ANALYSIS OF ASSOCIATIVE RULES OF SENSOR DATA BASED ON THE APRIORI ALGORITHM

Otabek Qadamboyevich Xo'jayev, Azizbek Dilshodovich Jumanazarov
Urgench Branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
otabek.hujaev@gmail.com, devdilshodovich@gmail.com

ABSTRACT

This study presents an analysis of associative rules of electrocardiogram (ECG) sensor data based on the Apriori algorithm. The objective is to identify frequent patterns and relationships within the ECG data, which can aid in the diagnosis and treatment of cardiovascular diseases. The methodology involves collecting ECG data from patients, applying data mining techniques to uncover hidden patterns and associations, and using the Apriori algorithm to identify frequent itemsets and association rules. The results show that the Apriori algorithm is an effective method for identifying relevant patterns and relationships within ECG sensor data. The findings provide valuable insights into the behavior of ECG sensors and their relationships with other variables, which can be used to optimize the diagnosis and treatment of cardiovascular diseases. Overall, this study demonstrates the potential of data mining techniques in the analysis of ECG sensor data, and highlights the importance of understanding the underlying patterns and associations in ECG data for effective decision-making in healthcare.

Keywords: Sensor data analysis, Apriori algorithm, Electrocardiogram (ECG) sensor data, Data mining techniques, Cardiotocography (CTG).

INTRODUCTION

Sensor data plays a crucial role in various fields such as medical diagnosis, industrial automation, and environmental monitoring. However, the large volume of data generated by sensors poses a challenge in extracting meaningful insights that can aid decision-making. To address this issue, the Apriori algorithm is commonly used to analyze sensor data by identifying associative rules between different sensor measurements. The purpose of this analysis is to uncover the hidden relationships between sensor measurements and provide insights that can optimize processes, detect anomalies, and predict future events. This study focuses specifically on the analysis of associative rules of sensor data based on the Apriori algorithm, with the aim of developing a

framework that can handle large-scale sensor datasets. Despite its limitations, the Apriori algorithm remains a popular choice for association rule learning due to its simplicity, ease of understanding, and ability to operate without labeled data. Moreover, the availability of various extensions tailored to different use cases further enhances its usefulness as a valuable tool for data analysis.

METHODS

In Wenjing Zhang's [1] study, the importance of data mining methods, specifically association rule mining, in discovering valuable knowledge from large datasets generated by various disciplines, including medicine, is discussed. The paper highlights the use of the classic Apriori algorithm for data mining analysis of medical data and the need to improve the algorithm to suit the characteristics of medical data. The results show that the improved algorithm can identify useful association relationships or patterns between large data item sets, which can be used in medical diagnosis. Overall, the paper suggests that using data mining methods to analyze medical data is a worthy research direction.

Mirpouya Mirmozaffari [2] conducted a study to predict heart disease by comparing various strong rules of the Apriori algorithm in data mining. They developed a unique model with one filter and evaluation methods and tested three strong rules and different evaluation methods to determine the best software. Through this comparison, they were able to introduce a high-performance software that can help physicians predict uncertain cases and offer advice accordingly. The study emphasizes the potential of using data mining methods, particularly the Apriori algorithm, in the medical field to predict heart disease and provide effective medical advice.

The study of Meng Chen and Zhixiang Yin [3] proposes a method for solving the intersection problem of suspicious data between health and pathological data. The method involves feature selection, multi-model prediction, and classification based on the Apriori algorithm. By dividing suspicious data into health and pathology, the accuracy of the classification of the entire dataset is significantly improved, and the proposed method has higher accuracy compared to other models. The study's results suggest that the feature extraction and model classification methods have a positive impact on clinical decision-making, healthy fetal development, and safe delivery of pregnant women. However, the study lacks real data to verify the prediction results for suspicious data, which can be addressed in future studies. The study also suggests that feature extraction and classification from

the perspective of CTG signal processing can be conducted to increase the study's authenticity.

DISCUSSION

This topic focuses on using the Apriori algorithm to analyze associative rules in sensor data. Associative rule mining can help identify patterns in sensor data that are associated with different conditions or outcomes, with applications in healthcare, manufacturing, and environmental monitoring. While the Apriori algorithm can handle large datasets, challenges such as data sparsity and noise remain. Further research is needed to improve scalability, accuracy, and interpretation of results. Overall, analyzing associative rules of sensor data can provide valuable insights for decision-making and lead to improvements in various domains.

Data Collection

This study uses a dataset from Kaggle [4] that includes various medical variables related to patients with suspected heart disease. The dataset includes information such as the age and sex of the patient, exercise-induced angina, the number of major vessels, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, and a target variable indicating the likelihood of heart attack. The dataset also includes explanations of the values for certain categorical variables such as chest pain type and resting electrocardiographic results.

Correlation matrix and heat map

A correlation matrix is a table Figure [1] showing the statistical relationship between a set of variables. The table shows the correlation coefficient between each pair of variables, which is a measure of how strongly the two variables are related. A heat map is a graphical representation Figure [2] of the correlation matrix, where the values are represented by colors. A heat map can help visualize the strength and direction of the relationship between the variables.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308	-0.116211	-0.398522	0.096801	0.210013	-0.168814	0.276326	0.068001	-0.225439
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032	-0.058196	-0.044020	0.141664	0.096093	-0.030711	0.118261	0.210041	-0.280937
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444	0.044421	0.295762	-0.394280	-0.149230	0.119717	-0.181053	-0.161736	0.433798
trtbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531	-0.114103	-0.046698	0.067616	0.193216	-0.121475	0.101389	0.062210	-0.144931
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294	-0.151040	-0.009940	0.067023	0.053952	-0.004038	0.070511	0.098803	-0.085239
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000	-0.084189	-0.008567	0.025665	0.005747	-0.059894	0.137979	-0.032019	-0.028046
restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189	1.000000	0.044123	-0.070733	-0.058770	0.093045	-0.072042	-0.011981	0.137230
thalachh	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567	0.044123	1.000000	-0.378812	-0.344187	0.386784	-0.213177	-0.096439	0.421741
exng	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665	-0.070733	-0.378812	1.000000	0.288223	-0.257748	0.115739	0.206754	-0.436757
oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747	-0.058770	-0.344187	0.288223	1.000000	-0.577537	0.222682	0.210244	-0.430696
slp	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894	0.093045	0.386784	-0.257748	-0.577537	1.000000	-0.080155	-0.104764	0.345877
caa	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979	-0.072042	-0.213177	0.115739	0.222682	-0.080155	1.000000	0.151832	-0.391724
thall	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019	-0.011981	-0.096439	0.206754	0.210244	-0.104764	0.151832	1.000000	-0.344029
output	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046	0.137230	0.421741	-0.436757	-0.430696	0.345877	-0.391724	-0.344029	1.000000

Figure 1. Correlation coefficient table

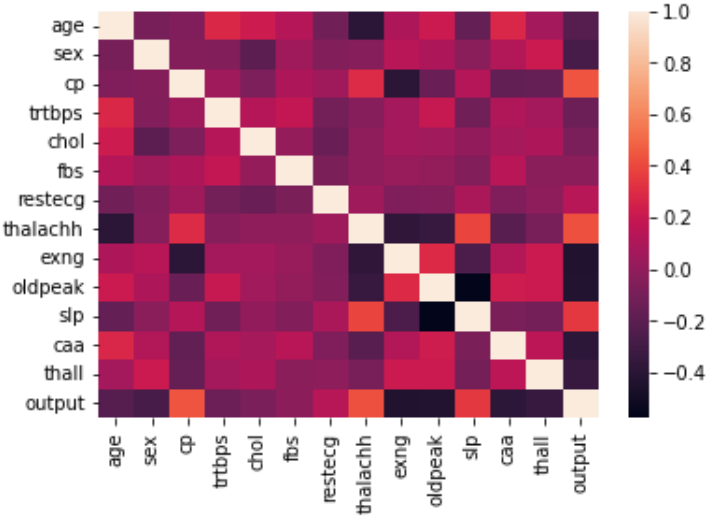

Figure 2. Correlation matrix visualization

RESULT

Data preprocessing

This report analyzes data preprocessing techniques and emphasizes their importance in the data analysis process. Data preprocessing involves cleaning and transforming raw data to a suitable format for further analysis. Proper data preprocessing enables accurate and meaningful analysis and the derivation of useful insights from the data. The report covers the different steps involved in data preprocessing, including data cleaning, transformation, and normalization.

In the following step, we replaced the outliers, one-hot encoding, and normalization.

	age	trtbps	chol	thalachh	oldpeak	sex	exng	caa	cp	fbs	restecg	slp	thall
0	0.952197	0.985791	-0.225460	0.004270	1.274983	1	0	0	3	1	0	0	1
1	-1.915313	-0.007210	0.155309	1.654943	2.429798	1	0	0	2	0	1	0	2
2	-1.474158	-0.007210	-0.875007	0.985751	0.408872	0	0	0	1	0	0	2	2
3	0.180175	-0.669211	-0.158266	1.253428	-0.168535	1	0	0	1	0	1	2	2
4	0.290464	-0.669211	2.484719	0.584236	-0.361004	0	1	0	0	0	1	2	2

Figure 3. Table form of analysis process

Association rule mining

The first step is transforming the data for rule mining using the Apriori algorithm. It reads in the dataset we provided and creates bins for each feature. It then generates a list of transactions by iterating through each row of the dataset and appending the bin labels for each feature to the transactions list. The Apriori algorithm is applied to the transactions list to generate rules that meet certain criteria.

	items	support	ordered_statistics
0	(chol(125.562, 272.0])	0.732673	[((), (chol(125.562, 272.0]), 0.7326732673267327, 1.0)]
1	(fbs-0)	0.851485	[((), (fbs-0), 0.8514851485148515, 1.0)]
2	(oldpeak(-0.0062, 2.067])	0.834983	[((), (oldpeak(-0.0062, 2.067]), 0.834983498349835, 1.0)]
3	(chol(125.562, 272.0], age(45.0, 61.0])	0.402640	[((age(45.0, 61.0]), (chol(125.562, 272.0]), 0.726190476190476, 0.991151866151866)]
4	(fbs-0, age(45.0, 61.0])	0.455446	[((age(45.0, 61.0]), (fbs-0), 0.8214285714285714, 0.9647009966777409)]

The resulting rules are displayed in a table format, and the top 10 rules that contain the output bin label are selected and sorted by support in descending order.

Figure 4. Associations table

Following the association rule mining using the Apriori algorithm, we have found 10 strong associations Figure [4]. Based on these associations, it can be concluded that patients with oldpeak values in the range of (-0.0062, 2.067], no exercise-induced angina (exng=0), and normal fasting blood sugar (fbs=0) may have

	items	support	ordered_statistics
58	(output-1, oldpeak(-0.0062, 2.067])	0.521452	[((output-1), (oldpeak(-0.0062, 2.067]), 0.9575757575757575, 1.1468199784405317)]
35	(output-1, exng-0)	0.468647	[((output-1), (exng-0), 0.8606060606060606, 1.2782531194295899)]
45	(fbs-0, output-1)	0.468647	[((output-1), (fbs-0), 0.8606060606060606, 1.0107117688513036)]
203	(fbs-0, output-1, oldpeak(-0.0062, 2.067])	0.452145	[((output-1), (fbs-0, oldpeak(-0.0062, 2.067]), 0.8303030303030302, 1.1647306397306396), ((fbs-0...
183	(oldpeak(-0.0062, 2.067], output-1, exng-0)	0.448845	[((output-1), (exng-0, oldpeak(-0.0062, 2.067]), 0.8242424242424242, 1.364729259811227), ((oldpe...
15	(chol(125.562, 272.0], output-1)	0.432343	[((output-1), (chol(125.562, 272.0]), 0.793939393939394, 1.0836199836199836)]
11	(output-1, caa-0)	0.429043	[((caa-0), (output-1), 0.7428571428571428, 1.364155844155844), ((output-1), (caa-0), 0.787878787...
75	(output-1, thall-2)	0.429043	[((output-1), (thall-2), 0.7878787878787877, 1.438116100766703), ((thall-2), (output-1), 0.78313...
241	(output-1, thall-2, oldpeak(-0.0062, 2.067])	0.415842	[((output-1), (oldpeak(-0.0062, 2.067], thall-2), 0.7636363636363636, 1.5122994652406416), ((tha...
149	(chol(125.562, 272.0], output-1, oldpeak(-0.0062, 2.067])	0.409241	[((output-1), (chol(125.562, 272.0], oldpeak(-0.0062, 2.067]), 0.7515151515151515, 1.24431197218...

a higher chance of experiencing a heart attack.

CONCLUSION

In conclusion, the Apriori algorithm is a popular algorithm for performing association rule mining. It works by identifying frequent itemsets, or sets of items that frequently co-occur in the dataset, and then using these itemsets to generate association rules. The algorithm is efficient and scalable, making it well-suited for analyzing large datasets.

In the context of heart health, the Apriori algorithm can be used to discover frequent itemsets and association rules that are associated with a higher or lower risk of heart attack. For example, we may find that patients with certain combinations of age, sex, and chest pain type are more likely to experience a heart attack. By identifying these patterns, we can develop targeted interventions and treatment plans to reduce the risk of heart disease and improve patient outcomes.

Overall, the Apriori algorithm is a powerful tool for analyzing complex datasets and identifying meaningful associations between variables. It can be used in a wide range of applications, including in healthcare, finance, and marketing.

REFERENCES

1. Medical Diagnosis Data Mining Based on Improved Apriori Algorithm. Wenjing Zhang, Donglai Ma, Wei Yao. College of Information Science & Technology, Agricultural University of Hebei, Baoding, China. Journal of Networks, vol. 9, no. 5, May 2014
2. Data Mining Apriori Algorithm for Heart Disease Prediction. Mirpouya Mirmozaffari, Alireza Alinezhad and Azadeh Gilanpour. Int'l Journal of Computing, Communications & Instrumentation Engg. Vol. 4, Issue 1 (2017)
3. Classification of Cardiotocography Based on the Apriori Algorithm and Multi-Model Ensemble Classifier. Meng Chen and Zhixiang Yin. School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai, China. May 2022. vol. 10. Article. 888859
4. Heart Attack Analysis & Prediction Dataset. A dataset for heart attack classification