K-Means and Apriori Algorithm for Pharmaceutical Care Medicine (Case Study: Eye Hospital of South Sumatera Province)

Luthfiah¹, Ken Ditha Tania²*

¹,² Departement of Information System, Faculty of Computer Science, Universitas Sriwijaya, Inderalaya 30662, Indonesia
Email: ¹luthfiahzahra4@gmail.com,²ken_tania@unsri.ac.id
*corresponding author

Abstract. The pharmacy at the Eye Hospital of South Sumatra Province is in charge of managing to range from planning, procurement, storage, preparation, compounding, direct services to patients up to medical supplies circulating and drugs circulating in the hospital and providing medicines for patients based on doctor's prescriptions. All transactions carried out by patients are stored in the hospital system database are put into use at the end of 2017. Transaction data that stored is so many and stacked in the database, data mining can be used to extract data and get information from big data. Data mining techniques which used are K-Means algorithm to divide patient and association with Apriori algorithm to determine frequent itemset on drug data. The result obtained from data mining technique applied to the Pharmacy is intended to fulfill the pharmaceutical care medicine in hospitals.

1. Introduction
The hospital pharmacist should be an expert on medicines who advises on prescribing, administering, and monitoring, as well as a supply manager who ensures that medicines are available through procurement, storage, distribution, inventory control, and quality assurance. The responsibility for establishing policies and procedures related to medication selection, procurement, distribution, and use often lies with the Drug and Therapeutics Committee (DTC) [1].

The pharmacy at the Eye Hospital of South Sumatra Province provides medicines to patients based on doctor's prescriptions, and all of those transactions are stored in the hospital system database are put into use at the end of 2017. Transaction data that stored is so many and stacked in the database, data mining can be used to extract data and get information from big data. The information obtained from data mining in the Pharmacy is intended to fulfill the pharmaceutical care medicine in hospitals.

According to Satibi (2013) planning of pharmaceutical care medicine is an activity that aims to get the type, amount of drug according to need and avoid emptiness based on the pattern of the disease and the frequency of the number of patient visits [2].

Data mining is defined as the process of finding hidden patterns and interesting information, relevant and useful from large data or databases. The data mining process consists of, selecting data, analyzing data, preparing, implementing, interpreting and evaluating results [3].

Clustering is a data mining technique that used to group data based on similarities in data objects or patterns and minimizes similarities to other clusters. Clustering can be used to classify data roughly when they are very complex and mixed or contain too many variables or too many dimensions. In the practice of cluster analysis, no classification variables are included to divide data. Cluster analysis is rarely used alone it is necessary to use other methods to understand the meaning of grouping [2], [4].

Association rule is used to find the relationship among large data sets and how the data can affect the other. Mining frequent patterns is an important aspect in association rule mining. It is useful for discovering relationships among items from large databases [5].

K-Means clustering algorithm is used to classify patient data according to their characteristic values into different clusters so there are several groups of patient data which are then processed using apriori algorithm where is used to identify medicine patterns in each patient category. Combining clustering
techniques using the K-Means algorithm and association using apriori algorithm is intended to produce more specific useful information of the data patterns [2] beside just using only one data mining method so that the utilization of information by stakeholders is more maximal.

2. Related Research
Data mining has been widely used in several fields, one of which is the health sector applied in the hospital industry. Clustering in research [3] is used in the health field to predict heart disease using the K-Means algorithm. Research [4] uses cluster analysis techniques because it has the ability to avoid uncertainty in the analysis of numerical-type data and as a grouping criterion, and to effectively segment data with different group characteristics. K-Means clustering is used to classify data as in research [2] which is used to classify hospital patient data based on the similarity of the characteristics of the data in order to see the types of patients from the hospital.

While apriori algorithms are used to search for frequent itemset, as in the study [2] using apriori algorithms on drug supply systems in hospital pharmacies. While research [6], [7] conducted surveys and development of apriori algorithms.

3. Proposed Method
The method used for this research is Cross Standard Industry Processing for Data Mining (CRISP-DM). This method is the method most referenced and used in the data mining practice methodology. The main objective of the founders of CRISP-DM is to create a standard process model that is not exclusive and freely available for data mining application engineering. CRISP-DM proposes an iterative process flow, with loops that are infinitely defined between phases, and the overall nature of the repetitive cycle of the DM project itself. The results of each phase determine which phase to do next. The six phases of CRISP-DM as follows [8]:

1. Business Understanding. In this phase, an understanding of the project objectives of data analysis and conversion of these requirements is carried out, from the perspective of the subject area, and the problems formulated into the definition of data mining problems. In this phase, the initial plan for achieving the goal is determined, defining the criteria for success.

2. Data Understanding. This phase starts with initial data collection and access to the data set. Data quality problems must be identified and made assuming the beginning of which dataset can be of interest to the next step.

3. Data Preparation. The data preparation phase includes all the activities needed for the final data set. The data preparation phase activity very dependent on the features and quality of the original raw data. Some of the characteristic tasks of data preparation involve selecting tables, projecting attributes and records, attribute transformation, classification, normalization, noise removal, and sampling.

4. Modeling. In this phase, the selection of modeling techniques, algorithms, or suitable combinations is done, wherein this research used a combination of two methods that it K-Means clustering that shows clusters on existing data and association rules with a priori algorithms to show frequent itemset of the cluster which exists.

5. Evaluation. Before the final step of the model, it is important to evaluate carefully and review of the model construction steps also ensure that the data mining extension for the purpose is truly achieved. The final results of this phase - the choice of whether the results of data mining can be used in practical settings.

6. Deployment. It is important that end users anticipate actions that need to be taken to get practical benefits from the data mining model produced. So, in this phase, the resulting modeling is displayed in a form that is easy to understand so that it can be used by interested parties.

3.1. K-Means
K-Means algorithm is one of non-hierarchical clustering method that well-known grouping tool applied in a variety of scientific and industrial applications. K-Means groups data according to their
characteristic values into different clusters. Data categorized into the same group has identical feature values and minimizes variation in one group. The number of groups must be determined first which is a positive integer [3].

The steps of clustering using k-means algorithm are as follows [2]:

1. Prepare training data.
2. Determine the number of Cluster k.
3. Determine the initial value of the centroid (center point of the cluster).
4. Calculate the distance between data and centroid using the Euclidean Distance formula below:

\[ d_{(i,j)} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + ... + (x_{ki} - x_{kj})^2} \]  

(1)

\[ d_{(i,j)} \]: data-i distance to the center of the cluster j
\[ x_{ki} \]: data-i on the attribute data-k
\[ x_{kj} \]: j center point on the attribute data-k

5. Data partitioning based on minimum values.
6. Iterate as long as the data partition is still moving (there are no more moving objects to another partition), if the data still moves then return to point 3.
7. If there is no data to move, then stop the iteration.
8. Data has been partitioned according to the final centroid value.

3.2. Apriori

The Apriori Algorithm is the most classic and important algorithm for mining frequent itemset, proposed by R.Agrawal and R.Srikant in 1994. Apriori is used to find all frequent itemset in a database. The Apriori Algorithm has a number of weaknesses although it is clear and simple. The main limitation is the waste of time storing large numbers of frequent itemset lots of candidates, low minimum support or large itemset [6], [7]. Support shows how much the dominant level of an item is in a transaction while confidence is a value that indicates whether or not the relationship between these items is strong.

To determine the support of each itemset candidate to show how much the dominance level of the entire transaction has used the equation:

\[ \text{Support}(A) = \frac{\text{number of transactions are containing A}}{\text{total transaction}} \]  

(2)

While the support value of the two items is obtained through the equation:

\[ \text{Support}(A \cap B) = \frac{\text{number of transactions are containing A and B}}{\text{total transaction}} \]  

(3)

After the high-frequency pattern is determined, be sought association rules that have greater than the minimum support value requirements by calculating the confidence value of association rules A to B obtained through the equation:

\[ \text{Confidence} = P(B|A) = \frac{\text{Support}(AB)}{\text{Support}(A)} \]  

(4)

4. Discussion

The combining clustering techniques using the K-Means algorithm and association using apriori algorithm is intended to produce more specific useful information of the data patterns [2] beside just using only one data mining method. The K-Means algorithm is used to separate the data into several clusters or groups that have the same characteristics as the patient data object, then apriori algorithm is used to see frequent itemset from the data that has been divided where the object is drug sales data based on the category of patients who have in the previous cluster.
After determining the objectives and targets to be achieved, data understanding phase and data preparation phase is carried out. The data that has been prepared will look like in Figure 1, data of patients have to through the transformation process by representing non-numeric data into numeric data as in Figure 2 so that the data is appropriate and can proceed to the next process.

4.1. K-Means
After all Eye Hospital Pharmacy data is collected and prepared, the data will be grouped using the K-means Clustering algorithm by measuring the distance using equation (1). It is necessary to take several steps to measure the distance of each object to the center point of the cluster and divide the data into several groups based on the smallest distance to the center point of the cluster. The number of clusters will be divided and evaluated using Davies-Bouldin Index (DBI) with the following formula:

$$DB = \frac{1}{K} \sum_{i=1}^{K} Ri, qt$$

Where $Ri,qt = \max j, j \neq 1 \left[ \frac{S_i + S_j}{d(i,t)} \right]$. $K$ is a lot of clusters. The results of the cluster evaluation in Table 1 show that the K-means algorithm with 4 clusters obtains with smaller DBI value of 0.06157. With smaller DBI values, the cluster scheme is more optimal.

| Algorithm | Clusters | DBI     |
|-----------|----------|---------|
| K-Means   | 3        | 0.07075 |
| K-Means   | 4        | 0.06157 |

Then determine the center point of the cluster by calculating quarter of the data and obtaining the center point which can be seen in Table 2. After determining the center point of the cluster, calculating the distance of each data to the center point of the cluster using equation (1). The data is decided to enter a particular cluster seen from which one has the closest distance to the center of the cluster.

| Column | Data | Poly |
|--------|------|------|
| No     | 1    | 2    |
| Reg    | RJ18021735 | 4 | 1 | C1 |
| Number | RJ18022106 | 5 | 3 | C2 |
| Gender | RJ18021927 | 7 | 1 | C3 |
| Age    | RJ18022088 | 5 | 1 | C4 |
Cluster 1 was dominated by elderly patients with the youngest age of 71 years and the oldest age of 97 years came from Lens Poly, General Poly, and Refraction Poly. Cluster 2 is also dominated by elderly patients but have different ages, namely 52-68 years from more diverse polys such as Lens Polymers, General, EED, Glaucoma and Refraction Poly. Cluster 3 was dominated by patients aged 31-49 years with most patients in cluster 3 coming from EED Poly. In cluster 4 the patients in this cluster are dominated by children up to adolescents aged 1 to 30 years with patients who come from diverse poly areas such as pediatric poly and EED.

4.2. Apriori
After clustering the data, analyzing the data with apriori algorithm on sales begins by selecting and cleaning the data to be analyzed, then searching for all types of medicine name items in the sales transaction list, then searching for the number of items in all transaction data sales based on previously divided clusters.

Table 3. Information on types of medicine in cluster 1 that meets minimal support

| Medicine                     | Support | Support (%) |
|------------------------------|---------|-------------|
| Polidemisin ED               | 3       | 43          |
| Cendo Lyteers ED             | 2       | 29          |
| Citicoline 500 mg Bernofarm  | 1       | 14          |
| Cendo Catarlent ED           | 2       | 29          |

Table 4. Information on types of medicine in cluster 2 that meets minimal support

| Medicine                     | Support | Support (%) |
|------------------------------|---------|-------------|
| Cendo Timol 0,5% ED          | 2       | 13          |
| Cendo Lyteers ED             | 4       | 25          |
| Cendo Protagent-A MDS ED     | 2       | 13          |
| Cendo Xytrol ED              | 2       | 13          |
| Cendo Eyefresh Mild ED       | 2       | 13          |
| Cendo Catarlent ED           | 5       | 31          |

Table 5. Information on types of medicine in cluster 3 that meets minimal support

| Medicine                     | Support | Support (%) |
|------------------------------|---------|-------------|
| Cendo LFX ED                 | 2       | 10          |
| Cendo Lyteers ED             | 4       | 20          |
| Cendo Cenfresh ED            | 3       | 15          |
| RG Cholin Tab                | 3       | 15          |
| Vigamox ED                   | 3       | 15          |
| Cendo Eyefresh Mild ED       | 4       | 20          |
| Cendo Protagent-A MDS ED     | 3       | 15          |
Table 6. Information on types of medicine in cluster 4 that meets minimal support

| Medicine              | Support | Support (%) |
|-----------------------|---------|-------------|
| Rg Cholin Syr         | 3       | 25          |
| Cendo LFX ED          | 2       | 17          |
| Cendo Vernacel ED     | 2       | 17          |
| Cendo Proagent-A MDS ED | 2   | 17          |

Formation of a combination pattern is based on the minimum support value. If the minimum support value is fulfilled and the itemsets combination pattern is more than one, then the itemsets combination pattern that follows can be formed, the minimum support value is 0.1 or equal to 10 percent (%). Table 3 shows the form of medicine data from cluster 1 which consists of all types of medicine that are in the transaction. In Table 3 all data items meet minimal support then continued the formation of frequency patterns with two itemsets formed from data items that meet the minimum support. The formation of a frequency pattern with two itemsets formed from data items in cluster 2, cluster 3, and cluster 4 does not meet the minimal support, so the formation of frequency patterns in cluster 2, 3, and 4 stops at one itemset.

4.3. Evaluation and Deployment
Data mining modeling uses a clustering technique with K-Means algorithm and association rule with apriori algorithm has been produced as in the previous discussion. Testing is carried out on modeling whether the modeling is in accordance with the objectives and the modeling is displayed in a form that is easy to understand. Here are some views that are the result of the deployment phase. The patient page displays a list of patients based on clusters or groups that have been previously calculated and the top medical page displays the results of the apriori algorithm that shows the most medicine sales based on the cluster, displayed in bars and tables as an explanation of the bars.

5. Conclusions
The Eye Hospital of South Sumatra Province has 4 groups or data clusters with different characteristics, the main difference can be seen from the age division where cluster 1 consists of patients aged 70 years and over, cluster 2 consists of patients aged 50 years to 68 years, then patients in cluster 3 are approximately 31 years to 49 years old and cluster 4 has patients aged 1 to 30 years. From the result, we can also see the disease patterns from patients in Eye Hospital of South Sumatra Province based on the polyclinic intended by patients, such as Pediatric Poly is only owned by cluster 4 patients or most of the patients in cluster 3 visit EED Poly.

The apriori algorithm results from the clusters also become more specific, it can be seen that the types of drugs that meet the minimum support of each cluster are different, e.g. Polidemisin ED with the support value of 43% is only in cluster 1 and not in other clusters. Also, cluster 4 is dominated by medicine with lower doses and only be used by children to adolescents.
References

[1] MSH, “Chapter 45: Hospital pharmacy management,” 2012.

[2] N. P. Dharshinni, H. Mawengkang, and M. K. M. Nasution, “Mapping of medicine data with k-means and apriori combinations based on patient diagnosis,” J. Phys. Conf. Ser., vol. 978, no. 1, 2018.

[3] B. Sundar V, T. Devi, and N. Saravanan, “Development of a Data Clustering Algorithm for Predicting Heart,” Int. J. Comput. Appl., vol. 48, no. 7, pp. 8–13, 2012.

[4] W. T. Lin, Y. C. Wu, J. S. Zheng, and M. Y. Chen, “Analysis by data mining in the emergency medicine triage database at a Taiwanese regional hospital,” Expert Syst. Appl., vol. 38, no. 9, pp. 11078–11084, 2011.

[5] M. Bhargava and A. Selwal, “Association Rule mining using Apriori Algorithm : A Review,” vol. 4, no. 2, p. 2013, 2013.

[6] A. Bhandari, A. Gupta, and D. Das, “Improvised apriori algorithm using frequent pattern tree for real time applications in data mining,” Procedia Comput. Sci., vol. 46, no. Icict 2014, pp. 644–651, 2015.

[7] K. S. Kumar and R. M. Chezian, “A Survey on Association Rule Mining using Apriori Algorithm,” Int. J. Comput. Appl., vol. 45, no. 5, pp. 47–50, 2012.

[8] O. Niakšu, “CRISP Data Mining Methodology Extension for Medical Domain,” Balt. J. Mod. Comput., vol. 3, no. 2, pp. 92–109, 2015.