Identify product families using cluster analysis: case study in Passenger Car Radial (PCR) tire product

Rere Nugrahita¹ and Istri Surjandari¹
¹Industrial Engineering Department, Faculty of Engineering, Universitas Indonesia, Kampus UI Depok, Indonesia

*E-mail: rere.nugrahita@ui.ac.id

Abstract. Manufacturing companies, such as tire manufactures are facing great challenges to cope with increased product variety which induced by customer demand. This variety lead to higher internal complexity in term of design and production. Thus, variety has to be well-managed in order to guarantee the positive outcome for company. One of the solution is to have a well-structured product family. In this research, products data are partitioned into clusters by applying cluster analysis for mixed-type data based on their general characteristic and component specification. Variants within cluster have similarities in term of characteristics and main product component used in production. By applying k-prototypes algorithm to handle these mixed type data, the data set is clustered and interpreted into eight different clusters using selected variables.

1. Introduction

Many manufacturing companies are increasing the number of product variety in order to stay competitive, tire manufacture is no exception. Increasing diversity in product is one of the effort to cover external factor such as, customer demand in term of product features, different needs in large number of market segments, certification and requirements due to regional issue [1]. High number of product variety induces opportunity and challenges at the same time. Product variety come with great potential for expanding market while increasing sales volume and revenues, but this best condition will only achieve if the product variety is well-managed by the company. In reality, offering more product variants to customer incurs more expenses to all stages of product including design, planning, production, inventory, selling and after sales services which lead to reducing acquired profit [1, 2]. Aside from that, product variety increase design and production complexity internally [1, 3, 4, 5], and also shorten product lifecycle which lead to the need of short period time in developing new product variants [3].

A product that seems simple such as tire, actually have broad product variety which is represented by thousands Stock Keeping Units (SKUs) number with hundreds of product lines, especially for Passenger Car Radial (PCR) tire. This variety needs to be well-managed to reduce complexity both internally and externally. One of essential prior knowledge which can support manufacturing companies in facing variety challenges is known as Product Variety Management (PVM). The main objective of PVM is to reduce complexity and its associated costs which induced by product variety [4]. Implementation of flexible manufacturing and advance information technologies has great impact in production system to deliver high level of variety in product at low cost as possible [1]. PVM strategies in acquire high flexibility classified into three main activities which are design, planning and
manufacturing ranging from component, parts and product to entire enterprise [1]. Main enabler of PVM success are identifying commonalities of product and parts and grouping into families [6]. Product families represent characteristic which only belong to one group of similar product. Unclear defined product family at design stage, identified as the root cause for complexity in production system [3].

In this paper, product family is identified and interpreted using cluster analysis based on component specification that consists of mixed data between numeric and categorical data. Clustering method itself is proven to be well-performed and has been widely applied in many application of PVM related to design product architecture, product family and product platform [5]. Many previous research related to PVM using cluster analysis using clustering algorithm such as k-MM, hierarchical agglomerative, hierarchical complete linkage and k-means [5, 7, 8, 9], but mostly use numeric data as its input. Other research apply combination of two or more methods when different type of data is mixed, such as association rule, fuzzy c-means clustering and classification [10]. Clustering algorithm called k-prototypes algorithm can directly handle mixed type data of numeric and categorical [11], extension to this algorithm also has been done recently [12]. By applying k-prototypes clustering algorithm, this research has aim to grouping product variants based on product geometry and component specification which consists of total 31 variables and 1443 observed product. The paper sections are organized as follow. Section 2 presents the data collection, data pre-processing and methods in determining cluster variables. Section 3 describes how k-prototypes was applied using selected cluster variables. Finally, the paper is concluded in Section 4.

2. Materials and Methods
2.1. PCR tire product geometry and component specification
Product families, especially in tire product, can be identified through product geometry and component specification. In term of product geometry, dimension of section height, steel belts, plies, height of bead filler and bead construction are considered as parameters which differentiate one product with another. Similar in component specification, such parameters are code (for different material), height, angle or width of semi-finished component including, steel belts, plies, sidewall, bead, bead filler, inner liner, and other supporting material.

2.2. Research methods
Generally, the research methods is proceed as in Figure 1, each process will be described in next sub-section.

![Figure 1. Research methods](image)

2.2.1. Data integration and pre-processing
As collected data are separated in multiple data sets, data integration phase is needed to combine the data into a coherent data set before continue to data pre-processing. In data pre-processing, data cleaning is done which relates to several activities, such as eliminate irrelevant data item, omit the missing value, ensure the structure of a variable only consist of one type of data and standardize the numeric data using Z standardization. Output of the data pre-processing stage is a data set consists of 31 variables of 1443 observed product variants.

2.2.2. Variables Selection
Variable selection is the stage that must be done carefully before continue to process the data using clustering algorithm to avoid bias result and complex interpretation of cluster. This research use mixed type of numerical and categorical data. In term of numerical data, it is encouraged to examine the variables used for substantial multicollinearity. If multicollinearity is found, it is advised to include only cluster variables that are not highly correlated [13]. One of the way to detect multicollinearity is
by examining correlation matrix of related variables and suppressed highly correlated variables. In term of categorical data, measuring the categorical level and approach by calculating permutation misclassification rate for each variable of the data using an existing tools in R knowingly as FeatureImpCluster, are done. Variable importance is interpreted by calculate the mean of misclassification rate over all iterations. The permutation misclassification rate of a variable is the number of wrong cluster assignments divided by the number of observations given a permutation of the variable.

2.2.3. K-Prototypes algorithm

An extension to k-means algorithm that can handle clustering process for both numeric and categorical data was developed by Huang [11] and widely known as k-prototype algorithm. The idea is integrate the k-means algorithm that works for purely numeric variables and k-modes algorithm for purely categorical variables. In k-prototypes algorithm, the cluster prototypes recalculate iteratively by the algorithm and reassign clusters. Clusters assignment aim to minimize cost function in Eq. 1 and use similarity measure $d(X,Q)$ in Eq. 5 between two mixed type objects [11].

$$P(W,Q) = \sum_{l=1}^{k} (P^r_l + P^c_l)$$

(1)

Where $P^r_l$ and $P^c_l$ describe for numeric and categorical variables respectively.

$$P^r_l = \sum_{i=1}^{n} w_{il} \sum_{j=1}^{p} (x^r_{ij} - q^r_{ij})^2$$

(2)

$$P^c_l = \lambda \sum_{i=1}^{n} w_{il} \sum_{j=p+1}^{m} \delta(x^c_{ij},q^c_{ij})$$

(3)

Subject to,

$$\sum_{l=1}^{k} w_{il} = 1, \quad 1 \leq i \leq n$$

$$w_{il} \in \{0,1\}, \quad 1 \leq i \leq n, 1 \leq l \leq k$$

(4)

Where $W$ is an $n \times k$ partition matrix and $Q$ is a set of objects in the same object domain, $Q=\{Q_1,Q_2,\ldots,Q_k\}$. Similarity measure,

$$d(X_i,Q_l) = \sum_{j=1}^{p} (x^r_{ij} - q^r_{ij})^2 + \lambda \sum_{j=p+1}^{m} \delta(x^c_{ij},q^c_{ij})$$

(5)

Where $p$ is number of numeric and $m-p$ is number of categorical variables which $V^r_1, V^r_2, \ldots, V^r_p, V^c_{p+1}, \ldots, V^c_m$. $x^r_{ij}$ and $q^r_{ij}$ are values of numeric variables and $x^c_{ij}$and $q^c_{ij}$ are values of categorical variables for object $i$ and the prototype of cluster $l$. The first term of Eq. 5 is the squared Euclidean distance measure on the numeric variables and the second term is the similarity matching measure on the categorical variables where,

$$\delta(a,b) = \begin{cases} 1, & \text{if } a \neq b \\ 0, & \text{if } a = b \end{cases}$$

(6)

The weight $\lambda$ is used to avoid over-emphasizing either type of variables. The balance between both terms can be controlled by the parameter $\lambda$ which has to be specified in advance, also for the number of clusters $k$. Higher values of lambda, $\lambda > 0$ will increase impact of the categorical variables, for $\lambda \approx 0$, the impact of the categorical variables decrease and vanishes at $\lambda=0$ which only numeric variables are taken into account, like in k-means clustering. An extension to estimate the weight $\lambda$ was introduced by dividing the average variance $\sigma^2$ of all numeric variables and the concentration $h_{cat}$ of all categorical variables [12]. In term of number of clusters, researchers’ perspective and qualitative
judgement play significance role in determining number of cluster, but the number $k$ can be estimated through scree plot, silhouette width, or index value [13, 14].

3. Result and discussion

3.1. Variables selection

In term of numeric variables, in order to detect multicollinearity, examining correlation matrix of related variables have a significance role. If any high correlated variables are found, either reduce the variables or use distance measure that takes into account the multicollinearity [13]. In this research, highly correlated variables are suppressed, hence numeric variables were reduced from 13 to 7 variables.

In term of categorical variables, level of factor is main concern for choosing variables to be included in clustering algorithm. It is recommended to choose categorical variables whose level number close to $k$ cluster which these variables will have strong contribution and play significant role in determining clusters [15]. In opposite, categorical variables whose too high in number of factor level will spread out per clusters and this condition have to be avoided. In this research, 4 categorical whose level number less than twenty are chosen. So, the total number of variables that used in clustering algorithm is 11 variables as the code can be seen in Table 1.

| Numeric Variables          | Categorical Variables          |
|---------------------------|--------------------------------|
| NSH (Nominal Section Height) | BEAD (Bead Construction)        |
| DW (Drum Width)            | CAPC (Tread Compound)           |
| BTB (Bead to Bead)         | MC (Machine Type)               |
| X1BW (First Belt Width)    | SI (Speed Index)                |
| BIC (Bead Inner Circle)    |                                |
| HBF (Height of Bead Filler)|                                |
| SWW (Sidewall Width)       |                                |

3.2. Cluster Result

Prior to running the algorithm, value of $\lambda$ and $k$ need to be determined. As explained before, $\lambda$ play significant role in balancing the effect of numeric and categorical to the cluster results. Value of $\lambda$ that estimated in extended version of k-prototypes algorithm [12] can be used as starting point to examine the impact of each type of variables. The $\lambda$ value is tuned to balance the effect of numeric and categorical variables. This research use $\lambda$ value equal to twice of estimated $\lambda$, which is 0.233, since it will give balance impact of numeric and categorical variables as illustrated in Figure 2.

![Figure 2. Visualization of balance impact of numeric and categorical variables using FeatureImpCluster in R](image)

Number of cluster can be estimated through silhouette width calculation that run for certain range of cluster number. Mainly, there are two things need to be considered while estimating the number of
cluster in clustering algorithm such k-prototypes which need value of k as an input. First is the silhouette width index which describe how close the distance between object inside the cluster and between clusters. Second, prior knowledge of researcher about the data also play significant role to ensure number of cluster will not too small or too high considering the difficulties in cluster interpretation stage. After calculating the silhouette index for k value ranging from 1 to 20 as illustrated in Figure 3, number of k=8 is considered as the most balance value in term of silhouette index and clusters representation of the data.

![Figure 3. Silhouette index of clusters](image)

3.3. Clusters interpretation
Examining the result of k-prototypes clustering algorithm, clusters can be interpreted as described in Table 2 below.

| Cluster Number | Interpretation                        | Description                                                                 |
|----------------|---------------------------------------|-----------------------------------------------------------------------------|
| Cluster 1      | Single middle series tire             | Middle series and high speed tire, using single wire construction and expensive tread compound material built in one stage |
| Cluster 2      | Small silica high series tire         | High series and high speed tire with small rim, using box construction and silica tread compound |
| Cluster 3      | LT high series tire                   | High series and low speed tire, close to light truck (LT) size, using single wire or big box construction and built in two stage |
| Cluster 4      | Low series tire                       | Low series and ultra-high speed tire with large rim, using single wire construction and expensive silica tread compound built in two stage machine |
| Cluster 5      | Box middle series tire                | Middle series and high speed tire, using box construction and carbon black tread compound material built in two stage |
| Cluster 6      | Small carbon high series tire         | High series and high speed tire with small rim, using box construction and carbon black tread compound built in one stage |
| Cluster 7      | SUV high series tire                  | High series and low speed tire, sport utility vehicle (SUV) size, using single wire or medium box construction and carbon black compound built in one stage |
| Cluster 8      | Large single middle series tire       | Middle series and high speed tire with large rim, using single wire construction and carbon black tread compound material built in two stage |

4. Conclusion and future research
This research has been analyzed data of PCR tire product then clustering the data set into eight different clusters using k-prototypes clustering algorithm and interpret it in order to identify product
families. It is well-proven that k-prototypes algorithm can handle clustering for mixed type data of numeric and categorical variables. In term of λ value, an existing developed package, FeatureImpCluster in R, is powerful in determining the suitable λ value through misclassification rate calculation and its boxplot visualization. For future research, this research can be extended as these clusters result can be used as an input for other research related to PVM, such as scheduling optimization, identify modular standardization and new variant design activities improvement in tire manufacture industry.

References
[1] ElMaraghy H, Schuh G, ElMaraghy W, Piller F, Schonsleben P, Tseng M and Bernard A 2013 CIRP Annals Manuf. Tech. 62 629–52
[2] Jhonson P 2003 Supply Chain Practice 5 54-67
[3] Schuh G, Rudolf S, Riesener M, Dölle C and Schloesser S 2017 Procedia CIRP 60 344–9
[4] Schuh G, Reinhart G, Prote J P, Sauermann F, Horsthofer J, Oppolzer F and Knoll D 2019 Procedia CIRP 81 874-9
[5] Hochdörffer J, Laule C and Lanza G 2017 Proc. Int. Conf. on Industrial Engineering and Engineering Management (Singapore: IEEE) pp 593-7
[6] Jiao J, Simpson T W and Siddique Z 2007 J. Intell. Manuf. 18 5–29
[7] Zhihuang D and Scott M J 2007 J. Intell. Manuf. 18 97-113
[8] Wallis R, Erohin O, Klinkenberg R, Deuse J and Stromberger F 2014 Procedia CIRP 23 178-83
[9] Bohnen F, Maschek T and Deuse J 2011 CIRP J. Manuf. Sci. Tech. 4 247-51
[10] Moon S and Simpson T W 2006 Proc. Int. Des. Eng. Tech. Conf. & Comput. and Inf. Eng. Conf., (Philadelphia, USA) pp 317-25
[11] Huang J Z 1998 Data Mining and Knowledge Discovery 2 283-304.
[12] Szepannek G 2019 The R journal 10 200-8
[13] Hair J F, Black W C, Babin B J and Anderson R E 2013 Multivariate Data Analysis (London: Pearson New International Edition) chapter 8 pp 415-474
[14] Wijayanto A, Suprapto Y K, and Wulandari D P 2019 International Seminar on Intelligent Technology and Its Applications (Surabaya, Indonesia) pp 210-5
[15] Khan S and Ahmad A 2013 Expert system with Applications 40 7444-56