Mapping of medicine data with k-means and apriori combinations based on patient diagnosis

N P Dharshinni*, H Mawengkang** and M K M Nasution***
1Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan 20155, Indonesia
2Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sumatera Utara, Medan 20155, Indonesia
3Department of Information Technology, Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan 20155, Indonesia

*Email: 1priyadharshinnainaidu@gmail.com, 2hmawengkang@yahoo.com, 3mahyuddin@usu.ac.id

Abstract. Medicine is one of the items needed by sick society, the high influence of medicine on service and the economy in hospitals, requires mapping and planning the optimal need for medicines according to the conditions, because 50% -60% of hospital income is sourced from medicine sales. The purpose of this study was to find patterns of doctor’s prescription medicine association with sales data using an apriori algorithm based on data grouping using a k-means algorithm. The results of the experiments show that medicine prescription data with medicine sales have significant differences so that the data can not be used as materials for medicine planning, this is due to some indication of one of the unavailability of medicine caused by mapping inaccuracy so that the planning of medicine requirements is not optimal. The results of this analysis can be used as input materials in decision making, so the planning needs of medicines can be in accordance with the development of patient disease patterns.

Keywords apriori, k-means, Medical Prescription, medicine

1. Introduction

Hospital pharmacy service is one of the support services and also serves as the center of the main income, because 90% of health care use of pharmaceuticals (medicine, chemicals, radiological materials, materials medical devices consumables, medical equipment, and medical gas ), and 50% of all income derived from the management of pharmaceutical supplies [1]. The high contribution to the service of the hospital pharmacy, requires a careful and full management responsibility in the medicine plan [2]. Planning pharmaceutical care medicine is an activity which aims to get the type, amount of medication as needed and avoid the emptiness that is based on patterns of disease and the frequency of the number of visits patients [3]. Medicine is one of the items needed by a sick society [3]. The determination of medicine based on the results of the doctor’s diagnosis that is written in the prescription, and all patient diagnostic data stored in the database, one method that can extract large amounts of data to obtain useful medicine information in planning based on patient’s medical data is to apply data mining.

Data mining is a multidisciplinary field useful in finding and defining patterns from data sets in databases into information[4]. Data mining is widely applied in the medical field, including medicine prediction, heart disease and the discovery of clinical and pathological data relationships. Data mining is an efficient method for detecting some types of problems to pinpoint important patterns in a set of data, one of which is Frequent Itemsets Mining (FIM). Frequent Itemsets play an important role in many data mining problems that try to determine important patterns from the database[5]. Most data mining problems are grouped into two, namely (1) candidate generation using Breadth First Search (BFS) algorithm and (2) pattern growth with Depth First Approach (DPA) approach[6]. Depth First Approach (DPA) is considered more complicated because it uses data structures in its research[7],
while Breadth First Search (BFS) is more widely used because it is more simple, one of them by using Algoritma Apriori[4].

Apriori algorithm is the most classic and quite important algorithm in Frequent Itemsets Mining (FIM). Although many similar algorithms are developed more efficiently, such as FP-Growth, LCM, and so on [8]. Apriori algorithm can reduce the number of candidates that must be calculated support by pruning, so it has good performance [6] but has a long period of computation due to the scanning process that must be done at every time iteration.

Clustering is one technique for grouping objects or patterns in a set (cluster) so that each cluster has similarities to each other [9]. FCM is one of the cluster algorithms that is quite optimal in the grouping but requires a long computation time [10]. FCM is very strong against noise [11] but is too sensitive to cluster initialization [12]. K-Means Clustering is an efficient algorithm for grouping large and fast datasets in computation [10] but is limited to numerical data [13].

In this research, an apriori method is proposed to conduct medicine mapping based on doctor's prescriptions sourced from diagnosis results, this doctor's prescribing as a source of mapping for the correlation of diagnosis result with medicine as useful information in medicine planning. This is because of the changing pattern of illnesses, there are still doctors who make prescriptions beyond standard set by the Pharmacy and Therapy Committee, so that the purchase of medicine to the outside pharmacy or unserviceable prescriptions, especially for inpatients and outpatients, due to the unavailability of the medicine. The analytical framework uses the data grouping method with the k-means algorithm to apriori, then we analyze the influence of grouping on the computation process and the accuracy of the mapping of medicine sales relation with doctor's prescription. This paper is presented as follows: Chapter 2 related research and Chapter 3 Proposed method. Chapter 4 discussion and Chapter 5 conclusion.

2. Related Research
Apriori data mining algorithm is the most commonly used method for extraction of data in large data with association rules in determining associational connectivity in item sets, each relationship can be measured using two different units of support value and confidence level. Any relationship that meets the minimum threshold is called the k-itemset, the level of processing in the apriori algorithm is determined the number of candidate sets with the set produced at each level so that the algorithm takes a high computation time, for optimization [14] proposes a combination of tree based frequent. Research of Li Wei et al [15], developed an apriori optimization algorithm for medical big documents based on MapReduce, WEN-JING ZHANG et al [12], apply an apriori algorithm for the health field with the purpose of breast cancer data classification so that doctors can be more efficient and accurate in diagnosing patients. V R Sadasivam et al [16], implemented an apriori algorithm for criminal pattern data mining, Xiaoyan Cui [17], applied an apriori algorithm for large data mining on medical and cloud computing with the Hadoop framework.

K. Zuhtuogullari [18], propose for the mining of health data to find the pattern of illness that often occurs in patients by seeking symptom relation with disorders in the medical database. Yan Yuguang et al [19], research developed a multi-model based on apriori algorithm at the hospital to extract data from the database to produce useful information in medical decision-making, Miao Wang et al [20], applies to the health field to discover the characteristics of headaches in traditional apriori-based treatments, making it easier for doctors to decide on prescriptions of various headache sufferers. Wang Renli [21], proposed the Apriori-BSO algorithm for the mining of health data, with the aim of finding data associations and symptoms of medical records, medicine combination data, and the relationship between symptoms and treatment.

3. Proposed Method
3.1. Apriori algorithm
The apriori algorithm consists of several stages of iteration in the database[4], each iteration will generate a calculated frequency pattern by scanning the database to obtain the support for each item,
items that have support above the minimum support selected into the frequency pattern high with one length or often called 1-itemset. K-itemset is a term for a set consisting of k items. In the second iteration process will produce 2-itemset which each set has two items[5]

3.2. K-Means Algorithm
The K-Means algorithm determines the cluster of objects based on the attributes/features of the object into the cluster (K). The groups are distinguished by their centroid. This algorithm is very sensitive to placing the initial value of the centroid. The determination of cluster number (k) partition is very important in this algorithm, but there is no provision to determine the number of clusters (k) formed [22]. The steps in the K-means algorithm as follows:
1. Prepare training data.
2. Set the value of K cluster.
3. Set the initial value of centroids.
4. Calculate the distance between the data and the centroid using the formula (Euclidean Distance).
5. Partition data based on the minimum value.
6. Then do the iteration during the data partition is still moving (no more objects moving to another partition), if still moving to another partition then go back to point 3.
7. If the data group is now the same as the previous data group, then stop the iteration.
8. Data has been partitioned according to the final centroid value.

3.3. Dataset
Dataset sourced from government hospitals and private hospitals with data of 2015 and 2016, these two sources of data will be used as a material for analysis and comparison.

3.4. Tool
In this research, we apply 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, an apriori algorithm is applied to locate an item set of a set of data from the database of medicines in the hospital, but after the data are grouped by k-means algorithm. In Figure 1 is an illustration of the research framework we are proposing. Based on the results of our observation, mapping medicine use based on sales data of general pharmacy or pharmacy BPJS, because of that, we will make this as a reference in this case. We used two approaches, the first approach to apply grouping sales data and then applied to the apriori algorithm, while the second approach we apply grouping in medical data of the patient and then applied to the apriori algorithm.

Figure 1. Research framework
In general, the process of medicine sales in hospitals sourced from the results of doctor's diagnosis written on the prescription and the patient brought the prescription to the pharmacy, so the doctor's prescription will be listed in pharmacies as medicine sales, but sometimes patients do not order at the hospital pharmacy because the availability of medicines at the pharmacy does not match the progression of the patient's illness. Our main objective of this study was to look for the relationship of patient prescription data to the sale of medicine, to serve as inputs in the medicine planning adjusted based on the patient's disease pattern.

4. Discussion
This study uses two different data sources, but there are linkages that need to be extracted to produce useful information in the decision making of medicine requirements planning. The first source is based on sales data with a total of 14,509 transactions consisting of 2,165 types of medicines and the second source of medical / prescription data totaling 14,509. Both of these sources will be applied to the apriori algorithm after the grouping of k-means data. in table 1 and table 2 is the result of grouping of two data sources.

Table 1. Cluster Sales Data

| #of Instance | ratio % |
|-------------|---------|
| Cluster₀    | 95      | 84%     |
| Cluster₁    | 76      | 16%     |

Table 2. Cluster patient medical data

| #of Instance | ratio % |
|-------------|---------|
| Cluster₀    | 101     | 32%     |
| Cluster₁    | 76      | 30%     |
| Cluster₂    | 54      | 13%     |
| Cluster₃    | 45      | 20%     |
| Cluster₄    | 39      | 5%      |

In Table 1 The first approach uses a sales data source with four dominant attributes, cluster₀ is a medicine sale for inpatients, while emergency patients are the most sources of medicine sales in this cluster. Cluster₁ patient age is dominant between 50-60 years old who is sourced from an outpatient with lung disease and tuberculosis, while patient age between 0-7 year old comes from children's health with outpatient status with cough disease. In table 3 the result of association rules first approximation.

Table 3. Association rules related to Disease data (Min Sup: 20%, Min Conf: 65%)

| Rules | Sup. | Conf. | Lift | Consequent  | Antecedent    |
|-------|------|-------|------|-------------|---------------|
| R1    | 0.45 | 0.81  | 1.35 | Emergency Services | Inpatient    |
| R2    | 0.31 | 0.79  | 1.47 | Diagnosis    | Cataract      |
| R3    | 0.23 | 0.85  | 1.23 | Diagnosis    | Age = 50-60 |
| R4    | 0.26 | 0.77  | 1.17 | Children's Health | Cough        |

In Table 3 the minimum limit of Support 20% and confidence 65% with four extraction results is called R1, R2, R3, and R4. The result of the analysis shows that emergency service is very dominant to conduct medicine sales transaction of pharmacies for inpatients, while outpatients who buy medicines are dominated from child health.
Table 4. Cluster General Model

| Attribute | Full Data (14,509) | Cluster0 (95) | Cluster1 (76) |
|-----------|-------------------|---------------|---------------|
| Group     | Emergency Services| Emergency Services | Children's Health |
| Diagnosis | Cataract          | Cataract      | Cough         |
| Age       | Age = 50-60       | Age = 50-60   | Age = 0-7     |
| Status of care | Inpatient       | Inpatient    | Outpatient    |

The results of the analysis in Table 4 of the dataset were sourced from sales with 14,509 data, in Cluster0 there were 95 data that were successfully grouped with the source of the patient from emergency services, the dominant diagnosis was a cataract, with patient age 50-60 and patient is an Inpatient. In Cluster1, the dominant patient is an outpatient with a 0-7 year-old age who is sourced from Child's Health with a Cough disease diagnosis. Therefore, in medicine planning, consideration should be given to increasing the stock of medicines with categories for children age 0-7 year-old, because the highest sales are sourced from the Child Health for outpatients. The second approach uses a patient medical data source with four rules used in sales data, the results are shown in Table 5.

Table 5. Associated rules related to patient medical data (Min Sup: 20% Min Conf: 65%)

| Rules | Sup. | Conf. | Lift | Consequent | Antecedent |
|-------|------|-------|------|------------|------------|
| R1    | 0.75 | 0.92  | 1.75 | Internal disease | Inpatient |
| R2    | 0.43 | 0.85  | 1.68 | Diagnosis   | THT        |
| R3    | 0.56 | 0.89  | 1.45 | Diagnosis   | Age = 55-75|
| R4    | 0.89 | 0.83  | 1.58 | Children's Health | Diarrhea |

In Table 5 the results of dataset extraction from doctor's prescriptions amounted to 14,509 prescriptions, a minimum limit of support 20% and 60% confidence. The results of this second approach analysis were dominated by inpatients from Internal Disease with age 55-75 years, the most dominant disease was THT, while outpatient was dominant from polyclinic of Child Health with a diagnosis of the dominant disease was Diarrhea.

Table 6. Cluster General Medical Model

| Attribute | Full Data (14,509) | Cluster0 (101) | Cluster1 (76) | Cluster2 (54) | Cluster3 (45) | Cluster4 (39) |
|-----------|-------------------|---------------|---------------|---------------|---------------|---------------|
| Group     | Internal Disease  | Internal Disease | Emergency Services | Children’s Health | Internal | Children’s Health |
| Diagnosis | THT               | THT           | THT           | Diarrhea      | Cataract     | Cough         |
| Age       | 50-60             | 50-60         | 50-60         | 0-7           | 60-75        | 0-7           |
| Status of care | Inpatient           | Inpatient     | Outpatient    | Inpatient     | Inpatient    | Outpatient    |

The second approach to the data sourced from the medical records of patients there are differences in the association rules as shown in Table 6. In Cluster0, the dominant disease group data is a group of internal diseases, with an THT diagnosis, the dominant age of the disease is between 50-60 and the status of inpatients. Cluster1 is sourced from emergency services, cluster2 child health is more dominant with age 0-7 is inpatient. Cluster3 includes cataract disease with patient age 60-75 is inpatient, while the cluster4 dominant health of children with cough disease is outpatient.
5. Conclusions

Optimal medicine planning based on the pattern of disease progression in the health sector is one of the things that greatly affect medical services and economics. The results of the mapping from the sales data and medical data of patients, we observe there are significant differences, where 35% of doctors research data on patients does not occur sales transactions, this is one good input for decision-making especially medicine planning. The application of k-means clustering method to apriori gives a significant influence in determining association rules and computation time. For further work In mapping data with data mining, it is necessary to compare with other methods to find more accurate methods in prediction, this will be useful in decision making.

References

[1] Rohmani S, Fudholi A and Hakim L 2016 Analisis Faktor Internal-Eksternal terhadap Pengolahan Obat di Instalansi Farmasi RSUD dr.Moewardi Surakarta Journal of Pharmaceutical Science and Clinical Research, 01, 10-20
[2] Hamid T B 2005 Elemen Pelayanan Minimum Farmasi di Rumah Sakit, Direktorat Jendral Pelayanan Kefarmasian dan Alat Kesehatan, Departemen Kesehatan RI, Jakarta
[3] Satibi 2014 Managemen Obat di Rumah Sakit, Fakultas Farmasi, Universitas Gadjah Mada Yogyakarta
[4] Rao S and Gupta P 2012 Improved Algorithm. Over APRIORI Implementing Data Mining Association Rule Algorithm IJCST Vol. 3, Issue 1. 489-493
[5] Goswami D N, Anshu C and Raghuvanshi C S 2010 An Algorithm for Frequent Pattern Mining Based On Apriori (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 04, 942-947
[6] Patel B, et al 2011 Optimization of Association Rule Mining Apriori Algorithm Using ACO. International Journal of Soft Computing and Engineering (IJSECE) Vol.1, Issue 1. 24-26. ISSN: 2231-2307
[7] Kumar S S, et al 2012 Aprori M. Improved Algorithm Based on a bottom up approach using probability and Matrix IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No. 3. 242-246
[8] Kumar K S and Chezian R M 2012 A Survey on Association Rule Mining using Apriori Algorithm International Journal of Computer Applications (0975-8887), Volume 45- 5. 47-50
[9] Izakian H and Abraham A 2011 Fuzzy C-Means and fuzzy swarm for fuzzy clustering problem Expert Systems With Applications, 38 (3), 1835-1838
[10] Ghosh S and Dubey K S 2013 Comparative Analysis of K-Means and Fuzzy C Means Algorithms. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 4, 2013
[11] Hamasuna Y and Endo Y and Miyamoto S 2009 On tolerant fuzzy c-means clustering and tolerant possibilistic clustering. Soft Computing, 14(5), 487–494
[12] Zhang W- J, Ma D- L and Dong B, et al 2012 The Automatic Diagnosis System Of Breast Cancer Based On The Improved Apriori Algorithm International Conference on Machine Learning and Cybernetics. Xi’an, 15-17. 63-66
[13] Bai L, Liang J and Dang C 2011 An initialization method to simultaneously find initial cluster centers and the number of clusters for clustering categorical data Knowledge-Based Systems, 24 (6), 785-795
[14] Yamuna N Y, Devi N and Devi J S 2016 A combinatorial tree-based frequent pattern mining Journal of Theoretical and Applied Information Technology 781-789. ISSN: 1992-8645. E-ISSN: 1817-3195
[15] Wei L, et al 2016 Optimization and Application in Medical Big Document-Data of Apriori Algorithm based on MapReduce International Conference on Computer Communication and Informatics 978-1-4673-6680-9
[16] Sadasivam V R, Duraisamy K and Bharathi R M. 2015 Association Rule Mining and Frequent Pattern Mining Applications on Crime Pattern Mining: A Comprehensive Survey International Journal of Innovative Research in Science, Engineering and Technology Vol. 4, Special Issue 6. 228-236. ISSN (Online): 2319-8753
[17] Cui X, Yang S, and Wang D 2016 An Algorithm of Apriori based on Medical Big Data and Cloud Computing. Proceedings of CCIS 361-365
[18] Zuhtuogullari K and Allahverdi N 2011 An Improved Itemset Generation Approach for Mining Medical Databases, IEEE. 39-43. 978-1-61284-922-5
[19] Yuguang Y, Chunyan W and Min L 2011 Application of the Model_Multi based on Apriori algorithm in a Supporting System of Medical Decision International Conference on Measuring Technology and Mechatronics Automation 566-569. 978-0-7695-4296-6, DOI 10.1109 / ICMTMA.2011.144
[20] Wang M, et al 2014 The Application Characteristics of Traditional Chinese Medical Science Headache Treatment on Data-Mining Based on Apriori Algorithm IEEE International Conference on Bioinformatics and Biomedicine 153-157
[21] Renli W, Yueming D and Liming D 2016 The Application Of Apriori-BSO algorithms in Medical Records Data Mining. 827-832
[22] Zhang C and Fang Z. 2013 An Improved K-means Clustering Algorithm Traditional K-mean Algorithm Journal of Information & Computational Science, 1, 193-199