نموذج تصنيف للتنقيب في البيانات المتدفقة مع انحراف المفهوم

مشاعل شعيل الثبيتي
د. منال عبدالله

المستخلص

تدفق البيانات هو عبارة عن الكم الهائل من البيانات التي يتم إنشاؤها في مجالات عدة مثل، العمليات المالية وأنشطة مواقع التواصل الاجتماعي وتطبيقات إنترنت الأشياء والعديد من المجالات الأخرى. لا يمكن معالجة هذا النوع من البيانات باستخدام خوارزميات التنقيب عن البيانات التقليدية، نظرًا لوجود العديد من القيود، بما في ذلك محدودية الذاكرة وسرعة البيانات والبيئة الديناميكية. يُعرف انحراف المفهوم بأنه القيد الرئيسي لخوارزميات التنقيب في تدفق البيانات، خصوصاً في التصنيف. حيث يعبر عن التغّير في تدفق البيانات خلال الوقت. وبالتالي، فإنه يؤدي إلى تدهور دقة نماذج التصنيف وينتج عنه التنبؤات الخاطئة. تعتبر رسائل البريد الإلكتروني العشوائي والتغييرات في سلوك المستهلك والانشطة العدائية أمثلة على انحراف المفهوم.

في هذا البحث، تم تقديم نموذج الكشف عن الانحراف المفهوم (CDDM)، حيث يعمل على مراقبة دقة نموذج التصنيف بافتراض أن الانخفاض في دقة النموذج يشير إلى حدوث انحراف. ايضاً تم تقديم نموذج محسن من نموذج CDDM يسمى ب W-CDDM.

تم تقييم كلا النموذجين باستخدام مجموعتي بيانات حقيقية وأربع مجموعات بيانات اصطناعية. أظهرت النتائج التجريبية للانحراف المفاجئ أن CDDM وW-CDDM يتفوق على النماذج الأخرى في مجموعتي البيانات ذات المئة ألف مثال ومليون مثال على التوالي. فيما يتعلق بالانحراف التدريجي، تفوقت W-CDDM من حيث الدقة ووقت التشغيل وتأخير الكشف في مجموعة البيانات ذات المئة ألف مثال. بينما في مجموعة البيانات ذات المليون مثال، حصلت CDDM على أعلى دقة باستخدام مصنف ال NB. علاوة على ذلك، يحقق W-CDDM أعلى دقة في مجموعات البيانات الحقيقية.

Classification Model For Data Stream Mining With Concept Drift

By

Mashail Shaeel Al-Thabiti

Dr. Manal Abdulaziz Abdullah

Abstract

Data stream is the huge amount of non-stop and high-speed data generated in various fields, including financial processes, social media activities, Internet of Things applications, and many others. Such data cannot be processed through traditional data mining algorithms due to several constraints, including limited memory, data speed, and dynamic environment. Concept Drift is known as the main constraint of data stream mining, mainly in the classification task. It refers to the change in the data stream underlying distribution over time. Thus, it results in accuracy deterioration of classification models and wrong predictions. Spam emails, consumer behavior changes, and adversary activates, are examples of Concept Drift.

In this thesis, a Concept Drift Detection Model is introduced, Concept Drift Detection Model (CDDM). It monitors the accuracy of the classification model over a sliding window, assuming the decline in accuracy indicates a drift occurrence. A modification over CDDM is proposed and named W-CDDM.

Both models have evaluated against two real datasets and four artificial datasets. The experimental results of abrupt drift show that CDDM, W-CDDM outperforms the other models in the dataset of 100K and 1M instances, respectively. Regarding gradual drift, the W-CDDM overtakes the rest in terms of accuracy, run time, and detection delays in the dataset of 100 K instances. While in the dataset of 1M instances, CDDM has the highest accuracy using the NB classifier. Moreover, W-CDDM achieves the highest accuracy on real datasets.