Analysis of Accuracy K-Means and Apriori Algorithms for Patient Data Clusters

To cite this article: N P Dharshini et al 2019 J. Phys.: Conf. Ser. 1230 012020

View the article online for updates and enhancements.
Analysis of Accuracy K-Means and Apriori Algorithms for Patient Data Clusters

N P Dharshinni*, Fadhillah Azmi, I Fawwaz, A M Husein, Saut Dohot Siregar
Faculty of Technology and Computer Science, Universitas Prima Indonesia, Indonesia
* priyadharshinni@unprimdn.ac.id

Abstract. The stacking data is useful to get a new information. Data mining is a method to determine the important pattern in Frequent Itemset Mining (FIM). Apriori is part of association rule that is used to determine the associative relationship in items combination. But apriori has a high computational time weakness because frequent itemset process searching must scan the database repeatedly for each itemset combination. This study aims to see the effect of the k-means clustering algorithm on the apriori algorithm by combining these two algorithms. The test results show that the combination of k-means and apriori algorithms produces more information detaily and faster time computing than the apriori algorithm with a total computing time of 21.93 minutes and a combination of k-means and apriori algorithms 17.41 minutes.

1. Introduction
Data Mining is a technique used to process very large amounts of data stored in a database so that it produces a knowledge. Data Mining is the core of the Knowledge Discovery process in Database (KDD), which involves algorithms that explore data, develop models and find patterns that were previously unknown. This model is used to understand phenomena from data, analysis and prediction. Data mining is widely applied in the medical field, including medicine prediction, heart disease and the discovery of the relationship between clinical and pathological data. Data Mining is divided into several groups including: association, classification, clustering, and sequence pattern [1].

Apriori algorithm is part of the association rule that is used to determine the associative relationship of an item combination. Apriori algorithm will be suitable to be applied if there are several relationship items to be analyzed [2]. Apriori algorithms are widely used by researchers for various aspects, among others, product arrangement [3], prediction [4]. However, the priori algorithm has a weakness of computational time which is quite high because the frequent itemset search process must scan the database repeatedly for each combination [5].

Clustering is a process to group data into several clusters or groups so that data in one cluster has the maximum level of similarity and data between clusters has a minimum similarity [6]. The clustering method is generally divided into two, namely hierarchical clustering and partitional clustering. Clustering which is included in hierarchical clustering is complete linkage clustering, single linkage clustering, average linkage clustering and centroid linkage clustering. While those included in partitional clustering are k-means and fuzzy k-means [7].

K-Means is the simplest and the most to used as clustering algorithm. It is caused k-means that has the advantage to be able the most group and relatively fast and efficient to compute time. However, k-means has the disadvantage to determine the first center clustering so that the results clustering are very depend on determining the value of the initial cluster center (centroid) given.
In this study focused on using 2 (two) approaches. The first approach applies apriori algorithm to get the associated association rules in extracting information from a set of data items. The second approach applies the k-means and apriori combinations. The two results of this approach will be analyze to determine the impact of grouping data on apriori algorithms. This paper is presented as follows: chapter 2 of the method used, chapter 3 described the proposed method. The results and discussion in chapter 4 and chapter 5 conclusions.

2. Related Research
N P Dharshinni et al [8], applied a combination of k-means and apriori algorithms to find patterns of prescription linkages with medicine sales data. So that the results obtained 35% of medicine prescription data given by doctors is not obtained in medicine sales transactions. M. Harahap et al [9], applies the k-means algorithm to clustering the 10 most dominant diseases and look for linkages between diseases and medicines with apriori algorithms. Ali Akbar Jafarzadeh et al [10], applies the fuzzy c-means (FCM) algorithm in accordance with the factors that influence forest fire risk divided into 3 groups with different membership levels, then apriori algorithm gets a strong relationship between forest fire events, distance from settlement, density population, distance from road, slope, standing dead wood, temperature, land cover and distance from agricultural land. Lakshmi K.S and G.Vadivu [11], applies an apriori association rule algorithm with multicriteria to determine the relationship between diseases, namely diseases and symptoms, diseases and medicines.

Potta Swathi and Bodapati Prajna [12], discussing that data mining techniques have been widely used in medical data and the process is effective in a priori algorithms on medical data. Visnja Istrat and Nenad Lalic [13], applies apriori algorithm to determine the buyer's behavior patterns in textile industry transaction data to improve decision-making models. Lismardiana et al [14], develops apriori algorithms for decision making, by comparing apriori algorithms and FP-Growth algorithms, and finds weaknesses in the apriori algorithm in terms of speed because it repeatedly scans the database compare to the FP-Growth algorithm which only scans the database only once. Kuswantoro, E. & Suprapto, Y.K [15], applied the K-Means algorithm to obtain a labor force level model in the Maluku province by generating a level of unemployment from the results of each cluster. Metisen, B.M. & Sari, H.L [16], applied the K-Means algorithm to classify product sales. This study produced 2 groups of data, namely low sales data and high sales data to find out the types of items that were selling well and not selling well.

3. Proposed Method
3.1 Apriori Algorithm
The work steps of a priori algorithm can be described as follows:
Determine minimum support; Calculate items from support (transactions that contain all items) by scanning the database for 1- item set; Perform the 2-itemset combination from the previous k-itemset; Then calculate support by scanning the database for 2-itemset. Itemset that meets the minimum support will be the frequency of the candidate; Set k-itemset value from support which meet minimum support from k-items; Perform the process for the next iteration until there are no more k-itemset that meet the minimum support.

The Formula to calculate Support, Confidence and Lift:

\[ \text{Support (A,B)} = \frac{\text{P (A} \cap \text{B)}}{\text{The Total Processing}} \]

\[ \text{Confidence (A} \rightarrow \text{B)} = \frac{\text{P (B | A)}}{\text{The Processing Contains of A and B}} \]

\[ \text{Lift (A} \rightarrow \text{B)} = \frac{\text{Support (A} \cap \text{B)}}{\text{Support (A) Support (B)}} \]
3.2 K-Means Algorithm

The concept of the K-Means algorithm uses K as a parameter, which divides n objects into clusters, to create relatively high similarities in clusters and relatively low similarities between clusters and minimizes the total distance between values in each cluster to the centroid. Each Centroid is a cluster average value. Similarity calculation is done with the average value of the cluster object. Measurement of similarity for algorithm selection is done by a reciprocating Euclidean distance which means that the closer the distance is, the greater the similarity between two objects and vice versa [17].

The Formula to calculate distance value using Euclidean distance:

$$d(x_i,c_j) = \sqrt{\sum_{j=1}^{n} (x_j - c_j)^2}$$

(4)

3.3 Dataset

Sources of data used in this study are patient diagnosis data in 2016 which came from one of the Private Hospitals in Medan City.

3.4 Tool

In this research, we apply the apriori algorithm and k-means in WEKA open source application because it supports several data mining methods such as preprocessing data, clustering, regression classification, visualization, and feature selection.

3.5 Research framework

In this study, the dataset was taken from one of the private hospital databases and continued by processing data at the data preprocessing stage by transforming data. Then the preprocessing data is applied to approach 1 and approach 2. In approach 1 the data will be managed using apriori algorithm, while in approach 2 the data is grouped first using the k-means clustering algorithm then the grouping result data is applied to the apriori algorithm. Both of these approaches will each produce the Rule Mining Association. The results of Approach 1 and Approach 2 will be analyzed to determine the impact of grouping data on apriori algorithms based on computation time comparisons.
4. Discussion
This research is used a dataset which is needed to extract to achieve useful information about the effect of k-means algorithm to apriori algorithm from computation time and rule achieved. The dataset used consists of 8243 disease diagnose data. Medical data variables consist of disease diagnosis, age group, gender, the status of care. The partial data used can be seen in Table 1.

**Table 1. Sample patient diagnosis data in 2016**

| No. | Disease Diagnosis | Age Cluster | Gender | Status of Care |
|-----|-------------------|-------------|--------|----------------|
| 1   | Observation of Febris | Baby        | Male   | Outpatient     |
| 2   | Observation of Febris | Baby        | Female | Outpatient     |
| 3   | Observation of Febris | Baby        | Male   | Outpatient     |
| 4   | Observation of Febris | Baby        | Female | Outpatient     |
4.1 Approach Algorithm apriori without grouping (Approach 1)
In the first approach, directly apply the apriori algorithm in the dataset to 4 input variables, namely disease diagnosis, age group, gender, the status of care in order to obtain confidence values, rules and computational time on apriori algorithms. The test results obtained from the Apriori algorithm can be seen in Table 2.

| Disease Diagnosis | Age Cluster | Gender | Status of Care | Confidence Percentage |
|-------------------|-------------|--------|----------------|-----------------------|
| Cataracts, not specified | Elder | Adult | Outpatient | 69% |
| Cataracts, not specified | Elder | Female | Outpatient | 76% |
| Cataracts, not specified | Elder | Female | Outpatient | 60% |
| Another allergic rhinitis | Child | Male | Outpatient | 66% |
| Another allergic rhinitis | Adult | Female | Outpatient | 69% |
| Cataracts, not specified | Elder | Female | Outpatient | 69% |
| Cataracts, not specified | Elder | Male | Outpatient | 69% |

As shown in Table 2 above, the rule information obtained in the Large Itemset 4 results in two rules, namely the diagnosis of another allergic rhinitis with the age group of female children and outpatient status. Then, the diagnosis of postoperative disease with the adult age group gender male and outpatient status with each confidence value of 69%. From these results, it can be seen that the information obtained from the Apriori algorithm is still lacking.

4.2 Approach Combination K-Means and Apriori Algorithm (Approach 2)
In the second approach applying the k-means clustering algorithm first on the dataset with 4 input variables, disease diagnosis, age group, gender, care status with the number k = 4. The grouping results obtained can be seen in Table 3.

| Cluster | #of Instance | ratio% |
|---------|--------------|--------|
| Cluster_0 | 2241 | 27% |
| Cluster_1 | 3976 | 48% |
| Cluster_2 | 630 | 8% |
| Cluster_3 | 1396 | 17% |

Table 3. Medical data cluster
In Table 3, dataset of disease diagnosis is used with four dominant attributes, Cluster$_0$ is data of disease diagnose for man’s oldest patient with outpatient status. Cluster$_1$ is woman’s oldest patient with outpatient status. Cluster$_2$ is the girl patient with outpatient status. Finally, Cluster$_3$ is the boy patient with outpatient status.

After the clustering process is carried out on the dataset, the clustering data will then be applied to the apriori algorithm. The test results of the K-Means algorithm + Apriori algorithm can be seen in Table 4.

| Cluster$_0$ | Cluster$_1$ | Cluster$_2$ | Cluster$_3$ |
|------------|------------|------------|------------|
| Disease Diagnose | Cataracts, not specified | Cataracts, not specified | Another allergic rhinitis | Post Operation |
| Age Cluster | Elder | Elder | Child | Adult |
| Gender | Male | Female | Female | Male |
| Status of Care | Outpatient | Outpatient | Outpatient | Outpatient |
| Confidence Percentage | 66 % | 66 % | 92 % | 93 % |

As shown in Table 4 above, the combination of the K-Means algorithm and the Apriori algorithm produces more complete and detailed information compared to the results obtained by the application of a priori algorithm only. Comparison of computational time from the Apriori algorithm and the combination of the K-Means + Apriori algorithm can be seen in Table 5.

| Apriori | K-Means + Apriori |
|---------|-------------------|
| Full Data | Computation Time |
| Cluster$_0$ | 21.93 minutes |
| Cluster$_1$ | 4.7 minutes |
| Cluster$_2$ | 8.36 minutes |
| Cluster$_3$ | 1.39 minutes |
| Total Time | 2.96 minutes | 17.41 minutes |

As shown in Table 5, the computation time of the combination of the K-Means algorithm and the Apriori algorithm is much faster than just the Apriori algorithm. It is caused the number of combinations of data in the combination of the K-Means algorithm and the Apriori algorithm is less than the Apriori algorithm. In the combination of the K-Means algorithm and the Apriori algorithm, the data has been focused on each cluster, so there is no excessive combination of data. The results of computational time comparisons can be described in graphical form as shown in Figure 2.
5. Conclusions

After comparing the apriori algorithm and the combination of the K-means + Apriori algorithm on the dataset then the computation time is calculated so that the results of the combination of the K-Means algorithm and the Apriori algorithm are more detailed and complete when compared with the results obtained by applying the Apriori algorithm only. Meanwhile, the computation time of K-Means and Apriori algorithms combinations are faster than the Apriori algorithm, where the total time from K-Means algorithm and Apriori algorithms combinations are 17.41 minutes while the total time of the Apriori algorithm is 21.93 minutes.

References

[1] Han J, Kamber M and Pei J 2012. Data Mining Concept and Techniques 3rd Edition. Morgan Kaufmann USA.
[2] Yanto R, and Khoiriah R 2015. Implementasi Data Mining dengan Metode Algoritma Apriori dalam menetukan Pola Pembelian Obat. Citec Journal (Vol.2) 2:102-113.
[3] Wulandari H N and Rahayu N W 2014. Pemanfaatan Algoritma Apriori untuk Perancangan Ulang Tata Letak Barang di Toko Busana. Seminar Nasional Aplikasi Teknologi Informasi (SNATTI).
[4] Fauzy M, W Saleh R K and Asror I 2016. Penerapan Metode Association Rule menggunakan Algoritma Apriori pada Simulasi Prediksi Hujan Wilayah Kota Bandung. Jurnal Ilmiah Teknologi Informasi Terapan (Vol.2) 2 : 221-227.
[5] Erwin 2009. Analisis Market Basket dengan Algoritma Apriori dan FP-Growth. Jurnal Generic (Vol. 4) 2 : 26-30.
[6] Tan P N, Steinbach M and Kumar V 2006. Introduction to Data Mining.
[7] Alfin T, Santosa B and Barakbah A R 2012. Analisa Perbandingan Metode Hierarchical Clustering K-Means dan Gabungan Keduanya dalam Cluster Data (Studi Kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS). Jurnal Teknik (Vol.1).
[8] Dharshinni N P, Mawengkang H and Nasution M K M 2018. Mapping of medicine data with k-means and apriori combinations based on patient diagnosis. International Conference on Computing and Applied Informatics. Vol 2, Series 978.
[9] Harahap M, et al 2018. Mining association rule based on the disease population for recommendation of medicine need. MECnIT. Vol 1.
[10] Jafarzadeh A K, Mahdavi A and Jafarzadeh H 2017. Evaluation of forest fire risk using the Apriori algorithm and fuzzy c-means clustering. Journal of Forest Science. Vol. 63, Issue No. 8 : 370-380.
[11] Lakshmi K S and Vadivi G 2017. Extracting Association Rules from Medical Health Records Using Multi-Criteria Decision Analysis. International Conference on Advances in Computing & Communications (ICACC). Vol. 7, No.115 : 290-295.
[12] Swathi P and Prajna B 2016. The Effective Procession of Apriori Algorithm Precribed Data Mining on Medical Data. IJCST. Vol. 7, Issue No. 3. ISSN : 0976-8491.
[13] Istrat V and Lalic N 2017. Association Rules as a Decision Making Model in the Textile Industry. FIBRES & TEXTILES in Eastern Europe. Vol.25, Issue No. 4(124):8-14.
[14] Lismardiana, Mawengkang H and Nababan E B 2015. Pengembangan Algoritma Apriori untuk Pengambilan Keputusan. Jurnal Teknologi Informasi dan Komunikasi Vol 4, No.2.
[15] Kuswantoro E and Suprapto Y K 2015. Pemodelan Tingkat Kerja dengan Algoritma K-Means. Jurnal Ilmiah NERO Vol.2, No.1
[16] Metisen B M and Sari H L 2015. Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila. Jurnal Media Infotama Vol.11, No.2.
[17] Yadav R and Sharma A 2012. Advanced Methods to Improve Performance of K-Means Algorithm: A Review. Global Journal of Computer Science and Technology (Vol.12) (9): 47-52.

Acknowledgments
This research was supported by Universitas Prima Indonesia and one of the private hospitals in Medan City, Sumatera Utara.