الجمهورية الشعبية الديمقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي 08 ماي 5491 بسيدي بلعباس
École Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbès

THESIS

To obtain the diploma of **Master**
Field: **Computer Science**
Specialty: **Information Systems and Web (SIW)**

Theme

In-Database Machine Learning (In-DB ML)

Presented by:
YOUSFI Ikram Oumelkheir

Submission Date: **September 24, 2024**
In front of the jury composed of:

Ms. BENCHERIF Khayra	President
Ms. KLOUCHE Badia	Examiner
Ms. ELOUALI Nadia	Supervisor
Ms. AMER-YAHIA Sihem	Co-Supervisor
Mr. BOUAROUR Nassim	Co-Supervisor

Academic Year : 2023/2024

Abstract

In-Database Machine Learning (In-DB ML) or Database Machine Learning (DBML) is a young, key area of research that embeds Machine Learning (ML) right into Database Management Systems (DBMS). This approach puts to use the intrinsic strengths of databases—data integrity, security, and concurrency control—while making data processing efficient and performing analytics close to the source. Basically, that would entail a reduction in data movement and other complexities or overhead that occur in traditional machine learning pipelines, where data must be exported to external environments in order to be processed.

In-DB ML run machine learning algorithms natively inside the database environment by using stored procedures or User-Defined Functions (UDFs) that can use SQL and other languages native to the database. This not only enhances performance by leveraging optimized database operations but remains within database access control mechanisms for improved security.

The practical advantages of In-DB ML include lower latency in insight generation, better data governance, and streamlined architecture for data analytics workflows. More importantly, this integration makes scalable machine learning operations over large datasets common in enterprise systems possible, thus making it relevant for use cases in finance, healthcare, and retail—industries for which data privacy and fast data processing are essential.

Research in this field is primarily focused on the development of new algorithms that are friendly to in-database execution, with extensions to database architectures that allow for rich analytic functionality and ease of use for data scientists and database administrators through better tooling and interfaces. As this field continues to grow and evolve, so will the potential significance it may have in the future with respect to a far greater, broader diffusion of machine learning, eventually making it much more accessible and efficient for a myriad of applications.

Keywords— In-Database Machine learning, Database management system, DBMS, ML

Résumé

L'apprentissage automatique dans les bases de données (In-DB ML) ou apprentissage automatique basé sur les bases de données (DBML) est un domaine de recherche jeune et clé qui intègre l'apprentissage automatique (ML) directement dans les systèmes de gestion de bases de données (SGBD). Cette approche exploite les forces intrinsèques des bases de données—intégrité des données, sécurité et contrôle de la concurrence—tout en rendant le traitement des données plus efficace et en effectuant des analyses près de la source. Concrètement, cela permet de réduire les mouvements de données et les autres complexités ou surcharges qui surviennent dans les pipelines traditionnels d'apprentissage automatique, où les données doivent être exportées vers des environnements externes pour être traitées.

L'In-DB ML exécute des algorithmes d'apprentissage automatique de manière native dans l'environnement de la base de données en utilisant des procédures stockées ou des fonctions définies par l'utilisateur (UDF) qui peuvent utiliser le SQL et d'autres langages propres à la base de données. Cela améliore non seulement les performances en exploitant des opérations optimisées de la base de données, mais reste également sous les mécanismes de contrôle d'accès de la base de données pour une sécurité renforcée.

Les avantages pratiques de l'In-DB ML incluent une latence réduite dans la génération d'informations, une meilleure gouvernance des données et une architecture simplifiée pour les flux de travail analytiques. Plus important encore, cette intégration rend possibles des opérations d'apprentissage automatique à grande échelle sur de grands ensembles de données, ce qui est fréquent dans les systèmes d'entreprise. Cela en fait une solution pertinente pour des cas d'utilisation dans des secteurs tels que la finance, la santé et la vente au détail, où la confidentialité des données et le traitement rapide des informations sont essentiels.

La recherche dans ce domaine se concentre principalement sur le développement de nouveaux algorithmes adaptés à l'exécution au sein des bases de données, avec des extensions des architectures de bases de données permettant des fonctionnalités analytiques riches et une utilisation simplifiée pour les data scientists et les administrateurs de bases de données grâce à de meilleurs outils et interfaces. Alors que ce domaine continue de croître et d'évoluer, son potentiel pourrait devenir de plus en plus significatif à l'avenir, permettant une diffusion beaucoup plus large de l'apprentissage automatique, le rendant ainsi plus accessible et plus efficace pour une multitude d'applications.

Mots clés— In-Database Machine learning, Database management system, DBMS, ML.

الملخص

تعلم الآلة داخل قواعد البيانات In-DB ML أو تعلم الآلة المعتمد على قواعد البيانات DBML هو مجال بحثي حديث ومهم يدمج تعلم الآلة ML مباشرة في أنظمة إدارة قواعد البيانات DBMS. تستفيد هذه المقاربة من القوة الكامنة في قواعد البيانات—سلامة البيانات، الأمان، والتحكم في التزامن—مع جعل معالجة البيانات أكثر كفاءة وإجراء التحليلات بالقرب من مصدر البيانات. بشكل أساسي، يساهم ذلك في تقليل حركة البيانات والتعقيدات أو الأعباء الإضافية التي تحدث في خطوط الأنابيب التقليدية لتعلم الآلة، حيث يجب تصدير البيانات إلى بيئات خارجية من أجل معالجتها.

يقوم تعلم الآلة داخل قواعد البيانات بتشغيل خوارزميات تعلم الآلة بشكل أصلي داخل بيئة قاعدة البيانات باستخدام الإجراءات المخزنة أو الدوال المعرفة من قبل المستخدم (UDFs) التي يمكنها استخدام SQL وغيرها من اللغات الخاصة بقاعدة البيانات. لا يعزز ذلك الأداء من خلال الاستفادة من العمليات المحسنة لقاعدة البيانات فحسب، بل يبقى أيضًا ضمن آليات التحكم في الوصول الخاصة بقاعدة البيانات لتحسين الأمان.

تشمل الفوائد العملية لتعلم الآلة داخل قواعد البيانات انخفاض زمن الوصول في توليد الرؤى، وتحسين حوكمة البيانات، وتبسيط البنية التحتية لتدفقات العمل التحليلية. والأهم من ذلك، أن هذا التكامل يجعل عمليات تعلم الآلة القابلة للتوسع على مجموعات بيانات كبيرة ممكنة في الأنظمة المؤسسية، مما يجعلها ذات صلة بحالات الاستخدام في القطاعات مثل المالية والرعاية الصحية والتجزئة، حيث تكون خصوصية البيانات وسرعة معالجة البيانات ضرورية.

تركز الأبحاث في هذا المجال بشكل رئيسي على تطوير خوارزميات جديدة ملائمة للتنفيذ داخل قواعد البيانات، مع توسيعات في بنى قواعد البيانات التي تسمح بوظائف تحليلية غنية وسهولة الاستخدام للعاملين في مجال البيانات ومديري قواعد البيانات من خلال تحسين الأدوات وواجهات الاستخدام. ومع استمرار نمو هذا المجال وتطوره، فإن أهميته المحتملة قد تزداد في المستقبل مع نشر أوسع بكثير لتعلم الآلة، مما يجعله أكثر سهولة وكفاءة في مجموعة متنوعة من التطبيقات.

الكلمات المفتاحية: تعلم الآلة داخل قاعدة البيانات, نظام إدارة قواعد البيانات, ML, In-DB ML.