استخدام مناهج التعلم الآلي للكشف عن الشذوذ في البيانات في إنترنت الأشياء

اسم الطالب: نسيبه رجاالله عياده الغانمي

المشرفين:
د. ريم متعب العتيبي
د. سيد محمد بخاري

المستخلص

تقنية انترنت الأشياء عبارة عن شبكة من الأجهزة أو أجهزة الاستشعار الموزعة و المتصلة عبر الإنترنت و التي تسمح بجمع البيانات ومشاركتها حيث تتأثر البيانات التي يتم تجميعها بواسطة هذه الأجهزة لبعض الشذوذ و ذلك لعدة أسباب مختلفة مثل مشكلات امنية أو تعطل الأجهزة و مع ذلك فإن أنظمة الكشف عن الحالات الشاذة الحالية القائمة على الوضع الخاضع للإشراف معتمدة على البيانات المُصنفة مع العلم إن البيانات التي تم جمعها من أجهزة إنترنت الأشياء عادة ما تكون غير مُصنفة ، والذي يعني أن تصنيف أو تسمية البيانات إلى بيانات طبيعية أو الشاذة غير معروف. والأهم من ذلك، فإن البيانات في إنترنت الأشياء تنمو بشكل متسارع والتي تحتاج إلى التنبؤ بتسمية و تصنيف البيانات المستقبلية.

تقترح هذه الدراسة HLMCC، وهو نموذج التعلم الآلي المُهجن الذي يستخدم كلاً من أساليب التصنيف والتجميع لتصنيف البيانات الياً و للكشف عن الشذوذ في بيانات إنترنت الأشياء. يتكون النموذج HLMCC من مرحلتين هما تصنيف البيانات الياً والكشف عن الحالات الشاذة. أولاً، تقوم HLMCC بتجميع البيانات الياً في مجموعات عادية وشاذة من خلال استخدام hierarchical affinity propagation (HAP). ثانياً ، يتم استخدام البيانات التي تم تصنيفها و الحصول عليها من مرحلة التجميع لتدريب نموذج decision trees (DTs) لتصنيف البيانات غير المرئية في المستقبل. تم تطبيق نموذج HLMCC على مجموعتي بيانات من إنترنت الأشياء هما : Labelled Wireless Sensor Network Data Repository (LWSNDR) و Landsat satellite datasets على التوالي.

تُظهر النتائج أن HLMCC قادر على تسمية و تصنيف البيانات الياً وتقليل التدخل البشري مقارنة بخوارزميات التجميع الأخرى. إضافة إلى ذلك أن HLMCC تفوق في الأداء على DTs المُطبقة على مجموعات البيانات المُصنفة مسبقاً والنماذج المستخدمة حالياً باستخدام عدة مقاييس تقييم و بناءً على متوسط الرتب . يُظهر HLMCC أعلى متوسط للرتب مقابل النماذج الأخرى من حيث false positive rate (FPR), recall, precision and the area under the precision-recall curve (AUCPR) بـ ١,٨ ، ١,٦ ، ١,٨ و ١,٨ على التوالي.

Using Machine Learning Approaches for Anomaly Detection in IoT

Student Name: Nusaybah Rajaallah Ayyadah Alghanmi

Supervised By
Dr. Reem Moteab Alotaibi
Dr. Seyed Mohamed Buhari

ABSTRACT

The Internet of Things (IoT) is a network of distributed devices or sensors connected via the Internet to allow gathering and sharing of data. The data generated by these devices is affected by anomalies or abnormal behaviour for different reasons, such as attacks or breakdown in devices. However, current anomaly detection systems based on the supervised mode rely on labelled data, while class labels for IoT data are usually unavailable. More importantly, the data in IoT grows fast, creating a need to predict the classification labels for the future data.

This study proposes a Hybrid Learning Model which uses both Clustering and Classification methods (HLMCC) to automate the labelling process and detect anomalies in IoT data. The model consists of two practical phases, automatic labelling and detecting anomalies. First, the HLMCC groups the data into normal and anomaly clusters by adopting hierarchical affinity propagation (HAP) clustering. Second, the labelled data obtained from the clustering phase is used to train the decision trees (DTs) and classify future unseen data. The HLMCC is applied to two existing IoT datasets, the Labelled Wireless Sensor Network Data Repository (LWSNDR) and Landsat satellite datasets, respectively.

The results show that the HLMCC is able to automate labelling of data, which can be beneficial to minimize human involvement. Moreover, the HLMCC outperforms the DTs on the originally labelled datasets and the state-of-the-art model over a wide range of evaluation metrics, based on the average ranks. The HLMCC shows the highest average ranks compared to other models in terms of false positive rate (FPR), recall, precision, and the area under the precision-recall curve (AUCPR) with 1.8, 1.6, 1.8 and 1.8, respectively.