الملخص

أحد الحقول الناشئة بسرعة هو حقل البيانات الضخمة. البيانات الضخمة هي كلمة صاخبة تصف البيانات التي تتميز بالحجم الكبير والسرعة والتنوع والصحة. تُستخدم تحليلات وأدوات البيانات الضخمة لاستخراج المعلومات والتنبؤات القيمة من هذه البيانات بطريقة فعالة. مجال الرعاية الصحية هو أحد المجالات الواعدة لتطبيق تكنولوجيا البيانات الضخمة. لسوء الحظ ، لم يتم تطبيق البيانات الضخمة حتى الآن على نطاق واسع في هذا المجال لإمكاناته القصوى. وينبع هذا من أسباب عديدة بما في ذلك تنوع مصادر البيانات التي تؤدي إلى بيانات غير متجانسة من أنواع مختلفة (منظمة وشبه منظمة وغير منظمة). تمثل معالجة هذه البيانات وتحليلها تحديًا مهمًا. واحدة من أكثر المشاكل الحرجة في الرعاية الصحية هي توقع احتمالية إعادة دخول المستشفى في حالة الأمراض المزمنة مثل مرض السكري لتكون قادرة على تخصيص الموارد اللازمة مثل الأسرة والغرف والمتخصصين والعاملين الطبيين للحصول على جودة مقبولة للخدمة. لسوء الحظ ، حاولت دراسات بحثية قليلة نسبيًا في الأدبيات معالجة هذه المشكلة ؛ تهتم غالبية الدراسات البحثية بالتنبؤ باحتمالية الإصابة بالأمراض نفسها. العديد من تقنيات التعلم الآلي مناسبة للتنبؤ. ومع ذلك ، هناك أيضًا نقص في الدراسات المقارنة غير الكافية التي تحدد التقنيات الأكثر ملاءمة لعملية التنبؤ. الهدف من هذه الرسالة هو جمع البيانات الضخمة للرعاية الصحية من مصادر مختلفة. ويلي ذلك دراسة تقنيات تحليل البيانات الضخمة المتاحة المناسبة لمعالجة هذه البيانات المعقدة وفهم كل تقنية. لتحقيق هذا الهدف ، اقترحت هذه الرسالة منهجية لتحليل البيانات الضخمة. أيضا ، يقدم دراسة مقارنة بين التقنيات الشائعة في الأدب للتنبؤ باحتمال إعادة دخول المستشفى في حالة مرضى السكري. تقوم مساهمة هذه الدراسة بتقييم إمكانية تحسين و / أو دمج تقنيات التعلم الآلي وتكييفها للحصول على معلومات وتنبؤات محسنة لتحسين الرعاية الصحية. هذه التقنيات هي أشجار القرار (DTs) والانحدار اللوجستي (LR) والتحليل التمييزي الخطي (LDA) والشبكات العصبية الاصطناعية (ANNs) وآلة ناقلات الدعم (SVM) و Naïve Bayesian (NB) والغابة العشوائية (RF) و AdaBoost و تعزيز التدرج (GB).

تستند الدراسة المقارنة إلى بيانات واقعية تم جمعها من عدد من المستشفيات في الولايات المتحدة. تم إجراء العديد من التجارب على تلك التقنيات حيث كشفت الدراسات المقارنة أن تقنيات التعلم القائمة على المجموعات (تعزيز وتعبئة) على سبيل المثال GB و RF و AdaBoost أظهرت أفضل أداء ، في حين أن مصنف NB وتحليل LR و LDA كانت الأسوأ.

ABSTRACT

One of the rapidly emerging fields is the big data field. Big data is a buzz word that describes data characterized by the large volume, variety, velocity, and veracity. Big data analytics and tools are used to extract valuable information and predictions from such data in an efficient way. One of the promising fields for the application of big data technology is the healthcare domain. Unfortunately, big data has not yet been applied extensively in this field to its extreme potential. This stems from many reasons including the variety of data sources leading to heterogeneous data of various types (structured, semi-structured and unstructured). Processing and analyzing such data are an important challenge. One of the most critical problems in healthcare is predicting the likelihood of hospital readmission in case of chronic diseases such as diabetes to be able to allocate necessary resources such as beds, rooms, specialists, and medical staff, for an acceptable quality of service. Unfortunately, relatively few research studies in the literature attempted to tackle this problem; the majority of the research studies are concerned with predicting the likelihood of the diseases themselves. Numerous machine learning techniques are suitable for prediction. Nevertheless, there is also a shortage of inadequate comparative studies that specify the most suitable techniques for the prediction process. The goal of this thesis is to collect healthcare big data from different sources. This is followed by studying available big data analytics techniques suitable for processing such complex data and understanding each technique. Towards this goal, this thesis proposed a methodology for big data analytics. Also, it presents a comparative study among common techniques in the literature for predicting the likelihood of hospital readmission in the case of diabetic patients. The contribution of this study is assessing the possibility of improving and/or integrating machine learning techniques and tailoring them for improved information and predictions for enhanced healthcare. Those techniques are decision trees (DTs), logistic regression (LR), linear discriminant analysis (LDA), artificial neural networks (ANNs), support vector machine (SVM), Naïve Bayesian (NB), random forest (RF), AdaBoost and gradient boosting (GB). The comparative study is based on realistic data gathered from a number of hospitals in the United States. Many experiments were conducted on those techniques where the comparative studies revealed that ensemble-based learning techniques (boosting and bagging) for example GB, RF and AdaBoost showed the best performance, while the NB classifier, LR analysis, and LDA were the worst.