كلية الهندسة المعلوماتية

السنة الخامسة

استكشاف المعرفة

بحث في طريقة معالجة القيم المحذوفة المقنعة

Disguised missing values

تقدمة الطلاب:

طوني بطرس عبد الرؤوف حسحس

عبد الله الزبداني

اسم البحث:

Efficiently identifying disguised missing values in heterogeneous, text-rich data

اسم المجلة الناشرة:

نشر هذا البحث في مجلة " Transactions on Large-Scale Data- and Knowledge-Centered Systems" ضمن العدد الخاص بإدارة البيانات - المبادئ، التكنولوجيات، والتطبيقات.

ملخص الورقة البحثية:

يتم إنتاج البيانات الرقمية في العديد من النماذج البيانية، بدءًا من الهياكل المنظمة بشكل كبير (عادة ما تكون علاقية) إلى النماذج شبه المنظمة (XML، JSON) وصولاً إلى تنسيقات الرسوم البيانية المختلفة (RDF، الرسوم الخاصة بالخصائص) أو النصوص. تحتوي معظم مجموعات البيانات الواقعية على كمية معينة من القيم الفارغة، التي تشير إلى البيانات المفقودة أو غير المعروفة أو غير القابلة للتطبيق. في حين أن بعض النماذج البيانية تسمح بتمثيل القيم الفارغة باستخدام رموز خاصة بها، فإن القيم المفقودة المقنعة (DMVs، بالاختصار) تواجه أيضًا بشكل متكرر: إذ تعتبر هذه القيم قيمًا ليست فارغة بصريًا، لكنها تشير، مع ذلك، إلى عدم وجود المعلومات أو عدم توفرها أو عدم قابليتها للتطبيق. في هذا العمل، نتعامل مع اكتشاف نوع معين من DMV: النصوص التي يدخلها المستخدمون بحرية. لم يتم التعامل مع هذه المشكلة من قبل طرق اكتشاف DMV المركزة على البيانات العددية أو التصنيفية. بالإضافة إلى ذلك،

تفلت هذه المشكلة من طرق اكتشاف DMV المعتمدة على تكرار القيم، لأن مثل هذه النصوص الحرة غالبًا ما تكون مختلفة عن بعضها البعض، وبالتالي فإن معظم DMVs فريدة. واجهنا هذه المشكلة داخل مشروع ConnectionLens حيث تتم دمج البيانات الغير متجانسة في رسوم بيانية كبيرة. نقدم اثنين من طرق اكتشاف DMV لمشكلتنا الخاصة: (أ) عن طريق استغلال استخراج المعلومات، المستخدمة بالفعل في رسوم ConnectionLens؛ و (ب) من خلال تضمينات النصوص والتصنيف. نوضح تضحيات الأداء والدقة على مجموعات بيانات العالم الحقيقي.

شرح طريقة الحل:

كان الهدف من البحث كشف القيم المفقودة المقنعة DMVs في البيانات النصية المتنوعة.
تناول طريقتين رئيسيتين:

- استغلال استخراج المعلومات: تعتمد هذه الطريقة على استخدام نظام ConnectionLens الذي يقوم باستخراج الكيانات المسماة من النصوص. يتم تحديد القيم المفقودة المتنكرة من خلال تحليل النصوص المستخرجة وتحديد ما إذا كانت تحتوي على الكيانات المتوقعة أم لا. على سبيل المثال، إذا كانت النصوص تتعلق بمعلومات حول الشركات، فإن النصوص التي لا تحتوي على أي كيان من نوع "شركة" تعتبر قيم مفقودة متنكرة.
يتم استخدام ملفات تعريف الكيانات (entity profiles) لتحديد الكيانات المتوقعة في النصوص المختلفة.

- التصنيف باستخدام تمثيلات النصوص: تعتمد هذه الطريقة على استخدام تقنيات التضمين النصي (text embeddings) مثل TF-IDF وSentence-BERT لتحويل النصوص إلى تمثيلات عددية.

يتم تدريب نموذج تصنيف (Random Forest) باستخدام مجموعة بيانات صغيرة مصنفة يدويًا (تم تصنيفها باستخدام الطريقة الأولى) لتحديد ما إذا كانت النصوص المماثلة هي قيم مفقودة متنكرة.
يتم استخدام النموذج المدرب لتصنيف النصوص الجديدة وتحديد القيم المفقودة المتنكرة بدون الحاجة إلى استخراج الكيانات بشكل كامل لكل نص.

مجموعات البيانات التي استخدمت للاختبار ونتائج الاختبار:

مجموعة بيانات HATVP --> 35 MB

تحتوي على بيانات عن 270000 شخص.
الفكرة أنه تضمنت القيم المفقودة المتنكرة مبالغ مالية مثل 0 والتي كانت قد تكررت بشكل كبير.
النتيجة: تم اكتشاف القيم المفقودة المتنكرة بنجاح .

مجموعة بيانات HATVP أصغر --> 2.1 MB

تتضمن معلومات عن 9,000 شخص.
تحتوي على قيم مثل "dispense" التي كانت تستخدم كقيم مفقودة متكررة.
تم اكتشاف القيم المفقودة المتنكرة بنجاح باستخدام الطريقة الأولى.
400,000 إشعار ببليوغرافي من PubMed:

تحتوي على عبارات تعارض المصالح (CoI) التي يتم ملؤها من قبل المؤلفين.
تضمنت القيم المفقودة المتنكرة عبارات مثل " The authors report "no conflict of interest".

تم اكتشاف معظم القيم المفقودة المتنكرة بنجاح باستخدام الطريقة الثانية، حيث أظهرت نتائج جيدة في تقليل وقت استخراج الكيانات بنسبة كبيرة.

النتائج:

حققت الطريقة الأولى دقة عالية في كشف DMVs ولكن بكلفة حسابية عالية.
قدمت الطريقة الثانية نتائج دقيقة وكفاءة في الأداء من حيث توفير الوقت المستغرق في استخراج الكيانات.
في اختبار عملي على مجموعة بيانات PubMed بدون تكرار، وفرت الطريقة الثانية حوالي 45% من وقت استخراج الكيانات مقارنة بالطريقة الأولى.

السلبيات أو القيود:

الطريقة الأولى (استخدام استخراج المعلومات):

تتطلب وقتًا طويلًا ومعالجة حسابية مكثفة بسبب ضرورة استخراج الكيانات من جميع النصوص.
تعتمد دقتها على دقة استخراج الكيانات، وأي أخطاء في استخراج الكيانات يمكن أن تؤدي إلى أخطاء في تحديد DMVs.
الطريقة الثانية (التصنيف باستخدام تمثيلات النصوص):

تعتمد على جودة ودقة نموذج التصنيف المدرب، مما قد يتطلب معالجة إضافية وتحسينات دورية للحفاظ على دقة عالية.
قد لا تكون فعالة بشكل كامل في البيئات التي تتغير فيها النصوص بشكل كبير ودوري، مما يتطلب إعادة تدريب النموذج بشكل منتظم.