تطوير نماذج التنقيب عن البيانات لتحليل البيانات الضخمة
دراسة حالة: (تحسين دقة التنبؤ بالعدوى المكتسبة في المستشفيات)

عمر سامي باعيسى

المستخلص

تستهدف هذه الرسالة تطوير نماذج التنقيب عن البيانات لتحليل البيانات الضخمة وذلك لتحسين دقة التنبؤ للعدوى المكتسبة في المستشفيات. فقد تم تطوير خوارزم جديد للتصنيف يقوم بالتنقيب عن البيانت وذلك لحل المعضلات في التقنيات الحالية. حيث أن نماذج التنقيب عن البيانات التقليدية لا تواكب احتياجات خصائص البيانات الضخمة. كما أن هذا النوع من البيانات يضيف العديد من التحديات والمعوقات لعملية التحليل. حيث يمتاز بحجم البيانات الضخم واختلاف أنواعها، كما تمثل سرعة وصول وتحليل البيانات أحد اهم تصنيفات البيانات الضخمة، ويضاف لذلك ضعف جودة البيانات بسبب تعدد مصادرها. وتتم عملية التنقيب بإتباع خطوات الإكتشاف المعروفة لتبدأ بفهم المشكلة والأهداف، ثم جمع البيانات وتنظيمها، ثم اختيار العوامل المستخدمة في التنقيب، ثم تطبيق خوارزميات التنقيب، وأخيراً فهم وتقييم النتائج. وقد تم ~~اختبار~~ مقارنة الخوارزم الجديد (خوارزم التصنيف باستخدام وزن القيمة المميزة) ~~ضد~~ مع أفضل خوارزم مستخدم للتنبؤ بحالات العدوى المكتسبة في المستشفيات (خوارزم نيف-بايز) حسب تقييم الأفضلية في البحث. حيث تمت عمل مراجعه لأكثر الخوارزميات المستخدمة في المجال الطبي بخلاف نوع التطبيق و من ثم حصر مميزات و عيوب كل خوارزم. و تم عمل مراجعة للخوارزميات المستخدمة للتنبؤ بالعدوى المكتسبة لإختيار افضلها للمقارنة. ثم تمت المفاضلة بين سبعة خوارزميات مختلفة لإختيار الافضل و من ثم مقارنته بالخوارزم الجديد. وقد أثبت خوارزم التصنيف باستخدام وزن القيمة المميزة امكانية التنبؤ بالعدوى قبل ثلاثة الى خمسة أيام قبل التأكيد بالتحاليل الطبية المطلوبة. تتيح هذه الأفضلية التدخل الطبي المبكر للحالات المرضية مما يساعد في التعجيل بالشفاء و حماية الممارسين الطبيين من الإصابة بالعدوى و تقليل مدة الإقامة في المستشفى وبالتالي تخفيض التكاليف. و قد تمت عملية الاختبارات و المقارنات باستخدام بيانات طبية حقيقية من مدينة الملك عبدالله الطبية بمكة. حيث احتوت البيانات المستخدمة على ما يقارب تسعة و عشرين ألف حالة مرضية لمدة خمس سنوات و معلومات مكونة من نظام المستشفى الالكتروني و نظام المختبر الطبي الالكتروني و نظام الاشعة الالكتروني و ملاحظات الأطباء الالكترونية. وقد اعتمد التقييم على عدة مقاييس أهمها دقة النتائج، والكفاءة الحاسوبية، وسرعة انجاز الكشف عن العدوى، وصحة المخرجات من حيث صحة الحالات الإيجابية وصحة الحالات السلبية المكتشفة. و من أهم العوامل الإيجابية للنموذج تسريع عملية الكشف عن المرض، والحصول على نتائج ذات دقه عالية، وتقليل اعتماد القرار النهائي على الخبرة الشخصية للأفراد.

Development of Data-Mining Models for Big Data Analytics Case Study: Improving Prediction Accuracy of Hospital Acquired Infection

Omar Sami Baeissa

Abstract

This research focuses on developing data-mining models for big data analytics to improve prediction accuracy for Healthcare-Associated Infections (HAIs) as a case study. A new data-mining classification algorithm was created to overcome some of the issues plaguing this endeavor. The new algorithm was assessed against the best data mining techniques defined by this study and was found to minimize prediction times of HAIs by three to five days. The early prediction enables immediate intervention by clinical staff, which speeds up the recovery time and minimizes harm to the patient. Traditional algorithms are inapplicable to the target domain analytic process. Big Data raises the bar as a result of using additional features. It is characterized mainly by a tremendous amount of data in different forms. It also deals with rapid data flow rates that are generated from multiple sources, and to top it off, the quality of the data is questionable. The research surveys data-mining algorithms in healthcare to define strengths and weaknesses. Accordingly, it defines the most suitable techniques and compares them to similar approaches. The proposed classification algorithm was evaluated with real patients' data from King Abdullah Medical City, Makkah. It contains more than 28,000 cases that consist of laboratory results, radiology findings, surgical histories, and physicians' notes. The outcomes prove that the developed classification algorithm is superior to the others.