معالجة البيانات المفقودة في السجلات الصحية الالكترونية لمرضى الأمراض المزمنة

مرام عبدالله باعطية

باشراف / د. أروى عبدالله جمجوم

المستخلص

تُعد أنظمة السجلات الصحية الإلكترونية EHR مهمة لتشغيل الممارسات الطبية. يتم استخدامها من قبل الأطباء وصناع القرار لتتبع جميع جوانب رعاية المرضى. عادةً ما يكون لدى مرضى الأمراض المزمنة مجموعات بيانات واسعة تخزن كمية كبيرة من البيانات المختلفة لتتبع التاريخ الطبي للمرضى. لذلك فإن وجود قيم مفقودة في أنظمة السجلات الصحية الإلكترونية هو أكبر عقبة تواجه الباحثين والطاقم الطبي. في الآونة الأخيرة يلعب تعليم الآلة "ML" دورًا مهمًا في مختلف القطاعات بالإضافة إلى تقنية الحوسبة، لتحسين الخدمات المقدمة وتحسين جودة البيانات. لذلك تمّ استخدام تقنيات ML لاحتساب البيانات الطبية المفقودة. ومع ذلك فقد وُجد أنه تم استخدام طرق مختلفة لعملية التعويض، ولكن تم تصميم كل منها لمعالجة قضايا معينة.

تهدف هذه الرسالة إلى اقتراح منهجية تعاودية هجينة، تُستخدم لحساب البيانات المفقودة في EHR لمرضى الأمراض المزمنة (مرضى السكري كدراسة حالة). يشار إلى هذه المنهجية باسم "C2I" وتعني (التجميع Clustering، التقليد Imitation، التعويض Imputation). تمّ فحص C2I باستخدام مجموعة بيانات مرض السكري بأثر رجعي من مستشفى الملك عبد العزيز الجامعي. في هذه الدراسة اختيرت المتغيرات المترابطة ذات الصلة لمرضى السكري. بعد ذلك تمّ تقليد أنماط ونسب القيم المفقودة في السجلات الكاملة من مجموعة بيانات مرض السكري. بعد ذلك تمّ حساب القيم المفقودة المقلدة باستخدام C2I. ثم قياس أداء C2I

بثلاث تقنيات مختلفة: (١) حساب مقاييس خطأ الحساب، و (٢) مقارنة الأداء بالطريقة شائعة الاستخدام "التعويض المتعدد MI"، و (iii) تطبيق C2I على مجموعة البيانات الكاملة.

في C2I، تم استخدام طرق ML خاضعة للرقابة وغير خاضعة للرقابة. ومن المثير للاهتمام أن C2I حقق نتائج إيجابية ملحوظة أفضل بنسبة ١٩٪ من معدلات الخطأ عبر MI. علاوة على ذلك، يقلل C2I التشتت في مجموعة البيانات ويحتفظ بنطاقات متغيراتها. كما تشير نتائج هذه الدراسة إلى أن نسبة فقد البيانات العالية للمتغير المرتبط تتسبب في معدل خطأ مرتفع للمتغير المفقود. المزيد من الدراسات التي تأخذ في الاعتبار المتغيرات المرتبطة سنحتاج إلى إجرائها بأساليب جديدة مثل التعلم العميق والشبكات العصبية.

Handling Missing Data in Electronic Health Records for Chronic Diseases Patients

Maram Abdullah Baatya

Supervised by: Dr. Arwa Abdullah Jamjoom

ABSTRACT

Electronic Health Records 'EHR' systems are crucial to the operations of medical practices. They are used by physicians and decision-makers to track all aspects of patients care. Chronic diseases patients usually have extensive data collections that store a huge amount of various data to track patients' medical history. Therefore, having missing values in EHR systems is quite the most substantial obstacle that faces researchers and medical staff. Recently, machine learning 'ML' plays an important role in different sectors in addition to computing technology, to improve the provided services and enhance the quality of data. Therefore, ML techniques have been used to impute clinical missing data. However, it has been found that various methods were used for the imputation process, but each of them tackles particular issues.

This thesis aims to propose a hybrid recursive imputation method to impute missing data in chronic disease patients' EHR (diabetes patients EHR as a case study). This method referred to as 'C2I' and it stands for (Clustering, Imitation, Imputation), C2I has been examined with a retrospective diabetes dataset from King Abdulaziz University Hospital and the most related variables to diabetes patients. Then, the missingness patterns and percentages of the diabetes dataset are imitated in the complete records. Later, the imitated missing values are imputed using C2I. The performance of C2I was measured with three different techniques: (i) calculate imputation error measures, (ii) compare the performance with the commonly used method Multiple Imputation 'MI', and (iii) apply C2I on the complete dataset.

In C2I, supervised and unsupervised ML methods were used. Interestingly, C2I produces better positive results by 19% of error rates over MI. Furthermore, C2I decreases the dispersion in the diabetes dataset. The results of this study indicate that a high missing percent of a related variable causes a high error rate. Further studies, which take the correlated variables into account, will need to be undertaken with new methods like deep learning and neural networks.