $R: X \to Y$, нам необходимо вычислить $p - value(a) = p - value(sup(XY))$ таблицы сопряжённости, полученной из каждого обобщения правила: $R': W \to Y$, где $W = X \backslash Z$ для $Z \subseteq X$. Если $p - value\big(sup(XY)\big) > \alpha$ для хотя бы одного такого сравнения, тогда мы можем отклонить правило $R: X \to Y$ как непродуктивное. С другой стороны, если $p - value\big(sup(XY)\big) \leq \alpha$ для всех обобщений, тогда R продуктивное. Заметим, что при $|X| = k$ существует $2^k - 1$ обобщений, и, чтобы избежать экспоненциальной сложности для больших антецедентов, мы, как правило, ограничиваемся только непосредственными обобщениями в следующей форме: $R': X \backslash z \to Y$, где $z \in X$ – одно из значений атрибута а антецеденте. Однако мы включаем тривиальное правило $\emptyset \to Y$, поскольку условная вероятность $P(Y|X) = conf(X \to Y)$ должна быть выше априорной вероятности $P(Y) = conf(\emptyset \to Y)$.

Проверка множества гипотез

Имея исходный датасет $\mathbf{D}$, можно получить экспоненциально большое количество правил, которые необходимо проверить на продуктивность. Так мы попадаем в проблему проверки множества гипотез, то есть, из-за большого количества гипотез некоторое количество непродуктивных правил пройдёт порог $p - value \leq \alpha$ со случайной вероятностью. Способ избежать этого – использовать поправку Бонферрони для уровня значимости, которая явно берёт в рассмотрение количество экспериментов, выполненных во время проверки гипотез. Вместо использования заданного порога α необходимо использовать скорректированный порог $\alpha' = \frac{\alpha}{\#r}$, где $\#r$ – количество (или его оценка) правил, которые необходимо протестировать. Такая поправка гарантирует, что частота обнаружения ложных правил ограничена α, где ложное обнаружение означает утверждение, что правило продуктивно, хотя это не так.

Проверка значимости перестановками

Проверка *перестановками или рандомизацией* позволяет определить распределение заданной статистики Θ путём многократных случайных изменений наблюдаемых данных, чтобы получить случайное количество датасетов, которые затем можно использовать для проверки значимости. В контексте оценки шаблонов, имея датасет $\mathbf{D}$, мы генерируем k случайных перемешанных датасетов $\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_k$. После этого мы можем произвести различные проверки значимости. Например, для заданного паттерна или правила мы можем проверить, являются ли они статистически значимыми, вычислив эмпирическую функцию распределения для тестовой статистики Θ путём вычисления её значения Θ_i в i-том датасете $\mathbf{D}_i$ для всех $i \in [1, k]$. Для этих значений мы можем сгенерировать выборочную функцию распределения:

$$\hat{F}(x) = \hat{P}(\Theta \leq x) = \frac{1}{k} \sum_{i=1}^{k} I(\Theta_i \leq x),$$

где I – индикаторная функция, принимающая значение 1, когда её аргумент – истина, и 0 в обратном случае.

Пусть Θ – значение тестовой статистики для входного датасета $\mathbf{D}$, тогда $p - value(\Theta)$, то есть, вероятность получить значение выше Θ, может быть вычислено так:

$$p - value(\Theta) = 1 - F(\Theta).$$

Имея заданный уровень значимости α, при значении $p - value(\Theta) > \alpha$ мы принимаем нулевую гипотезу о том, что правило/паттерн не являются статистически значимыми. С другой стороны, при $p - value(\Theta) \leq \alpha$ мы можем отклонить нулевую гипотезу и заключить, что паттерн является значимым, поскольку значение выше Θ крайне маловероятно. Проверка перестановками может быть применена для оценки всего набора правил и паттернов. Например, мы можем проверить список часто встречающихся наборов объектов, сравнивая число часто встречающихся наборов объектов в $\mathbf{D}$ с распределением числа часто встречающихся наборов объектов, полученных опытным путём из перемешанных датасетов $\mathbf{D}_i$. Мы можем также провести этот анализ как функцию $minsup$, и так далее.

Случайный обмен

Ключевой вопрос в генерации перемешанных датасетов $\mathbf{D}_i$ – какие характеристики входного датасета $\mathbf{D}$ нам необходимо сохранить. Имея датасет $\mathbf{D}_i$, мы случайным образом создаём k датасетов с теми же значениями пределов столбцов и рядов. Затем мы ищем часто встречающиеся паттерны в $\mathbf{D}$ и проверяем, отличается ли статистика шаблонов от полученных в перемешанном датасете. Если разница не значительна, мы можем заключить, что паттерны возникают исключительно из пределов ряда и столбца, а не из каких-либо интересных свойств данных.

Для заданной двоичной матрицы $\mathbf{D} \subseteq \mathcal{T} \times \mathcal{I}$ метод случайного обмена предлагает обменять две ненулевые ячейки матрицы посредством *перестановки*, которая оставляет пределы ряда и столбца неизменными. Чтобы показать, как работает такая перестановка, рассмотрим две любые транзакции $t_a, t_b \in \mathcal{T}$ и два объекта $i_a, i_b \in \mathcal{I}$ такие, что $(t_a, i_a), (t_b, i_b) \in \mathbf{D}$ и $(t_a, i_b), (t_b, i_a) \notin \mathbf{D}$, что соответствует подматрице $\mathbf{D}$ размера 2×2:

$$\mathbf{D}(t_a, i_a; t_b, i_b) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

После операции перестановки мы получим новую подматрицу:

$$\mathbf{D}(t_a, i_b; t_b, i_a) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

где мы обменяли элементы из $\mathbf{D}$ так, что $(t_a, i_b), (t_b, i_a) \in \mathbf{D}$ и $(t_a, i_a), (t_b, i_b) \notin \mathbf{D}$. Мы будем обозначать эту операцию следующим образом: $Swap(t_a, i_a; t_b, i_b)$. Отметим, что перестановка не изменяет пределы ряда и столбца, и таким образом мы можем сгенерировать перемешанный датасет с теми же суммами рядов и столбцов, что и в $\mathbf{D}$, последовательностью перестановок. Алгоритм 12.1 представляет собой псевдокод для генерации датасетов, перемешанных случайным обменом. Алгоритм выполняет i перестановок, выбирая две пары $(t_a, i_a), (t_b, i_b) \in \mathbf{D}$ случайным образом; перестановка успешна, только если $(t_a, i_b), (t_b, i_a) \notin \mathbf{D}$.

Алгоритм 12.1. Генерация датасета случайного обмена

- СлучайныйОбмен($t, \mathbf{D} \subseteq \mathcal{T} \times \mathcal{I}$):
- пока $t > 0$ выполнить:
- Выбрать случайную пару $(t_a, i_a), (t_b, i_b) \in \mathbf{D}$
- если $(t_a, i_b) \notin \mathbf{D}$ и $(t_b, i_a) \notin \mathbf{D}$ то:
- $D \leftarrow D \setminus \{(t_a, i_a), (t_b, i_b)\} \cup \{(t_a, i_b), (t_b, i_a)\}$
- $t = t - 1$
- вернуть $\mathbf{D}$