محاضرة تنقيب البيانات

قواعد الارتباط وخوارزمية Apriori

تنقيب البيانات (Data Mining) — IT401 ملزمة جامعة الحكمة — كلية العلوم والهندسة

١. تحليل سلة المشتريات

قواعد الارتباط تبحث عن العناصر التي تظهر معاً في المعاملات. المثال المعروف: متجر يريد معرفة المنتجات التي تُشترى معاً ليرتبها على الرفوف أو يقدّم عليها عروضاً.

تُكتب القاعدة بالشكل X → Y وتُقرأ: «من يشتري X يشتري Y غالباً».

٢. مقاييس القاعدة

المقياس	التعريف	المعنى
الدعم Support	عدد المعاملات التي تحوي X و Y ÷ عدد كل المعاملات	مدى شيوع القاعدة
الثقة Confidence	دعم (X و Y) ÷ دعم X	مدى صحة القاعدة عند تحقق X
الرفع Lift	الثقة ÷ دعم Y	أكبر من ١: ارتباط موجب حقيقي

٣. مثال

رقم المعاملة	العناصر
١	خبز، حليب
٢	خبز، تمر، عصير، بيض
٣	حليب، تمر، عصير، شاي
٤	خبز، حليب، تمر، عصير
٥	خبز، حليب، تمر، شاي

القاعدة {حليب، تمر} ← {عصير}: الدعم = ٢ ÷ ٥ = ٠٫٤، والثقة = ٢ ÷ ٣ = ٠٫٦٧.

٤. خوارزمية Apriori

عدد المجموعات الممكنة من n عنصراً هو 2^n، ففحصها كلها غير عملي. تعتمد الخوارزمية على خاصية بسيطة:

> **خاصية Apriori:** إذا كانت مجموعة عناصر غير متكررة، فكل مجموعة تحتويها غير متكررة أيضاً. لذلك نحذفها ولا نولّد منها شيئاً.

الخطوات

1. احسب دعم كل عنصر منفرد، واحتفظ بما يحقق الحد الأدنى للدعم ($L1$).
2. ولّد المجموعات المرشحة بحجم k من المجموعات المتكررة بحجم $k-1$.
3. احذف كل مرشح له مجموعة جزئية غير متكررة (التقليم).
4. امسح قاعدة البيانات لحساب دعم المرشحين الباقين، واحتفظ بالمتكرر (Lk).
5. كرر حتى لا تنتج مجموعات جديدة.
6. ولّد القواعد من المجموعات المتكررة، واحتفظ بما يحقق الحد الأدنى للثقة.

تطبيق على المثال (الحد الأدنى للدعم = ٣ معاملات)

المرحلة	المجموعات المتكررة
L1	{خبز}:٤، {حليب}:٤، {تمر}:٤، {عصير}:٣
L2	{خبز، حليب}:٣، {خبز، تمر}:٣، {حليب، تمر}:٣، {تمر، عصير}:٣
L3	لا شيء: {خبز، حليب، تمر} دعمها ٢ فقط

٥. توليد القواعد

من المجموعة المتكررة {تمر، عصير} نولّد قاعدتين ونحسب ثقة كل منهما:

القاعدة	الثقة	الرفع	القرار (حد الثقة ٧٠٪)
{عصير} ← {تمر}	٣ ÷ ٣ = ١٠٠٪	١٫٢٥	تُقبل
{تمر} ← {عصير}	٣ ÷ ٤ = ٧٥٪	١٫٢٥	تُقبل
{خبز} ← {حليب}	٣ ÷ ٤ = ٧٥٪	٠٫٩٤	تُقبل بالثقة، لكن الرفع أقل من ١

الثقة العالية لا تكفي وحدها. القاعدة {خبز} ← {حليب} ثقتها ٧٥٪ لكن الحليب موجود أصلاً في ٨٠٪ من المعاملات، فالرفع أقل من ١ ولا يوجد ارتباط حقيقي.

٦. عيوب الخوارزمية وتحسيناتها

- تمسح قاعدة البيانات مرة لكل حجم k، وهذا مكلف مع البيانات الكبيرة.
- عدد المرشحين قد يكون كبيراً جداً إذا كان حد الدعم منخفضاً.
- البديل: خوارزمية FP-Growth التي تبني شجرة مضغوطة وتمسح البيانات مرتين فقط.

٧. تمرين

بالبيانات نفسها وحد أدنى للدعم يساوي معاملتين: أوجد $L3$، ثم ولّد كل القواعد التي ثقتها ٨٠٪ فأكثر.