تجميع العملاء

متوسط | تجميع Clustering | Orange Data Mining | KNIME Analytics Platform | WEKA

وصف المشكلة

تقسيم العملاء إلى شرائح متشابهة السلوك (الإنفاق، الزيارات، الحداثة) دون تسميات مسبقة، واختيار عدد المجموعات بمؤشر Silhouette. البيانات اصطناعية.

الأهداف

- تطبيع الخصائص قبل التجميع
- اختيار k بمؤشر Silhouette
- تفسير ملامح كل مجموعة

البيانات

المجموعة	الحجم	المصدر	الترخيص
سلوك العملاء (اصطناعية)	300 صفاً × 4 عموداً، نحو 5.7 كيلوبايت	الأكاديمية (توليد اصطناعي)	ترخيص CC BY 4.0 من أكاديمية الذكاء الاصطناعي بدون برمجة (dlarabic.com)

الخوارزميات

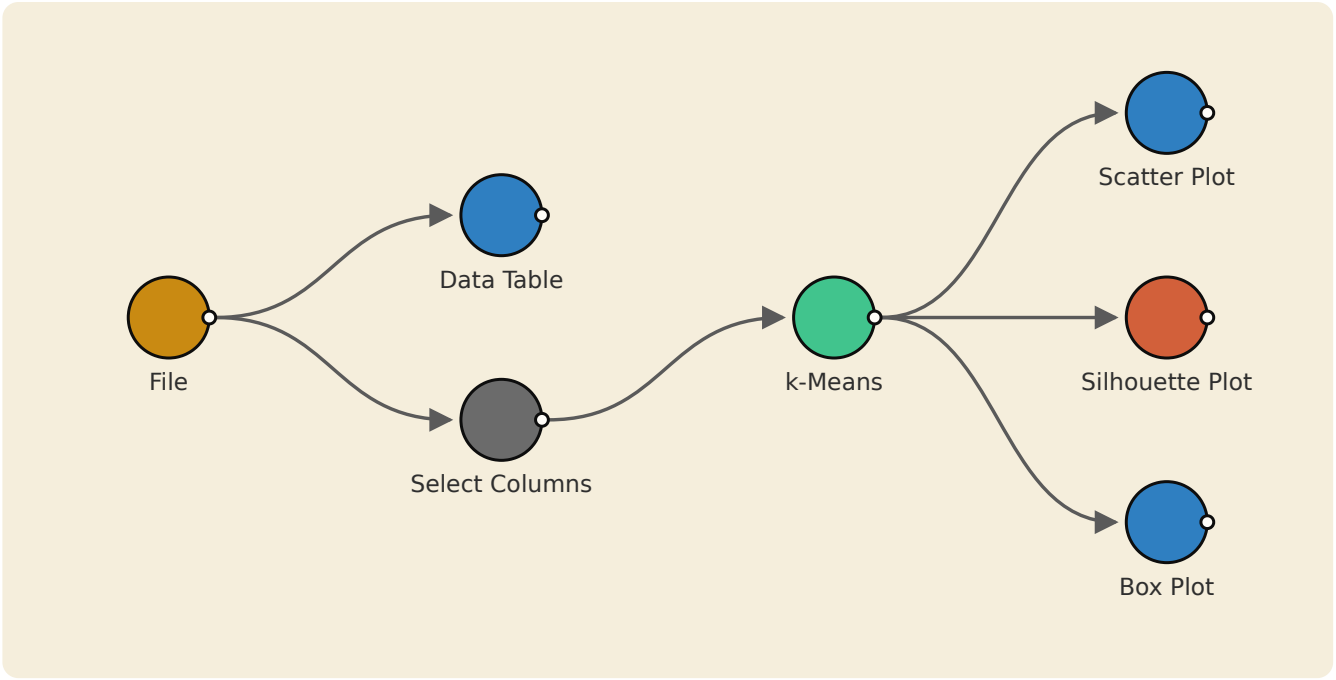
التجميع بالمتوسطات (k-Means Clustering)، التجميع الهرمي (Hierarchical Clustering)، التجميع بالكثافة (DBSCAN)

خطوات التنفيذ

1. افتح customer-segmentation.ows في Orange.
2. في k-Means اختر Normalize واجعل k من 2 إلى 7 وراجع Silhouette.
3. صل النتائج بـ Scatter Plot ولوّن حسب المجموعة.
4. استبعد customer_id من الخصائص (عنصر Select Columns أو دوره meta).

مخطط سير العمل

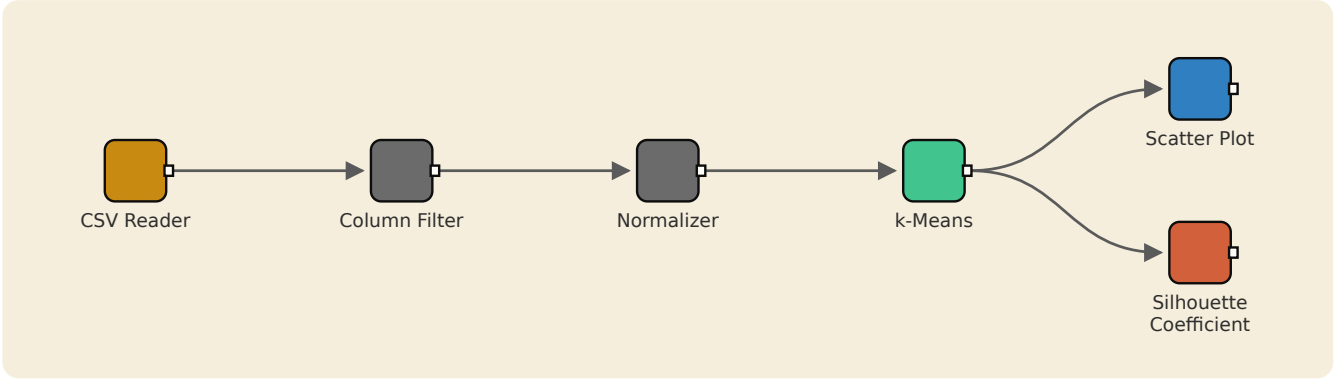
Orange



مرسوم من ملف سير العمل بصيغة OWS المرفق في الحزمة.

1. File → Data Table
2. File → Select Columns
3. Select Columns → k-Means
4. k-Means → Scatter Plot
5. k-Means → Silhouette Plot
6. k-Means → Box Plot

KNIME



مخطط توضيحي بأسماء عُقد KNIME لتبنيه بنفسك، وليس ملف KNWF.

1. CSV Reader → Column Filter
2. Column Filter → Normalizer
3. Normalizer → k-Means
4. k-Means → Scatter Plot
5. k-Means → Silhouette Coefficient

الملفات

الملف	المسار داخل الحزمة / الرابط

مجموعة البيانات CSV	customers-synthetic.csv
ملف WEKA بصيغة ARFF	customers-synthetic.arff
سير عمل Orange (.ows)	customer-segmentation.ows
البيانات بصيغة Orange (.tab) لسير العمل	customers-synthetic.tab
النتائج المرجعية المحسوبة (CSV)	customers-clusters.csv

ملفات KNIME Workflow بصيغة KNWF غير مرفقة لأنها لم تُبنَ وتُختبر داخل KNIME؛ مخطط عُقد KNIME مرسوم أعلاه لتبنيه بنفسك.

النتائج

نتائج فعلية محسوبة على الملفات المرفقة

k-Means (scikit-learn 1.9.1) على خصائص مطبّعة. Silhouette حسب k: k=2: 0.566، k=3: 0.652، k=4: 0.771، k=5: 0.638، k=6: 0.486، k=7: 0.383

أفضل k = 4. ملامح المجموعات (المتوسطات):

- المجموعة 0 (63 عميلاً): إنفاق 842.6، زيارات 2.4/شهر، آخر زيارة قبل 60.5 يوماً
- المجموعة 1 (82 عميلاً): إنفاق 2,583.3، زيارات 19.9/شهر، آخر زيارة قبل 3.3 يوماً
- المجموعة 2 (90 عميلاً): إنفاق 1,197.6، زيارات 3.8/شهر، آخر زيارة قبل 9.9 يوماً
- المجموعة 3 (65 عميلاً): إنفاق 5,230.4، زيارات 11.8/شهر، آخر زيارة قبل 4.5 يوماً

مقاييس التقييم

- Silhouette
- أحجام المجموعات وملامحها

تفسير النتائج

أربع شرائح واضحة: عملاء كبار الإنفاق، وعملاء كثيرو الزيارة، وعملاء عاديون نشطون، وعملاء خاملون لم يزوروا منذ شهرين (مرشحون لحملة استعادة). أرقام المجموعات تعسفية؛ المهم هو الملامح.

أسئلة للمناقشة

- ماذا يحدث لو لم نطبّع الإنفاق السنوي؟
- كيف تسمي كل شريحة لفريق التسويق؟
- متى تفضل DBSCAN على k-Means؟

البيانات الاصطناعية في هذه الحزمة مولّدة بالبرمجة لأغراض تعليمية (موسومة بكلمة synthetic) وليست بيانات حقيقية. الأرقام في قسم النتائج محسوبة فعلياً على الملفات المرفقة فقط حيث يُذكر ذلك. — dlarabic.com