مقارنة خوارزميات تعلم الآلة

متقدم | تصنيف Classification | Orange Data Mining | WEKA | KNIME Analytics Platform

وصف المشكلة

مقارنة منهجية لست خوارزميات تصنيف على بيانات Wine الحقيقية بنفس التقسيم والمقاييس، وفهم أثر التطبيع والإعدادات الافتراضية على النتيجة.

الأهداف

- تصميم مقارنة عادلة (نفس الطيات والمقاييس)
- قراءة الانحراف المعياري لا المتوسط فقط
- فهم أثر التطبيع

البيانات

المجموعة	الحجم	المصدر	الترخيص
أنواع النبيذ Wine	178 صفاً × 14 عموداً، نحو 13.0 كيلوبايت	UCI Machine Learning Repository (Aeberhard & Forina, 1992)، النسخة المرفقة بمكتبة scikit-learn	CC BY 4.0

الخوارزميات

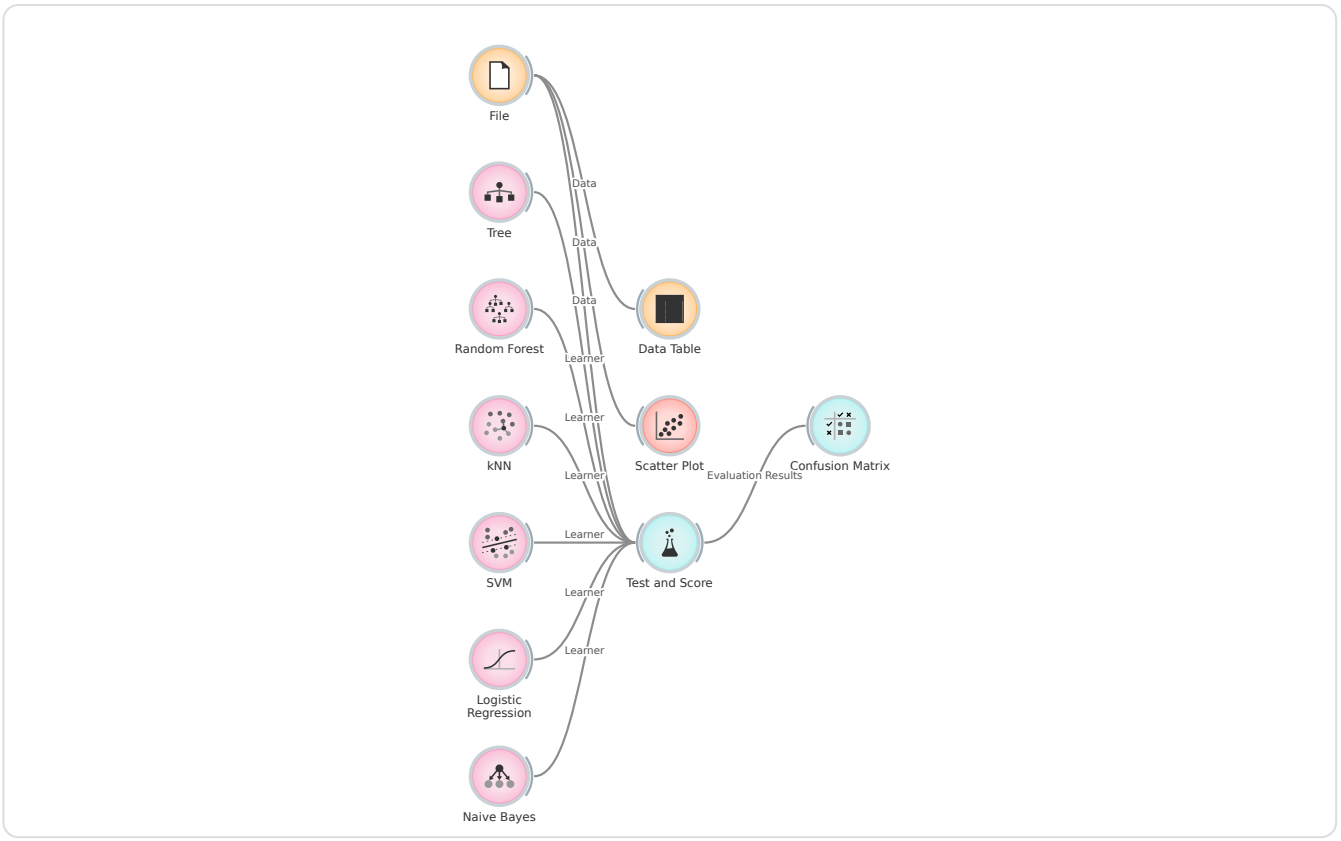
شجرة القرار (Decision Tree)، الغابة العشوائية (Random Forest)، أقرب الجيران (k-Nearest Neighbors (KNN))، آلة المتجهات الداعمة (Support Vector Machine (SVM))، الانحدار اللوجستي (Logistic Regression)، بايز الساذج (Naive Bayes)

خطوات التنفيذ

1. في KNIME: File ← Import KNIME Workflow... واختر algorithms-comparison.knwf، ثم Execute all.
2. افتح algorithms-comparison.ows في Orange: ست خوارزميات متصلة بـ Test and Score واحد.
3. في WEKA: Experimenter ← أضف wine.arff وست خوارزميات ← Analyse ← Run مع Paired T-Tester.
4. استخدم مختبر مقارنة النماذج في الأكاديمية لتجربة سريعة في المتصفح.

مخطط سير العمل

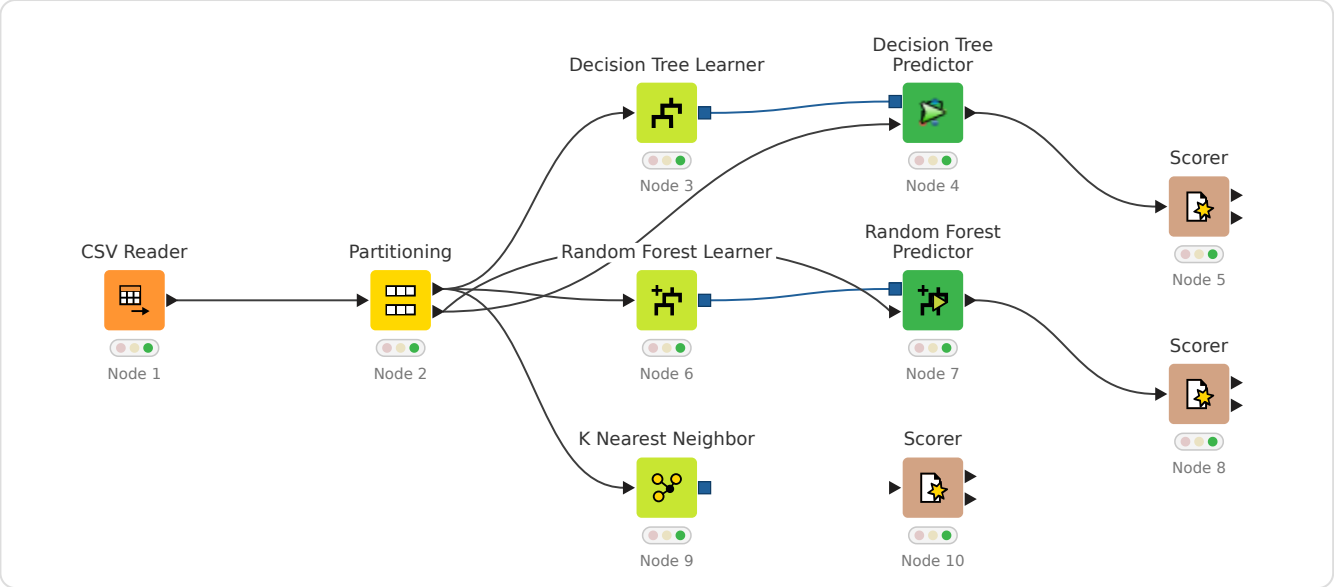
Orange



مرسوم من ملف سير العمل بصيغة OWS المرفق في الحزمة، بأيقونات العناصر الأصلية وألوان فئاتها وأسماء القنوات.

1. File → Data Table (Data)
2. File → Test and Score (Data)
3. Tree → Test and Score (Learner)
4. Random Forest → Test and Score (Learner)
5. kNN → Test and Score (Learner)
6. SVM → Test and Score (Learner)
7. Logistic Regression → Test and Score (Learner)
8. Naive Bayes → Test and Score (Learner)
9. Test and Score → Confusion Matrix (Evaluation Results)
10. File → Scatter Plot (Data)

KNIME



مخطط بأسماء عُقد KNIME وأيقوناتها الأصلية وألوان أنواعها، مطابق لملف KNWF المرفق في الحزمة (نُفّذ بنجاح في محرك KNIME 5.12).

1. CSV Reader → Partitioning
2. Partitioning → Decision Tree Learner
3. Decision Tree Learner → Decision Tree Predictor
4. Decision Tree Predictor → Scorer
5. Partitioning → Decision Tree Predictor
6. Partitioning → Random Forest Learner
7. Random Forest Learner → Random Forest Predictor
8. Random Forest Predictor → Scorer
9. Partitioning → Random Forest Predictor
10. Partitioning → K Nearest Neighbor
11. K Nearest Neighbor → Scorer

الملفات

الملف	المسار داخل الحزمة / الرابط
مجموعة البيانات CSV	wine.csv
ملف WEKA بصيغة ARFF	wine.arff
سير عمل Orange (.ows)	algorithms-comparison.ows
البيانات بصيغة Orange (.tab) لسير العمل	wine.tab
النتائج المرجعية المحسوبة (CSV)	wine-results.csv
سير عمل KNIME (.knwf): نُفّذ بنجاح في محرك KNIME 5.12 (لم يُفتح في الواجهة)	algorithms-comparison.knwf

ملف KNIME (.knwf) بُني على KNIME 5.12 ونُفّذت كل عُقده بنجاح في محرك KNIME دون فتح الواجهة؛ افتحه من File ← Import KNIME Workflow.

النتائج

نتائج فعلية محسوبة على الملفات المرفقة

مرجع scikit-learn 1.9.1، 10-fold stratified CV, shuffle, random_state=42:

- Decision Tree: Accuracy 90.0% (± 7.8%), F1-macro 0.902
- Random Forest: Accuracy 98.9% (± 2.2%), F1-macro 0.989
- kNN (k=5): Accuracy 97.2% (± 4.5%), F1-macro 0.973
- SVM (RBF): Accuracy 97.8% (± 3.7%), F1-macro 0.977
- Logistic Regression: Accuracy 98.3% (± 2.5%), F1-macro 0.983
- Naive Bayes: Accuracy 97.8% (± 3.7%), F1-macro 0.978

مرجع scikit-learn يطبّع الخصائص قبل kNN SVMg Logistic Regressiong.

ونتائج مكتبة Orange 3.41.0 (10-fold، كإعدادات Test and Score الافتراضية):

- Tree: CA 91.0%, F1 0.910, AUC 0.931
- Random Forest: CA 97.2%, F1 0.972, AUC 0.997
- kNN: CA 67.4%, F1 0.667, AUC 0.865
- SVM: CA 98.3%, F1 0.983, AUC 0.987
- Logistic Regression: CA 94.4%, F1 0.944, AUC 0.992
- Naive Bayes: CA 97.2%, F1 0.972, AUC 0.997

مقاييس التقييم

- Accuracy مع الانحراف المعياري
- F1-macro
- ROC AUC

تفسير النتائج

أغلب الخوارزميات تتجاوز 97% في مرجع scikit-learn، والفروق بينها داخل الانحراف المعياري، فلا يصح إعلان «فائز» دون اختبار إحصائي. لاحظ هبوط kNN في Orange إلى 67.4% لأن المسافات تحسب هنا على خصائص غير مطبّعة (البرولين بالمئات يطغى على البقية)؛ أضف Normalize ← Preprocess قبل kNN وأعد التجربة؛ تحققنا بمكتبة Orange نفسها أن إضافة التطبيع ترفع kNN إلى 96.1% بالطيات ذاتها.

أسئلة للمناقشة

- لماذا لا نحكم بالفائز من فرق 0.5%؟
- ما أثر التطبيع على kNN SVMg وعلى الأشجار؟
- كيف تصمم مقارنة لا تتسرب فيها معلومات الاختبار إلى التطبيع؟

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