التصنيف المتعدد للبيانات غير المتوازنة :تآزر التصنيف الثنائي مع اختيار الخصائص

[١] تهاني بنت سعد المشدق، [٢] أ.د صالح بن محمد الشمراني، [١] د .عياد بن أحمد البشري

[١] قسم علوم الحاسبات، كلية الحاسبات وتقنية المعلومات، جامعة الملك عبدالعزيز

جدة، المملكة العربية السعودية

tsalmoshadak@kau.edu.sa, aaalbeshri@kau.edu.sa

[٢] قسم نظم المعلومات، كلية الحاسبات وتقنية المعلومات، جامعة جدة

جدة، المملكة العربية السعودية

sshomrani@uj.edu.sa

المستخلص

يعد تصنيف البيانات غير المتوازنة من أكثر المشاكل تكرارًا في مجال تطبيقات التصنيف، كما يعد أحد التحديات؛ لأنه يؤدي إلى نتائج تصنيف خاطئة .إضافة إلى ذلك يمكن أن تتزامن مشكلة عدم توازن البيانات مع مشكلة تعدد الفئات مما يؤدي إلى المزيد من التعقيد؛ لأنه يمكن أن تكون هناك فئة ذات أقلية بالنسبة لبعض الفئات، وفي نفس الوقت ذات أكثرية بالنسبة لفئات أخرى؛ ومن هذا المنطلق اقترحنا نموذج تصنيف جديد أطلقنا عليه LFSC-OVO :لتحسين أداء تصنيف البيانات غير المتوازنة من ناحية رفع معدل دقة التصنيف، يعتمد تصميم هذا النموذج على دمج فكرة تقسيم المشكلة مع اختيار الخصائص، ويكمن تميّز هذا المقترح في مستوى تطبيق اختيار الخصائص خلال عملية التصنيف، حيث إنه لم يتم تطبيق اختيار الخصائص سابقًا بشكل منفرد لكل مشكلة ثنائية، بعد ذلك تم اختبار أداء LFSC-OVO على سبع مجموعات من البيانات غير المتوازنة التي تحتوي على فئات متعددة، وذلك باستخدام أنواع متعددة من المصنفات وأساليب التجميع، كما تم مقارنة أدائه مع طريقة أخرى مقترحة في الدراسات السابقة تطبق اختيار الخصائص بشكل عام، وكانت النتيجة تفوق أداء LFSC-OVO في كل الحالات المختلفة لجميع أنواع المصنفات وأساليب التجميع؛ بسبب

نقصان تأثير الفئات ذات الأكثرية على أداء المصنفات.

MULTI-CLASSIFICATION TASKS IN IMBALANCED DATASETS: ON THE SYNERGY BETWEEN ROBUST PAIRWISE LEARNING TECHNIQUES AND FEATURE SELECTION

[1]Tahani S. Al-Moshadak, [2]Saleh M. Al-Shomrani, [1]Aiiad A. Albeshri

[1] *Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
tsalmoshadak@kau.edu.sa, aaalbeshri@kau.edu.sa
[2] *Department of Information Systems, Faculty of Computing and Information Technology, University of Jeddah, Jeddah, Saudi Arabia*
sshomrani@uj.edu.sa

ABSTRACT

Classification in imbalanced datasets is one of the recurring problems in real-world applications of classification. It considered as a challenge since it needs to deal with uneven distribution of examples in the training datasets that lead to generate sub-optimal classification models. The presence of multiple classes implies an additional difficulty since the relations between the classes tend to complicated. One class can be a minority class for some, while a majority for others. So, we proposed a Local Feature Selection Classification model using OVO (LFSC-OVO) for multi-class imbalanced datasets, to improve the performance of the classification in terms of average accuracy. LFSC-OVO is constructed based on problem decomposition and feature selection. The novelty of the proposed work resides in the level of application feature selection in the classification procedure, since feature selection has not been previously used locally for each binary problem. LFSC-OVO is validated and tested by 7 multi-class imbalanced datasets from KEEL dataset repository using different base classifiers and aggregation methods. Then, a comparative study is conducted to compare the performance of LFSC-OVO versus another method in state-of-art. LFSC-OVO shows the best performance in all scenarios of using different base classifiers and aggregation methods as a result of decreasing the effect of the majority classes on the base classifiers.