匿名加工データにおける
商品カテゴライズ化による一般化加工手法
k-匿名化の安全性評価

菊池研究室

中村幸輝

研究背景（匿名加工とは）

匿名加工：データから個人が特定されないように加工すること
→なぜ必要なのか？

標準的な加工手法や安全性基準が明確に定まっていない

顧客	購入日	商品
1	{11/2;4/6}	[A;B;C]
2	{11/2;4/6}	[A;B;C]
3	{11/2;4/6}	[A;B;C]

研究背景（一般化とは）

一般化とは…元データの日付や単価を**区間**として加工することで

匿名加工をするものである。加工されたデータ自体、

虚偽の情報は含まれていない。

元データ

顧客	購買日	商品ID	単価	購買数量
中村	11/14	A	1	1
伊藤	11/2	B	2	2
堀米	4/6	C	3	10

匿名加工

一般化データ

顧客	購買日	商品ID	単価	購買数量
中村	[11/2;4/6]	{A;B;C}	[1;3]	[1;10]
伊藤	[11/2;4/6]	{A;B;C}	[1;3]	[1;10]
堀米	[11/2;4/6]	{A;B;C}	[1;3]	[1;10]

問題点・解決策

安全性(1/k)と
有用性を同時に策定し、
総合評価として
k=2～20で
どれが適しているのか
検証する

商品をカテゴライズ化
することで一般化を行う

k-匿名化プログラムの説明①

①ユーザを購買回数順でソート

3-匿名加工の場合
この3名で1組となる

顧客	購買日	商品ID	単価	購買数量
A	11/14	158	2	1
A	11/2	B20	2.5	2
A	4/6	891	3	10
A	8/2	291D	10	8

顧客	購買日	商品ID	単価	購買数量
B	11/4	324	1	3
B	10/10	521P	2	1
B	5/28	G402	2	9

顧客	購買日	商品ID	単価	購買数量
C	1/3	B20	2.5	2
C	4/2	521P	2	10

k-匿名化プログラムの説明②

②レコードごとに**単価、購買数量**でソート（降順）

顧客	購買日	商品ID	単価	購買数量
A	11/14	158	2	1
A	11/2	B20	2.5	2
A	4/6	891	3	10
A	8/2	291D	10	8

顧客	購買日	商品ID	単価	購買数量
B	11/4	324	1	3
B	10/10	521P	2	1
B	5/28	G402	2	9

顧客	購買日	商品ID	単価	購買数量
C	1/3	B20	2.5	2
C	4/2	521P	2	10

顧客	購買日	商品ID	単価	購買数量
A	8/2	291D	10	8
A	4/6	891	3	10
A	11/2	B20	2.5	2
A	11/14	158	2	1

顧客	購買日	商品ID	単価	購買数量
B	5/28	G402	2	9
B	10/10	521P	2	1
B	11/4	324	1	3

顧客	購買日	商品ID	単価	購買数量
C	1/3	B20	2.5	2
C	4/2	521P	2	10

k-匿名加工プログラムの説明③

③レコードの上から一般化していく（はみ出たレコードは削除）

購買日、単価、購買数量は区間に
商品IDは羅列することで一般加工

顧客IDを仮名化することで
再識別を行なっても3人より
特定することはできなくなった

顧客	購買日	商品ID	単価	購買数量
A	8/2	291D	10	8
A	4/6	891	3	10
A	11/14	158	2	1
B	5/28	G402	2	9
B	11/4	B24	1	2
C	1/3	B20	2.5	2
C	4/2	521P	2	10

顧客	購買日	商品ID	単価	購買数量
A	[1/3;8/2]	{291D;G402;B20}	[2;10]	[2;9]
A	[4/2;10/10]	{891;521P}	[2;3]	[1;10]
*	*	*	*	*
*	*	*	*	*

顧客	購買日	商品ID	単価	購買数量
B	[1/3;8/2]	{291D;G402;B20}	[2;10]	[2;9]
B	[4/2;10/10]	{891;521P}	[2;3]	[1;10]
*	*	*	*	*

顧客	購買日	商品ID	単価	購買数量
C	[1/3;8/2]	{291D;G402;B20}	[2;10]	[2;9]
C	[4/2;10/10]	{891;521P}	[2;3]	[1;10]

商品カテゴライズについて

元データは商品数が3253種類ある

→カテゴリーに分類することで一般加工できるのではないか

商品IDが5桁の乱数なので商品IDではクラスタリングできない！

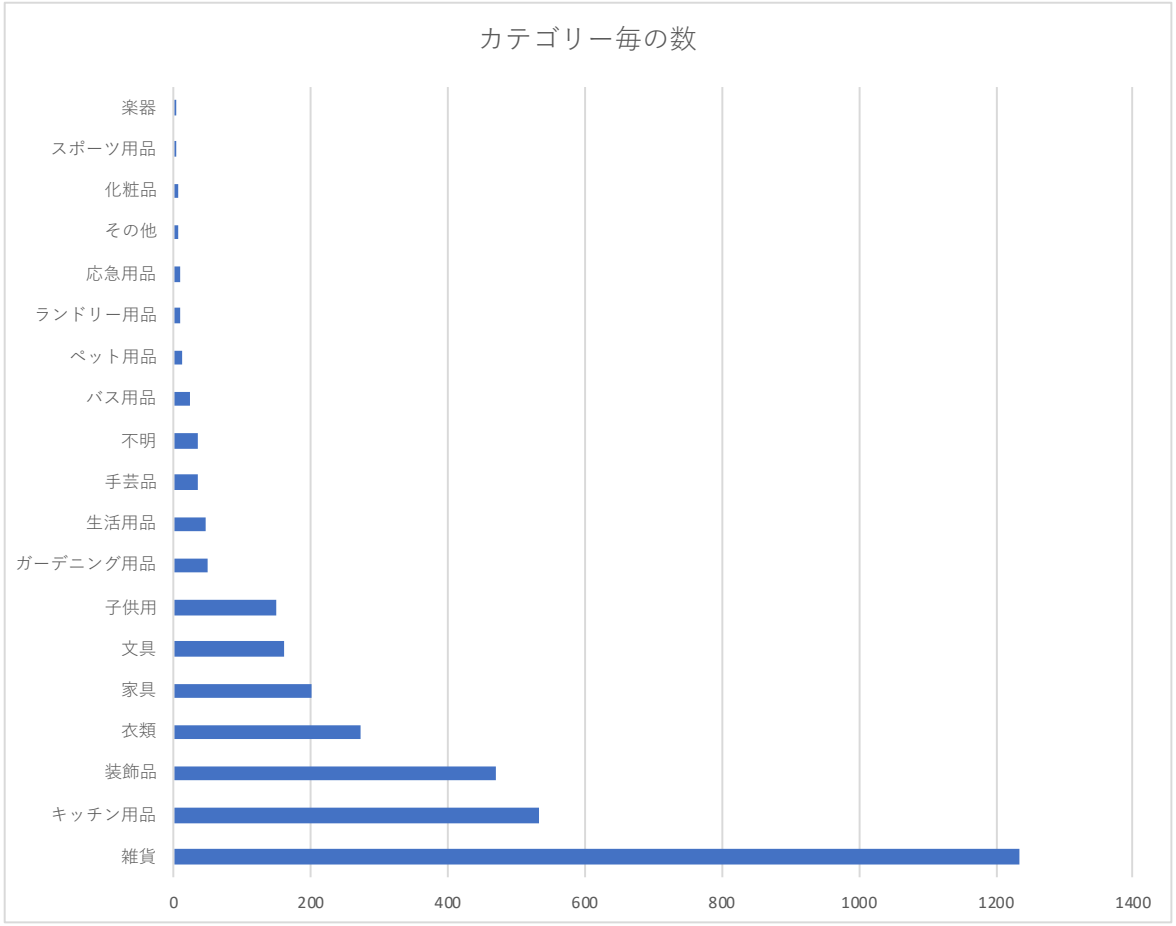
→**商品名から商品を独自にカテゴライズ**することで有用性を向上

3253種類→19種類

中村カテゴライズ

中村カテゴライズの種類数一覧

カテゴリー名	種類数
雑貨	1233
キッチン用品	534
装飾品	471
衣類	271
家具	201
文具	161
子供用	150
ガーデニング用品	50
生活用品	45
不明	34
その他	6



中村カテゴライズの例

例)「CERAMIC STRAWBERRY DESIGN MUG」
商品ID：22059
→「キッチン用品」

「PINK HEART SHAPE EGG FRYING PAN」
商品ID：84050
→「キッチン用品」

中村カテゴライズ後のレコード変化

③レコードの上から一般化していく（はみ出たレコードは削除）

顧客	購買日	商品ID	単価	購買数量
A	[1/3;8/2]	{291D;G402;B20}	[2;10]	[2;9]
A	[4/2;10/10]	{891;521P}	[2;3]	[1;10]
*	*	*	*	*
*	*	*	*	*

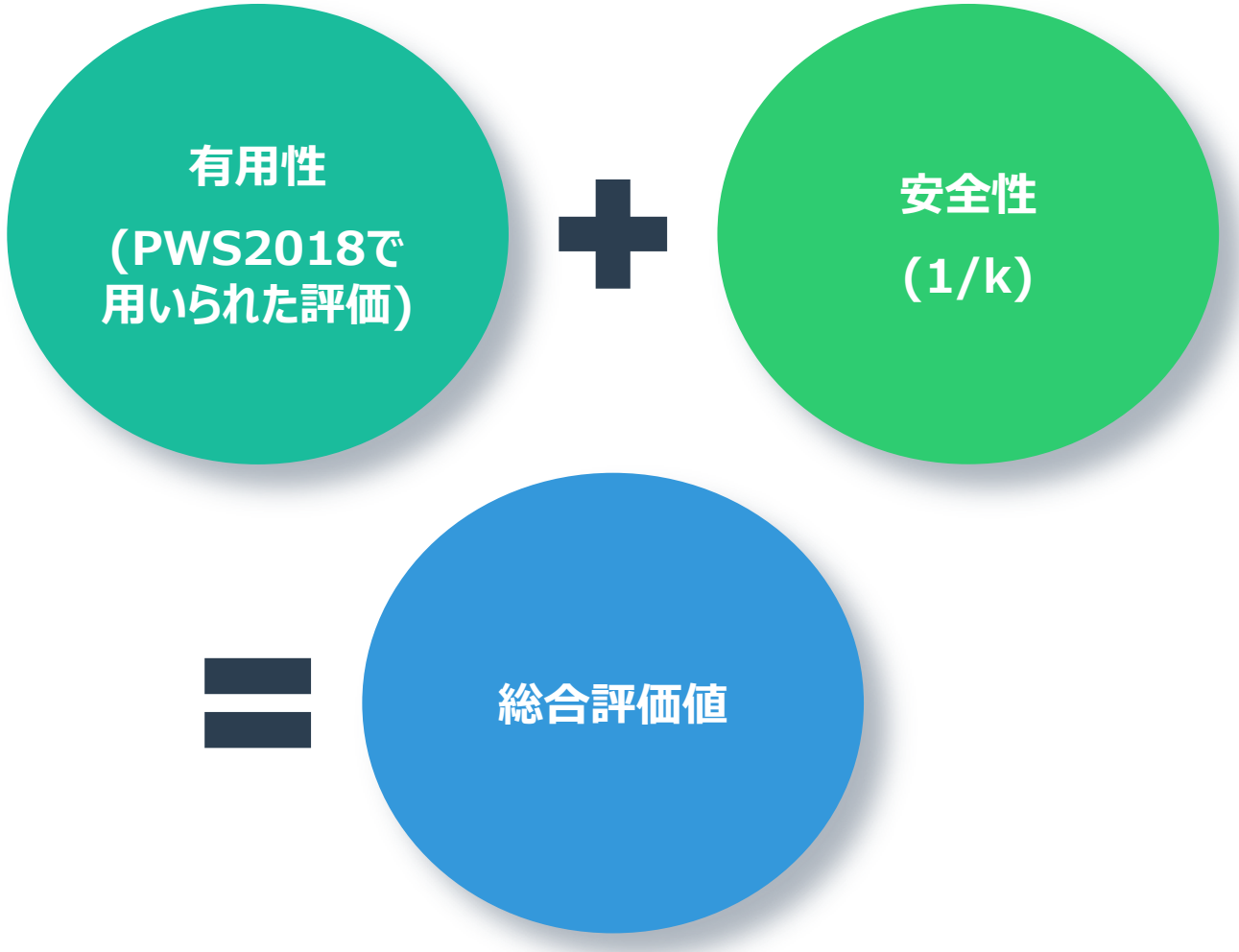
顧客	購買日	商品ID	単価	購買数量
A	[1/3;8/2]	{家具;文具}	[2;10]	[2;9]
A	[4/2;10/10]	{雑貨}	[2;3]	[1;10]
*	*	*	*	*
*	*	*	*	*

分類表

家具：291D, B20
文具：G402
雑貨：891, 521P

解決策のおさらい、総合評価の説明

安全性(1/k)と
有用性を
同時に策定し、
総合評価として
k=2～20で
どれが適しているのか
検証する



評価実験

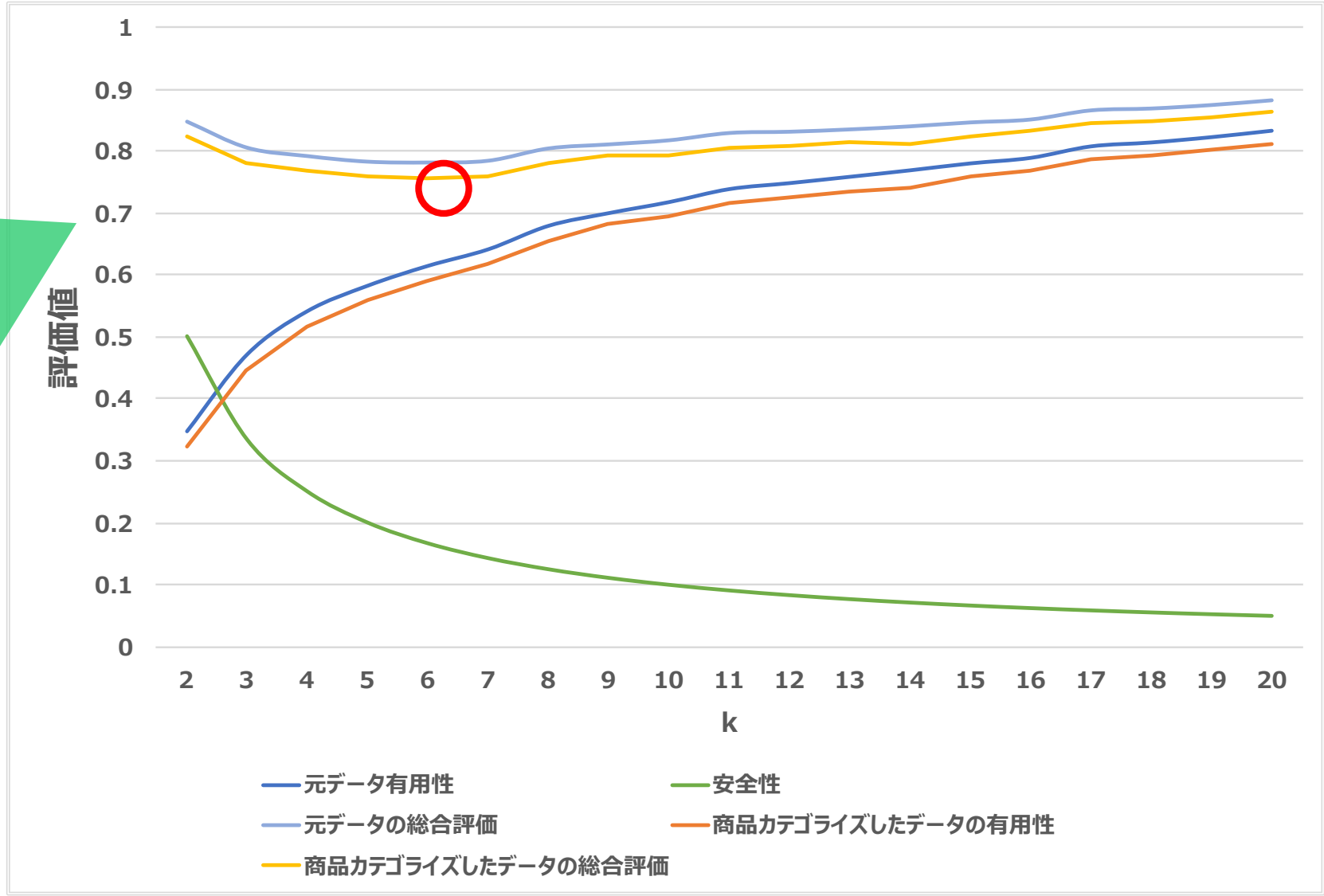
k=2~20-匿名加工データの有用性、安全性一覧

k=	元データ有用性	中村カテゴライズ後有用性	安全性	元データ総合評価	中村カテゴライズ後総合評価
2	0.348	0.324	0.500	0.848	0.824
3	0.472	0.447	0.333	0.806	0.781
4	0.542	0.517	0.250	0.792	0.767
5	0.583	0.558	0.200	0.783	0.758
6	0.615	0.590	0.167	0.782	0.757
7	0.642	0.617	0.143	0.785	0.760
8	0.680	0.655	0.125	0.805	0.780
10	0.718	0.694	0.100	0.818	0.794
12	0.748	0.725	0.083	0.832	0.808
15	0.780	0.758	0.067	0.847	0.824
20	0.833	0.812	0.050	0.880	0.862

評価実験

図からも見て取れるように
有用性値が全体的に
0.02~0.03ポイント向上した

また、総合評価値から、
最適なkの値は元データと同じく
k = 6 であることがわかった



まとめ

- 有用性、安全性の2点で比較したところ

  k=2~20匿名加工の中で最適な値はk=6である

- **商品カテゴライズ**によって商品数を減らし再検討した結果

  k=2~20匿名加工の中で最適な値はk=6である

- 実際のマーケティング分析では、商品をカテゴリーとして分類することで

  有用性が保てるのか（どこまで一般加工することで有用性が保てるのか）

  検証する必要がある

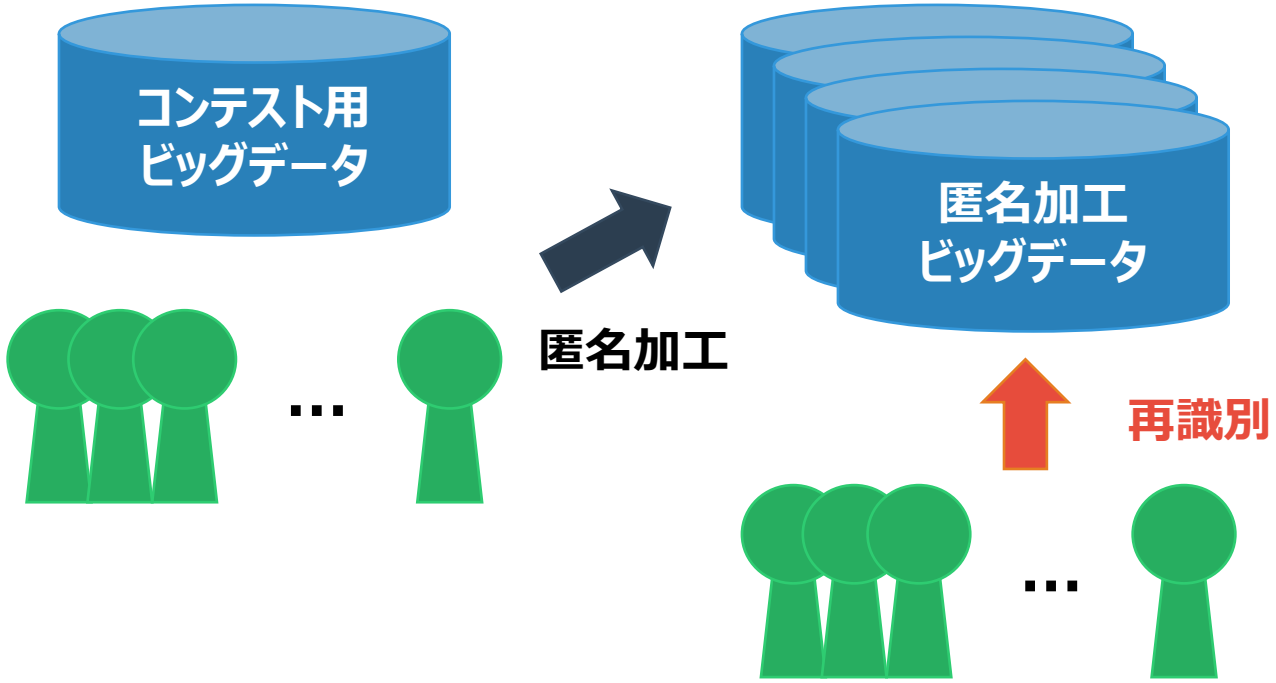
EoF

研究背景（PWSCUPとは）

MEIJI UNIVERSITY

匿名加工・再識別コンテストPWSCUP

匿名加工データの活用のために優れた加工手法や
評価指標を明らかにするコンテスト



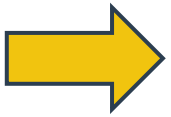
有用性評価

PWSCUP-2018の有用性指標は1つだけ.
データの2～5列目の属性値の誤差の平均値で評価する.

例

元データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/1	84879	1.69	32
13047	2010/12/2	22745	2.1	6
13047	2010/12/3	22748	2.1	6



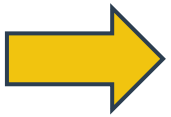
加工データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/3	10000	1.68	31
13047	2010/12/3	20000	2.0	5
13047	2010/12/3	30000	2.0	5

有用性評価(名義尺度)

元データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/1	84879	1.69	32
13047	2010/12/2	22745	2.1	6
13047	2010/12/3	22748	2.1	6



加工データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/3	10000	1.68	31
13047	2010/12/3	20000	2.0	5
13047	2010/12/3	30000	2.0	5

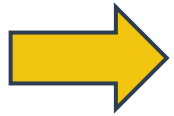
名義尺度の属性値間の誤差は，一致しているかどうかだけに注目する.
一致していたら誤差0，そうでなければ誤差1

この場合はすべて一致していないので，この属性の誤差の和は3

有用性評価(一般化されている場合)

元データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/1	84879	1.69	32
13047	2010/12/2	22745	2.1	6
13047	2010/12/3	22748	2.1	6



加工データ

顧客ID	購買日	商品ID	単価	数量
略	略	[84879,22745,22748]	略	略
略	略	[84879,22745,22748]	略	略
略	略	[84879,22745,22748]	略	略

加工データの属性値が一般化されている場合は，誤差の期待値を用いる.
※一般化された集合は一様分布であるとする.

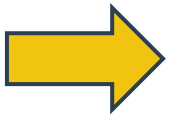
この場合，1レコード目に注目すると，誤差の期待値は

$0 * \frac{1}{3} + 1 * \frac{1}{3} + 1 * \frac{1}{3} = \frac{2}{3}$ である.

有用性評価(総合評価)

元データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/1	84879	1.69	32
13047	2010/12/2	22745	2.1	6
13047	2010/12/3	22748	2.1	6



加工データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/3	10000	1.68	31
13047	2010/12/3	20000	2.0	5
13047	2010/12/3	30000	2.0	5

この場合，誤差の総和は3+0.155+0.245+3.76=7.16

これをマス数(行数*対象属性数)で割り，平均値を求める.

7.16/(3*4)=0.60

↑

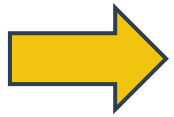
この加工データの有用性評価値

※値が削除されているときの誤差は1

有用性評価(比例尺度)

元データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/1	84879	1.69	32
13047	2010/12/2	22745	2.1	6
13047	2010/12/3	22748	2.1	6



加工データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/3	10000	1.68	31
13047	2010/12/3	20000	2.0	5
13047	2010/12/3	30000	2.0	5

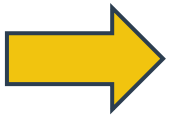
比例尺度の属性値間の誤差は，値の差を標準偏差σ_jで割ったものである.
誤差が小さいほど0に近づき，大きいほど1に近づく.

この場合単価，数量の標準偏差はそれぞれ0.193, 12.26なので，
誤差の和は順に(0.01+0.01+0.01)/0.193=0.155, (1+1+1)/12.26=0.245

有用性評価(間隔尺度)

元データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/1	84879	1.69	32
13047	2010/12/2	22745	2.1	6
13047	2010/12/3	22748	2.1	6



加工データ

顧客ID	購買日	商品ID	単価	数量
13047	2010/12/3	10000	1.68	31
13047	2010/12/3	20000	2.0	5
13047	2010/12/3	30000	2.0	5

間隔尺度の属性値間の誤差は，値を代表値からの差に置き換えてから，
比例尺度の時と同じように計算する.

この場合2010/12/1を代表値(=0)とすると，2010/12/2=1, 2010/12/3=2となる.
標準偏差は0.816なので，誤差の和は(2+1+0)/0.816=3.76