Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/77/2024MA0059

Implementing intelligent predictive models for patient disease risk in cloud data warehousing

Yadong Shi[1a,*], Jiaqiang Yuan[1b], Peiyuan Yang[2],Yufu Wang[3], Zhou Chen[4]

[1a]Computer Science, Fudan University, Shanghai, China
[1b]Information Studies, Trine University, Phoenix, USA
[2]Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA
[3]Computer Science & Engineering, Santa Clara University, Santa Clara, CA, USA
[4]Computer Science, Zhejiang University, Hangzhou, China

[*]Corresponding author. Email: lizengyi.zy@bytedance.com

Abstract. A data warehouse, which stores data after it has been extracted, processed, and organized into files and folders, is a cloud data warehousing solution for storing structured data from one or more sources. When data is stored in an organized format within files and folders, it becomes easily accessible and facilitates strategic, data-driven decision-making. This paper introduces the importance of cloud data warehousing in medical information management and discusses its application in a disease prediction model. By analyzing medical data and constructing predictive models, it can assist medical decision-makers in taking timely actions to enhance the quality and efficiency of medical services, providing patients with more personalized treatment plans.

Keywords: Cloud data warehouse, Disease risk prediction, Artificial intelligence, Machine learning model, Medical field.

1. Introduction

In the medical field, the application of cloud data warehouse is gradually becoming an important part of medical information management. With the rapid growth of medical data, traditional data storage and management methods have been unable to meet the needs of medical institutions and research institutions for data processing and analysis. [1]The cloud data warehouse is one of the preferred solutions in the medical field due to its flexibility, cost effectiveness and security. The flexibility and scalability of cloud data warehouses brings unprecedented convenience to healthcare organizations. The growth rate of medical data is often uncertain, and there can sometimes be sudden spikes. Traditional data warehouses often need to plan hardware equipment and storage space in advance, while cloud data warehouses can dynamically adjust resources according to actual needs, ensuring that medical institutions can process and store massive data in a timely manner without worrying about the lack or waste of resources.

In addition, cloud data warehousing can bring significant cost benefits. Medical institutions usually need to invest a lot of money to purchase and maintain traditional data warehouse equipment, while cloud data warehouses can realize on-demand payment, greatly reducing the operating costs of medical

DOI: 10.54254/2755-2721/77/2024MA0059

institutions.With the rapid growth of medical data and the increase in the degree of informatization, cloud data warehouse as a [2-3]flexible, cost-effective, secure and reliable data management and analysis solution has brought many advantages to medical institutions and research institutions. Through the cloud data warehouse, medical decision-makers can obtain the latest medical data in a more timely manner, and make predictions and decisions based on data analysis, thereby improving the quality and efficiency of medical services and providing patients with more personalized treatment plans.

2. Related work

2.1. Cloud data warehouse

The traditional warehouse is characterized by the integration of calculation and storage, that is, the calculation and storage are on the same machine. The cpu, memory, disk data partition, the entire data is broken up, each node is responsible for a part of the data, the data between each node is not related.

In this architecture, computing and storage must be combined. [4]Although the traditional warehouse can be expanded, it takes a long time. For example, after node4 is added in the following figure, data balancing operations may take a long time. In addition, new nodes cannot provide external services during capacity expansion, and resource charges have started, resulting in resource waste.

The modern cloud data warehouse is an analytical database, usually a relational database, which is created by two or more data sources, usually can store more than petabytes of historical data, and then rely on a large number of computing and memory resources to run complex query operations, and finally generate data reports. In addition, the data warehouse is the only path to direct data sources for business intelligence (BI) systems and machine learning.

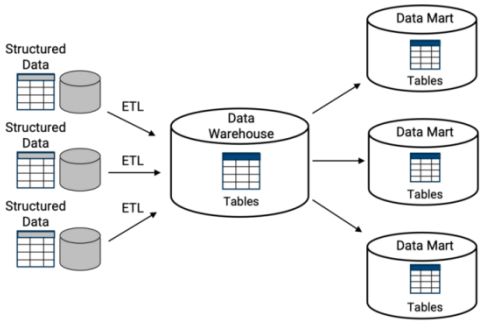

Figure 1. Warehouse structure diagram

As can see from the warehouse architecture diagram in Figure 1, the Databend is divided into four modules from top to bottom:

1. Data access: Data access supports [5]SQL, such as Mysql, Clickhouse, etc., which can easily use existing ecological connections to enter data. Databend abstracts a lot from this layer in the design of the architecture. Easy access to data through existing ecology.

2.Meta Service layer: Metadata and meta information are stored in this layer, which is equivalent to the brain of the Databend. This is an elastic multi-tenant isolated service.

3. Computing layer: Each computing node is independent and complete for an sql analysis. The compute node will parse the SQL, make logical and physical execution plans, and execute. Each compute node has its own Cache, index, and data layer access structure. These nodes can easily form a cluster, and users can define the services provided by the cluster.

4. Storage layer: The storage layer mainly utilizes cloud storage, local DFS, S3, Azure Blob, and other block storage solutions, forming the foundation for Databend based on shared storage.

2.2. Advantages of Cloud Data warehouse

First, a cloud data warehouse has a default schema, which means that it enforces data structures before or during data ingestion. This feature ensures that data analysis is easier to start, as the structured nature

DOI: 10.54254/2755-2721/77/2024MA0059

of the data makes the analysis work more efficient and accurate. Second, cloud data warehouses make data governance easier. This means that database architects, [6]SQL developers, and data analysts can all adapt more smoothly to the cloud data warehouse environment. Cloud data warehouses have a relatively low learning curve and often offer a quick option to migrate from a private architecture to the cloud. This enables organizations to take advantage of cloud data warehouses more quickly and accelerate the data analysis and decision making process.

In addition, a cloud data warehouse has many built-in features, such as a database sequence of proxy keys for data warehouse star or snowflake data modeling. These features provide a more convenient way to model and manage data, thereby improving the efficiency and reliability of the data warehouse. Finally, cloud data warehouses support [7]ACID (Atomicity, Consistency, Isolation, Persistence) database properties, which are critical for analytical reporting. Compared with the data lake architecture, the cloud data warehouse can more easily achieve multi-version control of data, enforce constraints and other functions, thus improving the quality and credibility of data.

2.3. Cloud data warehousing and healthcare

Data centers in the healthcare industry need to be built with components such as data exchange, security protection, databases, storage, server clustering, and disaster backup/recovery in mind. Cisco Healthcare Cloud Data Center solutions provide a unified solution for the healthcare industry, including the following three main aspects[8]:

First, unified data center construction. The solution is designed to build a sustainable, next-generation data center architecture for cloud computing. By pooling IT resources in a data center and providing resources on demand using an intelligent service scheduling mechanism, resource consumption is reduced and resource utilization efficiency is improved. Cisco provides end-to-end storage backup and disaster recovery solutions. Through optical fiber storage switches and optical transmission solutions, Cisco works with storage partners to provide efficient and reliable storage solutions to ensure secure backup and disaster recovery of medical data.

In order to ensure the security of system data, Cisco provides data center solutions with self-defense security system, effectively eliminate security risks, strengthen the data security level of the medical system, and ensure the normal operation of the medical system and the safe interaction of patient information.At the same time, the cloud data warehouse also provides medical institutions with data security, disaster recovery and other aspects of support to ensure the safety and reliability of medical data.

2.4. Disease intelligent prediction model

In recent years, artificial intelligence research in the field of disease prediction is hot, more and more researchers have applied artificial intelligence methods in disease prediction research, and achieved many results. In 2018, Ocampo et al. [9] established a model for lung cancer diagnosis based on Convolutional Neural Networks (CNN) [10], and the AUC of the model for lung cancer recognition reached 0.97. In 2018, Golatkar et al.On the heart disease data set of UCI, the accuracy of the two methods reached 84.5% and 85.1% respectively. In 2020, Lyngdoh et al.use a variety of machine learning methods to predict diabetes, among which K-Nearest Neighbor (KNN)has the highest accuracy, reaching 76%. Other machine learning algorithms, such as naive Bayes and support vector machines, have also achieved more than 70% accuracy.

These models are able to accurately predict an individual's future risk of disease based on their characteristics and historical data. For example, for chronic diseases such as cardiovascular disease and diabetes, the predictive model can analyze the patient's lifestyle habits, genetic factors, health indicators and other aspects of information, to provide medical institutions and doctors with timely warning and intervention measures to help patients reduce the risk of disease, improve lifestyle, so as to improve the overall health level. The application of intelligent prediction model not only contributes to individual health management, but also provides scientific basis for rational allocation of medical resources and

DOI: 10.54254/2755-2721/77/2024MA0059

formulation of disease prevention strategies, and promotes the development of health care and the improvement of social health.

3. Methodology

In the monitoring and management of patients with chronic hepatitis B (CHB), it is important to predict the risk of hepatocellular carcinoma (HCC). However, the baseline parameters of various HCC risk prediction systems are limited in number and accuracy.By integrating clinical data from medical institutions into cloud data warehouses, machine learning algorithms can more efficiently obtain the data they need for model training and optimization.This provides a basis for the establishment and real-time update of [11]HCC risk prediction model, and further improves the accuracy and practicability of the prediction model.

In this context, with the support of cloud data warehouse, medical institutions can more efficiently use machine learning algorithms to predict HCC risk, provide personalized monitoring and management services for patients, and improve the survival rate and quality of life of patients.

3.1. Prediction model

1. Random forest model

Random forest is an ensemble learning method, which is composed of multiple decision trees. Each decision tree is trained by randomly sampling the training data to form different sub-data sets. At the time of building each decision tree, the random forest improves the diversity of the model by making a random selection of features.

For each decision tree, the random forest uses the following two types of randomness to build the tree: (1) Randomly select feature subsets: On the nodes of each decision tree, a feature subset is randomly selected for partitioning, so that only a part of the features are considered in each decision tree. (2) Random division of nodes: For the division of each node, a division criterion is randomly selected to determine the best partition.

2.XGBoost (Gradient enhanced decision tree)

XGBoost, which stands for eXtreme Gradient Boosting, is an ensemble learning algorithm based on Gradient Boosting Decision Tree. Based on GBDT, the regularization term and second derivative information are introduced to improve the performance and generalization ability of the model.

The core idea of the [12]XGBoost model is to combine multiple weak classifiers (decision trees) into one strong classifier. Each decision tree is trained on the basis of the residual of the previous tree, and the residual is gradually reduced by iteratively optimizing the loss function. At the same time, the model reduces the overfitting risk by controlling the complexity and regularization terms of the tree.

3.Logistic regression model

Logistic regression model is a kind of probability model. It takes the probability P of an event occurring or not as the dependent variable and the factors affecting P as the independent variable to establish a regression model, and analyzes the relationship between the probability of an event occurring and the independent variable. It is a kind of nonlinear regression model.

3.2. Experimental design

In this study, three machine learning models were trained to predict the risk of liver cancer. They looked at various factors like age, sex, platelet count, and more. To make up for the low number of liver cancer cases in the data, they used a technique called [13]SMOTE to balance things out. The final risk score, which ranged from 0 to 1, combined the predictions of all three models. They used a cutoff of 0.5 to determine if someone was at risk of liver cancer based on their score.

3.3. Statistical analysis of data

Data were presented as mean ± standard deviation or as numbers and percentages. Differences between continuous variables were compared using the Student t-test or Mann-Whitney U test, while categorical variables were compared using the Chi-square test or Fisher exact test. Cumulative risk of liver cancer

DOI: 10.54254/2755-2721/77/2024MA0059

was calculated with Kaplan-Meier method and analyzed using logarithmic rank test. Cox regression analysis assessed the association between liver cancer risk and each variable, providing hazard ratios (HR) and 95% confidence intervals (CI). Time-dependent receiver operating characteristic (ROC) curves were constructed for each model to predict liver cancer development, with the area under the ROC curve (AUROC) calculated. Statistical analysis was performed using SAS and R software, with significance set at $p < 0.05$.

3.4. Data result

In the group of patients studied (n=960), 69 individuals (7.2%) developed liver cancer during a median follow-up period of approximately 59.3 months. The cumulative incidence rates were 1.1% at 1 year, 5.1% at 3 years, and 8.1% at 5 years. Factors such as age, male gender, presence of cirrhosis, HBV DNA level, ALT level, serum bilirubin level, platelet count, serum albumin level, PT INR, AFP level, and HBeAg positivity were significant predictors of liver cancer in initial analysis. Further analysis revealed that older age, male gender, presence of cirrhosis, lower ALT levels, lower platelet counts, and higher PT INR were associated with a higher risk of liver cancer in this patient cohort.

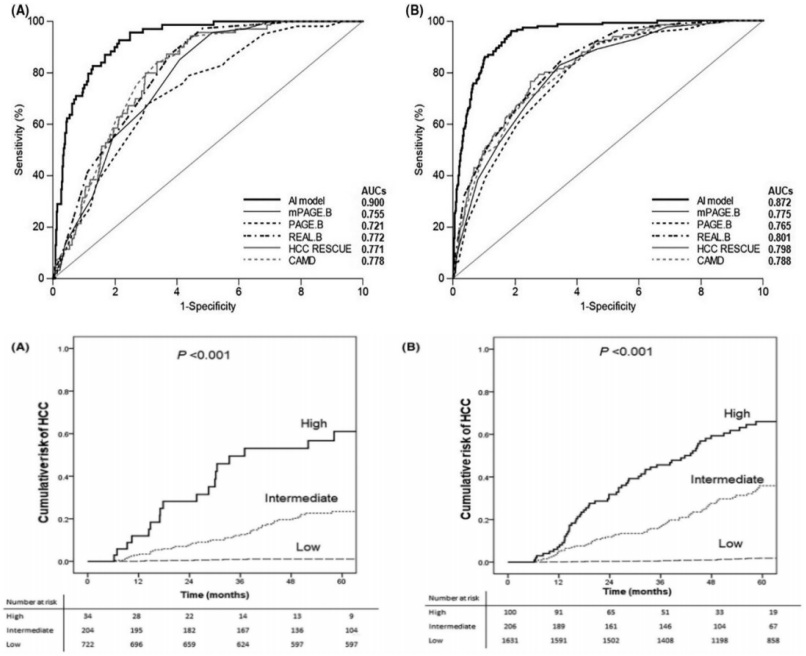

Figure 2. Training results curve

A machine learning (ML) model, incorporating random forest, XGBoost, and logistic regression, predicted the likelihood of developing liver cancer with a range of 0-1. Compared to existing models (mPAGE-B and CAMD), the ML model demonstrated superior performance, with an AUC value of 0.930 (95% CI 0.904 -- 0.956). Stratifying patients based on this ML-based risk prediction model revealed significantly different cumulative incidence rates among risk groups (<0.3, low risk; 0.3-0.5, medium risk; >0.5, high risk). For instance, the 1-year, 3-year, and 5-year cumulative incidence rates in the low-risk group were 0.1%, 0.8%, and 1.1%, respectively. Conversely, in the high-risk group, the rates were substantially higher, with 14.8%, 32.8%, and 54.9% observed at the same intervals. In the validation cohort, one patient classified in the very low-risk group (cutoff = 0.1) developed liver cancer.

3.5. Experimental discussion

The study's strengths lie in its inclusion of a diverse range of patients with chronic hepatitis B (CHB) undergoing antiviral therapy, ensuring broad applicability. Leveraging cloud data warehouses facilitates centralized storage and access to dispersed patient data, enhancing research convenience. A large

DOI: 10.54254/2755-2721/77/2024MA0059

training cohort ensures high prediction accuracy, validating the model's reliability. The machine learning (ML) model's ability to stratify patients into low, medium, and high-risk groups enables tailored monitoring strategies. [14]ML algorithms excel in handling large-scale data and complex relationships, facilitated by cloud data warehouses' storage and processing capabilities. Integration with electronic medical record systems enables practical implementation of automated calculators for routine clinical use. Overall, this study combines extensive patient data and advanced ML techniques to offer a more precise HCC risk assessment method for CHB patients, supported by cloud data warehouses for efficient data management and model optimization, driving personalized medical strategies.

4. Conclusion

This paper introduces the application of cloud data warehouse in medical field, and discusses its role in disease intelligent prediction model. With the rapid growth of medical data, traditional data storage and management methods have been unable to meet the needs of medical institutions and research institutions for data processing and analysis. Cloud data warehouses are the solution of choice in the healthcare sector due to their flexibility, cost-effectiveness and security. The flexibility and scalability of cloud data warehouses bring unprecedented convenience to healthcare organizations, enabling them to dynamically adjust resources according to actual needs, ensuring that healthcare organizations can process and store large amounts of data in a timely manner without worrying about lack of resources or waste.

In addition, the article highlights that cloud data warehouses can bring significant cost benefits, reducing operational costs for healthcare organizations. Medical institutions do not need to worry about updating and maintaining hardware devices, but can focus more on data analysis and application development to provide more support for medical research and clinical practice. In terms of medical prediction, the application of cloud data warehouse is also of great significance. By analyzing and mining medical data, healthcare organizations can better understand the pathogenesis of diseases, evaluate the effectiveness of treatment, and even predict the health risks of patients. In the future, with the continuous growth of medical data and the advancement of technology, the application of cloud data warehouse in the medical field will be further expanded. By combining advanced machine learning and artificial intelligence technology, it will be able to establish a more accurate and reliable disease prediction model, provide doctors with more auxiliary decision-making and personalized treatment plans, thereby promoting the intelligent and refined development of medical services, and improve the quality of medical care and the quality of life of patients

References

[1] Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. ACM Sigmod record, 26(1), 65-74.

[2] Damen, J. A., Hooft, L., Schuit, E., Debray, T. P., Collins, G. S., Tzoulaki, I., ... & Moons, K. G. (2016). Prediction models for cardiovascular disease risk in the general population: systematic review. bmj, 353.

[3] Kahn, Michael G., et al. "Migrating a research data warehouse to a public cloud: challenges and opportunities." Journal of the American Medical Informatics Association 29.4 (2022): 592-600.

[4] Dageville, B., Cruanes, T., Zukowski, M., Antonov, V., Avanes, A., Bock, J., ... & Unterbrunner, P. (2016, June). The snowflake elastic data warehouse. In Proceedings of the 2016 International Conference on Management of Data (pp. 215-226).

[5] Gupta, Himanshu. "Selection of views to materialize in a data warehouse." Database Theory—ICDT'97: 6th International Conference Delphi, Greece, January 8–10, 1997 Proceedings 6. Springer Berlin Heidelberg, 1997.

[6] Jostins, Luke, and Jeffrey C. Barrett. "Genetic risk prediction in complex disease." Human molecular genetics 20.R2 (2011): R182-R188.

DOI: 10.54254/2755-2721/77/2024MA0059

[7] Damen, Johanna AAG, et al. "Prediction models for cardiovascular disease risk in the general population: systematic review." bmj 353 (2016).
[8] Anderson, K. M., Odell, P. M., Wilson, P. W., & Kannel, W. B. (1991). Cardiovascular disease risk profiles. American heart journal, 121(1), 293-298.
[9] Vartak, M., Subramanyam, H., Lee, W. E., Viswanathan, S., Husnoo, S., Madden, S., & Zaharia, M. (2016, June). ModelDB: a system for machine learning model management. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics (pp. 1-3).
[10] Polson, Nicholas G., and Vadim O. Sokolov. "Deep learning for short-term traffic flow prediction." Transportation Research Part C: Emerging Technologies 79 (2017): 1-17.
[11] Çetiner, B. Gültekin, Murat Sari, and Oğuz Borat. "A neural network based traffic-flow prediction model." Mathematical and Computational Applications 15.2 (2010): 269-278.
[12] Ariyachandra, T., & Watson, H. J. (2006). Which data warehouse architecture is most successful?. Business intelligence journal, 11(1), 4.
[13] Ma, Haowei. "Automatic positioning system of medical service robot based on binocular vision." 2021 3rd International Symposium on Robotics & Intelligent Manufacturing Technology (ISRIMT). IEEE, 2021.
[14] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349.6245 (2015): 255-260.