UNCOVERING DISEASE PATTERNS: CLUSTERING AND ASSOCIATION RULE MINING OF PATIENT DATA IN ELECTRONIC HEALTH RECORD

Nguyen Ngoc Nhan, Dao Van Tuyet, Truong Hai Bang

Military Hospital 7A, The Saigon International University

daovantuyet@siu.edu.vn, truonghaibang@siu.edu.vn,

ABSTRACT: *In the medical field, understanding the structure of diseases plays a crucial role in improving treatment quality and public health care. This paper presents a study applying clustering and association rule mining techniques to analyze the disease structure at the outpatient department of Military Hospital 7A. The study utilizes the K-means algorithm to cluster patients based on shared characteristics, followed by the Apriori algorithm to discover association rules among diseases. The research results provide valuable insights into the disease structure, supporting decision-making in disease management and treatment, and contributing to improving the quality of healthcare services.*

keyword: *Clustering, Association Rule Mining, Data Mining, Disease Structure, K-means, Apriori, Chronic Diseases.*

I. INTRODUCTION

In the current digital age, data has become an invaluable resource, especially when mined and analyzed effectively. Data mining, also known as "Knowledge Discovery in Databases", is not only a powerful tool to gain a deeper understanding of the world around us, but it is also an important tool for creating new values, from predicting market trends to discovering hidden patterns behind data.

The healthcare sector is undergoing a significant shift in disease patterns, with chronic illnesses like cardiovascular diseases, diabetes, and hypertension on the rise. Understanding these patterns is crucial for improving treatment processes and predicting and preventing diseases. While previous research has used data analysis techniques like clustering and decision trees, there remains a gap in understanding the interplay of factors like patient age, location, and the combination of various diseases.

This paper presents a study that applies clustering and association rule mining techniques to analyze disease patterns at the outpatient department of Military Hospital 7A. The goal is to explore the structure of diseases, improve treatment processes, and predict and prevent diseases effectively. By using data mining techniques, we aim to provide valuable insights that can improve healthcare quality and patient outcomes.

II. RELATED RESEARCH AND RESULTS

Clustering Techniques: Clustering is a machine learning technique that groups similar objects into sets (clusters). The goal is to find hidden structures in data without human intervention. There are various clustering methods, each with its own strengths and weaknesses:

K-means clustering: This method partitions data into k clusters based on the nearest mean. It is simple, efficient for large datasets, but requires pre-defining the number of clusters and can be sensitive to the initial cluster center selection.

Hierarchical clustering: This method builds a hierarchy of clusters. It is useful for understanding data structures at various levels of granularity, but can be computationally expensive for large datasets.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): This method groups data points based on density. It can discover clusters of arbitrary shapes and handle noise, but requires careful parameter tuning.

GMM (Gaussian Mixture Models): This method assumes data is generated from a mixture of Gaussian distributions. It is flexible for complex data, but can be computationally expensive.

Spectral clustering: This method uses the spectrum of the similarity matrix of the data to perform dimensionality reduction before clustering. It can handle non-linearly separable data, but requires careful parameter tuning and can be computationally expensive for large datasets.

Association Rule Mining: Association rule mining finds interesting relationships or correlations among items in a dataset. A typical association rule is an implication of the form "If X, then Y", where X and Y are sets of items. Key components of association rules include:

Antecedent (IF part): The condition or itemset that triggers the rule.

Consequent (THEN part): The itemset that is predicted or implied by the antecedent.

Confidence: Measures how often the consequent occurs when the antecedent occurs.

Support: Measures how frequently an itemset occurs in the dataset.

Lift: Measures how much more likely the consequent is to occur when the antecedent occurs compared to when the antecedent does not occur.

Association rules are used in various domains, including market basket analysis, healthcare, and finance. In healthcare, they can be used to identify relationships between diseases, symptoms, and treatments. For example, "If a patient has disease X, they are likely to also have disease Y."

Apriori Algorithm: The Apriori algorithm is a classic algorithm for association rule mining. It efficiently identifies frequent itemsets in a dataset and generates association rules from them. The algorithm uses a "bottom-up" approach, starting with individual items and iteratively generating larger frequent itemsets. A key concept in Apriori is the "Apriori property": if an itemset is frequent, then all of its subsets must also be frequent. This property allows the algorithm to prune the search space and efficiently discover frequent itemsets**.**

Chronic Diseases: Chronic diseases are long-lasting health conditions that cannot be cured completely and often progress slowly. They are a significant public health concern, leading to disability, reduced quality of life, and increased healthcare costs. Common chronic diseases include cardiovascular diseases, diabetes, hypertension, arthritis, asthma, chronic obstructive pulmonary disease, Parkinson's disease, Alzheimer's disease, and various types of cancer.

Machine learning is increasingly used for predicting, diagnosing, and managing chronic diseases. By analyzing patient data, machine learning models can identify patterns and risk factors, leading to early diagnosis and personalized treatment plans.

ICD-10 Codes: ICD-10 (International Classification of Diseases, 10th Revision) is a standardized coding system maintained by the World Health Organization (WHO) to classify diseases and health problems. Each disease or health condition is assigned a unique alphanumeric code. ICD-10 codes are used globally for various purposes, including:

Clinical Diagnosis: To record and track patient diagnoses.

Research: To analyze disease patterns and trends.

Public Health Reporting: To monitor and report on the prevalence of diseases.

Billing and Reimbursement: To process healthcare claims.

Previous Studies and Research Gaps: Several studies have applied clustering and association rule mining in healthcare data analysis.

Bai et al. (2019) used clustering, classification, and logistic regression to predict diabetes.

Massi et al. (2020) used a two-step unsupervised clustering method to detect healthcare fraud.

Durairaj et al. (2013) reviewed the potential of data mining in healthcare to improve diagnosis, treatment, and disease management.

Huang et al. (2020) applied the Apriori algorithm to discover association rules in supermarket sales data.

Khedr et al. (2021) developed an efficient association rule mining algorithm to predict heart diseases.

Yang et al. (2015) used classification and logistic regression to analyze the correlation between clinical and pathological information in lung cancer diagnosis.

While these studies have made valuable contributions, there are still research gaps:

Limited use of temporal information: Many studies do not fully utilize the temporal dimension of patient data, such as the month of admission.

Insufficient consideration of patient demographics: Some studies do not adequately consider factors like age and location, which can significantly impact disease patterns.

Lack of combining multiple data analysis techniques: Many studies focus on a single technique, potentially missing insights that could be gained by combining multiple methods.

Our study aims to address these gaps by:

Clustering patient data based on various factors, including the month of admission, age, and location.

Using the Apriori algorithm to discover association rules among diseases.

Combining clustering and association rule mining to gain a more comprehensive understanding of disease patterns.

III. PROPOSED METHOD OR ALGORITHM

Research Process

This study investigates the disease structure at the outpatient department of Military Hospital 7A using clustering and association rule mining techniques. The research process includes the following steps:

1. **Data Collection:** Collect patient data from medical records at Military Hospital 7A, including information on main disease, comorbidities, age, location, and month of admission.
2. **Data Preprocessing:** Prepare the data for analysis by:
   - Removing missing values.
   - Encoding categorical variables, such as disease names and locations, into numerical representations.
   - Standardizing numerical variables, such as age, to have a mean of 0 and a standard deviation of 1.
3. **Clustering:** Apply the K-means algorithm to cluster patients based on shared characteristics, such as age, location, main disease, and comorbidities.
4. **Association Rule Mining:** Apply the Apriori algorithm to discover association rules among diseases and patient characteristics.
5. **Evaluation:** Evaluate the clustering results by analyzing the characteristics of each cluster and comparing them to domain knowledge.
6. **Interpretation:** Interpret the association rules and discuss their implications for disease management and treatment.

1. Data Collection

Data is collected from medical records at Military Hospital 7A. The collected data includes the following information for each patient:

- Month of admission (Month)

- Main disease (ICD10-Chính)
- Comorbidities (ICD10-Kèm theo)
- Age (Tuổi)
- Location (Địa chỉ)

2. **Data Processing:**

- **Convert disease names to ICD-10 codes:** This step involves mapping the literal names of diseases to their corresponding ICD-10 codes. ICD-10 codes are alphanumeric codes used for classifying diseases and health problems globally. This conversion ensures standardization and facilitates further analysis.
- **Remove records with missing values:** Records with missing values for any of the variables (month of admission, main disease, comorbidities, age, or location) are removed from the dataset. This step ensures that the analysis is performed on a complete and consistent dataset.
- **Encode categorical variables into numerical representations:** Categorical variables, such as location (Địa chỉ), are converted into numerical representations using appropriate encoding techniques. This conversion is necessary for applying the K-means algorithm, which requires numerical input data.
- **Standardize numerical variables:** Numerical variables, such as age (Tuổi), are standardized to have a mean of 0 and a standard deviation of 1. Standardization ensures that all variables have the same scale and prevents variables with larger values from dominating the clustering process.

3. **Data Clustering:**

Apply the K-means algorithm to cluster patients based on shared characteristics: The K-means algorithm is used to group patients into clusters based on their similarity across multiple dimensions, including age, location, main disease, and comorbidities.

Determine the optimal number of clusters using the Elbow method or Silhouette analysis: The Elbow method and Silhouette analysis are techniques used to determine the optimal number of clusters (k) for the K-means algorithm. The Elbow method looks for a point where the decrease in the within-cluster sum of squares (WCSS) starts to slow down, resembling an elbow. Silhouette analysis calculates the average silhouette score for different values of k, where a higher score indicates better-defined clusters.

Analyze the characteristics of each cluster: After clustering, the characteristics of each cluster are analyzed to identify common patterns and representative features of patients within each group.

4. **Association Rule Extraction:**

Apply the Apriori algorithm to discover association rules among diseases and patient characteristics: The Apriori algorithm is used to identify frequent itemsets (sets of diseases or characteristics that often occur together) and generate association rules.

Set the minimum support threshold to 6.5%: The minimum support threshold is set to 6.5% to ensure that only sufficiently frequent itemsets are considered for generating association rules. Support is a measure of how frequently an itemset occurs in the dataset.

Analyze the extracted association rules: The extracted association rules are analyzed to understand the relationships between diseases, comorbidities, and patient demographics.

- ## Ethical Considerations

The study adheres to ethical guidelines to protect patient data and privacy. The following measures are taken:

- **Compliance with Laws and Regulations:** The study complies with all relevant laws and regulations regarding patient data security and privacy.
- **Anonymization and Encryption:** Patient data is anonymized and encrypted to protect patient identities.
- **Confidentiality:** Patient information is kept confidential and is only used for the purpose of this study.
- **Informed Consent:** Patients are informed about the study and their data usage, and their consent is obtained before their data is included in the analysis.
- **Data Deletion:** Patient data is deleted when it is no longer needed for the study.

5. Evaluation of Clustering Results

The evaluation of the clustering results involves analyzing the characteristics of each cluster and comparing them to domain knowledge. This process aims to assess the quality and interpretability of the generated clusters.

- **Cluster Characteristics:** For each cluster, descriptive statistics and visualizations are used to identify the defining characteristics of patients within that cluster. These characteristics may include:
    - **Age:** The age range and distribution of patients within the cluster.
    - **Location:** The geographical distribution and prevalence of patients from specific locations.
    - **Primary Disease:** The most frequent primary diagnoses within the cluster.
    - **Comorbidities:** The most frequent comorbidities associated with the primary diagnoses within the cluster.
- **Comparison with Domain Knowledge:** The identified cluster characteristics are compared to existing medical knowledge and expert opinion to assess their validity and relevance. This comparison helps determine whether the discovered patterns align with established medical understanding or reveal novel insights.

6. Interpretation of Association Rules

The interpretation of association rules involves understanding the relationships between diseases and patient characteristics and discussing their implications for disease management and treatment.

- **Understanding Relationships:** The association rules are analyzed to identify meaningful relationships between diseases, comorbidities, and patient demographics. For example, a rule might indicate that patients with a specific primary disease are highly likely to have certain comorbidities or that a particular disease combination is prevalent in a specific age group or location.
- **Implications for Disease Management and Treatment:** The identified relationships can inform various aspects of disease management and treatment, including:
  - **Risk Assessment:** Identifying patients at higher risk of developing specific comorbidities or complications based on their primary disease and other characteristics.
  - **Treatment Planning:** Developing more targeted and effective treatment plans by considering the potential comorbidities and complications associated with a primary disease.
  - **Preventive Measures:** Implementing preventive measures for patients identified as being at risk of developing specific diseases or complications.
  - **Resource Allocation:** Optimizing the allocation of healthcare resources based on the prevalence and patterns of diseases in different patient populations.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Dataset Description: The experimental dataset consists of 602 patient records obtained from the outpatient department of Military Hospital 7A. Each record includes the following fields**.**

Field	**Description**
Month	Month of admission (1-12)
ICD10-Chính	ICD-10 code of the primary disease
ICD10-Kèm theo	ICD-10 code of comorbidities
Tuổi	Age of the patient
Địa chỉ	Location of the patient

The dataset includes 12 unique primary diseases and 27 unique comorbidities. The age of patients ranges from 38 to 93 years old.

Data Clustering Results. The K-means algorithm is applied to cluster patients based on the following criteria:

1. **Age:**
   - The optimal number of clusters is determined to be 4.
   - The clusters are characterized by different age ranges and prevalent diseases.
   - For example, patients aged 52-62 are predominantly diagnosed with hypertension (I10) and hypercholesterolemia (E78).
2. **Location:**

   - The optimal number of clusters is determined to be 5.
   - The clusters are characterized by different locations and prevalent diseases.
   - For example, patients from District 7 are predominantly diagnosed with chronic obstructive pulmonary disease (J44) and bronchitis (J20-J22).
3. **Primary Disease and Comorbidities:**
   - The optimal number of clusters is determined to be 7.
   - The clusters are characterized by different primary diseases and their associated comorbidities.
   - For example, patients with asthma (J45) are predominantly associated with sinusitis (J32).

Association Rule Mining Results

The Apriori algorithm is applied to discover association rules among diseases and patient characteristics. The minimum support threshold is set to 6.5%.

The extracted association rules reveal interesting relationships, such as:

- Patients with diabetes (E11) and hypercholesterolemia (E78) are predominantly aged between 40 and 60 years old.
- Patients with hypertension (I10) in District 5 are predominantly associated with hypercholesterolemia (E78).
- Patients with hypertension (I10) aged between 40 and 60 years old are predominantly associated with hypercholesterolemia (E78).

Comparison with Previous Studies

The research results are consistent with previous studies that have identified relationships between certain diseases, such as diabetes and hypercholesterolemia, and hypertension and hypercholesterolemia.

However, this study provides a more comprehensive analysis by considering multiple factors, including age, location, primary disease, and comorbidities.

Implications and Potential Applications

The research results have several implications for medical practice:

- **Improved Disease Management:** The identified disease patterns can help healthcare professionals better understand the relationships between diseases and develop more effective treatment plans.
- **Early Diagnosis and Prevention:** The association rules can help identify patients at risk of developing certain diseases, leading to early diagnosis and preventive measures.
- **Personalized Medicine:** The clustering results can help tailor treatment plans to specific patient groups based on their shared characteristics.

- **Resource Optimization:** The prediction of disease patterns can help optimize healthcare resource allocation and improve the efficiency of healthcare systems.

Limitations: The study has several limitations.

- **Limited Dataset:** The dataset is limited to 602 patient records from a single hospital. A larger and more diverse dataset would improve the generalizability of the results.
- **Data Availability:** The dataset only includes information on primary disease, comorbidities, age, location, and month of admission. Additional patient information, such as lifestyle factors and medical history, would enhance the analysis.
- **Algorithm Selection:** The study uses the K-means and Apriori algorithms. Other clustering and association rule mining algorithms could be explored to potentially discover additional patterns.

V. CONCLUSION

This study investigated the disease structure at the outpatient department of Military Hospital 7A using clustering and association rule mining techniques. The K-means algorithm was employed to cluster patients based on shared characteristics, while the Apriori algorithm was used to discover association rules among diseases and patient characteristics. The key findings of this study including: *Identification of distinct patient clusters* **-** The clustering analysis revealed distinct patient groups characterized by different age ranges, locations, primary diseases, and comorbidities. *Discovery of significant association rules* - The association rule mining analysis identified meaningful relationships between diseases, comorbidities, and patient demographics. *Insights into disease patterns* - The combined analysis of clustering and association rule mining provided valuable insights into the disease structure at the outpatient department.

This research contributes to a deeper understanding of disease patterns and has the potential to support: *Improved disease management and treatment* **-** By identifying high-risk patient groups and understanding disease relationships, healthcare professionals can develop more targeted and effective treatment plans. *Early diagnosis and prevention* - identified association rules can help predict potential comorbidities and complications, leading to early diagnosis and preventive measures. *Personalized medicine* **-** The clustering results can help tailor treatment plans to specific patient groups based on their shared characteristics. *Resource optimization* **-** The prediction of disease patterns can help optimize healthcare resource allocation and improve the efficiency of healthcare systems. Despite these contributions, the study has limitations: *Limited dataset* **-** The analysis was based on a relatively small dataset from a single hospital. *Data availability* **-** The dataset lacked detailed patient information, such as lifestyle factors and medical history. *Algorithm selection* **-**The study used specific clustering and association rule mining algorithms. Other algorithms might reveal additional patterns.

Future research directions include: *Expanding the dataset* **-** Analyzing larger and more diverse datasets to improve the generalizability of the findings. *Incorporating additional patient information* **-** Including more comprehensive patient data to enhance the analysis. *Exploring alternative algorithms* - Evaluating different clustering and association rule mining algorithms to potentially discover additional patterns. *Developing predictive models* - Building predictive models to forecast disease progression and treatment outcomes. *Evaluating the clinical impact-* Conducting clinical studies to assess the impact of the findings on patient outcomes and healthcare quality. By addressing these limitations and pursuing future research directions, this study can contribute to advancing healthcare knowledge and improving patient care.

REFERENCES

Vietnamese:

[1] Cục quản lý Khám Chữa Bệnh. (2024). *ICD-10*. https://icd.kcb.vn/#/icd-10/icd10

[2] Hà Siu. (2017). *Phân cụm dữ liệu và luật kết hợp ứng dụng trong phân tích dữ liệu công thức dược phẩm* (Luận văn thạc sĩ, Trường Đại học Công nghệ Thông tin, TP. Hồ Chí Minh).

[3] Tống Đức Phong. (2014). *Ứng dụng khai phá dữ liệu xây dựng hệ hỗ trợ chẩn đoán y khoa* (Luận văn thạc sĩ, Trường Đại học Hồng Bàng, TP. HCM).

[4] Nguyễn Đức Thuần. (2013). *Nhập môn khai phá dữ liệu và quản trị tri thức*. Hà Nội: Thông tin và truyền thông.

[5] Hoàng Thị Thanh Huyền. (2016). *Ứng dụng khai phá dữ liệu để xây dựng hệ thống chẩn đoán bệnh trầm cảm cho học sinh phổ thông* (Luận văn thạc sĩ, Trường Đại học Đà Nẵng, Đà Nẵng).

English:

[1] Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules in large databases*. In *Proceedings of the 20th international conference on very large databases* (pp. 488-499). Morgan Kaufmann Publishers.

[2] Bai, B. M., Nalini, B. M., & Majumdar, J. (2019). Analysis and detection of diabetes using data mining techniques—a big data application in health care. *Emerging research in computing, information, communication and applications*, 443-455.

[3] Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of personalized medicine*, *10*(2), 21.

[4] Bertsimas, D., Dunn, J., Velmahos, G. C., and Kaafarani, H. M. A. (2018). Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based predictive optimal trees in emergency surgery risk (POTTER) calculator. Ann. Surg. 268, 574–583.

[5] Phillips, J. L., & Currow, D. C. (2010). Cancer as a chronic disease. *Collegian*, *17*(2), 47-50.

[6] Durairaj, M., & Ranjani, V. (2013). Data mining applications in healthcare sector: a study. *International journal of scientific & technology research*, *2*(10), 29-35.

[7] Grant, R. W., McCloskey, J., Hatfield, M., Uratsu, C., Ralston, J. D., Bayliss, E., et al. (2020). Use of latent class analysis and k-means clustering to identify complex patient profiles. JAMA Netw. Open 3, e2029068.

[8] Hegland, M. (2007). The apriori algorithm–a tutorial. *Mathematics and computation in imaging science and information processing*, 209-262.

[9] Huang, M. J., Sung, H. S., Hsieh, T. J., Wu, M. C., & Chung, S. H. (2020). Applying data-mining techniques for discovering association rules. *Soft Computing*, *24*, 8069-8075.

[10] Khedr, A. M., Aghbari, Z. A., Ali, A. A., & Eljamil, M. (2021). An Efficient Association Rule Mining From Distributed Medical Databases for Predicting Heart Diseases. *IEEE Access*, *9*, 15320-15333.

[11] Ma, H., Ding, J., Liu, M., & Liu, Y. (2022). Connections between various disorders: combination pattern mining using apriori algorithm based on diagnosis Information from electronic medical records. *BioMed Research International*, *2022*.

[12] Massi, M. C., Ieva, F., & Lettieri, E. (2020). Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases. *BMC medical informatics and decision making*, *20*, 1-11.

[13] Miller, R. J., & Yang, Y. (1997). Association rules over interval data. *ACM SIGMOD Record*, *26*(2), 452-461.

[14] Löwe, B., Gräfe, K., Zipfel, S., Witte, S., Loerch, B., & Herzog, W. (2004). Diagnosing ICD-10 depressive episodes: superior criterion validity of the Patient Health Questionnaire. *Psychotherapy and psychosomatics*, *73*(6), 386-390.

[15] Segura-Delgado, A., Gacto, M. J., Alcalá, R., & Alcalá-Fdez, J. (2020). Temporal association rule mining: An overview considering the time variable as an integral or implied component. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(4), e1367.

[16] Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1-34.

[17] Yang, H., & Chen, Y. P. P. (2015). Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information. *Expert Systems with Applications*, *42*(15-16), 6168-6176.

[18] Patel, V., Chatterji, S., Chisholm, D., Ebrahim, S., Gopalakrishna, G., Mathers, C., ... & Reddy, K. S. (2011). Chronic diseases and injuries in India. *The Lancet*, *377*(9763), 413-428.

[19] Zhao, Y., & Karypis, G. (2005). Data clustering in life sciences. *Molecular biotechnology*, *31*, 55-80.