Journal Of Data Science
Volume 2, Number 02, 2024, DOI 10.58471/Science.v2i02
ESSN 3025-2792 (Online)
https://ejournal.seaninstitute.or.id/index.php/visualization

Application Of C4.5 Algorithm In Disease Classification

Sipra barutu
Universitas Panca Budi, Medan

Article Info

Keywords:
C4.5 algorithm, disease classification, diagnosis, accuracy, efficiency, decision tree, patient data, new patterns, quality of patient care.

ABSTRACT

In the modern era, information and communication technology (ICT) has a significant impact on the health sector, one of which is through the application of artificial intelligence (AI) for disease diagnosis. The C4.5 algorithm, one of the popular classification algorithms, shows great potential in helping doctors classify diseases more accurately and efficiently. Research shows that the C4.5 algorithm is able to achieve a high level of accuracy in classifying various types of diseases, such as diabetes mellitus, heart disease, and lung disease. Its advantages include ease of interpretation, resistance to data noise, and efficiency. However, its application also has several challenges, such as the availability of quality data, complex interpretation of results, and the potential for overfitting. Nevertheless, the C4.5 algorithm offers great potential to improve the quality of patient diagnosis and care. Further research is needed to overcome the challenges and improve the effectiveness of the C4.5 algorithm in disease classification, such as the development of anti-overfitting techniques, optimal attribute selection methods, and application to more types of diseases. With continued research and development, the C4.5 algorithm can become a valuable tool for doctors and other medical personnel in fighting disease.

Corresponding Author:
Sipra Barutu
Universitas Panca Budi, Medan
barutusipra@gmail.com

INTRODUCTION

In the modern era, information and communication technology (ICT) has developed rapidly and has had a significant impact in various fields, including health. One example is the application of artificial intelligence (AI) in supporting disease diagnosis. The C4.5 algorithm, one of the popular classification algorithms, has shown great potential in helping doctors and other medical personnel classify diseases more accurately and efficiently.

Disease is one of the major health problems faced by humans. Accurate and rapid diagnosis is the key to effective disease management. However, the process of diagnosing a disease is often complex and time-consuming, as doctors need to consider various factors such as symptoms, medical history, and physical examination results. This is where the C4.5 algorithm can play a role.

The C4.5 algorithm is able to build a decision tree based on historical patient data, which can be used to classify new patients with similar symptoms. This decision tree can provide a visual representation of the most important factors in determining a diagnosis, thus helping doctors make more informed decisions.

The application of the C4.5 algorithm in disease classification has several advantages. First, this algorithm can improve the accuracy of diagnosis by considering various factors simultaneously. Second, this algorithm can shorten the diagnosis time by providing clear guidance to doctors. Third, this algorithm can help in identifying new patterns in patient data, which can lead to new diagnostic discoveries.

Despite its many advantages, the application of the C4.5 algorithm in disease classification also has several challenges. One of the challenges is the availability of quality data. The C4.5 algorithm requires sufficient and accurate data to build a reliable decision tree. Another challenge is the interpretation of the classification results. The decision tree generated by the C4.5 algorithm can be complex and difficult to interpret, requiring special expertise to understand it.

Despite the challenges, the application of the C4.5 algorithm in disease classification offers great potential to improve the quality of diagnosis and patient care. With continued research and development, this algorithm can become a valuable tool for doctors and other health care professionals in the fight against disease.

METHOD

The application of the C4.5 algorithm in disease classification is generally carried out through the following steps:

Data collection

The first step is to collect patient data relevant to the disease to be classified. This data can be obtained from various sources, such as electronic medical records, laboratory test results, and patient interviews. The data collected must be complete, accurate, and consistent.

Data Pre-processing

Before the data is used to build a decision tree, data pre-processing needs to be done. Data pre-processing includes:

1. Missing Values Check: Checks whether there are any missing values in the data. Missing values can be deleted, filled with the mean, or imputed using other methods.
2. Categorical Data Handling: Converting categorical data into numeric data. This can be done using one-hot encoding or label encoding techniques.
3. Data Normalization: Normalizing numeric data to have the same scale. This can be done using min-max normalization or z-score normalization techniques.

Attribute Selection

The C4.5 algorithm uses the gain information technique to select the most important attributes in the classification. Gain information measures how much information is obtained by dividing the data based on a particular attribute. Attributes with high gain information will be selected first to build the decision tree.

Making a Decision Tree

The C4.5 algorithm recursively builds a decision tree by dividing the data based on selected attributes. This division is carried out until all data is classified or there are no more attributes that can be used to divide the data.

Model Evaluation

After the decision tree is built, it needs to be evaluated to determine its performance in classifying new data. Evaluation can be done using various metrics, such as accuracy, precision, and recall.

Interpretation of Results

The decision tree generated by the C4.5 algorithm can be interpreted to understand how the algorithm makes classification decisions. This interpretation can help doctors understand the factors that are most important in determining a diagnosis.

Software for Implementing the C4.5 Algorithm

Some software that can be used to apply the C4.5 algorithm in disease classification are:

a. WEKA (Waikato Environment for Knowledge Analysis):Open-source software that provides a variety of machine learning algorithms, including C4.5.
b. RapidMiner:Commercial software that provides a variety of tools for data mining and machine learning, including C4.5.
c. KNIME (Konstanz Information Miner): Open-source software that provides various nodes for data mining and machine learning, including C4.5.

Challenges and Considerations

The application of the C4.5 algorithm in disease classification has several challenges that need to be considered:

a. Data Availability:The C4.5 algorithm requires sufficient and accurate data to build a reliable decision tree.
b. Interpretation of Results:The decision trees generated by the C4.5 algorithm can be complex and difficult to interpret, requiring special expertise to understand them.
c. Overfitting:The C4.5 algorithm can suffer from overfitting, where the decision tree is too complex and does not generalize well to new data.
d. Attribute Selection:Proper attribute selection is essential for building an accurate decision tree. Wrong attribute selection can result in an unperformant decision tree.

Despite some challenges, the application of the C4.5 algorithm in disease classification offers great potential to improve the quality of diagnosis and patient care. With continued research and development, this algorithm can become a valuable tool for doctors and other health care professionals in fighting disease.

RESULTS AND DISCUSSION

Effectiveness of C4.5 Algorithm

Research on the application of the C4.5 algorithm in disease classification has shown promising results. Several studies have shown that the C4.5 algorithm can achieve a high level of accuracy in classifying various types of diseases, such as diabetes mellitus, heart disease, and lung disease.

Application Examples

a. Classification of Diabetes Mellitus:The C4.5 algorithm has been used to classify diabetes mellitus patients based on their symptoms and risk factors. The results show that the C4.5 algorithm can achieve an accuracy rate of up to 89%.
b. Classification of Heart Disease:The C4.5 algorithm has been used to classify heart disease patients based on their electrocardiogram (ECG) results. The results show that the C4.5 algorithm can achieve an accuracy rate of up to 95%.

c. Classification of Lung Diseases:The C4.5 algorithm has been used to classify lung disease patients based on their chest X-ray results. The results show that the C4.5 algorithm can achieve an accuracy rate of up to 87%.

Advantages of C4.5 Algorithm

The C4.5 algorithm has several advantages compared to other classification algorithms, including:

a. Easy to Interpret:The decision tree generated by the C4.5 algorithm is easy to interpret, allowing clinicians to understand how the algorithm makes classification decisions.
b. Robust against Data Noise:The C4.5 algorithm is quite robust against data noise, so it can work well even if the data used is not perfect.
c. Efficient:The C4.5 algorithm is efficient in terms of time and computation, so it can be used to process large amounts of data.

Challenges and Considerations

Despite having many advantages, the application of the C4.5 algorithm in disease classification also has several challenges, such as:

a. Data Availability:The C4.5 algorithm requires sufficient and accurate data to build a reliable decision tree.
b. Interpretation of Results:The decision trees generated by the C4.5 algorithm can be complex and difficult to interpret, requiring special expertise to understand them.
c. Overfitting:The C4.5 algorithm can suffer from overfitting, where the decision tree is too complex and does not generalize well to new data.
d. Attribute Selection:Proper attribute selection is essential for building an accurate decision tree. Wrong attribute selection can result in an unperformant decision tree.

CONCLUSION

The application of the C4.5 algorithm in disease classification offers great potential to improve the quality of diagnosis and patient care. With continued research and development, this algorithm can become a valuable tool for doctors and other medical personnel in fighting disease. Much research still needs to be done to improve the effectiveness of the C4.5 algorithm in disease classification. Some of the research that can be done are: Developing techniques to overcome overfitting. Developing methods to select attributes more optimally. Applying the C4.5 algorithm to more types of diseases.

REFERENCE

Sepharni, A., Hendrawan, I. E., & Rozikin, C. (2022). Klasifikasi Penyakit Jantung dengan Menggunakan Algoritma C4. 5. *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)*, *7*(2), 117-126.

Ridwan, A. (2022). Penerapan Algoritma C4. 5 Untuk Klasifikasi Penyakit Diabetes Mellitus. *Jurnal Bisnis Digital Dan Sistem Informasi*, *3*(2), 41-48.

Rofiani, R., Oktaviani, L., Vernanda, D., & Hendriawan, T. (2024). Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4. 5. *Jurnal Tekno Kompak*, *18*(1), 126-139.

https://en.wikipedia.org/wiki/C4.5_algorithm

Sofyan, F. M. A., Voutama, A., & Umaidah, Y. (2023). Penerapan Algoritma C4. 5 Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapidminer. JATI (Jurnal Mahasiswa Teknik Informatika), 7(2), 1409-1415.

Purwanto, A., Primajaya, A., & Voutama, A. (2020). Penerapan Algoritma C4. 5 Dalam Prediksi Potensi Tingkat Kasus Pneumonia Di Kabupaten Karawang. *JUSTIN (Jurnal Sistem dan Teknologi Informasi)*, *8*(4), 390-396.

Rohman, A., Suhartono, V., & Supriyanto, C. (2017). Penerapan algoritma c4. 5 berbasis adaboost untuk prediksi penyakit jantung. *Jurnal Cyberku*, *13*(1), 2-2.