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

توظيف آلة المتجه الداعم مع خوارزمية التحسين المستوحاة من الحمام في تصنيف البيانات غير المتزنة

رسالة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة ماجستير
علوم في الإحصاء

من قبل

آلاء عبدالعزيز قنبر شهاب حسن

بإشراف

أ.د. زكريا يحيى نوري يحيى

1446 هـ 2024 م

المستخلص

شهدت العقود الأخيرة تزايداً في كمية البيانات ونوعها وأصبحت العديد من التطبيقات مصدراً لتدفق هذه البيانات؛ إذ ان الزيادة والتراكم في حجم البيانات يحتاجان إلى ابتكار طرائق وأساليب لتلخيص هذه البيانات من اجل فهمها والاستفادة منها ودراستها وتنقيبها؛ إذ إن علم تنقيب البيانات والبحث عن المعرفة من العلوم الحديثة التي لا زالت في حالة تطوير مستمر خصوصا بعد ظهور العديد من الخوارزميات الذكائية Intelligent Algorithms التي أثرت بشكل ملحوظ في علم مجال التنقيب عن المعرفة البيانات ومعالجتها.

تضم العديد من التطبيقات ولاسيما التطبيقات الطبية المتعلقة بدراسة الجينات كمَّ هائلاً من البيانات المتعلقة بوصف هذه الجينات، وإن الحاجة إلى نتائج هذه البيانات دفعت الكثير من العلماء إلى البحث عن الطرائق التي تعمل على استخلاص تلك المعلومات بأفضل النتائج، وفي هذه الرسالة تم تحليل مجموعات متعددة من البيانات وكان الهدف الرئيس هو تحسين التصنيف للبيانات الضخمة غير المتزنة باستعمال آلة المتجه الداعم لتحليل البيانات الضخمة؛ إذ تعد آلة متجه الداعم من الطرائق المهمة وواسعة الاستعمال في التنصيف، فضلاً عما سبق فان هذه الطريقة تفقد دقتها بالتصنيف عندما تكون البيانات غير متزنة وضخمة في نفس الوقت وتحوي قيماً شاردة، ولغرض تحقيق الهدف تم توظيف أحد الخوارزميات المستوحاة من الطبيعة وهي خوارزمية الحمام في تشخيص القيم الشاردة وحذفها ومن ثم الحصول على دقة تصنيف باستعمال آلة المتجه الداعم.

تم استخدام أسلوب مونت– كارلو في المحاكاة لتوليد بيانات تتبع أنموذجاً تصنيفياً وتحتوي على قيم شاردة، لقد أظهرت نتائج المحاكاة بالاعتماد على معايير دقة التصنيف ومعايير دقة التشخيص تفوق الطريقة المقترحة مقارنة بطرائق أخرى، فضلاً عن ذلك فقد تم تطبيق توظيف الطريقة المقترحة على بيانات حقيقية عالمية في مجال علم الجينات، وقد بينت النتائج تفوق الطريقة المقترحة على باقي الطرائق الأخرى التقليدية حسب إعطائها دقة تصنيف عالية.

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

Employing Support vector machine with Improving pigeon nature algorithm for Imbalance data classification

**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 Master of Science
in
Statistics**

By

Alaa Abdulazeez Qanbar Shehab Hassan

Supervised by

Professor

Prof. Dr. Zakariya Yahya Nouri Yahya

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

Abstract

The last decades have witnessed an increase in the quantity and type of data, and many applications have become a source of the flow of this data, as the increase and accumulation in the volume of data requires the creation of methods and methods to summarize, study and mine this data in order to understand and benefit from it, as the science of data mining and the search for knowledge is one of the Modern science is still in a state of continuous development, especially after the emergence of many intelligent algorithms that have significantly influenced the science of knowledge mining and data processing. Many applications, especially medical applications related to the study of genes, include a huge amount of data related to the description of these genes, and the need for the results of this data has led many scientists to search for methods that work to extract this information with the best results. In this thesis, multiple sets of data were analyzed, and the main goal was to improve classification of unbalanced big data using the support vector machine for big data analysis. The support vector machine is considered one of the important and widely used methods in classification. However, this method loses its classification accuracy when the data is unbalanced and huge data at the same time and contains outlier values. In order to achieve the goal, one of the algorithms inspired by nature, which is the pigeon optimization algorithm, was employed to diagnose outlier values and delete them, thus obtaining classification accuracy using the support vector machine. The Monte-Carlo simulation method was used to generate data that follows a classification model and contains outlier values. The simulation results, based on classification accuracy criteria and diagnostic accuracy criteria, have shown the superiority of the proposed method compared to other methods. In addition, the proposed method was applied to real global data in the field of genetics. The results showed that the proposed method is superior to other traditional methods as it gives it high classification accuracy.