DATA ANALYTICS

UE20CS312

LITERATURE REVIEW OF RESEARCH PAPERS

Name: Purvik S Nukal
SRN: PES1UG20CS315

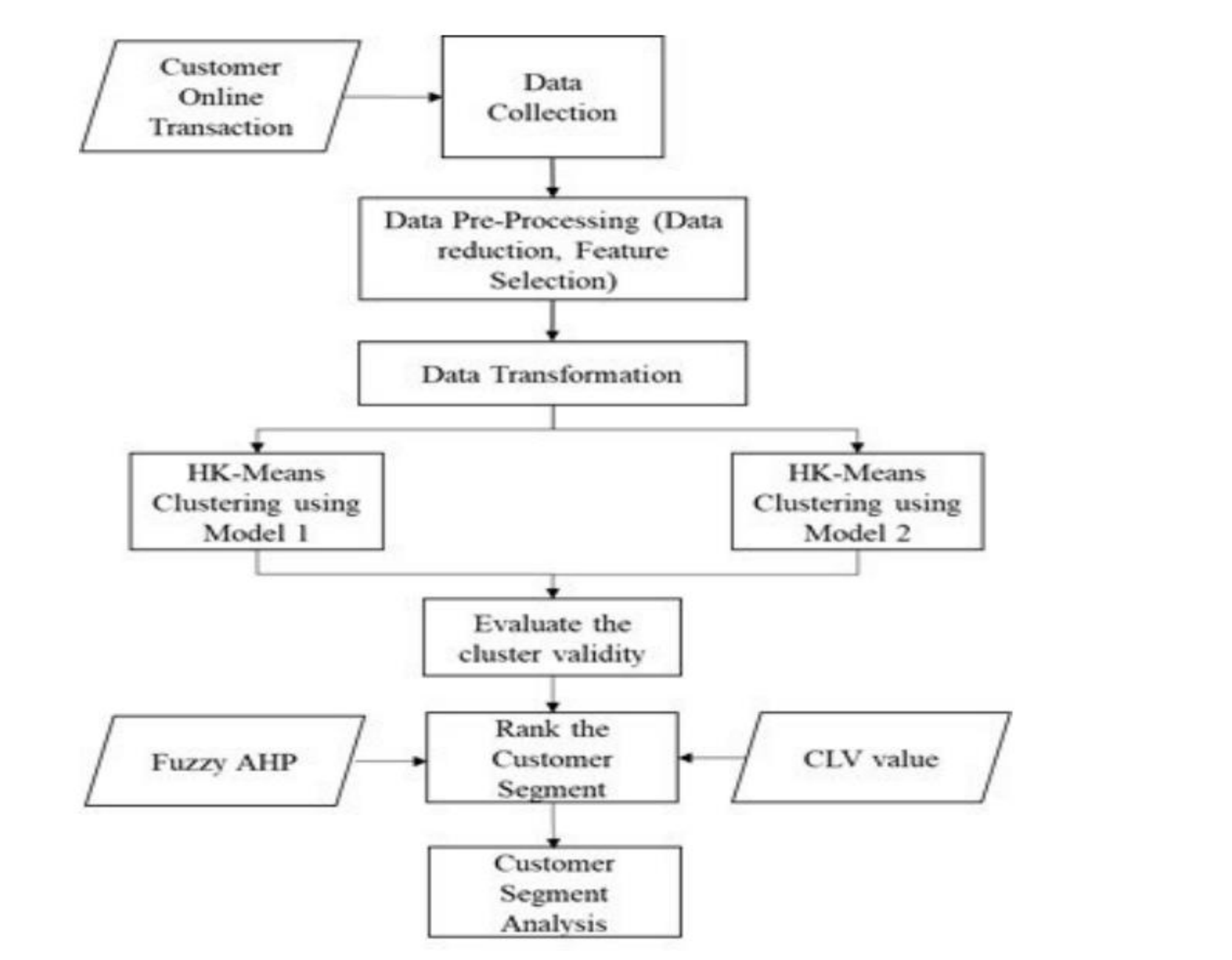
Name: Rahul Ranganath
SRN: PES1UG20CS316

Research Paper - 1

Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method

In order to cope with the competitive environment related to beauty industry sector in Indonesia, companies need to manage and evaluate customer interactions by enhancing Customer Relationship Management (CRM). This study aims to specify customer segment that has similar lifetime value with clustering method, hence company can conduct appropriate strategies to the right segment. Two-stage clustering method for segmenting customers is proposed in this study. Ward's method is used for choosing an initial number of cluster and K-Means method to perform clustering analysis. Two approaches using LRFM (Length, Recency, Frequency, Monetary) model and extended model called LRFM - Average Item (AI) variables in clustering process are compared by validity index to obtain the best result for customer segmentation. The ranking process based on Customer Lifetime Value (CLV) score is conducted using weighted LRFM model variables. Final weight score for all variables are obtained from Fuzzy AHP method. In summary, company also get several inferences such as customer characteristics of high and less potential customers. It can be a guideline for making the sale and marketing strategies.

- Customer Lifetime Value (CLV) concept is a part of Strategic Customer Relationship Management (CRM) that is widely used for predicting revenues that can be obtained from a customer by identifying their lifetime values
- retaining customers is important and more profitable than bringing in new customers or acquisition.
- Dividing heterogeneous customer into segments will be helpful for companies to execute their marketing strategies. Customer segmentation is one of an effective method to satisfy customer needs and preferences.


Customer Online Transaction
Data Collection
Data Pre-Processing (Data reduction, Feature Selection)
Data Transformation
HK-Means Clustering using Model 1
HK-Means Clustering using Model 2
Evaluate the cluster validity
Fuzzy AHP
Rank the Customer Segment
CLV value
Customer Segment Analysis

Results

The result shows that adding new variable in the extended LRFM model gives no difference in clustering results. For further analysis in KMeans clustering, segmentation process with LRFM model can be used as a parameter. Fuzzy AHP is proposed to define weight of each variable, the final score indicates that Frequency was the most significant variable in this study. The ranking process of CLV according to K-Means weighted centroid score for each cluster revealed that cluster 1 has the highest score, followed by cluster 2 and cluster 3, respectively. Customers in cluster 1 are potential and valuable to be gold customer. By having this CLV score, company can identify their customer characteristics and focus on those customers who bring the maximum benefit and loyalty. This study has some limitations that can direct to future researches. Since the segmentation is done based on LRFM variables, company can customize their marketing strategies only based on customer behaviors. Future work, study can use extended method to identify products which are bought frequently by the customer for each segment.

Acknowledgment

- P. P. Pramono, I. Surjandari and E. Laoh, "Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method," 2019 16th International Conference on Service Systems and Service Management (ICSSSM), 2019, pp. 1-5, doi: 10.1109/ICSSSM.2019.8887704.

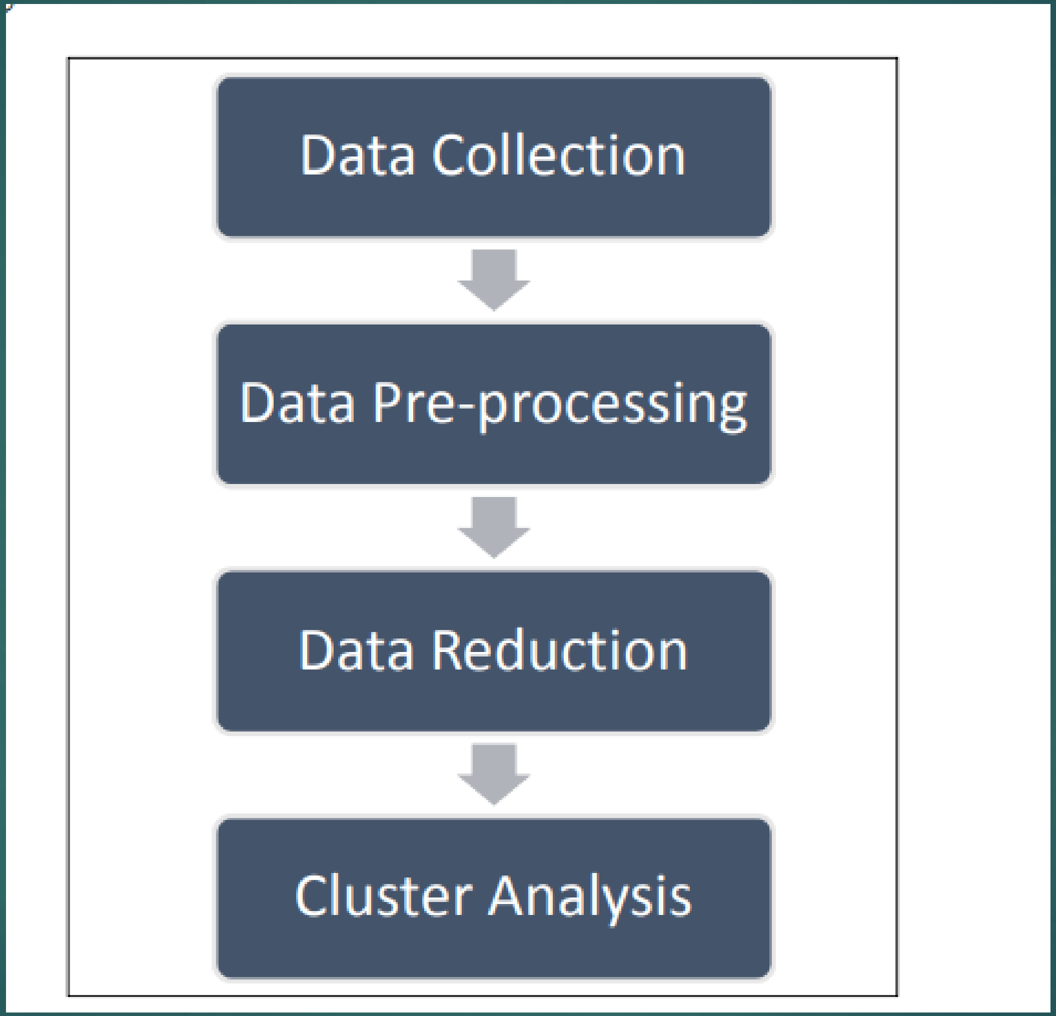
- URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8887704&isnumber=8887596

Research Paper - 2

Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products

E-commerce has become a crucial platform consists a large database of products with billions number of retailers and consumers. However, these products are placed into different categories according to the structure of different websites. A clustering analysis using K-Means Clustering algorithm helps in providing an insightful pattern on categories of clustered products. This analysis leads to an automatic classification model to classify the products efficiently. This paper presents a step by step cluster analysis using K-Means clustering to group e-commerce products from the online store website in Malaysia. The most frequent words in each cluster provided a useful insight on the category of the clustered products which were hair and face, oral and pets care products. Hence, K-Means clustering analysis able to group a large data set of e-commerce products effectively.

- There were 56176 products had been described according to 1596 features. The features are the count of words included in the corpus. After applying PCA, the dimensionality of the data was reduced to 1234 principal components as the features for clustering the data set.
- The number of clusters (k) that will be generated in the final output was determined using the average silhouette approach.
- For the first cluster, most of the frequent words related to hair and facial care products. Meanwhile, the second cluster may represent the cluster of pet care products. The third cluster seemed to relate with oral care products.

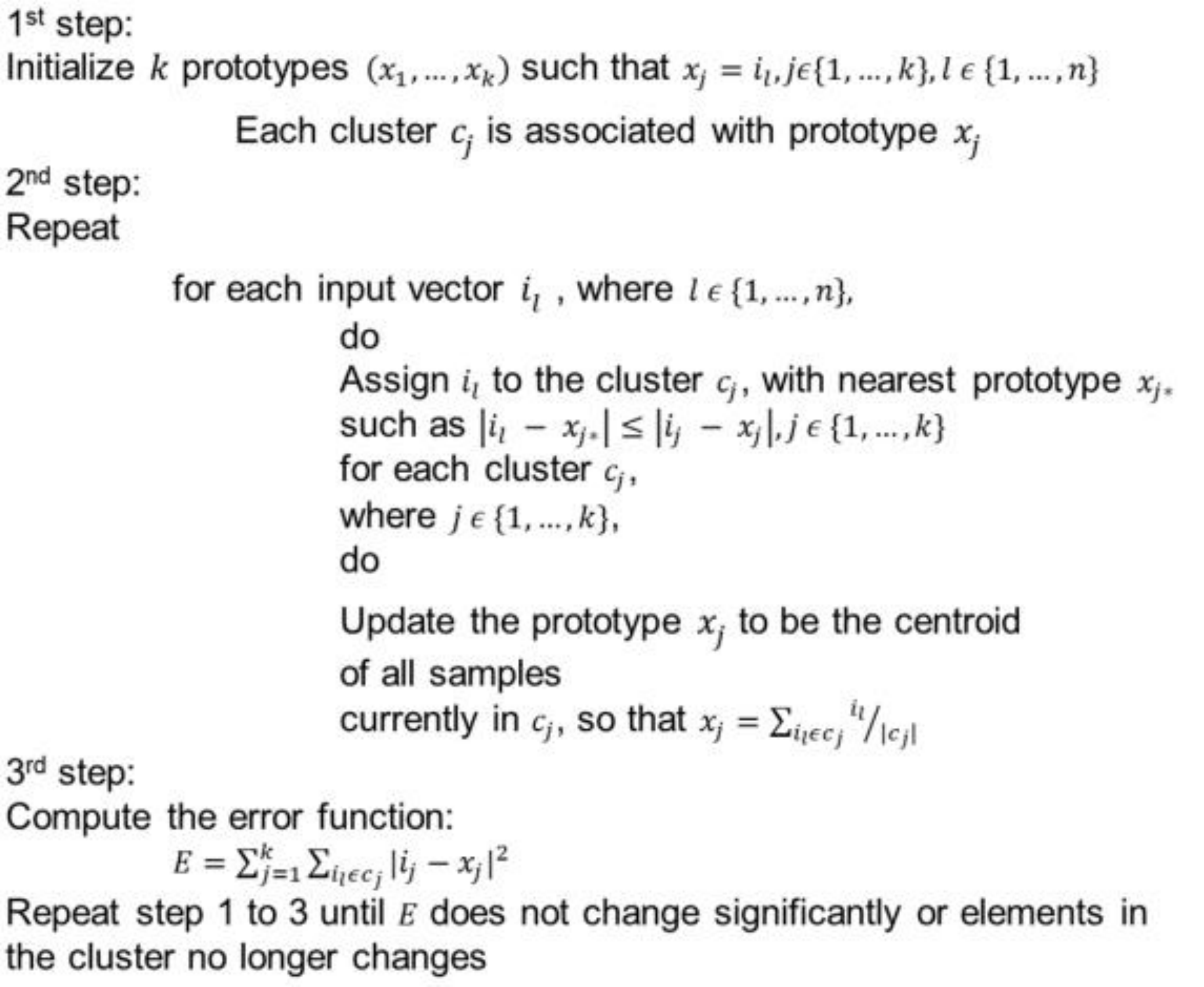


Data Collection

Data Pre-processing

Data Reduction

Cluster Analysis



1st step:

Initialize k prototypes $(x_1, \dots, x_k)$ such that $x_j = i_l, j \epsilon \{1, \dots, k\}, l \epsilon \{1, \dots, n\}$

Each cluster c_j is associated with prototype x_j

2nd step:

```
Repeat
    for each input vector $i_l$ , where $l \epsilon \{1, \dots, n\}$,
        do
        Assign $i_l$ to the cluster $c_j$, with nearest prototype $x_{j*}$
        such as $|i_l - x_{j*}| \le |i_j - x_j|, j \epsilon \{1, \dots, k\}$
        for each cluster $c_j$,
        where $j \epsilon \{1, \dots, k\}$,
        do
        Update the prototype $x_j$ to be the centroid
        of all samples
        currently in $c_j$, so that $x_j = \sum_{i_l \epsilon c_j} {}^{i_l}/_{|c_j|}$
```

3rd step:

Compute the error function:

$$E = \sum_{j=1}^{k} \sum_{i_l \epsilon c_j} |i_j - x_j|^2$$

Repeat step 1 to 3 until E does not change significantly or elements in the cluster no longer changes

Results

The experimental results show that the data can be clustered into three groups with distinct characteristics. This clustering technique able to act as a recommender system based on word occurrences in each cluster. In addition, PCA was beneficial in reducing the dimensionality of the data for a better performance of the cluster analysis. Future research may compare the performance of cluster analysis with different clustering algorithms to help categorising large text data sets.

Acknowledgment

- N. M. N. Mathivanan, N. A. M. Ghani and R. M. Janor, "Analysis of K-Means Clustering Algorithm: A Case Study Using Large Scale E-Commerce Products," 2019 IEEE Conference on Big Data and Analytics (ICBDA), 2019, pp. 1-4, doi: 10.1109/ICBDA47563.2019.8987140.

- URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8987140&isnumber=8986988

Research Paper - 3

Trend Analysis of Building Power Consumption Based on Prophet Algorithm

In this paper, a novel prophet algorithm was proposed to do a trend analysis of building power consumption. By comparing the prediction results of the Prophet algorithm and the ARIMA algorithm, the prophet algorithm has better prediction results in terms of prediction and can add the date and time that affect the building.

The simulation results show that the comprehensive power consumption of shopping malls is affected by holidays to some extent, but also by high-temperature weather to some extent. The comprehensive power consumption of office buildings is relatively less affected by holidays and high-temperature weather to some extent.

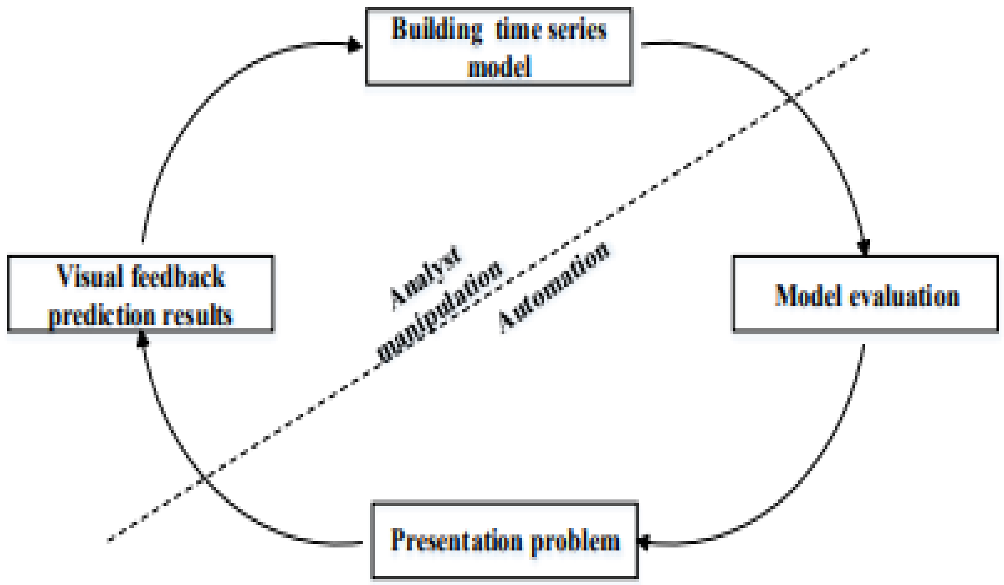


Figure 1. Prophet algorithm model.

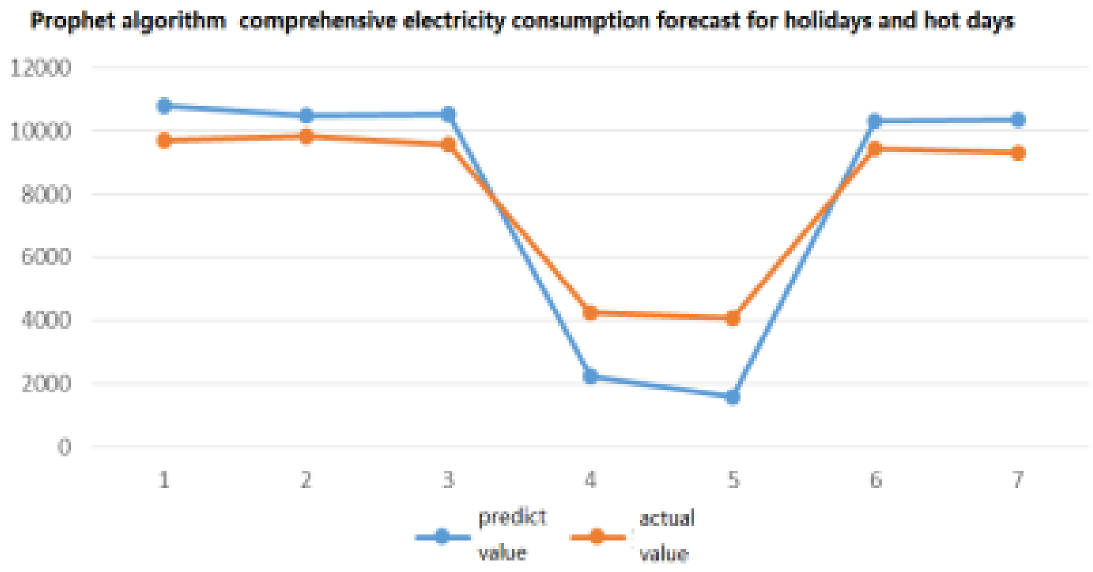


Figure 9. Forecast results of comprehensive office power consumption

In addition to electricity data, the author has also introduced some daily data that affects electricity consumption behavior, such as national legal holidays, high-temperature weather days, and other date-type data.
At the same time, the power consumption trend of these buildings is analyzed by using the prophet algorithm.
Predictions were made and comparison was done between the ARIMA and Prophet Model.

Results

The performance metric used in by the author is MAPE.

By comparing the prediction results of the Prophet algorithm and the ARIMA algorithm, the prophet algorithm has better prediction results in terms of prediction and can add the date and time that affect the building.

The results show that the comprehensive power consumption of shopping malls is affected by holidays to some extent, but also by high-temperature weather to some extent wherein in the case of office buildings, the power consumption is relatively less affected by holidays and high-temperature weather to some extent.

Acknowledgment

- I. Fernandez, C. E. Borges, and Y. K. Penya, “Efficient building load forecasting,” in Emerging Technologies & Factory Automation (ETFA), 2011 IEEE 16th Conference on, 2011.

- URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9121548