تقنية اختيار الخصائص المبنية على الحوسبة عالية الأداء

إعداد
عهود نايف الحربي

إشراف
الدكتور/ محمد ذهب

المستخلص

تشهد خوارزمية الفرع والحد (BB) نموًا كبيرًا في اختيار الميزة مع زيادة عدد الميزات ، الأمر الذي قد يتطلب في أسوأ الحالات ، استكشاف الشجرة بأكملها بحثًا عن الحل الأمثل. يقدم هذا البحث تحسينًا في خوارزمية BB لاختيار الميزات باستخدام دالة المعايير الرتيبة التقريبية والحوسبة عالية الأداء (HPC) .النسخة المتوازية المحسنة الفرعية دون المثلى المقترحه من خوارزمية BB التي تبحث عن الحل عن طريق قطع المسارات غير الواعدة وحذف ميزات متعددة في كل عقدة شجرة داخلية باستخدام متغير *tau* . يتم تحديد عدد العناصر المحذوفة في كل عقدة من خلال قيمة متغير *tau*، والذي يتم تحديده بشكل أفضل وفقًا للعلاقة بين الميزات. يتم استخدام مصنف K-Nearest Neighbors كدالة معيار J التي تحسب أعلى دقة لمجموعة جزئيه من الميزات. تم تطبيق التجربة على مجموعة بيانات مختلفة ومقارنتها بخوارزمية BB الأصلية والعديد من طرق الاختيار. على الرغم من أن الخوارزمية المقترحة لا تضمن اختيار المجموعة الجزئيه للميزات المثلى ، إلا أنها قادرة على الوصول إلى دقة عالية مقارنة بالإصدار الأمثل من خوارزمية BB الأصلية ، وربما أفضل من ذلك. تظهر النتائج نتائج واعدة من حيث الدقة والوقت المنقضي وحجم الشجرة.

Optimized Feature Selection Technique Based on High Performance Computing (HPC)

By

Ahood Naif Alharbi

Supervised by

Dr. Mohamed Dahab

Abstract

The Branch and bound (BB) algorithm undergoes an exponential growth in feature selection as the number of features increases, which may require in the worst cases, exploring the whole tree looking for an optimal solution. This research presents an enhancement in the BB algorithm for feature selection using an approximate monotonic criteria function and High Performance Computing (HPC). The enhanced parallel sub-optimal proposed version of the BB algorithm seeking for the solution by cutting off unpromising paths and deleting multiple features at each internal tree node using tau variable. The number of deleted features in each node is determined by the tau variable value, which best specified according to the correlation between the features. KNN (K-Nearest Neighbors) classifier is used as the J function that computes the highest accuracy of the subset of features. The experiment was applied to different datasets and compared to the original BB algorithm and numerous selection methods. Although the proposed algorithm does not guarantee the selection of the optimal features subset, it is able to reach a high accuracy comparable to the optimal version of the original BB algorithm, and perhaps better than it. The results show promising results in terms of accuracy, elapsed time, and tree size.