Analyzing Inventory Data Using K-Means Clustering

Michael Pigman+, Hung Le+, Urvish Bhagat+, Maxfield Thompson+, and Qingguo Wang*
College of Computing and Technology, Lipscomb University
Nashville, TN, USA 37204
*These authors contributed equally to this work
+Corresponding author: qingguo.wang@lipscomb.edu

Abstract—— Inventory control is a major aspect of profit margin maximization in retail organizations. In this paper, we use K-means clustering algorithm to analyze the inventory data of a global retail company, in an aim to help develop a forecasting system for the company in future. The data set we were provided for analysis is big and complex, consisting of 36-months of prior sales data, two hundred thousand different retail products. In this paper, we focus on pre-processing the data and then using K-means to cluster inventory items. Thorough analysis is also provided in the paper to verify the results of clustering.

Keywords— K-means clustering; data transformation; principal component analysis; inventory data analysis

I. INTRODUCTION

Inventory control is a major aspect of profit margin maximization in retail organizations. Inventory costs can encompass much more than the purchase price of an item. Factors such as product handling, storing, and depreciation all take their toll on a company’s bottom-line. As an example of poor inventory analysis and management, JCPenney lost approximately a billion dollars in 2012 from a rebranding flop that led to inventory liquidations with write downs as significant as 95% off some retail items [1].

Efforts have been made to study inventory management. In a recent study, Haines, Hough, and Haines investigated behavioral causes of the bullwhip effect [2]. The interaction of metacognitive behaviors with ordering in retail stores was also explored in an earlier study [3]. In 2014, Williams et al. showed the difficulties in predicting orders based upon point-of-sale (POS) data and that order history and more traditional forecasting methods could be used to mitigate the impacts of inaccurate ordering forecasts [4]. A number of computational methods have been applied to inventory and sales data to help design and implement accurate predictive models [5-12]. However, adopting a new process is large financial undertaking for retailers. Unless the return on investment is great enough, retailers would hesitate to devote resources to expand upon new methods.

As a service to their distribution centers, many retailers provide POS data to track sales, in order to aid in the replenishment of goods available to be sold. The interest of retailers in POS data has also heightened for its relation to human behavior when reordering sale items. Additional information allows retailers more effectively control the balance of inventory and reduce forecasting errors during the ordering process [4][9][12]. Williams et al. in the aforementioned study showed simultaneous uses of both retailer order history and POS data to predict retailer orders to suppliers outperformed approaches based singularly on order or POS data by up to 125% [4].

In this study, we analyze inventory data using K-means clustering. The rationale for clustering products into categories is that each cluster can be used to create a forecasting model for the products in this category. Various clustering methods have been applied to study inventory data [6][7]. For instance, Kartal et al. employed multi-criteria decision making to predict classification of each inventory item [6]. As this work is still in early stage, this paper focuses on data clustering, in an aim to find clusters that can be used for forecasting systems in future.

The rest of the paper is organized as follows. First, we describe our data and how we preprocess it. Then, we present our K-means clustering algorithm and the results of the clustering. Next, we provide discussion and suggest direction for future development.

II. DATA COLLECTION AND PREPROCESSING

A. Data Collection

The data used in this project was collected from a global retail company. The data is in the form of CSV file that was exported from the retailer’s Oracle database management system. The data consists of two sets: product description dataset and sales history dataset. The sales data along with stock keeping unit (SKU) data were collected through retailer’s order management system and data warehousing system. This data contains approximately two hundred thousand SKUs, which are described using seventeen attributes each. The sales data spans a duration of three years and consists of over four and a half million records.

B. Data Preparation

Preparing data, which contains noise and missing information, for analysis is an important and challenging step in data analytics. Poor data quality can lead to false results and misleading predictions, causing stakeholder’s financial losses. To maintain the integrity of the data during the process of data cleaning, we follow a three-step procedure: (1) merging the two datasets, i.e. product description dataset and sales history dataset based on SKU number; (2) using selected features to filter the dataset to decrease file size; (3) transforming the data

---

This work was supported in part by Faculty Summer Grant by Lipscomb University Provost's Office.

XXX-X-XXXX-XXXX-X/XX/XXX.00 ©20XX IEEE

into an analysis-ready form. Detailed descriptions of these three steps are given next.

In step (1), we merge product sales data with product description data by SKU. The original data from the retailer had limited features in their product descriptions. It is difficult to use it directly for clustering. By merging product sales data with product description data, we combine the descriptive features of the two types of data to allow clustering products with similar sales features / histories. The merged dataset contains 43 columns, 13 of which are from the product description dataset and the rest from the product sales dataset.

In step (2), we perform feature selection. Five of the product description columns were removed from further analysis, as they did not provide useful information. For the rest of the columns, we select only the features related to particular analysis, such as vendors, journeys, targeted gender of the products (e.g. men’s shoes - men’s shoes is a division and department of the retailer), etc.

This step effectively trim down the data into a manageable amount. After filtering, the size of the data is decreased from several million rows to some thousand rows, reducing the cost and time of analysis. This step also decreases data dimensions. The obvious but unwanted clusters in the data, such as the binary separation between genders and vendor differences, are eliminated.

In step (3), we transform the data into an analysis-ready form. Some columns of the dataset are discrete categorical strings, e.g. “vendor_number”, “retail_sub_department”, “edi_color_code”, “size_system”, and “product_group”. Using R programming language, we convert such type of values to factors and then transform them into numeric factors.

For quantitative features, K-means requires them to follow a normal (or normal-like) distribution. This, however, is not the case for most continuous features in the dataset. The following Fig. 1 shows the distribution of total sales of all products in June, which is obviously not normal.

Thirty features in the dataset have similar skewed distribution as in Fig. 1. Because they typically follow a power law distribution, we normalize them by applying a log transformation to their values. For the total sales in Fig. 1, its log-transformed distribution is plotted in Fig. 2, which is now normal like after the log transformation.

![Fig. 2. Distribution of total sales in June after a log transformation.](image)

C. Principal Component Analysis

To test the effect of the log transformation, principal component analysis (PCA) was applied to the data before and after the log transformation. The resultant distribution of the points (each representing a product) was provided in Fig. 3 and 4 below.

![Fig. 3. Visualization of the first two principal components of PCA analysis of the non-log-transformed data.](image)
Fig. 3 visualizes the first two principal components (PC) of the PCA analysis of the non-log-transformed data. It shows a large density of points close to the origin, with two distinct tails leading away from the origin and outliers following the tails out. This result indicates a binary trend in the pre-transformed data that split the trendy data into two halves: tail one and tail two, or the positive and negative aspect of PC 2. This binary separation only closely follows PC 2, i.e., positive values in PC 2 form the first tail, and negative values in PC 2 are grouped into the second tail.

To understand this phenomenon, we examined the factor loadings (also called component loadings) for PC 2 in Fig. 3. Factor loadings are the correlation coefficients between the features of products and the factors. Interestingly, most of the range sales values (and some month sales data, e.g., sales in February, July, October, as well as December) caused the positive aspect of PC 2, whereas most of the general sales values (with some range sales values) generated the negative aspect of PC 2. Although the overall principal component analysis of these features is too messy to allow drawing concrete conclusion from it, this general pattern is obvious: i.e., products with a large range of sales in a certain month form one tail in Fig. 3 and products with high general sales, but low range sales, form another separate tail.

In Fig. 3, x-axis, i.e. PC 1, is in a range from -150 to 0, and y-axis, i.e. PC 2, from -20 to 40. The absence of any loadings in the positive direction of PC 1 indicates that the PCA can’t evenly represent some features across PC 1. This skewed nature of the data can be a concern, as it can distort downstream clustering analysis. But this skewness is not a surprise, because it is a consequence of power law distribution, thus in high dimensional space, most of the data points are clustered together into a high dimensional blob, with tails of outliers spanning out from it. Modeling the variance of this high dimensional data in low dimensional space is relatively easy for PCA because there is little variance with the dense high dimensional blob. Log transformation, however, ‘opened up’ the blob into a high dimensional shape with consequently more informational variance in the scaled sales data. Hence, log transformation increases the variance overall and accordingly lowers explained variance in the first two principal components of PCA.

III. K-MEANS CLUSTERING AND RESULT ANALYSIS

PCA further reduced the dimension of the pre-processed data. On the output data of PCA, we then applied K-means clustering to identify clusters of the products. K-means requires a ‘k’, predetermined number of clusters, as an input. To determine the value of k, we evaluated a number of techniques [18-20] and they all suggested 2 or 3 to be the optimal number of clusters. One of the techniques we tried is Hubert statistic (or Hubert index) [20], a graphical method for finding the optimal number of clusters. The result of Hubert statistic is provided in Fig. 5.

In Fig. 5 (a), x-axis denotes the number of clusters and y-axis Hubert statistic. In Fig. 5 (a), a knee at x=3 suggests that three is the optimal number of clusters. In Fig. 5 (b), y-axis is the second derivative of Hubert index. The peak in Fig. 5 (b) falls at x=3, the same location of the knee of the red curve in Fig. 5 (a), reaffirming three as the optimal number of clusters. Moreover, there is a local minimum in Fig. 5 (a) and (b) at x=7 to 9, which is clearly inferior to the point x=3 and hence will not be discussed further.

Using k=3 as input, we then run K-means clustering algorithm on the pre-processed sales data. Fig. 6 shows the result of the K-means. In Fig. 6, each dot represents a product and each color a cluster of products. To verify the clustering, we randomly selected and examined some products from each cluster. The analysis of these products and discussion about the three clusters were presented below.
Fig. 6. Visualization of the pre-processed data run through PCA and then clustered using K-means (k=3). Each color represents a different cluster.

Firstly, we randomly selected two products from the cluster colored red in Fig. 6. For simplicity, we denoted them as products 1 and 2, respectively. Product 1 has an index ID 403 in the dataset and a coordinate at (-5.25, -0.53) in the PCA plot in Fig. 6, while product 2 has an index ID 1592 in the dataset and a coordinate at (-5.04, -0.42) in Fig. 6.

Fig. 7 shows the history of their general sales in the past year, while Fig. 8 gives their ranges of sales by month. The largest observable difference between the sales of the two products is the range sales in October, in which the range sales of product 1 is 7 times of that of product 2. Despite the significant difference in October’s range sales, however, from the two figures it is clear that overall these two products follow a similar trend of sales: they both have high sales in January, progressively lower sales into spring and summer, and then slightly increased sales in the fall.

Next, we randomly picked two products from the cluster colored black, the coordinates of which are (3.57, 3.67) and (4.14, 4.07) respectively in Fig. 6. We denoted them as products 3 and 4. Their overall sales history is provided in Fig. 9 and ranges of monthly sales in Fig. 10.
Fig. 9. Monthly sales of the two products randomly selected from the cluster in black color in Fig. 6.

Fig. 10. The range in month sales for the two products in Fig. 9.

From Fig. 9 and Fig. 10, we can see that the sales pattern of products 3 and 4 is very different from that of products 1 and 2: not a single item of products 3 and 4 was sold before September and the two products were only sold in the last quarter of the year. Additionally, the magnitude of the sales of the two products differs remarkably: the sales of product 4 is >100 times higher than that of product 3. With the product 4’s sales in the fourth quarter significantly more than product 3’s, the data was presented in log-scale in Fig. 10, which shows clearly both products follow the same pattern of sales, despite the difference in their magnitude. This explains why they were grouped into the same cluster by K-means.

Lastly, we examined the third cluster that is in green color in Fig. 6. From this cluster, we also selected randomly two products, denoted as products 5 and 6 respectively. The coordinate of product 5 is (4.98, -3.80) in Fig. 6 and that of product 6 is (5.32, -3.42). Fig. 11 and Fig. 12 visualize the sales of the two products, from which we can see neither products 5 nor 6 saw any sales in the second half of the year. The majority of the sales of product 6 happened in January. But even its January sales is small compared to other products discussed above. Their pattern of little sales is identical in this regard.

Although further systematic investigation is required, our preliminary analysis of the six products so far seemed to indicate that K-means correctly grouped items with similar sales trends together. Moreover, our results show the pattern of sales that each cluster represents.

IV. DISCUSSION

With a massive number of products, it is a challenge for the retailer to figure out the sales patterns of the products and the similarity between them. The results reported in this paper can not only help the company group items sharing obvious similarity in sales, but also help organize products with non-intuitive similarity. For example, the product 4 discussed in the paper is likely a big seller for this business, pulling in hundreds of thousands of sales dollars a month. It overshadows product 3 because it brings in less than one hundredth of the sales of product 4. Our clustering analysis linked them together in a way that may be overlooked without comparing their sales trends. With them in the same cluster, the strategy employed for product 4 might be applicable to product 3.

In addition, the clusters discovered using K-means could be useful for business in two ways. Firstly, the products with similar sales trends could benefit from similar marketing strategies. Secondly, from an inventory management perspective, the products that sell in similar trends over time may need similar inventory practices.

Finally, the results of this work can be potentially useful for the retailer to develop a forecasting system. A future development is therefore to apply predictive models to each cluster of products to predict future purchase, so as to optimize inventory management for the retail company.
ACKNOWLEDGMENT

We thank Dr. Todd Gary for his guidance and Cory McDermaid, Cindy Sandefur, and Brian Waters from the retailer for their help with the data and business questions. We also express our gratitude toward Damien Mingle for his insightful input to this work. Lastly, this work was supported in part by Faculty Summer Grant, Lipscomb University Provost’s Office.

REFERENCES