ملخص

حل مشكلة اختيار المعالم باستخدام خوارزمية بحث الوقواق الثنائية المعدلة ونظرية مجموعات الاستقراب

مشكلة اختيار المعالم في التصنيف هي عملية مهمة لايجاد أقل عدد ممكن من المعالم من مجموعة البيانات الأصلية عن طريق إزالة المعالم التي لا تحوي أو تقدم أية معلومات هامة لخوارزميات التصنيف. هذه العملية تهدف إلى تحسين دقة نموذج التصنيف وتبسيطه وتقليل الوقت المطلوب لبنائه من قبل خوارزميات التصنيف. نظرية مجموعات الاستقراب هي إحدى الطرق الفعالة في اختيار المعالم، ولكنها تستخدم البحث الشامل للبحث في كل الحلول الممكنة، وايضا تستخدم درجة الاعتمادية لتقويم هذه الحلول. ولكن طريقة البحث الشامل مكلفة وغير مناسبة لمجموعات البيانات الضخمة، اضافة الى ذلك فإن درجة الاعتمادية فعالة فقط لمجموعات البيانات من النوع nominal. لذلك فإن خوارزميات الاداء العليا وخصوصا الخوارزميات المستوحاه من الطبيعة اصبحت تستخدم بشكل واسع لتحل محل البحث الشامل في طريقة مجموعات الاستقراب، بالاضافة الى عوامل اخرى مثل القيم المتكررة التي تستخدم مع درجة الاعتمادية لتحسين اداء مجموعات الاستقراب في مختلف انواع مجموعات البيانات.

هذه الاطروحة تهدف الى تقديم طريقة جديدة لمعالجة مشكلة اختيار المعالم في التصنيف من خلال تطوير خوارزمية الوقواق الثنائية وتطوير دالة هدف جديده تعتمد على نظرية مجموعات الاستقراب والقيم المتكرره لاختيار اقل عدد ممكن من المعالم بوقت قليل ومن دون تقليل واضح في كفاءة اداء خوارزميات التصنيف، وان يكون فعالا على انواع مختلفة من مجموعات البيانات المتنوعة في الخصائص.

في مرحلة التقييم، قمنا بتقويم عملنا على 16 مجموعة بيانات ماخوذة من UCI. ثم قمنا بمقارنة عملنا بثلاثة طرق اخرى معروفة من نفس الفئة: خوارزمية الوقواق الثنائية ونظرية الاستقراب قبل تعديلهما(الطريقة الاساسية). وخوارزمية اسراب الطيور والخوارزمية الجينية. ايضا تم استخدام خوارزميات التصنيف (خوارزمية شجرة اتخاذ القرار وخوارزمية naïve Bayes) من اجل تقويم اداء التصنيف في هذه الطرق الاربعة.

اظهرت النتائج ان طريقتنا الجديده حققت افضل النتائج على مستوى عدد المعالم المختارة بالاضافة الى قيم اداء التصنيف مقارنة بالطرق الاخرى التي استخدمت في التجارب في معظم مجموعات البيانات. ايضا الطريقة الجديدة مقارنة بالطريقة الاساسية احتاجت وقت اقل في كل مجموعات البيانات المستخدمة في التجارب.