ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse69922020.pdf
https://doi.org/10.30534/ijatcse/2020/69922020

Hybrid Recommendation System with Clustering and Classification Method Based on Case of Clinic XYZ

Febrillian Pemisindo[1], Nilo Legowo[2]
[1] Information Systems Management Department, Binus Graduate Program-Master of Information Systems Management, Bina Nusantara University, Jakarta Indonesia 11480, febrillian.pemisindo@binus.ac.id,
[2] Information Systems Management Department, Binus Graduate Program-Master of Information Systems Management, Bina Nusantara University, Jakarta Indonesia 11480, nlegowo@binus.edu

ABSTRACT

With the development of IT systems in many areas, all data that generated hold potential value that waiting to be unlock, one way is to do product recommendation system based on it. Clinic XYZ is one of the largest beauty aesthetic services in Indonesia. After implementing its ERP systems, Clinic XYZ wants to gain more benefit by doing cross sell of their product to existing customer. Product recommendations system is the solution that developed for it. Product recommendation system is an information-filtering system that handles overload of information and give a recommendation to user a recommendation-based preference, interest of user behavior that observed form those data. The common use algorithm for product recommendation system is collaborative filtering and done directly to whole data. This research applies 2 steps of algorithm, clustering, to separate the customer into different cluster based on the master data that available and classification to each cluster for product recommendation in the clinic to the existing customer. The result show that the 2 steps method generate better result.

Key words: classification; clustering; data mining; product recommendation; recommendation system

1. INTRODUCTION

XYZ Clinic is a clinic whose line of business is specializes in beauty aesthetic services. The clinic has just migrated its ERP and CRM system to a new version. All transaction and customer information already recorded using the new system. Historical transaction also migrated. The ERP and CRM system also already used further by doing some analysis leveraging integrated BI system. As the variety of products available at the XYZ Clinic is many, the management of the clinic want to do sell the other relevant product on the existing customer to raise the transactions, or what is known as cross sell. One of the solutions to do cross selling effectively would be using data mining technique to do the product recommendation on the targeted customer.

Data mining technique has been used in many e-commerce application to do product recommendation to customer, one of it is product recommendation system [1]. Product recommendation system provide recommendation to the customer about the product that maybe suit their requirement [1], and in a sense act like an individual one-to-one marketing strategy [2]. Product recommendation system is an information filtering system that handles the overload of information by filtering important information generated from large volumes of information that is generated automatically based on preferences, interests, or user behavior observed about items [3].

Recommendation system already used in many area such as videos recommendation [4], movies recommendation [5], e-commerce product recommendation [6], online food stores recommendation [7], music recommendation [8], or cosmetics recommendation [9]. These recommendation systems keep being implemented because they are beneficial for service providers and customers [10]. The usage of recommendation system make the transaction cost of buying process in online shop environment less and also help to improve the process and the quality of decision making [3]. Recommendation systems generally have several benefits, one of them is increasing the cross-selling likelihood; produce customer loyalty and meeting customer requirement by recognizing goods they may be interested in buying [11].

Many methodologies already used for recommendation systems. The common used one are collaborative filtering (CF) or content-based filtering[11]. Another research use classification method as the base for its recommendation system and show that its usage is better result that using collaborative filtering [12]. In addition, one of research is trying a different approach to this by doing clustering before applying the association rules [9]. This research develops the recommendation method using classification method combined with clustering as the first step.

2. LITERATURE REVIEW

As mentioned previously, recommendation system already used in many area such as videos [4], movies [5], e-commerce [6], online food stores [7], music [8], or cosmetics [9] and the benefit that come from this system implementation is good, either for user or company. Many methods can be used for this type of system, with content-based filtering and collaborative filtering are the common one used. [3]

Content-based Filtering (CBF) technique is a algorithm that domain-dependent and consider more on the analysis of the item attribute to do the prediction. [3] .This method provides recommendations by matching profiles of customers with features of content (e.g. products' attributes). [11]. This method works well when recommend web pages, publication or news. In addition, because this method emphasizes on the product's attribute, the profile of the user is not needed to make the recommendation decision. Because all of recommendation is based on the product. By leveraging this, theoretically this method will work well even the data does not contain any user information. And also when there's changes on the user information, the adjustment of this method will relative faster than other method that consider user information in its recommendation. [3] The shortage of this method is that, it will require detailed information about the products as much as possible, if not, the recommendation would not be accurate, this is called limited content analysis [3]. But if the content is too much and too specialize, CBF also can't use it well [13]. Amazon is sample of the one that used this method [11]

Collaborative filter (CF) is a prediction technique that for product or items that cannot explained by only metadata, like movies or music [3] The CF method uses preference ratings given by various customers to determine recommendations based on the opinions of other similar customers for a target client. This technique divided into two categories: memory-based and model-based. Memory based can be achieved by two ways, user based or item-based techniques. Used based will compare the similarity based on the user and product profile. While the item based, will compare product to product, and not the similarity of the user. The model based will use the ratings or transactions that already happened to learn a model that will make the prediction performance better. The process of building the model can be done by leveraging the technique of data mining or machine learning. These techniques can recommend some group of item because they utilize model that pre-computed, and it also proved that the recommendation results that has been generated are like the neighborhood-based recommender techniques. [3].
CF has some advantage over CBF, especially regarding the not needed information of the product attributes. Also when there's difficulties to understand the products, such as news or opinions. But CF also has disadvantages, like cold-start problem, when this techniques can't give accurate prediction if there's no records / transaction yet [3]. Some sample that use this algorithm are Ringo and MRS for music recommender and siteseer for web page recommendation [11].

There's research has develop a recommender system that built for cosmetic business [9]. The research done by developed a personal recommender system that leverage collaborative filtering, content-based, and data mining techniques. Another research combines hybrid approach to recommendation system [11]. Another research also use classification method as the base for its recommendation system and show that its usage is better result that using collaborative filtering [12]. In addition, one of research is trying a different approach to this by doing clustering before applying the association rules. Based on those, this research does the recommendation method using CF techniques, specifically the decision tree method combined with clustering as the first step. This research does analysis of the customer at first step by consider customer master data attribute before doing the classification. This research also compares both results.

3. RESEARCH METHODOLOGY

3.1 Dataset Collection

Historical data is used in this research. All customer master data and transaction data is in the system and extracted by using some SQL procedures. Total of 11.498 data is collected based on 1-year transactions that is combination of customer master data and transactional data.

The first process, which is clustering the customer, will be using customer master data without considering the transaction history. The second process will use both customer master data and transaction data as consideration. After taking out sensitive information from the customer master data, the data is used for clustering process.
Table **1** describe the customer master data that used in this research.

Table 1 : Customer Master Attribute Description

Attribute	Description	Used in Model
ID Global	ID customer	No
ID Klinik	ID Clinic	Yes
Name	Name of the Customer	No
Gender	Gender of the customer	Yes
Membership Class	Membership level of customer	Yes

Frequency Transaction	how frequent the customer makes a transaction	Yes
Iklan/Referal Info	from which channel customer know the clinic	Yes
Age	Age of the customer	Yes
Kewarganegaraan	citizenship of customer	Yes

The transaction data would be consisting of the name of the products that has been purchased by customer. Every purchased product would have "Yes" status. Vice versa, product that have not purchased would have status "No". Table 2 **:** Product Transaction List would be show the fields that used in this research.

Table 2 : Product Transaction List

Product Code	Product Description
Product1	Medis ultrasound rejuvenation
Product2	Medis ultrasound eye
Product3	Medis ultrasound acne
Product4	Medis ultrasound body
Product5	Medis radio frequency rejuvenation
Product6	Medis laser and light body
Product7	Ampul rejuvenating
Product8	Moisturizer rejuvenation
Product9	Moisturizer acne
Product10	Masker whitening
Product11	Facial rejuvenating
Product12	HDS
Product13	Obat racik
Product14	Body care
Product15	Topical acne

3.2 Recommendation methods

The product recommendation system is build based on the historical data and will be processed by 2 algorithm, first is clustering to segment the customer based on the attributes, and the second one J48, to predict the possibility of the customer buy the product based on the customer attribute and the products that previously the customer buys. The process flow of the recommendation method shown in Figure 1

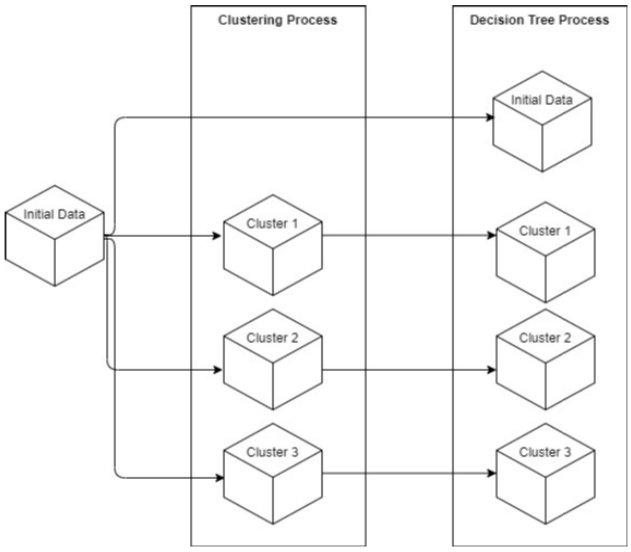

Figure 1 : Process Flow of Hybrid Recommendation Method

3.2.1 Clustering – K Means

Clustering algorithm divide a group of data into some sub-groups cluster of data to get the meaningful information that exist in it [14]. When the groups already divided, the assessment of the customer in a group can be calculated and can be utilized to give recommendation for each individual data in the group. In an collaborative based algorithms, this technique can also be utilize to lessen the candidate set [3].
One of the most utilized among other clustering methods is k-means. K-means will receive a parameter as an input, and will divide the data into certain K number of clusters [3]. There's a research that the k-means used to divide the cluster based on the available attribute, using k-means, underlying patterns to determine the clusters with maximum result [15]. Another research used k-means to do document clustering and the result show the cluster that generated by k-means produce good result[16]. Based on those, this research will use K-means as the clustering method.

3.2.2 Classification – Decision Tree

Classification is a method to generate a mode or function that can define and differentiate concepts or classes. The model that generated is expected to can be utilized to forecast the objects or classes that didn't have label yet. The generated model always generated based on the training data analysis (i.e data of classes or objects that the labels already known). This generated model usually visualized in different formats, like decision trees, classification rules, artificial neural network or mathematical formulas. In many cases, users want to predict data values that are not available or missing (not the label of the class). In this case usually the value of the data to be predicted is numeric data. This case is often referred to as prediction. In addition, predictions emphasize the identification of trends from distributions based on available data. [17]

One of the methods that commonly used in classification is decision tree. In Weka, the algorithm is developed based on C4.5 algorithm by Ross Quinlan. In Weka, this method called as J48. The decision that trees produced with this algorithm can be utilized to do classification. One of the research use decision tree to classify the soil type based on the composition in it [18]. Another research compares the method of naïve bayes and decision tree. This research uses the method to predict the chance of loan payment successfully paid. The result show that decision tree has the best result, although take more time to process [19]. Based on those, this research will use decision tree to do the prediction.

4. RESULT AND DISCUSSION

This research used WEKA as the application for data mining. WEKA is an application that based on JAVA. Weka consists of some number of machine learning algorithms that can be leveraged to generalize / formulate a set of sampling data. This algorithm can be directly used into a dataset or can also use another java code for custom process. Weka has built-in tools for many processes for data mining. This is including such process as data pre-processing, regression, clustering, association rules, classification, and even some visualization. They organize classes into packages and each class in the package can reference other classes in other packages. The reason to use WEKA is that no other additional package needed.

The process flow that done shown in Figure 1. The data that we have, is processed by WEKA and given the clustering process first to produce 3 cluster. For each cluster, including the initial data, is give the decision tree process which produce the recommendation prediction of each cluster. The data also check using WEKA default test flow, which separate the 10% of the data as the data that used as the confirmation. This process is done for all 4 groups of data. And the decision tree done by combining the data of the customer master dan the transactional.

4.1 Clustering – K Means

Clustering algorithm divide a group of data into some sub-groups cluster of data to get the meaningful information that exist in it [14]. By using WEKA built in k-means algorithm, the customer master data is processed and divided into 3 cluster. The clustering divided the data into 3 custer; Cluster 1 has 3378, cluster 2 has 3771 and cluster 3 has 4349.

The cluster made based on the field of customer master data that shown in
Table **1** which are:

- ID Klinik
- Gender
- Membership Class
- Frequency Transaction
- Iklan/Referal Info
- Age
- Citizenship of customer

The centroid of this cluster mostly is on age and frequency transaction, which shown in Table 3. This make the cluster mostly differ in age and frequency transaction.

Table 3 : Centroid Cluster

Attribute	Full Data	Cluster 1	Cluster 2	Cluster 3
ID Klinik	Cust-00012	Cust-00012	Cust-00019	Cust-00013
Gender	Female	Female	Female	Female
Membership Class	Registered	Registered	Registered	Registered
Frequency Transaction	Dormant	Dormant	Monthly	Dormant
Iklan/Referal Info	Internet	Internet	Marketing	Friends
Age	39.279	39.279	25.8099	38.0736
Citizenship	WNI	WNI	WNI	WNI

4.2 Classification – Decision Tree

After the clustering done, each of cluster that based on master data, combined again with the data transactional. This groups of data again each of them processed using the WEKA. This time using the Decision Tree J48 algorithm as the method that do the prediction. Additional from the clusters, whole data also processed using the same decision tree method. The models will be tested using 10% of the data automatically by WEKA as part of the function of it. The result of each cluster models will also be validated using the confusion matrix that generated in WEKA.

For the purpose to check the effectiveness between the 2 steps methods and with the one directly doing the prediction, a comparison done by comparing each cluster result with the result of J48 algorithm directly to whole data without clustering.

Table **4** is showing this comparison.

Table 4 : Result Comparison

Group	Cluster 1	Cluster 2	Cluster 3	All Data without clustering
Correctly Classified Instances	3148	3433	3900	10247
Incorrectly Classified Instances	228	341	448	1251
Accuracy	93.25%	90.96%	89.70%	89.12%

Mean Absolute Error	0.0206	0.0304	0.0388	0.0391

The result shown that for all cluster, the accuracy of the prediction if better than the subset data that did not processed using cluster. This result makes us confident to put the product recommendation system to be use as production system in the clinic.

5. SUMMARY AND CONCLUSION

5.1 Summary

In this paper, we have proposed a combination of clustering and classification method to do product recommendation to customer. This research done based on case study on clinic XYZ. From the comparison of the result, it showed that the clustering done make the classification result is better than directly do the classification on overall data without clustering. The good accuracy of the prediction might also cause by the good quality of data that already prepared previously. Another set of data might show different result, but based on this research result, we recommend that for further development of product recommendation should do clustering to separate the data into relevant groups, before applying the method use to do the recommendation. The production system is now live in the clinic and the review show good result.

5.2 Limitation

We do acknowledge that the accuracy of the classification is not too high (not more than 95%). We think that this caused by the parameter that provided. Not all attribute that provided is the driver for the classification. We think that additional data would make the result better.

5.3 Conclusion and Future Research

Future research can be done by applying the methods to other subset data. Other improvement can also be done by changing the second method used. Either by using other classification algorithm or change to other type of method.

REFERENCES

1. J. Ben Schafer, J. A. Konstan, and J. Riedl, "**E-commerce recommendation applications**," *Data Min. Knowl. Discov.*, vol. 5, no. 1–2, pp. 115–153, 2001, doi: 10.1007/978-1-4615-1627-9_6.
2. P. A. Schneider, D. Peppers, and M. Rogers, "**The One to One Future: Building Relationships One Customer at a Time**," *J. Mark.*, 1995, doi: 10.2307/1252334.
3. F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "**Recommendation systems: Principles, methods and evaluation**," *Egypt. Informatics J.*, vol. 16, no. 3, pp. 261–273, 2015, doi: 10.1016/j.eij.2015.06.005.
4. W. Hill, L. Stead, M. Rosenstein, and G. Furnas, "**Recommending and evaluating choices in a virtual community of use,**" in *Conference on Human Factors in Computing Systems - Proceedings*, 1995, doi: 10.1145/223904.223929.
5. B. N. Miller, I. Albert, S. K. Lam, J. A. Konstan, and J. Riedl, "**MovieLens unplugged: Experiences with an occasionally connected recommender system**," in *International Conference on Intelligent User Interfaces, Proceedings IUI*, 2003.
6. G. Linden, B. Smith, and J. York, "**Amazon.com recommendations: Item-to-item collaborative filtering**," *IEEE Internet Comput.*, 2003, doi: 10.1109/MIC.2003.1167344.
7. M. Svensson, K. Höök, J. Laaksolahti, and A. Waern, "**Social navigation of food recipes**," in *Conference on Human Factors in Computing Systems - Proceedings*, 2001, doi: 10.1145/365024.365130.
8. H. C. Chen and A. L. P. Chen, "**A music recommendation system based on music data grouping and user interests**," in *International Conference on Information and Knowledge Management, Proceedings*, 2001, doi: 10.1145/502624.502625.
9. Y. F. Wang, Y. L. Chuang, M. H. Hsu, and H. C. Keh, "**A personalized recommender system for the cosmetic business**," *Expert Syst. Appl.*, vol. 26, no. 3, pp. 427–434, 2004, doi: 10.1016/j.eswa.2003.10.001.
10. P. Pu and L. Chen, "**A user-centric evaluation framework of recommender systems**," in *CEUR Workshop Proceedings*, 2010.
11. D. R. Liu and Y. Y. Shih, "**Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences**," *J. Syst. Softw.*, vol. 77, no. 2, pp. 181–191, 2005, doi: 10.1016/j.jss.2004.08.031.
12. S. Dhawan, K. Singh, and Jyoti, "**High Rating Recent Preferences Based Recommendation System**," in *Procedia Computer Science*, 2015, doi: 10.1016/j.procs.2015.10.085.
13. T. Zhang and V. S. Iyengar, "**Recommender Systems Using Linear Classifiers**," *J. Mach. Learn. Res.*, 2002, doi: 10.1162/153244302760200641.
14. D. McSherry, "**Explaining the pros and cons of conclusions in CBR**," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2004, doi: 10.1007/978-3-540-28631-8_24.
15. C. C. Escolar-Jimenez, K. Matsuzaki, K. Okada, and R. C. Gustilo, "**Enhancing organizational performance through employee training and development using k-means cluster analysis**," *Int. J. Adv. Trends Comput. Sci. Eng.*, 2019, doi: 10.30534/ijatcse/2019/82842019.
16. M. Karthikeyan, A. Arivarasan, and D. Kumaresan, "**Performance assessment of various text document features through k-means document clustering approach**," *Int. J. Adv. Trends Comput.*

Sci. Eng., vol. 8, no. 5, pp. 1969–1977, 2019, doi: 10.30534/ijatcse/2019/21852019.

17. K. J. Atmaja, I. B. G. Anandita, and N. K. C. Dewi, "**Penerapan Data Mining Untuk Memprediksi Potensi Pendonor Darah Menjadi Pendonor Tetap** Menggunakan Metode Decision Tree C.45," *S@CIES*, 2018, doi: 10.31598/sacies.v7i2.284.
18. S. S. Baskar, L. Arockiam, and S. Charles, "**Applying Data Mining Techniques on Soil Fertility Prediction**," *Int. J. Comput. Appl. Technol. Res.*, 2013, doi: 10.7753/ijcatr0206.1005.
19. M. J. Christ, R. N. P. Tri, W. Chandra, and T. Mauritsius, "**Lending club default prediction using Naïve Bayes and decision tree**," *Int. J. Adv. Trends Comput. Sci. Eng.*, 2019, doi: 10.30534/ijatcse/2019/99852019.