Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Improving Machine Learning Performance for Imbalanced Data Mining

A Thesis Submitted to the Council of the College of Computer Science and Mathematics University of Mosul as a Partial Fulfillment of Requirements for the Degree of Doctor of Philosophy in Computer Science

By

Shaymaa Ahmed Razooqi Mohammed

Supervised by

Asst.Prof.Dr. Ghayda Abdulaziz Majeed Altalib

2024 A.D. **1446 A.H.**

ABSTRACT

One significant challenge in data mining applications is data imbalance which is the subject of this dissertation; this is where the dataset is controlled by certain class (majority class) and other class which is poorly represented (minority class). Under these circumstances of imbalance, the standard machine learning classifiers will produce unsuccessful performance on the minority class; consequently, the distribution of samples in classes is a significant constituent in machine learning classification.

Many of the existing standard machine learning classifiers assume that the distribution of classes is evenly and the effect of classes error is equal in the training of classifier. However, the distribution of classes in a datasets in the real world is considerably imbalanced and the effect of misclassification is not equal of different classes. On the other hand, the minority class is more interesting for some application such as fraud detection and medical diagnosis.

The methods used to solve the imbalance problem can be depend on using resampling methods which fall into data level solutions. It is crucial to employ resampling methods to improve imbalanced data classification but it is not clear that which one achieves best improvement to machine learning classifiers. Others methods depend on improving the classification algorithm itself, it fall into the imbalance solutions at the algorithm level, as they work to develop a new learning model.

The ensemble learning methods are considered from the best techniques used to solve problem results in classifying imbalanced data. The need to find new methods increased when dealing with large scaled imbalanced data.

Given the challenges posed by imbalanced data, three methods are proposed in this dissertation to improve the performance of machine learning classifiers in imbalance big dataset; two of the methods are dealing with data level solutions while the third solution under the algorithm level.

The first proposed model uses a suggested majority coding and reduction method to coding the majority samples based on the Euclidean and Manhattan distance between the sample and class center – depending on the values of all features on the dataset - which are used in reduction process to reduce the majority class size and rebalancing the train dataset before classification process. The majority reduction process in the model imposed two stopping conditions which are : the new majority size must

not be less than 50% from the original majority size, and it must not be less than the minority size in the same time.

The second proposed model enhances the first model by adding a stage of feature selection which uses a suggested composite feature selection method and certifies it in the majority coding – depending on the values of selected features only - which then affects on the majority reduction process and data rebalancing before classification. The composite feature selection method consist of four methods which are (Information Gain, ANOVA , and two of Permutation Based Feature Important). The first and the second proposed models are used to improve the performance of machine learning when classifying imbalanced data.

The third proposed model uses the ensemble learning to build a multilayer ensemble classifier consisting of three estimator chains which consist of a three types of boosting algorithms(AdaBoost, HistGBoost, and XGBoost). The estimator chains size depends on the nature of the dataset to be classified and its imbalanced ratio.

The classification results of the three proposed models showed an increase in the AUC and G-mean performance metrics which used instead of accuracy when classifying imbalanced data. Time results also show enhancement when using the proposed models. Five dataset selected from Kaggle repositories which are (Higgs, VIS-Prediction, UNSW15, Covtype, and Creditcard) used to evaluate the performance of proposed models.

The results of the proposed methods based on data level showed that reducing the majority size from dataset results after majority coding and reduction process does not lead to information loss and it successfully improve the classification performance. Furthermore, the results of the third proposed model in algorithm level showed that the multilayer ensemble classifier achieved a stability in AUC and G-mean performance when classifying datasets with different value of imbalance ratio. It also showed a AUC and G-mean performance improvement reach to 20% when classifying the used datasets, and it surpassed the performance of the proposed data level model.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

تحسين أداء تعلم الالة للتنقيب في البيانات غير المتوازنة

اطروحة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دكتوراه فلسفة في
علوم الحاسوب

من قبل

شيماء احمد رزوقي محمد

بإشراف
أ.م.د. غيداء عبدالعزيز مجيد الطالب

١٤٤٦ هـ ٢٠٢٤ م

الخلاصة

احد التحديات المهمة في تطبيقات تنقيب البيانات هو عدم توازن البيانات وهو موضوع هذه الاطروحة، وذلك عندما تكون مجموعة البيانات الغالب عليها من قبل فئة معينة (فئة الاغلبية) بينما تكون الفئة الاخرى ممثلة بشكل ضعيف (فئة الاقلية). في ظل ظروف عدم الموازنة هذه فان طرق تعلم الالة القياسية سوف تعطي اداءً ضعيفاً عند التعامل مع فئة الاقلية، وبناء على ذلك فان توزيع العينات ضمن فئات البيانات يمثل عنصرا اساسيا عند التصنيف باستخدام تعلم الالة.

الكثير من مصنفات تعلم الالة القياسية المتوفرة تفترض ان توزيع الفئات يكون منتظم وان تأثير الخطأ للفئات يفترض ان يكون متساوي عند تدريب المصنف. ولكن توزيع الفئات في مجموعات البيانات في العالم الحقيقي كثيراً ما يكون غير متوازن وتأثير خطأ التصنيف ليس متساوي لمختلف الفئات. من جهة اخرى فان فئة الاقلية غالبا ما تكون هي الاكثر اهمية لبعض التطبيقات مثل اكتشاف الاحتيال والتشخيص الطبي.

الطرق التي تستخدم لحل مشكلة عدم التوازن يمكن ان تعتمد على اعادة توزيع البيانات (استخدام طرق اختيار العينات). طرق اختيار العينات امرا حاسما لتحسين نتائج تصنيف البيانات غير المتوازنة، لكنه من غير الواضح اي من تلك الطرق تعطي افضل النتائج في تحسين اداء مصنفات تعلم الالة. والطرق الاخرى تعتمد على تطوير خوارزمية التصنيف نفسها، لذلك فهي تندرج تحت مسمى الحلول لعدم التوازن على مستوى الخوارزمية والتي تعمل على توليد نموذج تعلم جديد.

طرق التعلم التجميعي تعد من افضل التقنيات المستخدمة لحل المشاكل الناتجة عند تصنيف البيانات غير المتوازنة، كما تزداد الحاجة الى ايجاد طرق جديدة عند التعامل مع البيانات الضخمة غير المتوازنة.

نظرا للتحديات التي تفرضها البيانات غير المتوازنة، تم اقتراح ثلاثة طرق في هذه الاطروحة لتحسين اداء مصنفات تعلم الالة في تصنيف مجموعة البيانات الضخمة غير المتوازنة، اثنين من تلك الطرق تتعامل مع الحلول على مستوى البيانات فيما يكون الحل الثالث على مستوى الخوارزمية.

استخدم النموذج الاول طريقة مقترحة لترميز عينات فئة الاغلبية استناداً إلى المسافة بين العينة ومركز الفئة – باعتماد قيم كافة السمات في مجموعة البيانات – والتي يتم استخدامها في عملية التقليص لتخفيض حجم فئة الاغلبية واعادة موازنة مجموعة البيانات التدريب قبل اجراء عملية التصنيف. تفرض عملية تقليص عينات الاغلبية في النموذج المقترح شرطين للتوقف وهما :

ان حجم بيانات الاغلبية الجديد يجب ان لا يقل ٥٠ % من الحجم بيانات الاغلبية الاصلي وأن لا يقل في الوقت نفسه عن حجم بيانات الاقلية.

النموذج الثاني يعمل على تحسين النموذج الاول بإضافة مرحلة لاختيار الميزات تستخدم طريقة مركبة مقترحة لاختيار الميزات يتم اعتمادها في عملية ترميز الاغلبية – باعتماد قيم السمات التي تم اختيارها فقط– والتي تؤثر بدورها على عملية تقليص فئة الاغلبية واعادة موازنة البيانات قبل تنفيذ التصنيف. تتكون طريقة اختيار الميزة المركبة من أربع طرق وهي (اكتساب المعلومات، وتحليل التباين، واختبار الميزة المهمة القائم على التباديل). يتم استخدام النموذجين الاول الثاني المقترحين لتحسين اداء خوارزميات تعلم الالة عند التصنيف البيانات غير المتوازنة.

النموذج الثالث المقترح يستخدم التعلم التجميعي لبناء مصنف تجميعي متعدد الطبقات يتكون من ثلاث سلاسل تخمين كل منها يتكون من نوع مختلف من خوارزميات التجميع(AdaBoost, HistGBoost, and XGBoost). يعتمد حجم سلاسل التجميع على طبيعة مجموعة البيانات المراد تصنيفها ودرجة عدم التوازن فيها.

نتائج التصنيف للنماذج الثلاث المقترحة تبين زيادة في الاداء للمقاييس AUC و G–mean التي تستخدم بدلا من الدقة عند تصنيف البيانات غير المتوازنة. كما تم تحسين وقت التصنيف ايضا عند استخدام النماذج المقترحة. خمس مجموعات بيانات غير متوازنة تم اختيارها من مستوعبات البيانات Kaggle والتي هي (Higgs و VIS–Prediction و UNSW15 و Covtype6 و Creditcard) استخدمت لتقييم النماذج المقترحة.

اظهرت نتائج الطرق المقترحة المعتمدة على مستوى البيانات ان تقليل حجم الاغلبية من مجموعة البيانات الذي ينتج من عملية ترميز و التقليص عينات الاغلبية لم يؤدي الى فقدان في المعلومات وعمل على تحسين اداء. كما اظهرت نتائج النموذج الثالث المقترح ضمن مستوى الخوارزمية ان المصنف متعدد الطبقات حقق استقرار في اداء تعلم الالة عند تصنيف مجموعات البيانات بمختلف نسب عدم التوازن، كما اظهرت تحسين في مقياسي الاداء (AUC andG–mean) يصل الى ٢٠% في تصنيف مجموعات البيانات المستخدمة. وتفوق على الاداء العام لنماذج مستوى البيانات المقترحة.