IRAQI Academic Scientific Journals

KHAZAYIN OF ECONOMIC
AND ADMINISTRATIVE SCIENCES (2026) (Vol. 03) (No.02): P(175-183)

خزائن للعلوم الاقتصادية والإدارية
KHAZAYIN OF ECONOMIC AND ADMINISTRATIVE SCIENCES
ISSN: 2960-1363 (Print)
ISSN: 3007-9020 (Online)

Methods for high-quality statistical data analysis: An objective review

Gheadaa Ibrahim Shehab
Diyala University,Baqubah,Iraq
gh.gheadaa@uodiyala.edu.iq

Abstract. This research explores "Big Data" as a pivotal asset in the digital economy, focusing on the computational and statistical challenges posed by its volume, velocity, and variety. The study analyzes contemporary algorithms, highlighting the necessity of transitioning from traditional statistical methods due to their declining explanatory power and computational inefficiency when facing high complexity. The findings emphasize critical issues such as data heterogeneity, noise accumulation, and spurious correlations. The paper concludes that ensuring "Data Veracity" is the primary challenge; without specialized statistical expertise, the risks of inaccurate inference can transform data from a valuable resource into a significant institutional liability.

Keywords: High-dimensional statistical analysis, spurious correlations, statistical computing, intelligent algorithms.

DOI: 10.69938/Keas.26030215

أساليب التحليل الإحصائي للبيانات عالية الجودة: مراجعة موضوعية

غيداء ابراهيم شهاب
جامعة ديالى، كلية الإدارة والاقتصاد، قسم الاحصاء ،بعقوبة ، العراق
gh.gheadaa@uodiyala.edu.iq

المستخلص: تتناول هذه الورقة طبيعة "البيانات الضخمة" بوصفها مادة خام ذات قيمة اقتصادية واجتماعية، متجاوزةً مفهوم الحجم لتشمل تعقيد الأساليب الحسابية والإحصائية. يستعرض المقال التحولات الجوهرية في الخوارزميات لمواجهة تحديات السرعة، التنوع، وموثوقية البيانات. وتخلص النتائج إلى أن الأساليب الإحصائية التقليدية لم تعد كافية نظراً لقصور كفاءتها الحسابية، وظهور مشكلات هيكلية مثل عدم التجانس، تراكم الضوضاء، والارتباطات الزائفة. توصي الورقة بضرورة تطوير نماذج إحصائية مرنة توازن بين استخلاص القيمة وضمان جودة البيانات لمواجهة مخاطر التضليل الإحصائي في بيئات البيانات عالية الأبعاد.

الكلمات المفتاحية: التحليل الإحصائي عالي الأبعاد، ، الارتباطات الزائفة، الحوسبة الإحصائية، الخوارزميات الذكية

Corresponding Author: E-mail: gh.gheadaa@uodiyala.edu.iq

1 المقدمة

الانتشار الواسع لتقنيات المعلومات الحديثة في جميع مجالات المجتمع، وتوسع نطاق وصول الناس إلى الإنترنت، بالإضافة إلى زيادة عدد الخدمات المقدمة عبر شبكة الإنترنت، والتطبيق الواسع للتكنولوجيا الرقمية في مجالات الصحة، وعلم الفلك، والمعلوماتية الحيوية، والنقل، والإدارة العامة، وغيرها، أدى إلى زيادة تدفق البيانات، بما في ذلك ظهور ظاهرة "البيانات الضخمة". ويُستخدم هذا المصطلح الجديد لتحديد البيانات التي لا يمكن معالجتها من حيث الحجم وتعقيد أساليب

الإدارة الحالية، ومن خلال التحليل الذكي. تخزين البيانات الضخمة، تُشكّل إدارة البيانات وخلق القيمة مشكلةً خطيرة. تُولّد هذه المشكلة تلقائيًا ومستمرًا من مصادر مختلفة لمعالجة البيانات في الوقت الفعلي، بالإضافة إلى عدم منطقية حلول تكنولوجيا المعلومات الحالية. يسمح لنا الوضع الحالي بالتعبير عن "سهولة إنشاء البيانات وصعوبة معالجتها وقد دُرست على نطاق واسع الوضع الحالي، ومجالات التطبيق، والفرص، والمشاكل، وما إلى ذلك، لتقنيات البيانات الضخمة. تُبشّر البيانات الضخمة بمستوى جديد من الاختراعات العلمية والقيمة الاقتصادية. وبالتالي، فإن الإنجازات العلمية القائمة على إدارة البيانات بشكل متزايد، تجعل الباحثين أكثر استهلاكًا. تشمل تقنيات البيانات الضخمة عمومًا عددًا من التخصصات: الإحصاء، واستخراج البيانات، والتعلم الآلي، وتحليل الشبكات الاجتماعية، وطرق التحسين، وأساليب التصور، وغيرها. تُعد الأساليب الإحصائية بالغة الأهمية لتحليل البيانات الضخمة. تُخصّص هذه المقالة تحديدًا لتفسير هذه الأساليب. من ناحية أخرى، نظرًا لحجمها وكثافتها وتعقيدها، تتجاوز البيانات الضخمة قدرات المعايير القياسية.. تُسبب هذه البرمجيات مشاكل في تحليل البيانات باستخدام الأساليب الإحصائية التقليدية. لا يُمكن تحليل هذا النوع من البيانات الضخمة باستخدام الأساليب الإحصائية الحالية على جهاز حاسوب عادي. في هذه الحالة، تظهر مشاكل مثل تثبيت البيانات الضخمة واسعة النطاق على ذاكرة حاسوب عادية، وتباطؤ سرعة العملية، مما يتطلب وقتًا أطول. لذا، من الضروري اللجوء إلى تقنيات الحوسبة والحوسبة السحابية، والبنى المتوازية والموزعة. في الوقت نفسه، تتطلب الإمكانيات الحالية للبنى المتوازية والموزعة لتخزين وإدارة البيانات توافق الأساليب الإحصائية مع نموذج البيانات الضخمة. كما أن زيادة حجم البيانات وتعقيد بنيتها يُلزمان بتطبيق أساليب جديدة، ويفرضان متطلبات كبيرة على المنهجيات الإحصائية الحالية.

2. غرض الدراسة وأهدافها

الهدف الرئيسي لهذه الورقة هو تحليل البيانات الضخمة التي لا يمكن تحليلها وتحويلها إلى اتجاه علمي رئيسي بمساعدة الأدوات المتاحة. وفقًا للهدف المحدد، تم تحديد أهداف البحث التالية:

١. مراجعة مجالات تطبيق البيانات الضخمة.

٢. دراسة مشكلات التحليل الإحصائي في هذه المجالات.

3.تحليل البيانات المنشورة وبيان المشكلة

تُتيح البيانات الضخمة فرصًا جديدة للمجتمع المعاصر وتُمثل مشاكل لعلماء البيانات. ومن أهم خصائص البيانات الضخمة أن الأساليب الإحصائية، التي تُحسن أداءها على مجموعات البيانات صغيرة الحجم، عادةً ما تُنفذ بشكل سيء في بيئات البيانات الضخمة. لذا، ليس من السهل التطلع إلى أداء تلك الأساليب الإحصائية في حل مشاكل البيانات الضخمة. قد يُؤدي الحجم الكبير والسرعة العالية إلى عدم التجانس، وتراكم الضوضاء، والعلاقات المُزيفة، والتداخل الداخلي العشوائي، مما يُسبب مشاكل في جدوى الحوسبة واستقرار الخوارزميات. يُمكن التغلب على هذه العقبات إما باستخدام منهجيات إحصائية مُطورة حديثًا و/أو منهجيات حاسوبية. يُؤدي التنوع الكبير إلى أنواع بيانات غير تقليدية أو حتى غير مُهيكلة، مما يتطلب طرقًا جديدة ومُبتكرة لإعادة تنظيم البيانات. تحديد بنية البيانات، بل وطرح أسئلة دراسية ذكية. تُطرح البيانات الضخمة حاليًا قضايا تتجاوز بكثير مجال الإحصاء الكلاسيكي، وتتطلب تعاونًا بين فريق من الخبراء ذوي المعرفة المتخصصة، ومهارات الحوسبة، والتفكير الإحصائي. هناك العديد من الخوارزميات التي تم تحسينها مؤخرًا، وهي مناسبة للاستدلال الإحصائي للبيانات الضخمة، ويمكن استخدامها على الأجهزة المتوازية، بما في ذلك خوارزمية التمهيد الصغيرة ، ومعادلة التقدير المجمعة ، وخوارزميات التقسيم والغزو، وخوارزمية التقريب العشوائي القائمة على أخذ العينات الفرعية. علاوة على ذلك، تُستخدم الخوارزميات المتكررة على نطاق واسع في مجتمع الحوسبة العلمية الحالي. تشمل أمثلة هذه الخوارزميات التكرارية خوارزميات مونت كارلو لسلسلة ماركوف (MCMC) ، وخوارزمية EM والتي تتطلب عادةً عددًا كبيرًا من التكرارات ومسحًا كاملاً لمجموعة البيانات الكاملة لكل تكرار. ومع ذلك، لا تُعدّ التقنيات الإحصائية القياسية مناسبةً للتحكم في البيانات الضخمة، وقد اقترح العديد من الباحثين توسعاتٍ في التقنيات الكلاسيكية أو أساليب جديدة تمامًا. اقترح الباحثون خوارزمية تقريبية فعّالة للانحدار الرتيب متعدد المتغيرات واسع النطاق ، والإحصاءات المتوازية، والعديد من خوارزميات الإحصاءات المتوازية. لكن في الواقع، لا يمكن تحليل هذا النوع من مجموعات البيانات الضخمة على حاسوب واحد، لأن أحجامها كبيرة جدًا بحيث لا تتسع في الذاكرة، أو أن معالجتها تستغرق وقتًا طويلًا عند استخدام الأساليب الإحصائية الحالية. لتجاوز هذا العائق، قد يضطر المرء إلى تطبيق بنى متوازية وموزعة، حيث توفر منصات الحوسبة السحابية إمكانية الوصول إلى مئات أو آلاف المعالجات. مع الحجم المزدهر، عادةً ما يأتي تعقيد متزايد في هياكل البيانات،

وفي أنماط البيانات، وفي النماذج اللازمة لمراعاة هذه الأنماط أيضًا. لقد فرضت البيانات الضخمة تحديًا كبيرًا على المنهجية الإحصائية الحديثة. البيانات الضخمة في البحث العلمي. مع تطور علوم الحاسوب، أصبحت العديد من المجالات العلمية تعتمد بشكل كبير على البيانات. على سبيل المثال، تعتمد علوم الفلك، والأرصاد الجوية، والحوسبة الاجتماعية، وعلم المعلومات الحيوية، وعلم الأحياء الحاسوبي على كميات هائلة من البيانات بأنواع مختلفة، تم إنشاؤها أو إنتاجها في هذه المجالات العلمية. وقد أدى ظهور البيانات الضخمة إلى ظهور نموذج بحثي جديد. في عام 2007، وصف جيم جراي، الحائز على جائزة تورينج، نموذجًا رابعًا لحجم بيانات البحث العلمي. وأعرب عن اعتقاده بأن هذا النموذج هو السبيل الوحيد لمعالجة بعض أصعب التحديات العالمية التي نواجهها اليوم. يتعلم الفيزيائيون خصائص الجسيمات من خلال تصادمها مع جسيمات أخرى في تجارب عالية التقنية. وتكمن تطبيقات البيانات الضخمة الأخرى في مجالات علمية عديدة، مثل علم الفلك، وعلوم الغلاف الجوي، والطب، وعلم الجينوم، وعلم الأحياء، والكيمياء الحيوية الجيولوجية، وغيرها من الأبحاث العلمية المعقدة ومتعددة التخصصات.

4. مجالات تطبيق البيانات الضخمة

أشار معهد ماكينزي في تقرير صدر مؤخرًا إلى الإمكانات التحويلية للبيانات الضخمة في خمسة مجالات: الرعاية الصحية؛ إدارة القطاع العام؛ تجارة التجزئة؛ التصنيع العالمي؛ وبيانات الموقع الشخصي. في الأقسام الفرعية التالية، سنستعرض بإيجاز بعض تطبيقات البيانات الضخمة في مجالات التجارة والأعمال، والرعاية الصحية، والاقتصاد والمالية، والبحث العلمي. البيانات الضخمة في البحث العلمي. أصبحت العديد من المجالات العلمية تعتمد بشكل كبير على البيانات مع تطور علوم الحاسوب. على سبيل المثال، تعتمد علوم الفلك، والأرصاد الجوية، والحوسبة الاجتماعية، وعلم المعلومات الحيوية، وعلم الأحياء الحاسوبي على كميات هائلة من البيانات بأنواع مختلفة، تم إنشاؤها أو إنتاجها في هذه المجالات العلمية . أدى ظهور البيانات الضخمة إلى ظهور نموذج بحثي جديد. في عام 2007، وصف جيم غراي، الحائز على جائزة تورينج، نموذجًا رابعًا لحجم بيانات البحث العلمي. وأعرب عن اعتقاده بأن هذا النموذج هو السبيل الوحيد لمعالجة بعض أصعب التحديات العالمية التي نواجهها اليوم . يتعلم الفيزيائيون خصائص الجسيمات عن طريق تصادمها مع جسيمات أخرى في تجارب عالية التقنية. وتكمن تطبيقات البيانات الضخمة .

وعلوم الغلاف الجوي، والطب، وعلم الجينوم، وعلم الأحياء، والكيمياء الحيوية الجيولوجية، وغيرها من الأبحاث العلمية المعقدة ومتعددة التخصصات. وتستضيف منظمة الأبحاث النووية الأوروبية CERN واحدة من أكبر التجارب المعروفة في العالم، وهي مثالٌ بارزٌ على البيانات الضخمة. لأكثر من 50 عامًا، عانت منظمة الأبحاث النووية الأوروبية CERN من التدفق المتزايد للبيانات الناتجة عن تجاربها التي تبحث في الجسيمات الرئيسية والقوى التي تتفاعل بها. يتكون مصادم الهدرونات الكبير LHC من حلقة طولها 27 كيلومترًا من مغناطيسات فائقة التوصيل، مزودة بعدد من الهياكل المُسرّعة لتعزيز طاقة الجسيمات على طول الطريق. يحتوي الكاشف على 150 مليون مستشعر، ويعمل ككاميرا ثلاثية الأبعاد، حيث يلتقط صورًا لأحداث تصادم الجسيمات بسرعة 40 مليون مرة في الثانية، ويمكنه توليد 60 تيرابايت من البيانات يوميًا. يمكن للأنماط الموجودة في هذه البيانات أن تمنحنا فهمًا غير مسبوق لطبيعة الكون..

البيانات الضخمة في الاقتصاد والتمويل. وفقًا لتقرير معهد ماكينزي، فإن الاستخدام الفعال للبيانات الضخمة له فوائد جوهرية لتحويل الاقتصادات، وتحقيق موجة جديدة من النمو الإنتاجي. على مدار السنوات العشر الماضية، تتبنى المزيد من الشركات نهجًا قائمًا على البيانات لتقديم خدمات اكثر استهدافا وتقليل المخاطر ورفع الاداء. وتُجري هذه الشركات برامج تحليل بيانات مُتخصصة لجمع وتخزين وإدارة وتحليل مجموعات البيانات الضخمة من مجموعة من المصادر لتحديد رؤى الأعمال الرئيسية التي يُمكن استخدامها لدعم اتخاذ قرارات أفضل. من الصعب تحليل مجموعة كبيرة من البيانات الاقتصادية والمالية. على سبيل المثال، كأداة مهمة في تحليل التطور الشائع للسلاسل الزمنية للاقتصاد الكلي، لا يتضمن نموذج الانحدار الذاتي للمتجهات VAR عادةً أكثر من 10 متغيرات، نظرًا لأن عدد المعلمات يتزايد تربيعيًا مع حجم النموذج . ولكن، يحتاج خبراء الاقتصاد القياسي حاليًا إلى تحليل السلاسل الزمنية متعددة المتغيرات التي تحتوي على أكثر من مئات المتغيرات. سيؤدي تضمين جميع المعلومات في نموذج الانحدار الذاتي للمتجهات إلى الإفراط في المطابقة وسوء أداء التنبؤ. يُعد اللجوء إلى احتمالات الندرة، التي طُوّرت بموجبها أدوات إحصائية جديدة، أحد الحلول . ومن الأمثلة الأخرى تحسين المجلد وإدارة المخاطر . في هذه المسألة، يلعب تقييم مصفوفات التباين والتباين العكسي

لعوائد الأصول في المحفظة دورًا مهمًا. لاحظ الباحثون في أن الخطأ التراكمي لتقدير المصفوفة بأكملها قد يكون كبيرًا في ظل معايير المصفوفة، حتى لو تمكنا من تقدير كل معلمة على حدة بدقة، وهذا يتطلب

إجراءات إحصائية جديدة. البيانات الضخمة في التجارة والأعمال. لطالما جُمعت بيانات معاملات الشراء من المواقع التجارية. ولكن الآن، تُولّد المواقع التجارية أنواعًا جديدة من البيانات الضخمة. ووفقًا للتقييمات، يتضاعف حجم بيانات الأعمال عالميًا، بين الشركات، تقريبًا كل 1.2 عام. يمكن اكتشاف هذا الكم الهائل من البيانات باستخدام البيانات الضخمة، التي تساعد، على سبيل المثال، في تحسين العمليات التجارية، واكتشاف الأنماط، وفهم العملاء بشكل أفضل، وتوقع سلوكياتهم، واحتياجاتهم، ونواياهم. تدريجيًا، تبني الشركات قراراتها على البيانات. تحتاج الشركات إلى موظفين لجمع بيانات المنتج الملائمة وتحليلها في سياق الصناعة. لتحديد أنواع التحسينات أو المنتجات الجديدة التي يجب إدخالها لتلبية احتياجات عملائهم، ينظر المحللون إلى بيانات الشراء وتقييمات العملاء. على سبيل المثال، لمعرفة أنواع المنتجات التي يشتريها العملاء ومتى يشترونها، قد يدرس الموظفون بيانات المعاملات من بطاقات ولاء المتاجر. كما يمكن للبيانات الضخمة مساعدة الشركات على العمل بكفاءة أكبر. يطبق المحللون بيانات سلسلة التوريد للتحكم في المخزونات. كما يكتشفون الأخطاء من خلال فحص بيانات الإنتاج الآنية . يمكن لتحليل البيانات الضخمة واستخراجها أيضًا أن يحمي الأمن العام بشكل فعال، ويكافح الجرائم الجنائية والاقتصادية، وينظم الاتصالات، ويجمع
البيانات.

5. مشاكل تحليل البيانات الضخمة

أدت أحدث التطورات في مجال التكنولوجيا الرقمية الحديثة إلى انتشار واسع النطاق لجمع البيانات على نطاق واسع. بيانات المناخ، وبيانات شبكات التواصل الاجتماعي، والهواتف الذكية، وبيانات الحالة الصحية، وبيانات النصوص غير المنظمة، وسلاسل البيانات الزمنية لوسائل التواصل الاجتماعي والمالية، وبيانات التجارة الإلكترونية، ومذكرات عقود البيع بالتجزئة، وفيديوهات المراقبة، بما في ذلك هذه البيانات. وتتمتع التحليلات الإحصائية لهذه البيانات الضخمة بأهمية حاسمة. تتميز البيانات الضخمة بخصائص فريدة لا تتفوق عليها البيانات التقليدية. تتميز البيانات الضخمة بحجمها الكبير، وحجم العينات المرتفع، ومعدل نموها المرتفع. وكقاعدة عامة، يتم جمعها من مصادر متعددة باستخدام تقنيات مختلفة في أوقات معينة. وتؤدي هذه الميزات إلى مشاكل مثل جمع الضوضاء، والارتباطات الخاطئة، والتعقيد الحسابي، وعدم استقرار الخوارزميات. كما أنها تخلق مشاكل غير متجانسة، وتؤدي إلى أخطاء في التحليل التجريبي والإحصائي . وتخلق هذه الميزات مشاكل كبيرة في تحليل البيانات وزيادة الأساليب الإحصائية. وفي الوقت نفسه، تُصعّب تطبيق الإجراءات الإحصائية التقليدية. على سبيل المثال، من المستحيل تطبيق العديد من الأساليب التقليدية المناسبة لجمع البيانات متوسطة الحجم على البيانات الضخمة. وفي الوقت نفسه، تواجه الأساليب الإحصائية الممكنة للبيانات ذات الحجم الأصغر تحديات كبيرة أثناء تحليل البيانات عالية الأبعاد. مع تزايد حجم البيانات، يصبح من المستحيل حفظها في آلة عد ومعالجتها بفعالية باستخدام أساليب التحليل الإحصائي التقليدية.
وبالتالي، قد يواجه التحليل الإحصائي للبيانات الضخمة بعض
المشكلات الرئيسية:
- حجم كبير وتنوع كبير في البيانات الضخمة؛
- عدم القدرة على جمع كميات كبيرة من البيانات باستخدام خوارزميات التحليل الإحصائي الشائعة أو سرعة معالجة منخفضة للغاية.
—اختلاف بنية البيانات؛
- قضايا حماية البيانات والخصوصية؛
- ضيق الوقت؛
- عدم وجود ممارسات إحصائية لاستخراج المعلومات.
لذا، ليس من المنطقي توقع فعالية الأساليب الإحصائية التقليدية في حل مشكلات البيانات الضخمة. تُعد الأساليب الإحصائية والحسابية الجديدة مهمة لحل مشكلات البيانات الضخمة.

6. الأساليب الإحصائية للبيانات الضخمة

عام 2013 هو السنة الدولية للإحصاء. وهو تكليفٌ يهدف إلى تسليط الضوء على دور البيانات والتحليل الإحصائي في المجتمع . وقد حقق الإحصاء نجاحًا ملحوظًا في مجال أبحاث البيانات الضخمة أيضًا. هناك العديد من الخوارزميات التي طُوّرت مؤخرًا، وهي قابلة للتطبيق في الاستدلال الإحصائي للبيانات الضخمة، ويمكن استخدامها على الآلات المتوازية، بما في ذلك خوارزمية التمهيد الصغيرة، ومعادلة التقدير المجمعة،

وخوارزمية التقريب العشوائي القائمة على أخذ العينات الجزئية . تستند معادلة التقدير المجمعة إلى نفس فكرة التقسيم و، ولكنهما تُركّزان على أنواع مختلفة من

المشكلات ، فالأولى لتقدير المعاملات، والثانية لاختيار متغيرات نماذج الانحدار.
يمكن تقسيم المنهجيات الحديثة للبيانات الضخمة إلى ثلاث فئات إعادة العينة، والتحديث المتسلسل.

تتضمن الطريقة القائمة على أخذ العينات الفرعية، والمقترحة للتحليل الإحصائي للبيانات الضخمة، ثلاثة مناهج :
1) اساليب التمهيد الصغيرة
2) متوسط احتمالية اللوغاريتم
3) الاستفادة.
قدّم كلاينر وآخرون نهج "أساليب التمهيد الصغيرة" BLB للتحليل الإحصائي للبيانات الضخمة. ويُتيح هذا النهج توفير تقييمات منفصلة، مثل التشتت أو فترات الثقة، ومؤشرات الجودة. وهو مزيج من أخذ العينات الجزئية، ونهج التمهيد من خلال -out-n.m، ونهج التمهيد لتحقيق الكفاءة الحسابية
تتكون خوارزمية BLB من الخطوة التالية أولاً، تُؤخذ عينات فرعية s ذات أبعاد m من بيانات أولية ذات أبعاد n. يتم الحصول على مؤشرات التقييم المنفصلة والجودة (مثل فترات الثقة) من عينات إعادة تمهيد ذات أبعاد n مأخوذة لكل عينة فرعية s. بعد ذلك، يتم الحصول على تقييمات منفصلة عامة ومؤشرات جودة من مزيج من تقديرات نقطة إعادة تمهيد s (حسابات) ومقاييس الجودة (على سبيل المثال، بالمتوسط).
باختصار، تتكون خوارزمية BLB من إجراءين مترابطين ببعضهما البعض:
١) يُطبَّق الإجراء الداخلي للتحميل الأولي (إعادة التشغيل) للعينة الفرعية؛
٢) يتضمن الإجراء الخارجي العديد من تقييمات إعادة التشغيل (التحميل الأولي).
على الرغم من أن إجراء إعادة التشغيل الداخلي يُولِّد نظريًا بيانات متعددة مُعاد أخذ عينات منها بحجم n، إلا أن ما يلزم فعليًا في التخزين والحساب هو عينة بحجم m مع متجه وزن. على عكس أخذ العينات الفرعية وإعادة التشغيل m-out-of-n، لا توجد حاجة في هذا النهج إلى توسيع من مقياس فترات الثقة حتى النتيجة النهائية للارتباط التحليلي.
يُسهّل إجراء BLB الحساب الموزع من خلال إتاحة الفرصة لمعالجة كل عينة فرعية ذات بُعد m في معالجات منفصلة. أثبت كلاينر وزملاؤه ديناميكية BLB ودقتها العالية. أظهرت دراسات النمذجة التي أجروها دقة عالية، وسرعة تقارب، وكفاءة حسابية ملحوظة . تُعدّل حقيبة أدوات التمهيد الصغيرة أداة التمهيد العادية لجعلها مناسبة لمجموعات البيانات واسعة النطاق. على الرغم من تسهيل مشكلة معالجة عينات التمهيد الضخمة، إلا أن حساب التقديرات لعدد كبير منها مكلف للغاية في.
يُعد BLB غير عملي للمقدرين الحديثين ذوي التعقيد العالي بشكل عام. من ناحية أخرى، فإن استخدام مقدر LS البدائي في مخطط BLB الأصلي لا يوفر إجراء تمهيد قويًا إحصائيًا .
طريقة التمهيد السريع والمتين (FRB) التي اقترحها شهاب بصيري وآخرون قابلة للتطوير ومتوافقة مع بنى الحوسبة الموزعة وأنظمة التخزين، ومتينة في مواجهة القيم المتطرفة، وتوفر نتائج دقيقة باستمرار وبمعدل أسرع بكثير من طريقة BLB الأصلية .
—احتمالية لوغاريتم المتوسط
قدم ليانغ وآخرون نهج تقريب عشوائي يعتمد على اختيار التكرار مع تطبيق البيانات الجيو إحصائية. تستخدم الطريقة متوسطات مونت كارلو المحسوبة من عينات فرعية لتقريب الكميات اللازمة للبيانات الكاملة .
يتم الحصول على حل معادلة متوسط الدرجات من إجراء تقريب عشوائي، حيث يتم تحديث التقدير الحالي في كل تكرار بناءً على عينة فرعية بحجم m مأخوذة من البيانات الكاملة. ونظرًا لأن m أصغر بكثير من n، فإن الطريقة قابلة للتطوير لتشمل البيانات الضخمة.
قام ليانغ وآخرون بإنشاء التسلسل والتوزيع الطبيعي المقارب للمقدر الناتج في ظل ظروف غير مستقرة. في بحث محاكاة، كانت سرعة تقارب الطريقة مستقلة تقريبًا عن n، حجم عينة البيانات الكاملة . قام ليانغ وكيم بتوسيع متوسط احتمالية اللوغاريتم إلى خوارزمية متروبوليس-هاستينغز المُعاد تدويرها في سلسلة ماركوف مونت كارلو MCMC. في خوارزمية متروبوليس-هاستينغز، يتم تدوير نسبة احتمالية الاقتراح والتقدير الحالي مع النسبة التقريبية من متوسط احتمالية اللوغاريتم بناءً على k عينة مُعاد تدويرها بحجم m. يمكن تنفيذ الخوارزمية باستغلال البنية المتوازية غير الملائمة، مما يمنع إعادة مسح مجموعة البيانات بأكملها في التكرارات.

-أساليب الاستفادة
في طرق الاستفادة المقترحة من ما وسون، يتم أخذ عينات من نسبة صغيرة من البيانات ذات أوزان محددة (عينة فرعية) من العينة الكاملة. ثم تُجرى العمليات الحسابية المقصودة للعينة الكاملة باستخدام العينة الفرعية الصغيرة كبديل. يكمن سر نجاح طرق الاستفادة في تحديد الأوزان، واحتمالات أخذ العينات غير المنتظمة .
تختلف طرق الاستفادة عن أخذ العينات الفرعية التقليدي أو إعادة التشغيل m-out-of-n في :
1) تُستخدم لتحقيق حسابات قابلة للتحقيق حتى مع توفر نتائج تحليلية بسيطة؛
2) تُمكّن من تصور البيانات عندما يكون تصور العينة الكاملة غير ممكن؛
3) عادةً ما تستخدم احتمالات أخذ عينات غير متساوية لبيانات أخذ العينات الفرعية.
يتميز هذا النهج بتصميم فريد للغاية، إذ يسمح بالوصول الشامل لاستخراج المعلومات من كميات هائلة من البيانات دون الحاجة إلى حوسبة عالية الأداء.

ب. تجزئة وسد
تتكون هذه الخوارزمية من ثلاث مراحل:
1) تقسيم مجموعة بيانات كبيرة إلى K كتلة؛
2) معالجة كل كتلة على حدة (ربما بالتوازي)؛
3) تجميع الحلول من كل كتلة لتشكيل حل نهائي للبيانات الكاملة .
للتقييم غير الخطي العام، عُرضت معادلات تقريب خطي لمعادلات التقدير باستخدام توسع تايلور عند الحل في كل كتلة. أظهر لين وشي أن المقدر المجمع له نفس حد المقدر من البيانات الكاملة . في الخطوة 2، يُطبق الانحدار المعاقب على كل كتلة على حدة باستخدام دالة جزاء مُحفزة للتشتت تُلبي شروط انتظام معينة. يمكن أن يؤدي هذا النهج إلى اختيار متغير تفاضلي بين الكتل، حيث قد تؤدي كتل البيانات المختلفة إلى تقديرات معاقب عليها بمعاملات انحدار مختلفة غير صفرية.
في المرحلة الثالثة، تُجمع نتائج كتل K لإنشاء تقييم مشترك. في هذه المرحلة، تُجمع النتيجة المشتركة للتقييمات الناتجة عن دمج فكرة توزيعات الثقة في التحليل التلوي.
ج. التحديث التسلسلي لبيانات التدفق
في بعض الحالات، تأتي البيانات في تدفقات أو أجزاء كبيرة، ويُفضل إجراء تحليل مُحدّث تسلسليًا دون الحاجة إلى مساحة تخزين. يُوسّع شيفانو وآخرون نطاق عمل لين وشي بعدة طرق مهمة.

أولًا، في هذه الطريقة، تُدخل تقديرات تباين من نوع "التقسيم والتغلب" لمعلمات الانحدار في النموذج الخطي وإعدادات معادلة التقدير . يتيح تقييم التباين هذا للمستخدمين استخلاص استنتاجات حول معلمات الانحدار الحقيقية بناءً على تقديرات نقطة "التقسيم والتغلب" المُطوّرة مسبقًا لمعلمات الانحدار. ثانيًا، يقومون بتوسيع خوارزميات التقييم التكراري والاستدلالات الإحصائية للنماذج الخطية ومعادلات التقدير (للتحديث مع وصول بيانات جديدة) .
ثالثًا، أثناء التعامل مع كتل البيانات، تُعالج مسألة أوجه القصور المحتملة في الرتب وخصائص التفرد للمقدرين المُركّبين والتراكميين أثناء استخدامهم لمعكوس معمم. بالإضافة إلى ذلك، قدّم المؤلفون مقدّرًا جديدًا مُحدّثًا عبر الإنترنت لمعاملات الانحدار، ومقدّرًا مُقابلًا للخطأ المعياري في إعداد معادلة التقدير، مستفيدًا من المعلومات الواردة من البيانات السابقة.
استُخدمت خوارزمية تعظيم التوقعات (EM) على نطاق واسع في الحوسبة العلمية لتقدير المعاملات في حال وجود بيانات مفقودة. يثير نجاح الخوارزميات التكرارية في الحوسبة العلمية الحديثة اهتمامًا بتطوير بعض الخوارزميات التكرارية المُجدية للبيانات الضخمة .
أثبتت طرق مونت كارلو التكرارية، مثل MCMC، والتقريب العشوائي، وEM، أنها أدوات فعّالة جدًا لتحليل البيانات الإحصائية. ومع ذلك، فإن طبيعتها التي تتطلب استخدامًا مكثفًا للحاسوب، والتي تتطلب عادةً عددًا كبيرًا من التكرارات ومسحًا كاملاً لمجموعة البيانات الكاملة لكل تكرار، تمنع استخدامها لتحليل البيانات الضخمة. في مراجعة لأحدث تطورات أساليب مونت كارلو التكرارية لتحليل البيانات الضخمة. تُظهر الأبحاث أنه بناءً على مجالات التطبيق، تُستخدم أساليب إحصائية مختلفة في معالجة البيانات الضخمة. ولكن، عادةً ما تكون التقنيات الإحصائية القياسية. لا تتوافق هذه البيانات بشكل جيد مع قواعد البيانات الضخمة، وقد اقترح العديد من الباحثين توسعات في التقنيات الكلاسيكية أو أساليب جديدة تمامًا. على سبيل المثال، قام سي. أنجيليني وزملاؤه مؤخرًا بتطوير أساليب بايزية وظيفية مصممة خصيصًا لبيانات المصفوفات الدقيقة ذات المسار الزمني. تتعامل هذه الأساليب بنجاح مع مختلف الصعوبات التقنية التي تنشأ في هذا النوع من التجارب، مثل العدد الكبير من الجينات، وعدد قليل من الملاحظات، وفترات أخذ العينات غير المنتظمة،

والبيانات المفقودة أو المتعددة، والتبعية الزمنية بين الملاحظات لكل جين. يقترح بيرك وآخرون نموذجًا وظيفيًا للتأثيرات المختلطة لتقدير النمط الزمني لكل جين، والذي يُفترض أنه دالة سلسة . الإحصاء هو العلم الذي يجمع البيانات ويرتبها ويشرحها. مع تزايد مشاكل "البيانات الضخمة"، أصبحت الحوسبة الإحصائية والتعلم الإحصائي مجالًا حيويًا للغاية في العديد من المجالات العلمية، بالإضافة إلى التسويق والتمويل وغيرها من مجالات الأعمال. الإحصاءات الحسابية، أو الحوسبة الإحصائية، هي الواجهة بين الإحصاءات وعلوم الكمبيوتر . يقدم المؤلف تنبؤًا حول اتجاه تطوير الحوسبة الإحصائية في . يتمثل تطورها في أن التكنولوجيا ستؤثر على الحوسبة الإحصائية أكثر من العوامل الأخرى.

يستند هذا التنبؤ إلى الملاحظات الحديثة للمجال على مدى الأربعين عامًا الماضية. لا تشمل التكنولوجيا التي تقود هذا التنبؤ الأجهزة فحسب، بل تشمل أيضًا البرامج التي تضمن البنية التحتية للتواصل الفردي والمجتمعي مع أجهزة الكمبيوتر. سيعتمد تولي علماء الكمبيوتر لهذا المجال في نهاية المطاف على مدى فعالية مشاركة الإحصائيين. سيتعين على الإحصائيين المهتمين بالحوسبة الإحصائية وتطبيقاتها المستقبلية الانخراط في أبحاث تعاونية مع علماء الكمبيوتر لمواصلة التأثير.
يشير التعلم الإحصائي إلى مجموعة من الأدوات لنمذجة وفهم مجموعات البيانات المركبة. يشمل هذا المجال العديد من الأساليب مثل الانحدار الخطي والهش، وأشجار التصنيف والانحدار، وزيادة آلات المتجهات وتعزيزها. يتضمن التعلم الإحصائي بناء نموذج إحصائي للتنبؤ، أو تقدير، مخرجات بناءً على مُدخل واحد أو أكثر. وتظهر تحديات من هذا النوع في مجالات متنوعة مثل الأعمال والطب والفيزياء الفلكية والسياسة العامة . يُصوّر هذا الكتاب، الذي قدّمه تريفور هاستي وآخرون، الأفكارَ المهمة في مجال الإحصاء، واستخراج البيانات، والتعلم الآلي، والمعلوماتية الحيوية، في إطار مفاهيمي عام. يتناول هذا الإصدار الجديد الرئيسي العديد من المواضيع، بما في ذلك النماذج الرسومية، والغابات العشوائية، وأساليب التجميع، وانحدار الزاوية الصغرى، وخوارزميات المسار للاسو، وعوامل المصفوفات غير السلبية، والتجميع الطيفي. طوّر هذا الكتاب نماذج جمعيّة معممة، والعديد من برامج وبيئات النمذجة الإحصائية في R/S-PLUS، وغيرها .
يُعدّ حل مشاكل قابلية التوسع المتوازي أحد التحديات المهمة لأي تحليل إحصائي مُصمّم للبيانات واسعة النطاق. باستخدام سلسلة من الصيغ التي تسمح بتحديثات ثنائية متوازية وتزايدية، ذات مسار واحد، ومع ذلك متينة عدديًا، لكل من اللحظات الإحصائية المركزية ذات الترتيب العشوائي واللحظات المشتركة، بنى بينيت ج. وزملاؤه إطار عمل إحصائي متوازي مفتوح المصدر يُجري تحليل المكونات الرئيسية (PCA) بالإضافة إلى حساب الإحصاءات الوصفية والارتباطية ومتعددة الارتباطات . يُعد الانحدار الرتيب (MR) أداة فعالة لتقييم الدوال الرتيبة فيما يتعلق بمتغيرات الإدخال. لمشكلة الانحدار الرتيب العديد من التطبيقات في التشغيل
تُناسب جميع الطرق تقريبًا التي تُنمذج الخصائص الإحصائية لمعاملات المويجات مجموعات بيانات صغيرة فقط. يرتبط معظمها بإزالة الضوضاء، وتحليل الملمس، وخوارزميات التجزئة والتصنيف والاسترجاع. بالنسبة لمجموعات البيانات الضخمة، مثل مجموعات بيانات الاستشعار عن بُعد الزماني والمكاني، تحتاج أيضًا إلى نمذجة معاملات المويجات الخاصة بها للكشف عن الاتجاهات المتغيرة، واكتشاف الآليات الجوهرية، وتمثيل قواعد عملية تطورها.
اقترح ل. وانغ وآخرون في استخدام نموذج الموجي العام (GMM) لتوقيع الخصائص الإحصائية لمعاملات المويجات لمجموعة بيانات ضخمة للاستشعار عن بُعد. تهدف هذه المساهمة إلى حساب معلمات نموذج مجموعة بيانات ضخمة باستخدام جوانب أو أبعاد مختلفة مثل الوقت، والنطاقات الطيفية، والمقاييس، والملمس. قدّم لينغ سونغ طريقةً جديدةً للكشف عن العناقيد متعددة الدقة (MCD) لمعادلة العناقيد غير المنتظمة في الفضاء، واستنبط إحصائية اختبار متعددة المقاييس على خلية واحدة بناءً على إحصائية نسبة الاحتمالية لمتتالية برنولي، ومتتالية بواسون، والمتتالية الطبيعية. بالمقارنة مع طرق الاختبار أحادية المقياس التي تتحكم في معدل الاكتشاف الخاطئ، وإحصائيات المسح المكاني باستخدام بيانات المحاكاة والتصوير بالرنين المغناطيسي الوظيفي، تبيّن أن هذه الطريقة أكثر فعاليةً في اكتشاف العناقيد غير المنتظمة. ولا يتطلب تطبيق طريقته المقترحة حساباتٍ معقدة، مما يجعلها ملائمةً للكشف عن العناقيد للبيانات المكانية الكبيرة .
قدّم هونغتو تشو وآخرون عدة فئات من نماذج الانحدار المكاني، بما في ذلك نماذج المعاملات المتغيرة مكانيًا، ونماذج العمليات الغوسية التنبؤية المكانية، ونماذج الانحدار الموتر، ونماذج الانحدار الخطي الوظيفي كوكس، وذلك للتحليل التعاوني لبيانات التصوير العصبي الضخمة والبيانات السريرية والسلوكية. وتراعي نماذجهم الإحصائية بوضوح بعض الخصائص المميزة لبيانات التصوير العصبي: توافر مناطق انزلاقية متعددة القطع ذات حواف وقفزات غير معروفة، وارتباطات مكانية أساسية. كما حسّنوا بعض إجراءات التقييم السريع لتقدير دوال المعاملات المتنوعة والارتباطات المكانية في آنٍ واحد. كما استكشفوا بشكل منهجي الخصائص المقاربة (مثل الاتساق والتوزيع الطبيعي المقارب) لتقديرات المعاملات التكيفية

متعددة المقاييس. وقد أكدت محاكاة مونت كارلو وتحليل البيانات الحقيقية الأداء المثالي لنماذجهم في تطبيقات مميزة . تُسبب البيانات الضخمة مشاكل حتى في التحليل الإحصائي الأولي، مما يُؤدي إلى عوائق في ذاكرة الحاسوب ووقت الحوسبة. عادةً ما يكون حاجز ذاكرة الحاسوب تتم معالجتها بواسطة اتصال قاعدة بيانات يستخرج البيانات في أجزاء للمعالجة. تُدار هذه البيانات باستخدام الحوسبة المتوازية، والتي غالبًا ما تُسرّعها وحدات المعالجة الرسومية. تُساعد حزم R مفتوحة المصدر، مثل bigglm و bigmemory، على تجاوز حدود ذاكرة الحاسوب . نظرًا لحجمها وتعقيدها، تُشكّل مجموعات البيانات الضخمة العديد من التحديات الحسابية للتحليل الإحصائي. يُعدّ التغلب على قيود الذاكرة وتحسين الكفاءة الحسابية للطرق الإحصائية التقليدية من

بين هذه المشكلات يُقسّم التجميع الإحصائي مجموعة البيانات بأكملها إلى مجموعات فرعية أصغر، ويضغط كل مجموعة فرعية في إحصاءات ملخصة منخفضة الأبعاد، ويُجمّع هذه الإحصاءات لتقريب الحساب المطلوب بناءً على البيانات بأكملها. يجب أن تكون النتائج متكافئة تقريبيًا، والتي يتم الحصول عليها من التجميع الإحصائي. يُعد التجميع الإحصائي مفيدًا بشكل خاص لدعم التحليلات الإحصائية المتطورة للمعالجة التحليلية عبر الإنترنت في مكعبات البيانات.

7. الخاتمة

نعيش حاليًا في مجال واسع الانتشار في العلوم والهندسة والتكنولوجيا، مما أتاح إنتاج تدفقات بيانات ضخمة قابلة للقياس، تصل إلى إكسابايت وزيتابايت. تبشر البيانات الضخمة بمستويات حديثة من الوحي العلمي والقيمة الاقتصادية. أصبحت البيانات الضخمة أحد اتجاهات البحث الحالية والمستقبلية.
تجدر الإشارة إلى أن البيانات الضخمة ستُحدث ثورة في العديد من المجالات، بما في ذلك الأعمال والرعاية الصحية والبحث العلمي والإدارة العامة، وغيرها. كما أن التطبيق الناجح لهذه التقنية متاح. فقد تم إنشاء عدد كبير من مستودعات بيانات المصفوفات الدقيقة لدراسة التعبير الجيني، وأصبحت الكاميرات المتطورة منتشرة في كل مكان، مما يُنتج كميات هائلة من البيانات المرئية للمراقبة. ويجري حاليًا بناء تلسكوب المصفوفة الكيلومترية المربعة لأبحاث الفيزياء الفلكية، ومن المتوقع أن يُنتج عدة بيتابايتات من البيانات الفلكية سنويًا. تحتوي جميع مجموعات البيانات، التي تُسمى البيانات الضخمة، على عدد كبير من الأبعاد (السمات) وتشكل تحديات بحثية كبيرة للتحليل الإحصائي واستخراج البيانات. تُمثل البيانات الضخمة تحديات تتجاوز بكثير فرص الإحصاءات التقليدية، وتتطلب قوى عاملة مشتركة مع معرفة متخصصة، وتقنيات حوسبة، وأساليب إحصائية. تكمن أهمية النقاط التالية لفهم المشكلة بشكل صحيح:
1. تعقيد البيانات (التي جُمعت من مصادر وتنسيقات مختلفة).
2. تحدي البيانات المشوشة: تتضمن البيانات الضخمة عمومًا أنواعًا مختلفة من أخطاء القياس، والقيم المتطرفة، والقيم المفقودة.
3. تحدي البيانات التابعة: في أنواع مختلفة من البيانات الحالية، مثل السلاسل الزمنية المالية وما شابه.
4. يتطلب الأمر معالجة أساليب وخوارزميات إحصائية أكثر فعالية في حل المشكلات، بحيث تكون متينة في مواجهة تعقيد البيانات، والضوضاء، والاعتماد على البيانات. كما أظهرت الدراسات، فقد طُوّرت أساليب إحصائية تقليدية كافية في هذا المجال (مثل: خوارزميات إعادة التشغيل، وخوارزميات التجزئة والغزو، والأساليب القائمة على أخذ العينات الفرعية، وغيرها)، وعالج الباحثون خوارزميات جديدة (مثل: طريقة الكشف عن المجموعات متعددة الدقة (MCD)، ونموذج التأثيرات المختلطة الوظيفية، وغيرها). ويبدو أن هذا المجال سيظل موضوعًا للبحث الذي يطبقه الباحثون باستمرار.

المراجع

1. Alguliyev R. M. “Big Data” phenomenon of Information Technology. – 2014. – Vol. 2. – P. 3–16.: Challenges and Opportunities [Text] / R. M. Alguliyev, M. S. Hajirahimova // Problems

2. Philip, C. L. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data [Text] / C. L. Philip, C.-Y. Zhang // Information Sciences. – 2014. – Vol. 275. – P. 314–347. doi: 10.1016/j.ins.2014.01.015
3. Soares, S. Big Data Governance. An Emerging Imperative. 1st edition [Text] / S. Soares. – MC Press Online, LLC, 2012. – 368 p.
4. Kambatla, K. Trends big data analytics [Text] / K. Kambatla, G. Kollias, V. Kumar, A. Grama // Parallel and Distributed Computing. – 2014. – Vol. 74, Issue 7. – P. 2561–2573. doi: 10.1016/j.jpdc.2014.01.003
5. Fan, J. Challenges of Big Data analysis [Text] / J. Fa, F. Han, H. Liu // National Science Review. – 2014. – Vol. 1, Issue 2. – P. 293–314. doi: 10.1093/nsr/nwt0326. Jordan, J. M. Statistics for big data: Are statisticians ready for big data? [Text] / J. M. Jordan, D. J. Lin // International Chinese Statistical Association Bulletin. – 2014. – Vol. 26. – P. 59–66.
6. Yu, B. Let us own data science [Text] / B. Yu // IMS Bulletin Online. – 2014. – Vol. 43, Issue 7.
7. Chun W. C. A Survey of Statistical Methods and Computing for Big Data [Electronic resource] / W. C. Chun, M. H. Chen, E. Schifano, J. Wu, J. Yan. – 2015. – Available at: http://de.arxiv.org/abs/1502.07989v1
8. Lin, N. Aggregated estimating equation estimation [Text] / N. Lin, R. Xi // Statistics and Its Interface. – 2011. – Vol. 4, Issue 1. – P. 73–83. doi: 10.4310/sii.2011.v4.n1.a8
9. Chen, M. H. Statistical and Computational Theory and Methodology for Big Data Analysis [Electronic resource] / M. H. Chen, R. Craiu, F. Liang, C. Liu. – Available at: https://www.birs.ca/workshops/2014/14w5086
10. Liang, F. A resampling-based stochastic approximation method for analysis of large geostatistical data [Text] / F. Liang, Y. Cheng, Q. Song, J. Park, P. Yang // Journal of the American Statistical Association. – 2013. – Vol. 108, Issue 501. – P. 325–339 doi: 10.1080/01621459.2012.746061
11. Liang, F. A bootstrap Metropolis–Hastings algorithm for Bayesian analysis of big data. Tech. rep. [Text] / F. Liang, J. Kim. – Texas A & M University, 2013.
12. Liang F. Advanced Markov chain Monte Carlo methods: learning from past samples [Text] / F. Liang, C. Liu, R. J. Carroll. – Wiley, New York, 2010. – 378 p.