كلية الهندسة المعلوماتية - قسم الذكاء الصنعي
عملي استكشاف المعرفة 2023-2024
أسماء الطلاب: احمد الفريحان – معاذ الحلبي عبد الهادي رضا
اسم المجلة الناشرة: HAL open science

Disguised Missing Values

الطريقة :

DISGUISED NULL DETECTION THROUGH EMBEDDING AND CLASSIFICATION

يمكننا استخراج الكيانات من جزء صغير من البيانات ثم نقوم بانشاء نموذج ML وتدريبه لجعله قادرا على التعرف على القيم المحذوفة المقنعةDisguised missing values بالطبع قمنا باستخدام خوارزمية K-Means لانشاء عينات صغيرة للتدريب وتقوم خوارزمية التجميع K-Means السابقة بتجميع القيم المشابه وعند وضع عدد المجموعات 10 نستطيع ملاحظة ذلك ان القيم المقنعة قد تجمعت لوحدها لكن هذه الخوارزمية لا تعطينا فصل جيد بين القيم لذلك قمنا باستخدام نموذج تعلم الي ليضمن لنا نتائج اكثر دقة

تجميع القيم :(Clustering) هو عملية تقسيم البيانات إلى مجموعات بناءً على التشابه. في حالةDMVs ، نحاول تجميع القيم التي قد تكون معدومة متخفية في مجموعة واحدة.

ونعني بمصطلح استخراج الكيانات بان نقوم بتحديد العناصر الهامة في النصوص على سبيل المثال الأسماء او الأماكن او تواريخ الحوادث كما مر معنا في Chicago datasets
اتضح ان هذه الخوارزمية هي اسرع من قرينتها (*entity profiles*)

والان دعنا نتناول كيف يتم ذلك .. الطريقة الأكثر شيوعا هي عن طريق تقسم البيانات للتدريب والاختبار على النحو التالي 80% للاختبار و 20% للتدريب (الهدف هنا هو تقليل حجم مجموعة التدريب لتوفير الوقت)
واتضح ان العملية الأكثر استهلاكا للوقت عملية تصنيف القيم :يتم استخدام تقنية استخراج الكيانات (entity extraction technique) لتصنيف مجموعة التدريب، وهي عملية تستغرق وقتًا طويلاً.

أي وضع labels او مسميات لمجموعات التدريب ع حسب الكيان المستخرجة منه وكما هو موضح قمنا بتغيير حجم بيانات التدريب كل مرة على ثلاث مراحل 16.477(20%) و 8.238 (10%)و 823 (1%)
واكتشفنا انه يمكننا الحصول على نتائج دقيقة بحجم بيانات تدريب صغير …

مجموعة البيانات :

قمنا بتحميل مجموعات بيانات

1- HATVP XML وسنسميها DS1
2- وأخرى اصغر (2.1 MB) HATVP CSV واسمها DS2
3- PubMed bibliographic notices واسمها DS3
4- DDS3 هي نفسها DS3 لكن بإزالة أي قيم مكررة ملاحظة الاختبارات ستكون ع الرابعة

الاختبارات والنتائج

بناء بيان او ConnectionLens Graph
مصدر البيانات :تم جمع البيانات منPubMed ، وهي قاعدة بيانات للأبحاث العلمية والطبية

الأداة المستخدمة ConnectionLens :، وهي أداة لإنشاء الرسوم البيانية واستخراج البيانات

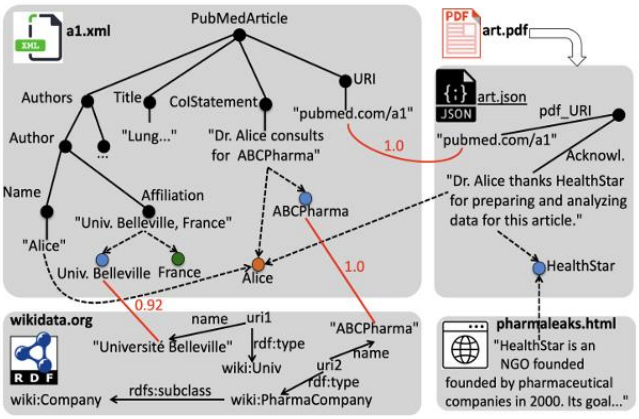

Figure 1: Sample ConnectionLens graph.

استغرق DDS3 سابقا حوالي 11,000 ثانية.
والان باستخدام طريقتنا (embedding and classification) مع مجموعة التدريب المكونة من 1% من القيم (823 قيمة) .
تستغرق من الوقت الآن للتنبؤ بالقيم التي يتعين علينا تطبيق المستخرج عليها (125
ثانية)، نضيف إليها الوقت اللازم لاستخراج القيم المهمة . لقد وجدنا في مجموعة البيانات الخاصة بنا أن هناك حوالي 45.000 قيمة مهمة. ونحن بحاجة إلى 5.900 ثانية لاستخراج تلك. وهذا يقودنا إلى إجمالي 6,000 ثانية لبناء الرسم البياني الخاص بنا بدلاً من 11,000 ثواني سابقة
(عند تطبيق طريقة ENTITY PROFILE) دون فقدان الكثير من المعلومات.
وهذا يرينا مدى التحسن على الصعيدين الدقة والوقت حيث انه باستخدام مجموعة تدريب صغيرة 1% تم تقليل الوقت من 11000 الى 6000 ثانية وأيضا الدقة بلا شك هي جيدة جدا والاسترجاع يكون اقل حساسية لتقليل حجم مجموعة التدريب مما يعني ان النموذج لايفقد الكثير من المعلومات الهامة

Values in Training-set	16.477	8.238	823
Extraction Times (s)	2.153	1.075	108
Training Times (s)	54	23	6
Prediction Times (s)	10	10	11
Total Times (s)	2.217	1.108	125
Precision	0,939	0,933	0,885
Recall	0,948	0,946	0,942
F1-score	0,943	0,940	0,913

الشرح (او المختصر المفيد:)
وقت التنبؤ :النموذج يتنبأ في 125 ثانية بالقيم التي تحتاج لاستخراج الكيانات.
استخراج الكيانات للقيم الهامة : تم استخراج من 45000 قيمة هامة فقط واستغرقت العملية 5900 ثانية
والمجموع هو 5900+125 = تقريبا 6000 ثانية

القيود والسلبيات

القيود :

الحاجة إلى مجموعة تدريب موسومة:

الوقت المستغرق :عملية وسم مجموعة التدريب باستخدام تقنية استخراج الكيانات تستغرق وقتًا طويلًا. حتى لو تم تقليل حجم مجموعة التدريب، فإن عملية الوسم تظل عملية مستهلكة للوقت.
الحجم الصغير لمجموعة التدريب :في حال كان حجم مجموعة التدريب صغيرًا جدًا، قد يؤثر ذلك على دقة النموذج في التعرف على القيم المفقودة المتخفية.(DMVs)
تعقيد النموذج المستخدم:
التدريب على نماذج معقدة :استخدام نماذج مثل "Random Forests" قد يكون بطيئًا في التدريب مقارنة بنماذج أبسط، مما قد يكون غير عملي في حالات البيانات الكبيرة جدًا.

السلبيات:

الحساسية لانخفاض الدقة:
الدقة مقابل الاسترجاع :على الرغم من أن النموذج قد يظهر دقة جيدة حتى عند تقليل حجم مجموعة التدريب، إلا أن الدقة تكون أقل حساسية من الاسترجاع. هذا يعني أن النموذج قد يفوت بعض القيم المفقودة المتخفية المهمة، مما قد يؤدي إلى فقدان معلومات قيمة.
استخراج الكيانات من النصوص:
تكلفة الحساب :عملية استخراج الكيانات تظل باهظة التكلفة من حيث الوقت والحسابات، حتى لو كانت تُجرى على جزء صغير من البيانات.
التعقيد الزمني :استخدام تقنيات مثل "BERT" لاستخراج التضمينات النصية يعتبر معقدًا وبطيئًا، مما يعارض هدف تقليل الوقت المستهلك.

المصادر:

[الورقة البحثية]

HAL OPEN SCIENCE :
inria.hal.science/hal-03347947