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

التكييف المعلمي لخوارزمية تخفيض الأبعاد متعددة العوامل للأنماط الظاهرية الترتيبية

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

من قبل
محمد ابراهيم عثمان محمد

بإشراف
م.د. زيد طارق صالح عباوي

١٤٤٥هـ ٢٠٢٣م

المستخلص

تشير الدراسات السريرية إلى العلاقة الوثيقة بين بعض الأمراض ووجود تداخلات محددة بين العوامل الجينية وأن كشف التداخلات الجينية ذات التأثير الكبير على ظهور الأمراض الوراثية يحتاج إلى تحليلات احصائية مستفيضة. وبسبب الحجم الهائل للبيانات الجينية في الجنس البشري، فكان لا بد من تطوير طرائق إحصائية مكيفة للتعامل مع البيانات الضخمة. تعد خوارزمية تخفيض الأبعاد متعددة العوامل Multifactor Dimensionality Reduction (MDR) إحدى الخوارزميات اللامعلمية الرائدة في هذا المجال. تعمل على تخفيض أبعاد البيانات الجينية للحصول على أهم تداخل ذا تأثير مباشر على زيادة احتمالية ظهور الأمراض الوراثية. وتعتمد الخوارزمية في تكوينها على مجموعة من الإجراءات اللامعلمية لتشخيص التداخل الجيني الأعلى تأثيراً على متغيرات الاستجابة الثنائية حصراً. وكأي طريقة إحصائية، فإن هذه الخوارزمية لا تخلو من نقاط الضعف والمحددات التطبيقية، لذا كان لا بد من تطوير الخوارزمية لتجاوز المعوقات. أحد نقاط الضعف في هذه الخوارزمية هي عدم إمكانيتها من التعامل مع البيانات التي تحتوي على متغير استجابة من النوع الترتيبي. وقد طور بعض الباحثين تعميماً لخوارزمية تخفيض الأبعاد متعددة العوامل لتمكينها من التعامل مع البيانات الترتيبية. مع ذلك فإن الخوارزمية المعممة أكثر تعقيداً من الخوارزمية الأصلية. لذلك اقترحنا تطوير الخوارزمية الأصلية تطويراً بسيطاً وذلك بتوظيف الانحدار اللوجستي الترتيبي في تصنيف الافراد في العينة، مع الإبقاء على جميع خطوات الخوارزمية الأصلية دون تغيير. ومن ناحية أخرى، فإن خوارزمية MDR تعتمد اسلوباً لا معلمياً للتحقق من معنوية التداخلات المرشحة في الخوارزمية. بُني هذا الاجراء اللامعلمي على فكرة الاختبارات التباديلية، وهو يستهلك وقتاً زمنياً طويلاً جداً مقارنة بالإجراءات المعلمية المعتمدة على الأساليب النظرية. اقترح بعض الباحثين استخدام توزيع القيمة العظمى المعمم للتحقق من المعنوية الإحصائية للتداخلات المرشحة، لكن لم يرد استخدام هذا الأسلوب الا مع المتغيرات المعتمدة المستمرة والثنائية. تم في هذا البحث توظيف الاسلوب النظري المعتمد على توزيع القيمة العظمى المعمم بدلا من الاختبارات التباديلية المعتمدة في الخوارزمية وذلك عندما يكون متغير الاستجابة من النوع الترتيبي. وأظهرت نتائج المحاكاة فاعلية الخوارزمية المعدلة في كشفها للتداخلات الحقيقة المؤثرة على تفاقم المرض. وتم إجراء التطبيق العملي على بيانات تتعلق بمستويات المرونة الادراكية Cognitive Resilience (CR) وعلاقتها ببعض العوامل الجينية المرتبطة بظهور

مرض الزهايمر الذي يصيب نسبة محددة من البشر. حيث تمت محاكاة بيانات أصلية باستخدام أسلوب الـ bootstrap وتقنية Synthetic Minority Over-sampling Technique SMOTE لتوليد عينة بحجم 1000. في حين احتوت البيانات على 12 عامل جيني ثنائي الأليل، وأظهرت النتائج بأن أكثر تداخل مؤثر في تدهور المرض هو التداخل بين العامل الجيني APOE والعامل الجيني PSEN2.

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

Parametric Adaptation of The Multifactorial Dimensionality Reduction Algorithm for Ordinal Phenotypes

**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

Mohammed Ibrahim Othman Mohammed

Supervised by

Lecturer Dr. Zaid Tariq Saleh Abawi

1445 A.H. **2023 A.D.**

ABSTRACT

Clinical studies indicate the close relationship between some diseases and the existence of specific interactions between genetic factors, and that detecting genetic interactions that have a significant impact on the emergence of genetic diseases requires extensive statistical analyses. Because of the enormous volume of genetic data in the human race, it was necessary to develop statistical methods adapted to deal with high-dimensional data. Multifactor Dimensionality Reduction (MDR) is one of the leading nonparametric algorithms in this field. It works to reduce the dimensions of genetic data to obtain the most important interaction that has a direct impact on increasing the likelihood of genetic diseases appearing. In its composition, the algorithm relies on a set of nonparametric procedures to diagnose genetic interaction with the highest impact exclusively on binary response variables. Like any statistical method, this algorithm is not devoid of weaknesses and application limitations, so the algorithm had to be developed to overcome the obstacles. One of the weaknesses of this algorithm is that it cannot handle data that contain an ordinal response variable. Some researchers have developed a generalization of the multi-factor dimensionality reduction algorithm to enable it to deal with ordinal data. However, the generalized algorithm is more complex than the original algorithm. Therefore, we proposed a simple development of the original algorithm by employing ordinal logistic regression to classify individuals in the sample, while keeping all steps of the original algorithm unchanged. On the other hand, the MDR algorithm adopts a non-parametric method to verify the significance of the nominated interactions in the algorithm. This nonparametric procedure is built on the idea of permutational testing, and it consumes a very long time compared to parametric procedures based on theoretical methods. Some researchers have suggested using the generalized maximum value distribution to verify the statistical significance of proposed interactions, but this method has only been used with continuous and binary dependent variables. In this research, the theoretical method based on the generalized maximum value distribution was employed instead of the permutational testing adopted in the algorithm when the response variable is of the ordinal type. The simulation results showed the effectiveness of the modified algorithm in detecting the real interactions affecting the aggravation of the disease. The practical application was carried out on data related to levels of cognitive

resilience (CR) and its relationship to some genetic factors associated with the emergence of Alzheimer's disease, which affects a specific percentage of people. Original data were simulated using the bootstrap method and the Synthetic Minority Over-sampling Technique (SMOTE) to generate a sample of size 1000. While the data contained 12 bi-allelic genetic factors, the results showed that the most influential interaction in the deterioration of the disease is the interaction between the APOE genetic factor and the PSEN2 genetic factor.