جامعة القاهرة
كلية الاقتصاد والعلوم السياسية
قسم الاحصاء

استخدام أساليب المعاينة في تحليل البيانات الضخمة وبيانات البث المباشر

إعداد

عز الدين أحمد محمد فتوح

إشراف

د. ليلى الزيني

الأستاذ بقسم الإحصاء

كلية الاقتصاد والعلوم السياسية

جامعة القاهرة

2025

مدخل إلى البيانات الضخمة وبيانات البث المباشر وأساليب معالجتها باستخدام المعاينة

في عصرنا الحالي، زادت وتنوعت مصادر البيانات بشكل كبير. كانت المصادر قديمًا تتمثل بالأساس في التعدادات، والسجلات الإدارية، والاستبيانات، بالإضافة إلى الكتب والصُحُف. حديثًا، أضيفت إليهم مصادر أخرى -بفضل التطور التكنولوجي- مثل وسائل التواصل الإجتماعي ومستشعرات "إنترنت الأشياء" (Internet of Things)، وهو ما أدى بدوره لزيادة كبيرة في حجم وتدفق البيانات، فيما يعرف بـ"البيانات الضخمة" (Big Data) وبيانات البث المباشر (Streaming Data).

ما هي **البيانات الضخمة (Big Data)؟**

تنوع كمي ونوعي ضخم من البيانات؛ نتيجة تسارعها وإتاحتها آليًا، ومن ثَم تتطلب معالجة واسعة، والتنقيب عنها بأدوات تكنولوجية غير تقليدية.

تتميز البيانات الضخمة بـثلاث خواص رئيسية يُشار إليها بـ3Vs، وهم:

1. **الحجم (Volume)**: كمية هائلة من البيانات، عادةً ما تكون بالملايين أو المليارات.
2. **السرعة (Velocity)**: تولُّد وتدفُّق البيانات بسرعة عالية.
3. **التنوع (Variety)**: البيانات تأتي بأشكال مختلفة مثل النصوص، الصور، ومقاطع الفيديو.

وفي بعض الحالات، يُضاف خصائص أخرى مثل:

- **الصحة/الدقة (Veracity)**: مدى موثوقية البيانات.
- **القيمة (Value)**: مدى الفائدة التي يمكن استخراجها منها.

ماهي **بيانات البث المباشر (Streaming Data)؟**

هي بيانات تُنتَج بحجم كبير وبشكل مستمر وتدريجي، بهدف معالجتها بزمن تأخير منخفض (Low Latency). تمتلك المؤسسات الآلاف من مصادر البيانات التي تُرسل عادةً رسائل أو سجلات أو بيانات يتراوح حجمها من بضع بايتات إلى عدة ميغابايتات في نفس الوقت.

تشمل البيانات المتدفقة بيانات المواقع، وبيانات الأحداث، وبيانات المستشعرات، والتي تعتمد عليها الشركات في التحليلات اللحظية والحصول على رؤية مباشرة لمختلف جوانب أعمالها.

من التعريفات السابقة، يتضح أنه من الصعب استخدام طرق تحليل البيانات التقليدية في التعامل مع ذلك الكم الكبير من البيانات، خاصةً في حالة بيانات البث المباشر التي لا يتم الاحتفاظ بها لتحليلها لاحقًا، لكن يتم تحليلها بشكل لحظي. لذا، قد يكون استخدام أساليب المعاينة مفيدًا لإدارة ومعالجة البيانات بأسلوب كفء.

أساليب المعاينة التي تصلح مع البيانات الضخمة وبيانات البث المباشرة عديدة، نكتفي بذكر ثلاثة منها، وهم:

1. **المعاينة العشوائية البسيطة (Simple Random Sampling):** هي تقنية أخذ عينات أساسية تستخدم في الإحصاء لضمان حصول كل فرد من أفراد المجتمع على فرصة متساوية للاختيار. تعد هذه الطريقة بالغة الأهمية للحصول على بيانات غير متحيزة وتمثيلية، وهو أمر ضروري للتحليل الإحصائي الدقيق.
2. **المعاينة العشوائية الطبقية (Stratified Random Sampling):** هي طريقة أخذ عينات تتضمن تقسيم السكان إلى مجموعات فرعية مميزة، تُعرف باسم الطبقات، والتي تشترك في خصائص مماثلة. تضمن هذه التقنية تمثيل كل مجموعة فرعية بشكل كافٍ في العينة، وبالتالي تعزيز دقة وموثوقية التحليل الإحصائي.
3. **المعاينة باستخدام الخزان - المعاينة الاحتياطية (Reservoir Sampling)**: هي طريقة معايَنة عشوائية قائمة على الحصص، تُستخدم للحصول على حجم عينة معين عندما لا تعرف حجم المجتمع (أي عندما تتعامل مع تدفق بيانات ذو طول غير معروف). يمكن أيضًا استخدامها لإنشاء عينة من مجموعات بيانات كبيرة جدًا.
يُطلق عليها "المعاينة باستخدام الخزان" لأن العناصر المحددة يتم وضعها في خزان (أي مجموعة احتجاز). مع استقبال كل عنصر من تدفق البيانات، يتم تحديث الخوارزمية بشكل ديناميكي. يمكن تحديث الخزان مع الإرجاع أو بدون إرجاع.

تحليل بيانات وسائل التواصل الاجتماعي

في العصر الرقمي الحديث، أصبحت وسائل التواصل الاجتماعي مصدرًا ضخمًا ومتجددًا للبيانات. الصناعات التي تعتمد على هذه المنصات – مثل التسويق، والعلاقات العامة، والإعلانات – تستفيد من تحليل هذه البيانات لفهم سلوك المستخدمين، وقياس فعالية الحملات، واستشراف الاتجاهات المستقبلية.

نظرًا للكميات الهائلة من البيانات المتولدة باستمرار، لا يكون من العملي أو الممكن تحليل كل البيانات. وهنا يأتي دور أساليب المعاينة لتوفير تمثيل مناسب للبيانات دون الحاجة لتحليلها بالكامل.

- **المعاينة العشوائية البسيطة:** تُستخدم لاختيار عينة عشوائية من المنشورات، التعليقات، أو التفاعلات على منصات مثل فيسبوك، تويتر، وإنستجرام. تساعد هذه الطريقة في إجراء تحليلات مثل تحليل المشاعر (Sentiment Analysis) تجاه منتج أو حملة معينة، أو لتحديد الأنماط والاتجاهات العامة في آراء المستخدمين.
- **المعاينة باستخدام الخزان:** نظرًا لتدفق البيانات الحي والمستمر من وسائل التواصل الاجتماعي، تُعد المعاينة باستخدام الخزان مثالية لاختيار عينة ممثلة من بيانات البث المباشر، مثل تغريدات تويتر أو منشورات فيسبوك. يتم ذلك دون الحاجة لمعرفة الحجم الكامل للبيانات، وتُستخدم العينات الناتجة في تحليل المشاعر اللحظي، أو رصد الأحداث الجارية، أو الكشف المبكر عن الأزمات (مثل شكاوى العملاء أو ردود الفعل السلبية المفاجئة).

مثال واقعي:
شركة متخصصة في تحليل البيانات تعتمد على **المعاينة العشوائية البسيطة** لاختيار عينة من التغريدات تتعلق بإطلاق منتج جديد، بهدف قياس مدى تفاعل الجمهور وتحليل انطباعاتهم. في حالة الحملات المباشرة أو الأحداث الجارية، تستخدم الشركة **المعاينة باستخدام الخزان** لتجميع تغريدات في الوقت الحقيقي، مما يتيح لها تقديم تقارير فورية حول أداء الحملة أو ردة فعل الجمهور.

أمثلة تطبيقية باستخدام لغة بايثون (Python)

المعاينة العشوائية البسيطة (Simple Random Sampling):

```
import dask.dataframe as dd

# Read the large CSV file in parallel with Dask
df = dd.read_csv('large_data.csv')

# Simple Random Sampling (10% of the data)
sample = df.sample(frac=0.1, random_state=42)

# Trigger computation (Dask operations are lazy, so you need to call
compute())
sample_result = sample.compute()

print(sample_result)
```

في المثال السابق، نستخدم حزمة `dask` لقراءة البيانات على بايثون على شكل إطار من البيانات (DataFrame)، ثم نقوم بأخذ عينة عشوائية بسيطة قدرها 10% من إجمالي البيانات، ثم نستخدم طريقة `compute()`. لتنفيذ أسلوب المعاينة [بما أن حجم البيانات كبير، يستخدم `dask` أسلوب "العمليات الكسولة" (Lazy Operations)، وهو ما يعني عدم تنفيذ العمليات حتى إعطاء الأمر بشكل مباشر، ما يساعد على إدارة الموارد، خاصةً الذاكرة العشوائية، بشكل أكثر كفاءة في التعامل مع البيانات الضخمة عن الحزم الأخرى، مثل `pandas`]. أخيرًا، تُعرض النتائج باستخدام أمر `print()`.

المعاينة العشوائية الطبقية (Stratified Random Sampling):

```
import dask.dataframe as dd

# Read the large CSV file in parallel with Dask
df = dd.read_csv('large_data.csv')

# Stratified Random Sampling (10% of each group in 'target_column')
Stratified_sample = df.groupby('target_column').apply(lambda x:
x.sample(frac=0.1, random_state=42), meta=df)

# Trigger computation
stratified_sample_result = stratified_sample.compute()

print(stratified_sample_result)
```

في المثال السابق، يتم قراءة البيانات على شكل إطار بإستخدام `dask` مرة أخرى، لكن هذه المرة نقوم بتقسيم البيانات إلى طبقات بناءًا على المتغير محل الدراسة باستخدام `df.groupby()`، وباستخدام طريقة `.apply()` مع دالة `lambda` –لكتابة دالة تستخدم مرة واحدة فقط، على نقيض استخدام `def()`– يتم أخذ عينة قدرها 10% من كل طبقة، ثم استكمال بقية العمليات مثل المثال السابق.

المعاينة باستخدام الخزان - المعاينة الاحتياطية (Reservoir Sampling):

```
from streaming import Reservoir

# Create a Reservoir for size 5 (adjust size based on your use case)
reservoir = Reservoir(5)

# Simulate streaming data: Imagine this as incoming data from a stream
for i in range(100):  # Simulating 100 pieces of incoming data
    reservoir.add(i)  # Add each new data point to the reservoir

# Get the final sampled reservoir
sampled_data = reservoir.sample()

print(f"Sampled data from the stream:{sampled_data}")
```

في المثال السابق، استُخدمت حزمة `streaming` لاستخدام المعاينة الاحتياطية. أُنشئت خمسة خزانات حتى تحتوي تدفق البيانات، ثم تمت محاكاة بيانات البث المباشر عن طريق (`for` loop) وإضافة البيانات المتدفقة إلى الخزانات. أخيرًا، تُعرض العينة المسحوبة من التدفق.

المراجع

1. شعبان، الحسن. (2022). البيانات الضخمة: ماهيتها وأهميتها وعناصرها. المجلة العربية الدولية لإدارة المعرفة. *1(2). 114*. تاريخ الدخول: [20 أبريل 2025]. البيانات الضخمة : ماهيتها وأهميتها وعناصرها
2. Amazon Web Services. (بدون تاريخ). ما هي بيانات البث المباشر؟ *Amazon Web Services*. تاريخ الدخول: [20 أبريل 2025]. What is Streaming Data?
3. إحصائيات بسهولة. (بدون تاريخ). ما هي العينة العشوائية البسيطة؟ شرح مفصل. إحصائيات بسهولة. تاريخ الدخول: [20 أبريل 2025]. ما هي العينة العشوائية البسيطة: شرح مفصل
4. إحصائيات بسهولة. (بدون تاريخ). ما هو: شرح العينة العشوائية الطبقية. إحصائيات بسهولة. تاريخ الدخول: [20 أبريل 2025]. ما هو: شرح العينة العشوائية الطبقية
5. StatisticsHowTo. (بدون تاريخ). المعاينة باستخدام خزان العينات. *StatisticsHowTo*. تاريخ الدخول: [20 أبريل 2025]. Reservoir Sampling - Statistics How To