DATA AGGREGATION FOR CAPACITY MANAGEMENT

A Thesis

by

YONG WOO LEE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

May 2003

Major Subject: Industrial Engineering

DATA AGGREGATION FOR CAPACITY MANAGEMENT

A Thesis

by

YONG WOO LEE

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Approved as to style and content by:

V. Jorge Leon (Chair of Committee)	César O. Malavé (Member)
John E. Mayer, Jr. (Member)	Brett A. Peters (Head of Department)

May 2003

Major Subject: Industrial Engineering

ABSTRACT

Data Aggregation for Capacity Management. (May 2003)
Yong Woo Lee, B.E., Inha University
Chair of Advisory Committee: Dr. V. Jorge Leon

This thesis presents a methodology for data aggregation for capacity management. It is assumed that there are a very large number of products manufactured in a company and that every product is stored in the database with its standard unit per hour and attributes that uniquely specify each product. The methodology aggregates products into families based on the standard units-per-hour and finds a subset of attributes that unambiguously identifies each family. Data reduction and classification are achieved using well-known multivariate statistical techniques such as cluster analysis, variable selection and discriminant analysis. The experimental results suggest that the efficacy of the proposed methodology is good in terms of data reduction.

ACKNOWLEDGEMENTS

I would like to thank Prof. Dr. V. Jorge Leon who supervised this work and gave me a motivation for this research.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

1. INTRODUCTION

For the past decades, the enterprise environment has changed rapidly such that business emphasis moved from seller-centered to customer-centered enterprises. Accordingly in some customer-oriented industries the products have been more diversified, the lot size has shrunk, the production processes have been more complicated, the demand has been more uncertain, and the life cycle has shortened to meet the customers' dynamic need. Under these circumstances the amount of production information produced has exploded exponentially as well.

Consider a company that manufactures a very large number of products through complex production processes, and suppose they frequently introduce new products and drop obsolete ones or modify old ones. In such a situation, challenges arise such as capacity planning and manufacturing data management. The decisions in the latter revolve around the following two questions: "Is there enough capacity to satisfy customer's demand?" and "How properly should the vast amount of data be managed?" This thesis focuses on the development of a data aggregation methodology to cope with the problems since a data aggregation methodology provides a good data management scheme for manufacturer to deal efficiently with capacity planning problems through suppressing irrelevant information.

In the proposed approach the products are put together into families and the many attributes associated with the products are cut down to a minimal subset of the attributes that unambiguously specifies each family. The methodology is devised based on well known multivariate data analysis techniques, such as clustering analysis, variable selection for data aggregation and reduction of the given product data and discriminant analysis for classification of newly introduced or revised product items. These techniques have been widely applied in the social, medical, and agricultural sciences as well as in

---

This thesis follows the style and format of *Computer and Industrial Engineering.*

economics, since many researchers need to explain phenomena in ways, where factors are dependent, which allowed them to sort data into some groups according to a variety of criteria. However, to the knowledge of the author, this thesis suggests a new and unique methodology for a production capacity data management.

This thesis is organized as follows: Section 2 describes a modeling framework for the problem. Section 3, presents a mathematical formulation of the problem. Section 4 reviews the relevant literature on related topics. Section 5 presents the proposed methodology. The experimental results are shown in section 6, and the conclusions follow in section 7.

2. MODELING FRAMEWORK

Consider a huge number of products, n, each associated with its corresponding production rate or unit per hour (UPH), u_i, where i = 1, 2 …, n and the products are uniquely identifies by p attributes (e.g., p = 3, {thickness, color, material}). Under this assumption the maximum number of products uniquely coded with these attributes is $\prod_{j=1}^{p} t_j$, where attribute j is represented by t_j values or levels. For instance, a semiconductor packaging operation may require about 20 attributes and 3 levels of each attribute; hence for this example, $t_j = 3$ for all j's and the possible product variety is $3^{20} \approx$ 3.5 billion products. If the value of an attribute is continuous quantity, then the maximum number is infinite. To manage such a large volume of product data possible coded, first, data reduction or aggregation must be achieved in such a way that n products are aggregated into, fewer, g product families, In order for the grouping to be adequate for capacity management, products in a family must have similar u. The production rate of family k is, then, defined as $\overline{u}_k$ that is mean value of u of the products in a particular family. A good scheme for capacity management will minimize the error incurred during the data aggregation process. Specifically, the inclusion of product i into family k induces an error equal to the difference between u_i and $\overline{u}_f$ when calculating capacity requirements associated with product i. The proposed model assumes that the aggregation error in this is less than a given value, e_o, for all products assigned to a family. Therefore, given family k, denoted as G_k, and the set of products in the family, the proposed model assumes:

$$| u_i - \overline{u}_k | \leq e_o, \forall i \in G_k \tag{1}$$

A related problem is that of coding the resulting families. Coding refers to the unambiguous specification of the resulting families such as a one-to-one mapping between the original product specification and a family. For instance, each product is

specified using p attributes, the n products are grouped into g families where g is less than n. Considering the characteristics of the families some attributes could be redundant for the coding, subsequently there exists the p', the fewer number of attribute combination able to identify the families, so that the smaller, p' attributes are required to store u's. The problem is to find the minimum number of attributes, p', less than p, specifying the families, which must be consistent with the data aggregation strategy.

Another problem is that of classification. When a new product is introduced or an old product is revised, it is asked to assign the product to one of the pre-defined families, which has the closest similarity to the introduced product. The problem is finding a way to separate the distinct families, measure the similarities between the introduced product and the families, and assign the product to a proper family. For example, the most significant attribute to identify the families is thickness of the products. If the thickness of a new product is 3mm, then the product will be assigned to a family represented by thickness closest to 3mm.

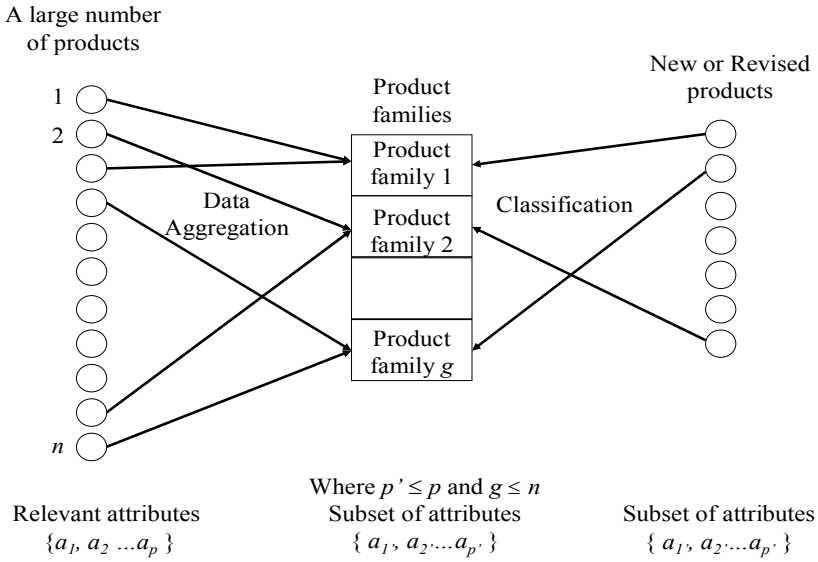


Fig. 1. The proposed data aggregation and classification model

Fig. 1 illustrates the proposed product data aggregation and classification model. In the model *n* products with *p* attributes are aggregated into *g* families specified using *p'* attributes and new or revised product is assigned into a family the closest similar to the corresponding product.

3. PROBLEM STATEMENT

This section formalizes the problems to be solved in this thesis. The problem at hand consists of two interdependent problems, namely: Data Reduction Problem (DRP) and Product Classification Problem (PCP).

As assumed, there are n products and each product has u and the corresponding values of the p attributes specifying the product. Then, the problem can be described as shown in Fig. 2.

Products	UPH	Relevant Attributes
Product 1	u_1	$\|a_{11}, a_{12}, \ldots, a_{1p}\|$
Product 2	u_2	$\|a_{21}, a_{22}, \ldots, a_{2p}\|$
	...	
Product i	u_i	$\|a_{i1}, a_{i2}, \ldots, a_{ip}\|$
	...	
Product n	u_n	$\|a_{n1}, a_{n2}, \ldots, a_{np}\|$

Fig. 2. The problem data structure

Consider the following notation:

$|\bullet|$: be the size of a set,

u_i: the standard unit-per-hour for product i for $i = 1, 2, \ldots n$,

G_k: a mutually disjoint family for $k = 1, 2, \ldots g$,

P: a subset of p attributes and the size is $p' \leq p$, where $|P| = p'$,

$\overline{u}_k$: the average of u's for G_k,

a_{ij}: the jth attribute of product i for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots p$,

$a^*_{kj'}$: the average of j'th attribute of P for G_k for $j' = 1, 2, \ldots p'$,

e_o: the given error for forming families where $e_o \geq 0$,

F: F-statistics

F^*_j: a partial F value and defined as $\frac{\textit{Mean Square of Between Group } (j|P)}{\textit{Mean Square of Within Group } (j+P)}$

, where $j \in P^c$.

The problem is, first, to group products with similar u together into the mutually disjoint family G_k, where $G = |\cup G_k|$, $|G| \le n$, and then to find a subset of attributes P, where $|P|$ $(= p')$ $\le p$. Given $e_o \ge 0$, DRP is formulated as follows:

Find the set of families, G and the subset of attributes, P to

$$\begin{aligned} &\text{Minimize } |G||P| \\ &\text{Subject to} \\ &\quad |u_i - \bar{u}_k| \le e_o,\ i \in G_k,\ \forall\ G_k \\ &\quad \text{and} \\ &\quad F^*_j \le F_{0.95,\,(g-1),\,(n-p'-g),} \text{ where } j \in P^c \end{aligned} \tag{2}$$

The first constraint forces products with similar u to be put together into the same family, and the second constraint, a partial F value of attribute j based on the selected subset of attributes is less than the given critical value, enables P well to account for almost as much variance of u between families as the original attribute set does, namely, P is enough to unambiguously specify each family.

The objective is to minimize $|G||P|$, or equivalently to minimize the amount of data storage required for capacity management calculations. As a result, instead of the original data elements required (n products × p relevant attributes + n production rate values), the resulting data elements required are just $g \times p' + g$. This is illustrated in Fig. 3.

Families	Average UPH	Subset of Attributes
Family 1	$\overline{u}_1$	$\|a^*_{11}, a^*_{12}, ..., a^*_{1p'}\|$
	...	
Family k	$\overline{u}_k$	$\|a^*_{k1}, a^*_{k2}, ..., a^*_{kp'}\|$
	...	
Family g	$\overline{u}_g$	$\|a^*_{g1}, a^*_{g2}, ..., a^*_{gp'}\|$

Fig. 3. The goal data structure

The PCP entails discrimination of the distinct families and classification of new products based on the pre-defined families. Conceptually, the discrimination model describes the difference between objects from the pre-defined families and serves as a classifier for the introduced product into the families. Thus, PCP is finding a way to contrast the previously defined families and formulated as follows:

$$f(P) = \text{Maximize } \left(\frac{\text{variation}_{\text{between-group}}}{\text{variation}_{\text{within-group}}}\right) \tag{3}$$

and the model will be used as classifier of future products to assign them to families.

4. LITERATURE SURVEY

This section provides a survey of topics related to this study. A variety of topics are broadly surveyed focusing on Data Aggregation problem. Data aggregation (DA) has played an important role in many research areas in order to resolve a large size problem such that the given massive problem information are simplified without loss of valuable characteristics of the original problem and one then solves the problem using the solution obtained from the aggregation process. In general DA can be said as any process to manipulate primary data into an aggregate to express data in summary form for further analyses or problem solving.

DA has been considered a major problem in a variety of areas such as Economics, Computer Science, Medical Science, Statistics, and Industrial Engineering as well. For example, consider a transaction database maintained by a special consumer goods retailer. There is a large volume of customers' information identified by many attributes of customers such as sex, residential region, age, education level, occupation, and so on. The huge amount of customers' information is aggregated by customer group defined by a special purpose, some important characteristics of customer group are found and the aggregated information is used to build a marketing strategy or advertisement plan.

DA is an important part of Knowledge Discovering in Database (KDD). KDD is to process large quantities of raw data, identify the most significant and meaningful patterns, and present this knowledge which is appropriate for achieving user's goals. Typically KDD encompasses more than data mining. Elmasri and Navathe (2000) and Piatetsky-Shapiro *et al.* (1996) described general KDD process for extracting useful knowledge form large volume data. Elmasri and Navathe (2000) described that the general process of KDD consists of 6 phases, data selection, data cleaning, data enrichment, data transformation, data mining, and the reporting the discovered data. The first four steps from data selection to data transformation extract essential or relevant data set from a massive raw data stored in data base such that data selection

selects subset of relevant data for KDD processing, data cleaning eliminates or correct erroneous data, data enrichment enhances the data additional source of information, and then data transformation reduces the amount data using aggregation or other methods. The result from the first four steps is used next steps, data mining, to search for some patterns or rules in the data and, finally, the mined data are interpreted and reported. Hu and Cercone (1994) developed an attribute-oriented rough set approach for discovery of decision rule in relation database. The proposed model identifies minimal relevant attributes from all attributes in the data base and automatically generates very concise and more accurate discover decision rules. The methodology is composed of two main steps, data generalization step generalize data base to the desirable level, which is called the general relationship and secondary, in data reduction step, Rough set theory, introduced by Pawlak (1982) is applied to the generalized relation such that the relationships are analyzed and the non-relevant or essential attribute to discovery task are eliminated without losing information about the original database system.

The termed Data Clustering (DC) is considered as a kind of DA techniques. Jain *et al*. (1999) defined that DC is the organization of a collection of patterns (observations, data items, or feature vectors) into clusters based on similarity. Clearly, patterns within a valid cluster are more similar to each other than they are to a pattern belong to a different cluster. The variety of clustering techniques focus on representing data, computing similarity between data elements, and grouping data element and producing information about clusters. Clustering techniques have been applied to a variety of fields such as data mining, document retrieval and pattern classification.

DA has been also played a major role in Database Management System. As the volume of data has grown, many DA techniques have been employed to represent information in a simple format, mange large volume data and efficiently store the data. Semantic Data Models (SDM) were introduced to represent knowledge of data efficiently in database system and Data Abstraction such as classification, aggregation

and association is used to represent more meaning of data in SDM. The Data Abstraction involves aggregating data and identifying common important properties of the aggregates while suppressing irrelevant information and creates a meaningful higher level object. Meanwhile Data Compression was developed to minimize the amount of data to be store or transmitted. It has been studied a long time and very important issue in data base management since 1940's by Debra and Daniel (1987). Data Compression is that data (e.g. string of characters, a set of attributes) is transformed into simple code that contains same information but the physical storage size is as small as possible. DC enables one to reduce the amount of data to be transmitted and communication cost. Graefe (1993) mentioned benefits from DC in database query. First, data abstraction/ or materializing output records is faster because records are shorter. Second, memory can be saved and DC enables one to implement querying activity within given memory space.

Aggregate production planning has been a big concern in production planning model because batch type production of a large number of items and a complex production process lead to very large scale problem hard to be solved. Axäster(1981), Axäster and Jönsson(1984) developed a method to aggregate product data in hierarchical production planning, which was able to transform a complicated MRP system to an efficient aggregated model more relevant to capacity constraints and achieve cost reduction. Bitran Hass and Hax(1981) introduced a aggregate production planning approach using a linear programming model in which the aggregated variables were used instead of a large number of individual variables. Krzysztof *et al.* (1993) devised a methodology to solve a lot size scheduling problem also in which the decision variables were aggregated based on the assumption of similar products using a mixed integer optimization problem.

Some multivariate statistics techniques are frequently employed for data aggregation problem, such as Principal Component Analysis (PCP), Factor Analysis (FA), Variable Selection Method (VSM) and Clustering Analysis (CA). Those are well known multivariate statistical techniques for data reduction. Nickerson and Sloan (1998) used

PCP and FA for analysis of benchmarking data. They studied the methodologies for reduction of performance information and decision variables and analyzed the relationship between the bench marking data using those multivariate statistical techniques.

The followings are, based on the proposed methodology, describing some statistical techniques that were applied to this study. Cluster analysis for product aggregation, variable selection for extraction of relevant product attributes and discriminant analysis for classification of product families and assignment of new products.

Cluster analysis is processed on the basis of similarities. The inputs are required to be similarity measures from which similarities can be computed. Once the measures are determined, a similarity matrix representing between pairs of objects across the measurements is computed. The proper clustering algorithm is then adopted to assign the objects into subgroups on the basis of these criteria. The objective of clustering is to find well contrasted clusters in order to maximize the between-group variation relative to the within-group variation. The uncovered clusters can be differentiated in terms of their mean values on these measure or other characteristics.

Once the groups have been clustered, there might be a between-group spread with attributes of primitive data which is associated with the product. The interest here is finding a way simply and unambiguously to identify the groups. Variable selection enables one to find out how much each attribute or a subset of the attributes contributes to the overall variation between groups. Obviously, it is thought of as a statistical technique to find select significant variables that specify the groups. Several variable selection methods are available, such as the forward method, backward method, stepwise method, and others. Of these, the stepwise method is recommended in this research because it has the advantages of both the forward and backward methods and is simple. The stepwise procedure is processed on MANOVA (Multivariate Analysis of Variance), which is a method to test the equality of means of predictor variables for groups. The conditional F value is adopted as a criterion to determine the important

variable for the stepwise method in this study. The conditional *F* values are concerning interrelationships among the variables and are confounded by variables that contribute redundantly to the overall discrimination of the groups. Utilizing the conditional *F* value, one determines the important variables such that places them into the subset of variables and removes the non-significant variables away from the subset of variables by turns. The procedure stops when there are no more significant variables to enter into the subset. Then, conclusively, the subset of relevant variables to characterize a group spread can be obtained via variable selection method.

Through this research classification of new or revised product is one of important part, which is separating distinct sets of objects and allocating new object into previously defined groups. The two aspects are reciprocal because discriminating may serve as allocator and classifying rule may suggest a discriminatory procedure. Discriminant analysis is a statistical technique for classifying individuals into mutually exclusive groups on the basis of a set of independent variables. In this study Fisher's discriminant function (Fisher, 1938) was introduced to derive a classification model that is linear combinations of independent variables that will discriminate groups in a way that maximizes the between-groups variance relative to the within-group variance, subject to the constraint that each uncovered linear combination must be uncorrelated with previous extracted combinations. Figure 4 illustrates how the discriminant function (linear combinations of primitive variables), *Y*, allows better to discriminate between groups *A* and *B* when compared to discrimination based on the original dimensions a_1 and a_2.

As shown in Figure 4 an applying discriminant function is a kind of data reduction. Only with one dimensioned axis, *Y*, can the two groups be represented. The variables of groups are projected onto discriminant functions and the centroids of groups on it, $\bar{y}_A$ and $\bar{y}_B$, are obtained, as in Fig. 4.

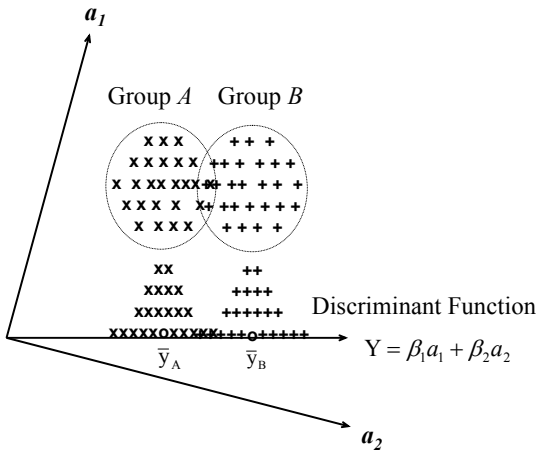


Fig. 4. Basic depiction of discriminant analysis

Next describes how to classify using Fisher's discriminate function. Obtained are the centroids of groups on the new dimension defined by the discriminant functions. If a new observation is added we should determine which previously defined group it belongs to. The measure to determine the allocation or similarities is the Euclidian distance between the group centroids and the scored value of the new observation transformed by the discriminant functions. The group with the shortest distance will be the predicted group for the new observation. Fig. 5 shows how the classification works.

Fig. 5 is a graphical display on the discriminant function Y_1 versus Y_2 plane. The graph displays the inter-group boundaries as well as the centroids, $\bar{y}_A, \bar{y}_B$ and $\bar{y}_C$, associated with each region, and y_{new} represents scored data of the new observation. The centroid of group A has the shortest distance from the new observation. The new object is assigned to group B.

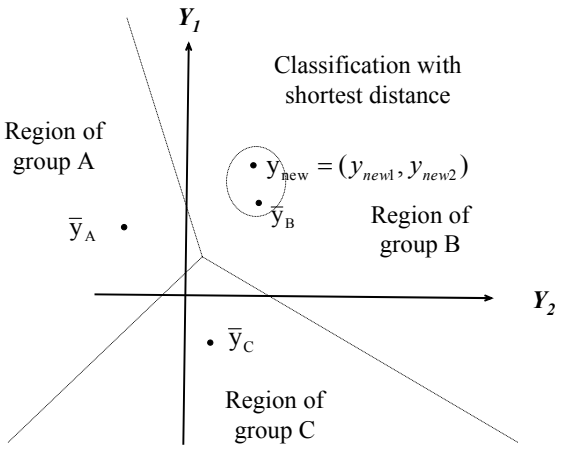


Fig. 5. Classification of new observation

5. SOLUTION APPROACH

This section shows the solution approach for the proposed methodology. The solution approach consists of two problems, DRP and PCP. Those are interdependent and consequential. The general solution approach is presented first and followed by a detailed description of each step of the methodology.

5.1. General Solution Approach

The DRP is divided into two sub-problems: minimizing the number of families (DRP 1) and the subset of attributes specifying the families subject to the constraints (DRP 2), the products assigned to a family must have similar u and the attribute subset can represent variation of u between groups and PCP is also divided into two parts: developing a function to discriminate the pre-defined families (PCP 1) and classifying a future product into a correct family (PCP 2).

According to the problem formulation (2), the objective is minimizing $|G|$ and $|P|$. The problem can be divided into two sequential sub-problems, minimizing $|G|$ and minimizing $|P|$. Then, DRP 1 and DRP 2 are defined as follows.

DRP1: DRP1 minimizes $|G|$ by solving

$$\begin{aligned} &\text{Minimize } |G| \\ &\text{Subject to} \\ &\qquad |u_i - \bar{u}_k| \le e_o \\ &\quad \text{for } k = 1, 2, \ldots, g, \\ &\quad i = 1,2 \ldots n_k \text{ and } i \in G_k \\ &\quad n_k \text{ is number of products belonging to } G_k \end{aligned} \tag{4}$$

From (4) we can obtain unique solution G satisfying the constraint. The procedure will be just sorting the data, dynamic computing the current average of u and comparing the difference between the error and the given criterion, e_o.

DRP 2: DRP 2 minimizes $|P|$ by solving

$$\begin{aligned} &\text{Minimize } |P| \\ &\text{Subject to} \\ &F^*_j \le F_{0.95,\,(g-1),\,(n-p'-g)} \quad j \in P^c \end{aligned} \tag{5}$$

The DRP 2 must be solved consecutively since P is obtained based on G attained from (4). Obviously, different G can cause different P. DRP 2 finds the minimal subset of attributes that explains most of the u variability between families. Once the families have been formed, through DRP 2 we identify the families with the minimum subset of attributes, excluding irrelevant attributes. The obtained subset is enough to refer to a variation between families.

Next concerned problem is discrimination of obtained families for classification (or assignment) of newly introduced product. Namely, the problems are how to separate the distinct families and how to assign a future product, new or revised product, into a previously defined family. First, the discrimination problem is defined as followings.

PCP 1: PCP1 finds a function of P to discriminate G.

$$f(P) = \text{Maximize } \left(\frac{\text{variation}_{\text{between-group}}}{\text{variation}_{\text{within-group}}}\right) \tag{3}$$

The function f will be made of P, discriminate G, and be served as an allocator of the future product. Clearly, the function will generate similarity measures to determine

which family is the closest to a future product. After finding the function f the classification problem, PCP 2 is defined as follows:

PCP2: PCP 2 finds a family k that is the most similar to the future product

$$k = \arg\min_{k \in G} \{s_{ik} \mid i \text{ is a new or revised product}\} \tag{6}$$

Fig. 6 summarizes the general solution approach; the methodology begins by grouping products into families and searches for a significant attribute subset to u variation, using the subset and aggregated families builds a discriminant model to contrast the families, and assigns future product into the most similar family based on the discriminant model obtained in PCP 1.

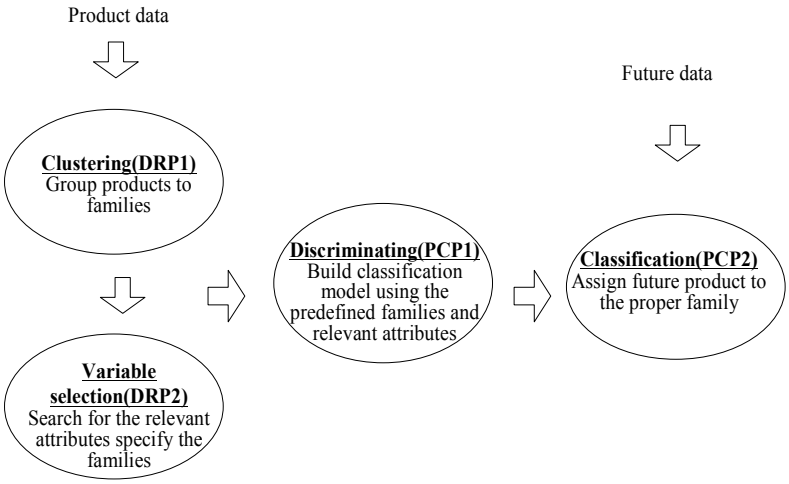


Fig. 6. General solution approach

5.2. Solving DRP 1: Clustering Analysis

The assignment of products into product families is the first step in the methodology since P is a subset of significant attribute to variance of G. The problem data structure described in Fig. 2 can be expressed again in matrix form as follows.

$$
\begin{aligned}
&[u_i]_{n\times 1} = \text{function of } [a_{ij}]_{n\times p} \\
&\quad \text{where } i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, p \\
&\quad u_i : \text{the standard unit per hour for product } i \\
&\quad [u_i]: u \text{ matrix } (n\times 1) \\
&\quad [a_{ij}]: \text{attributes matrix } (n\times p)
\end{aligned} \tag{7}
$$

The constraint in (4) means that the error of product i which is assigned to a family must be less than e_o. The clustering procedure begins by sorting products by u in increasing order, investigates every u one after another. At the every investigation check the error between either the current product or the first product of the current family, and the average u of the current family. If either of those two errors is greater than $e_{0,}$ then a family is formed and the investigation is repeated until all products are assigned to g families. The procedure is illustrated in Fig. 7.

After clustering, n products are mapped into g families and the initial problem shown in (8) is separated into set of matrix by family as follows:

$$
\begin{aligned}
&[u_i]_{n1\times 1} = \text{function of } [a_{ij}]_{n1\times p} \text{ for } i \in G_1 \\
&\quad \ldots \\
&[u_i]_{nk\times 1} = \text{function of } [a_{ij}]_{nk\times p} \text{ for } i \in G_k \\
&\quad \ldots \\
&[u_i]_{ng\times 1} = \text{function of } [a_{ij}]_{ng\times p} \text{ for } i \in G_g
\end{aligned} \tag{8}
$$

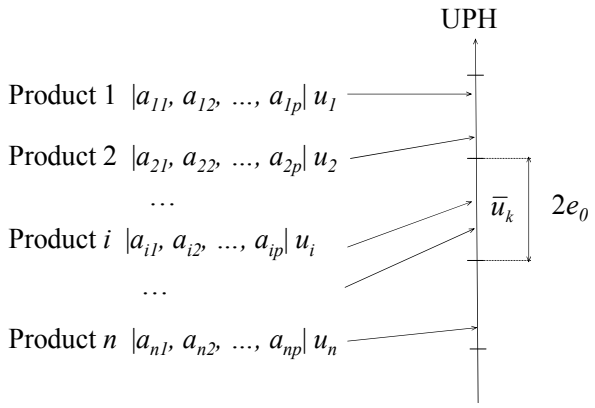


Fig. 7. The clustering procedure

In this step each family is identified by the corresponding $\overline{u}_k$ that is the average of u of the kth family and the mean vector of corresponding attributes, $\{\overline{a}_{k1}, \overline{a}_{k2}, \ldots, \overline{a}_{kp}\}$.

5.3. Solving DRP 2: Variable Selection (Stepwise Method)

There could be significant attributes of all the relevant attributes relative to variation of u between families since the clustering is formed based on u. So, the next step is to find the least number of attributes specifying the previously defined families subject to a constraint that the subset, P is enough to unambiguously specify each family in order to remove the irrelevant attributes. The subset of attributes to be obtained can have good characteristics to differentiate the families. For finding the subset variable selection method is applied. By the variable selection we can obtain a relatively small subset that would contain as much information as the original information composed of all the attributes. The main idea of variable selection method is searching for a set of variables to maximize between-group variation respective to within-group variation.

There are several methods for variable selection, such as forward method, backward method, all possible subsets method, and so on. Of these, the stepwise method is

employed because it is simple and has the advantages of both the forward and backward methods.

The approach starts by computing the sum of squares of within-groups for each family, $\boldsymbol{W_i}$, and the total sum of squares, $\boldsymbol{T}$. The matrix of the sum of squares for ith family, $\boldsymbol{W_i}$, is defined as (9). The following formulas (9)~(10) are well known for the test statistic in a MANOVA and those are brought from the reference (Dillon and Goldstein, 1984)

$$\boldsymbol{W}_i = \sum_{j=1}^{n_i} (\boldsymbol{a}_{ij} - \bar{\boldsymbol{a}}_i)(\boldsymbol{a}_{ij} - \bar{\boldsymbol{a}}_i)^T$$

$\bar{\boldsymbol{a}}_i$: the average attributes vector of family i (9)

$\boldsymbol{a}_{ij}$: jth attribute vector of famiy i

n_i : the number of products in family i

The matrix of the total mean corrected sum of squares for the attributes matrix, which is formulated as

$$\boldsymbol{T} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{a}_{ij} - \bar{\boldsymbol{a}})(\boldsymbol{a}_{ij} - \bar{\boldsymbol{a}})^T \quad (10)$$

$\bar{\boldsymbol{a}}$: total average attribute vector

The total within-group sum of squares for the whole system, $\boldsymbol{W}$, can be computed to add up all $\boldsymbol{W_i}$, $\boldsymbol{W} = \boldsymbol{W_1} + \boldsymbol{W_2} + ... + \boldsymbol{W_g}$. The between-group sum of squares then is attained by $\boldsymbol{B} = \boldsymbol{T} - \boldsymbol{W}$. Utilizing the conditional F value, one determines the important variables such that places them into the subset of variables and removes the non-significant variables away from the subset of variables by turns. The conditional F values are concerning interrelationships among the variables and are confounded by variables that contribute redundantly to the overall discrimination of the groups. The

criteria for a variable to enter into subset and to remove from subset are given. The criteria for entering is the minimum F value to determine if the variable can be entered into the subset, and the criteria for removing is the maximum limit to remove a non-proper variable from the subset. To select the first entering variable, the univariate F value for every variable, then, is computed as follows:

$$F_i = \frac{b_{ii}}{w_{ii}} \tag{11}$$

b_{ii}: ith diagonal element of $\boldsymbol{B}$

w_{ii}: ith diagonal element of $\boldsymbol{W}$

The univariate F ratios are proportional to the ratio of the between-group sum of squares and the within-group sum of squares on uncorrelated term. The attribute with the largest F value is chosen to be first entered into the subset of attributes, P. In successive steps, attributes are added and removed on the basis of their conditional F value. Of the entering candidate attributes, choose the largest one which is greater than the criteria to enter. Remove the lowest one less than maximum F value from the selected subset and put it in the entering candidate attributes. To compute the conditional F value for a variable to be entered and be removed, the concept of sweep operator is introduced to facilitate the calculation for the procedure (Dempster, 1969). The following formulas, (12)~(15), are from the reference. Suppose q attributes are included in P. The within-group sum of squares matrix, $\boldsymbol{W}$, can be partitioned as follows:

$$\begin{bmatrix} \boldsymbol{W}_{11} \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} \boldsymbol{W}_{22} \end{bmatrix} \tag{12}$$

Where $\boldsymbol{W}_{11}$ is $q \times q$. then the matrix $\boldsymbol{W}^*$ is formulated as

$$\boldsymbol{W}^{*} = \begin{bmatrix} -\boldsymbol{W}_{11}^{-1} & \boldsymbol{W}_{11}^{-1}\boldsymbol{W}_{12} \\ \boldsymbol{W}_{21}\boldsymbol{W}_{11}^{-1} & \boldsymbol{W}_{22} - \boldsymbol{W}_{21}\boldsymbol{W}_{11}^{-1}\boldsymbol{W}_{12} \end{bmatrix} = \begin{bmatrix} \boldsymbol{W}_{11}^{*} & \boldsymbol{W}_{12}^{*} \\ \boldsymbol{W}_{21}^{*} & \boldsymbol{W}_{22}^{*} \end{bmatrix} \tag{13}$$

At this stage, $\boldsymbol{W}_{11}^{*}$ is the inverse covariance matrix for selected attributes, which will be used to compute the F value to be removed, and $\boldsymbol{W}_{22}^{*}$ is an adjusted within-group covariance matrix used for the F value to be entered. The F to be entered and the F to be removed are computed as follows.

$$F \text{ to be removed } F_i = \frac{(w^{*}_{ii} - t^{*}_{ii})(n-q-g+1)}{t^{*}_{ii}\,(g-1)} \tag{14}$$

$$F \text{ to be entered } F_i = \frac{(t^{*}_{ii} - w^{*}_{ii})(n-q-g)}{w^{*}_{ii}\,(g-1)} \tag{15}$$

The procedure is repeated until there are no more attributes to be entered. Finally, we can obtain the subset of significant attributes specifying families with the stepwise method.

5.4. Solving PCP 1: Discriminant Analysis

The next task in demonstrating the methodology is how to build a model that identifies the families using the attained attribute subset and families in order to classify new or revised products into the most proper family in the later.

Although through the variable selection the significant attribute subset has been obtained but the chosen attributes might have different contributions to the variation between groups. So, there exists a better way to represent the variation, such as linear combination of the attribute subset, which is more reasonable, balanced, and reducible. As a result, through this step measures will be generated to compute similarity between the future product and the previously defined families.

Fisher's discriminant function (Fisher, 1938) is employed to build the model which consists of linear combinations to characterize the previously defined families. The *g* formed families are treated as individual populations and the *p'* selected attributes of product are used as independent random variables in this step.

The idea of the method is to find a way to maximize the ratio of the variation between groups to the variation within groups. The method generates functions to contrast the pre-defined families. First, compute the total sum of squares and the within-group sum of squares in the same way of (9) and (10). The procedure to find discriminant functions, *Y'*s, can be presented as: Maximize the between-group variation relative to the within-group variation such as maximizing ($\hat{\lambda} = \frac{\hat{\mathbf{b}}^{\mathrm{T}}\mathbf{B}\hat{\mathbf{b}}}{\hat{\mathbf{b}}^{\mathrm{T}}\mathbf{W}\hat{\mathbf{b}}}$). Next, do first-order derivative with respect to $\hat{\boldsymbol{b}}$ and let it be zero. Then $(\boldsymbol{W}^{-1}\boldsymbol{B} - \hat{\lambda}\boldsymbol{I})\hat{\boldsymbol{b}} = 0$ is obtained and this is an eigenvalue solving the problem. The obtained $\hat{\lambda}$'s are eigenvalues and $\hat{\boldsymbol{b}}$'s are eigenvectors of the problem. The *r* discriminant functions are presented as a linear combination of the selected *p'* attributes such that $Y_i = \hat{b}_{i1'}a_{1'} + \hat{b}_{i2'}a_{2'} + ... + \hat{b}_{ip'}a_{p'}$

where i = 1, 2 …r are obtained in the direction to maximize the ratio. The number of discriminant functions must be less than the smallest value either g -1 or p'.

The next task is to determine which discriminant functions to retain; that is, how many discriminant functions can contain a variation of the whole problem dominantly. The Bartlett's test (Batlett, 1947) is applied to solve the concern. Bartlett's test is for testing equality of variance. When applying the Bartlett's test to this problem, its χ^2 approximation can test the significance of the eigen values, $\hat{\lambda}$, of the $\boldsymbol{W}^{-1}\boldsymbol{B}$ matrix because the eigen value represent variance of the corresponding discriminant function. Then the hypothesis of Bartlett's test is

$$H_0 : \hat{\lambda}_1 = \hat{\lambda}_2 = ... = \hat{\lambda}_r$$
$$H_1 : \text{above not true for at least one} \lambda_i$$

If the attributes are multivariate normal within each group with equal variance-covariance matrices, then the significance of the *r* discriminant functions can be assessed by computing a logarithmic function of Wilks' lambda, Λ, (Wilks 1932) such as (17). The modified formulas, (16) ~ (19), for this study are brought from [16,21]

$$V = -\{(n-1) - \frac{1}{2}(p'+g)\} \ln \Lambda \tag{16}$$

The Λ is a Wilks' lambda variable and shown as

$$\Lambda = \prod_{j=1}^{r} (1 - \hat{\lambda}_j)^{-1} \tag{17}$$

Thus, since $\ln a = \ln(1/a)$, the Bartlett's statistic can be written again as

$$V = -\{(n-1) - \frac{1}{2}(p'+g)\} \sum_{j=1}^{r} \ln(1 + \hat{\lambda}_j) \tag{18}$$

The statistic *V* is approximately distributed as a χ^2 random variable with $n + g - 2j$ degree of freedom. Because the discriminant functions are uncorrelated, the additive components of *V* are each approximately variates. Therefore, the significance of the *j*th eigen value, $\hat{\lambda}_j$, can be computed individually as

$$V_j = -\{(n-1) - \frac{1}{2}(p'+g)\} \ln(1 + \hat{\lambda}_j) \tag{19}$$

Successive tests can be done by cumulatively subtracting the individual test statistic,

$V_1, V_2, ...V_r$ from the total test statistic, V such that

Subtracting the first discriminant function

$$V - V_1 \quad with\ (p'-1)(g-2)df.$$

Subtracting the first two discriminant functions

$$V - V_1 - V_2 \quad with\ (p'-2)(g-3)df.$$

...

Bartlett's test can choose the *r'* reduced discriminant functions which have as large a variation as the *r* original functions have.

5.5. Solving PCP 2: Classification

Through discriminant analysis one can obtain new axes to discriminate families, which are called discriminant functions. The method also provides a way to classify a new product, and map it onto a previously defined family. The main idea of the classification is that newly added products can be allocated onto the groups with the shortest Euclidian distance between the scored point of the new product and centroids of the families in the dimensions defined by the discriminant functions. As mentioned, the discriminant function is defined as $Y = \hat{\mathbf{b}}\mathbf{a}$ (e.g., $\mathbf{a}$ is the subset attribute vector and $\hat{\mathbf{b}}$ is the eigen vector). Then the center of each family on discriminant function domain can be computed such as

$$\overline{Y}_k = \hat{\boldsymbol{b}}\overline{\boldsymbol{a}}_{\boldsymbol{k}} \ \ for\ \forall\ k = 1, 2, \cdots, g \tag{20}$$

and the location of a new product on the discriminant dimension can be computed as

$$\hat{Y} = \hat{\boldsymbol{b}}\boldsymbol{a}_{\boldsymbol{new}} \tag{21}$$

In comparing the Euclidian distance between the point, $\hat{Y}$ and each of the center of families, $\overline{Y}_k$, the family that has the smallest Euclidian distance is then selected for mapping the new product.

5.6. Algorithm

This section shows the steps of the proposed data aggregation This section is also classified into two parts for DRP and PCP. For clear comprehension some notations are first defined as follows.

Notation

n: the number of products

n_k: the number of products of family k

p: the number of attributes

p':the number of selected attributes

g: the number of families

e_o: clustering criteria

$\boldsymbol{U_o}$: UPH matrix of initial problem ($n\times1$)

$\boldsymbol{A_o}$: attribute matrix of initial problem ($n\times p$)

$\boldsymbol{U}$: sorted UPH matrix ($n\times1$)

$\boldsymbol{A}$: sorted attribute matrix in order of $\boldsymbol{U}$ ($n\times p$)

$\boldsymbol{a_i}$: ith row vector of $\boldsymbol{A}$

$\boldsymbol{U_k}$: UPH matrix of family k ($n_k\times1$)

$\boldsymbol{A_k}$: attribute matrix of family k ($n_k\times p$)

$\boldsymbol{T}$: total mean corrected sum of square matrix for $\boldsymbol{A}$ ($p\times p$)

$\boldsymbol{W_k}$: sum of square matrix for each individual $\boldsymbol{A_k}$ ($p\times p$)

$\boldsymbol{W}$: within-group sum of squares matrix ($p\times p$)

$\boldsymbol{B}$: between-group sum of squares matrix ($p \times p$)

Note: All matrices marked by ' are made on basis of the selected attributes

$\hat{\lambda}_j$: eigen value for $\boldsymbol{W^{-1}B}$

$\hat{\boldsymbol{b}}_j$: eigen vector for $\boldsymbol{W^{-1}B}$.

Y_j: discriminant functions such as $Y_j = \hat{b}_{j1'}a_{1'} + \hat{b}_{j2'}a_{2'} + ... + \hat{b}_{jp'}a_{p'}$

V: Bartlett's statistic

E: set of attributes to be entered for variable selection

R: set of attributes to be removed for variable selection

F_i : ith F ratio

$\chi^2_{(x)}$: criteria for χ^2 distribution with x df

$F_{(x)(y)}$: criteria for F distribution with x and y df

5.6.1. DRP Algorithm

In the beginning the problem has n products and p attributes, so the problem is formed as matrix type data, u matrix, $\boldsymbol{U_o}_{n\times 1}$, and attribute matrix, $\boldsymbol{A_o}_{n\times p}$.

Initialization:

Set $\boldsymbol{U}$ such as sorting $\boldsymbol{U_o}$ in increasing order, and set $\boldsymbol{A}$ such as arranging the rows of $\boldsymbol{A_o}$ according to the order of $\boldsymbol{U}$.

Set $k = 1$, $i = 1$, $s = 1$, $n_o = 0$, and n = the number of total products

Set e_0

Step1. Clustering based on u (DRP 1)

Set $\boldsymbol{U_k} = \{\Phi\}$

While $i \le n$

Put u_i in $\boldsymbol{U_k}$, $\boldsymbol{U_k} = \boldsymbol{U_k} + u_i$

Compute $\overline{u}_k$, the average of u in $\boldsymbol{U_k}$.

If $|u_i - \overline{u}_k|$ or $|u_s - \overline{u}_k| \geq e_o$

Set $n_k = i - n_{k-1} - 1$

Set $k = k + 1$ and $s = i$

Set $\boldsymbol{U_k} = \{\Phi\}$

Put u_i in $\boldsymbol{U_k}$, $\boldsymbol{U_k} = \boldsymbol{U_k} + u_i$

Put $\boldsymbol{a_i}$ in $\boldsymbol{A_k}$, $\boldsymbol{A_k} = \boldsymbol{A_k} + \boldsymbol{a_i}$

Else

Set $i = i + 1$

End If

END While

Set $g = k$

Step2. Select significant attributes (DRP 2)

Step 2.1 Generate MANOVA table

Compute $\boldsymbol{W_k}_{\ p\times p}$, the sum of square matrix for each individual $\boldsymbol{A_k}$ for all $k = 1, 2 \ldots g$ as shown (9)

Compute $\boldsymbol{T}_{p\times p}$., the total mean corrected sum of square matrix for $\boldsymbol{A}$ as shown (10)

Compute $\boldsymbol{W}_{p\times p}$, the total within group sum square matrix such that $\boldsymbol{W} = \boldsymbol{W_1} + \ldots + \boldsymbol{W_g}$

Compute $\boldsymbol{B}_{\boldsymbol{p\times p}}$, the between group sum of squares matrix such that $\boldsymbol{B} = \boldsymbol{T} - \boldsymbol{W}$.

Step 2.2 Variable selection (Stepwise method)

Set $E = \{\Phi\}$ and $R = \{1, 2..., p\}$

Set $q = 1$

Compute F_i, the univariate F value as shown (11) for $i \in$ R

If the largest $F_i \geq F_{0.95, (g-1), (n-g)}$

Put the attribute into E subtracting from R, $E = E + i$ and $R = R - i$

Else

Stop Step 2.2

While

Partition $\boldsymbol{W}$ and $\boldsymbol{T}$ as shown (12), $\boldsymbol{W_{11}}_{q\times q}$.

Compute $\boldsymbol{W^*}$ and $\boldsymbol{T^*}$ as shown (13)

Compute F_i to be entered for $i \in R$ as shown (15)

If the largest $F_i \geq F_{0.95,\,(g\text{-}1),\,(n\text{-}q\text{-}g)}$

Put i into E subtracting from R, $E = E + i$ and $R = R - i$

Else

Stop While

END If

Compute F_j to be removed for $j \in E$ as shown (14)

If the smallest $F_j \leq F_{0.90,\,(g\text{-}1),\,(n\text{-}q\text{-}g+1)}$

Put j into R subtracting from E, $R = R + j$ and $E = E - j$

END If

END While

5.6.2. PCP Algorithm

Initialization:

Set $\boldsymbol{A'}_{n\times p'}$ such as taking the selected attribute data columns.

Step1. Build classification model (PCP 1)

Step 1.1 Partition $\boldsymbol{A'}$ into $\boldsymbol{A'_k}$

Partition $\boldsymbol{A'}$ into individual $\boldsymbol{A'_k}$ according to Step 1.1

Step 1.2 Finding discriminant functions (Fisher's discriminant function)

Compute $\boldsymbol{T'}_{p'\times p'}$., the total mean corrected sum of square matrix for $\boldsymbol{A'}$

Compute $\boldsymbol{W'_k}_{p'\times p'}$, the sum of square matrix for each individual $\boldsymbol{A'_k}$ for all k =

1, 2.., g

Compute $\boldsymbol{W'}_{p' \times p'}$, the total within group sum square matrix such that $\boldsymbol{W'} = \boldsymbol{W'_1} + \ldots + \boldsymbol{W'_g}$

Compute $\boldsymbol{B'}_{p' \times p'}$, the between group sum of squares matrix as $\boldsymbol{B'} = \boldsymbol{T'} - \boldsymbol{W'}$

Compute eigen value, $\hat{\lambda}_j$, and eigen vector, $\hat{b}_j$, for $\boldsymbol{W^{*-1}B^{*}}$.

Set r discriminant functions such as $Y_j = \hat{b}_{j1'} a_{1'} + \hat{b}_{j2'} a_{2'} + \ldots + \hat{b}_{jp'} a_{p'}$ where j = 1, 2... r in decreasing order of eigen value

Step 1.3 Screen out significant discriminant functions (Bartlett's test)

Compute V, Bartlett's statistic, for all discriminant functions as shown in (18)

Compute V_j, Bartlett's statistic for individual discriminant function for j = 1, 2..., r as shown (19)

Set i =1, $x = p$ and $y = g$

While ($i \leq n$)

If $V - V_i \leq \chi^2_{0.01, (x-1)(y-1)}$

Stop While

Else

Set $x = x$ -1 and $y = y$ -1

END If

END While

Set the screened discriminant functions as classification model.

Step 2 Assign a new product to proper family (PCP 2)

Compute centroid of each family on basis of the classification model generated Step 1.3

Compute score of new product on basis of the classification model

Compare distance from each family's centroid and assign the one to a family with smallest distance.

6. EXPERIMENTAL STUDY

In this section some experimental results are shown, and the efficiency of the proposed methodology is demonstrated. The objective of the experiment for DRP is to evaluate a data reduction, how efficient the proposed method can group products into families without exceeding a given error and identify the subset of attributes that explains the u-based families. In the other an efficiency of classification is assessed for PCP, how well the method can assign new or revised products into correct family.

The experimental scenario is designed such that the problem is randomly generated that has 10 attributes and 100 products where attribute values are normally distributed and are standardized to convert the values in standard unit in the later analysis. The controllable factors are, in the experiments, A: the given number of families, B: the percentile of relevant attributes, and C: the type of u generating function made of the relevant attributes. The given number of families (A) is related to the constraint in the problem formula, $| u_i - \overline{u}_k | \leq e_o$. As described in Figure 7, $2 \times e_o$ represents a interval of a family in u, so the number of families can be approximately expected by e_o. The percentile of relevant attributes (B) is used to determine how correct the methodology screens out the given relevant attributes into the subset of attributes. To give significance to the given attribute, the u generating functions (C) are created only using the given attributes and how the different types of the functions influence the performance is shown in this section.

In addition, the coefficient's (β_i) variability, C in the u generating function is given such that ~U(1,10) and ~U(1, 1000) to validate if there is any influence incurred by the coefficient's variability in the u generating function. During the experiment $3 \times 3 \times 5 \times 2 = 90$ problem types and 30 replications per problem type are executed, totaling 2700 runs. The following Table 1 shows the controllable factors mentioned:

Table 1. The experimental factors

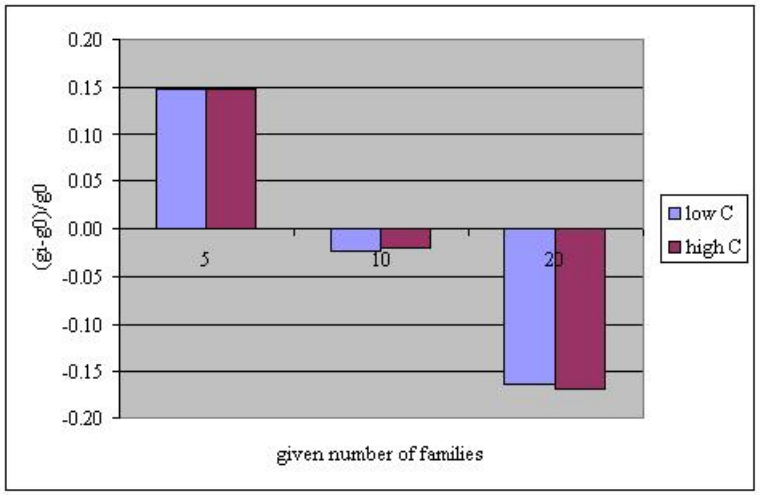
No.	Controllable factors	Used quantities
1	Given number of families	5,10 and 20
2	Percentile of relevant attributes	30%, 50% and 70%
3	Type of u generating function	Type1. Linear: $u = \beta_1 a_1 + \beta_2 a_2 + \cdots$ Type 2. Exponential: $u = \beta_1 e^{a1} + \beta_2 e^{a2} + \cdots$ Type 3. Quadratic: $u = \beta_1 a_1^2 + \beta_2 a_2^2 + \cdots$ Type 4. Hybrid: 1) + 2) + 3) Type 5. Random: No relationship
4	The coefficients in the u generating function	Randomly generated ~U(1,10) and ~U(1, 1000)

The statistics used for determining efficiency of DRP are: 1) Deviation from given number of families = $(g - g_o)/ g_o$ (g is the number of families obtained and g0 indicates the given number of families determined by e_o). 2) Deviation from a given number of significant attributes = $(a - a_o)/ a_o$ (a is the number of the obtained significant attributes and a_o indicates the given number of significant attributes) and proportion of correct variable selection

The results in Table 2 and Fig. 8 show the test statistics for the average of (gi – go)/ go in the experiment. It is shown that the number of generated families is a little different from the given number of families and as the given families (go) increase, the number of created families (gi) drops off. In case the number of given families are 5 and 10, the clustered families are just one less than the given number. In case of 20 families, the difference is less than 2 families. In view of data reduction as the number of families increases the method performs better. It is also known that there is not any obvious difference in the coefficient's (βi) variability, C from the result.

Table 2. Test result for deviation from given number of families

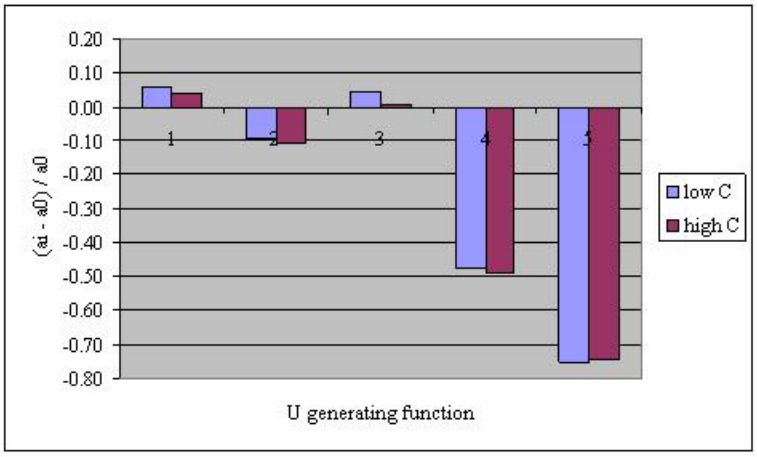
Given number of families	Average of $(g_i - g_o) / g_o$	
	Low variability of C	High variability of C
5	0.1471	0.1471
10	-0.0236	-0.0200
20	-0.1639	-0.1693



Note: C means the coefficients variability

Fig. 8. Deviation from given number of families

Table 3 and Fig. 9 below demonstrate that the attribute dimension reduction is well carried out. The proportion of correct variable section was tested in Table 4 as well. As illustrated in Table 3 and Fig. 9, in case of a non-combined u generating function such as 1, 2 and 3 the variable selection is relatively accurate on both the deviation and the rate but the others are not. In case 1~3, the obtained a_i are at most approximately one less than a_o, which means a_i are mostly close to a_o. In case 4, the combined function, some terms, particularly like exponential, can contribute much more to make up u, thus a large reduction took place. As expected, when the u is generated randomly the variable selection doesn't work in function type 5 because there is no relation between attribute and u. Table 4 shows the variable selection is relatively inaccurate in the case of hybrid function type. The reason is the same as the result right before. Some dominant attributes like exponential terms are selected as significant attributes.



Note: C means the coefficients variability

Fig. 9. Deviation from given number of significant attributes

Table 3. Deviation from given number of significant attributes

U generating function	Average of $(a_i - a_0) / a_0$		Function type
	Low variability of C	High vaiability of C	
1	0.06	0.04	linear
2	-0.09	-0.11	exponential
3	0.05	0.01	polynomial
4	-0.48	-0.49	combined
5	-0.75	-0.74	no relation

Table 4. The proportion of correct variable selection

U generating function	The correct variable selection for low variability of c	
	Proportion of correct variable selection	Standard deviation for the rate
1	0.96	0.054
2	0.91	0.080
3	0.97	0.046
4	0.68	0.079
5	0.51	0.098

The next concern to be considered is the ability to classify future products for PCP: namely, when new or revised products come in how well the ones can be classified to

their correct families. In addition to the proposed method using discriminant analysis, several classification methods are employed for comparison.

In the proposed model the scored data by discriminant functions are used to compute similarity, the distance between the new product and centroids of families, and then the families with the smallest distance is the predicted family for the new one. Some of the other models can be considered such that non-scored data, significant attribute data filtered by variable selection, are used as the basis for calculating distances or similarity. In other words, the distances are estimated on an attribute dimension, not on a discriminant function dimension. And as mentioned, the eigen values, $\hat{\lambda}$'s have information for variation, and eigen vector coefficients, $\hat{\beta}$'s show the contribution of the corresponding attribute on the corresponding discriminant function. Also, the F value indicates the matching variables' variation to between-group spread. So, the ratio's individual value to the total sum of the values is multiplied when computing distances. Table 5 shows a variety of classification methods concerned in this section, including the proposed method.

A total of eight methods are considered for classification including the proposed model. For this experiment test data are generated in the same way as before, but in the former experimental result the coefficients' variability didn't mean much, so the coefficients' variability in u generating function ~U(1,10) is taken. During the experiment 3×3×5 = 45 problem types and 30 replications per problem type are executed, totaling 1350 runs.

The relevant measures for the performance of classification are the proportion of correct classification (i.e. the number of test products correctly classified / the total number of test products) and error deviation incurred by classification (i.e. $(e_0 - |u_{new} - \overline{u}_{classified\_family}|)/e_0$). Table 6 and Fig.10 illustrate the test results for the correct classification rate. In case the given number of families is 5, the best success classification rate is a little more than 40%. Although the rate is greater than 20% that is the random assignment for 5 families' case. The result indicates poor classification.

Table 5. The applied classification methods

Method	Data for distance computing	Weight	Distance with *k*th family	Description
1	Scored data	None	$\sqrt{\sum_{j=1}^{r'}(y_{new\,j}-\bar{y}_{kj})^2}$	Euclidian distance of the scored value between new product and centroid of family (The proposed model)
2	Significant attribute data	None	$\sum_{j=1}^{p'}\lvert a_{new\,j}-\bar{a}_{kj}\rvert$	Absolute difference of attribute value between new product and centroid of family
3	Scored data	None	$\sqrt{\sum_{j=1}^{p'}(a_{new\,j}-\bar{a}_{kj})^2}$	Euclidian distance of the attribute value between new product and centroid of family
4	Significant attribute data	$w_j=\frac{F_j}{F_{total}}$	$\sqrt{\sum_{j=1}^{p'}(w_j\times(a_{new\,j}-\bar{a}_{kj}))^2}$	Similar to 3 but F value ratio is weighted for using the property that each attribute has different significance.
5	Scored data	$w_j=\frac{\hat{\lambda}_j}{\hat{\lambda}_{total}}$	$\sqrt{\sum_{j=1}^{r'}w_j\times(y_{new\,j}-\bar{y}_{kj})^2}$	Similar to 1 but eigen value ratio is timed because λ ratio means variation contribution for the corresponding discriminant function.
6	Significant attribute data	$w_j=\frac{\sum_{i=1}^{r'}\hat{b}_{ij}}{\sum_{i=1}^{p'}\sum_{l=1}^{r'}\hat{b}_{il}}$	$\sqrt{\sum_{j=1}^{p'}(w_j\times(a_{new\,j}-\bar{a}_{kj}))^2}$	Similar to 4 but sum of coefficients of eigen vector ratio is used as weight.
7	Scored data	$w_j=\frac{\sum_{i=1}^{p'}\hat{\lambda}_{ij}\hat{b}_{ij}}{\sum_{i=1}^{r'}\sum_{l=1}^{p'}\hat{\lambda}_{il}\hat{b}_{il}}$	$\sqrt{\sum_{j=1}^{r'}(w_j\times(y_{new\,j}-\bar{y}_{kj}))^2}$	Similar to 5 but eigen value×eigen value ratio is timed.
8	Sort the attributes in decreasing order of univariate F value, compute absolute difference of the first ranked attribute between the new product attribute and the centroid of families (i.e. $\lvert a_{new}-\bar{a}_k\rvert$) and then take as many families with short distance as the number of the selected attributes. Next , compute the distances for next ranked attribute and put the worst one out. Repeat it until only one family remains.			

Table 6. The proportion of correct classification

Given number of families	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8
5	0.42	0.40	0.44	0.43	0.35	0.19	0.20	0.32
10	0.25	0.24	0.27	0.27	0.20	0.10	0.10	0.20
20	0.15	0.15	0.17	0.17	0.12	0.05	0.06	0.12

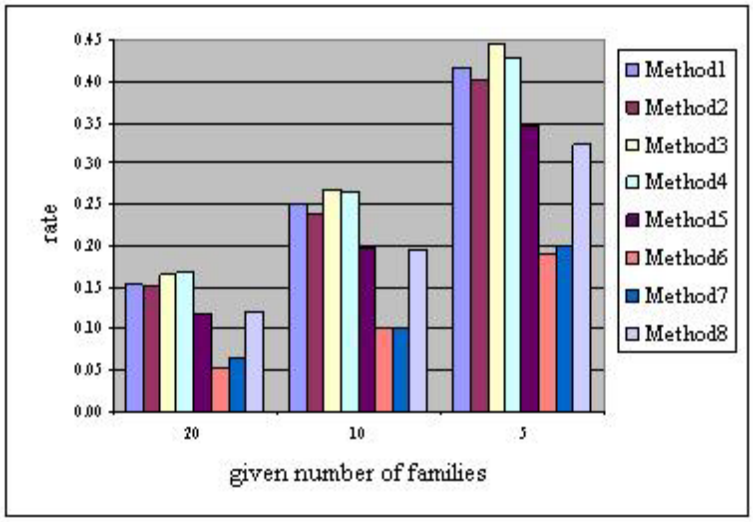


Fig. 10. The proportion of correct classification

The first 4 methods were demonstrated alike but method 3 and 4 are little better than the proposed method.

Another measurement is considering an error caused by classification, which shows the deviation between e_o and an error incurred by classification. The incurred error is a difference between the average of u of the assigned family, $\overline{u}_{classified\_familyi}$ and corresponding u_{new} computed by u generating function. As known, $2\times e_o$ indicates u interval of a family in Fig. 7, so the measurements means how far the allocation is from the real value of u. The result is illustrated in Table 7 and Fig. 11. The value in the table is an average of the error deviation, and the below value in blankets is standard deviation

of the error deviation. In case the given number of families is 5, the averages are around 1.5 and the standard deviations are little less than 2. The most wrong classification is assigned onto an adjacent family because the error is one and half times as much as u_o. Clearly, the poor results in classification is cause by the fact that the centroids of some families in classification model are so close to each other that wrong allocations happen in a neighboring area. If the centroids representing each family are spread the classification model would allocate new product more correctly.

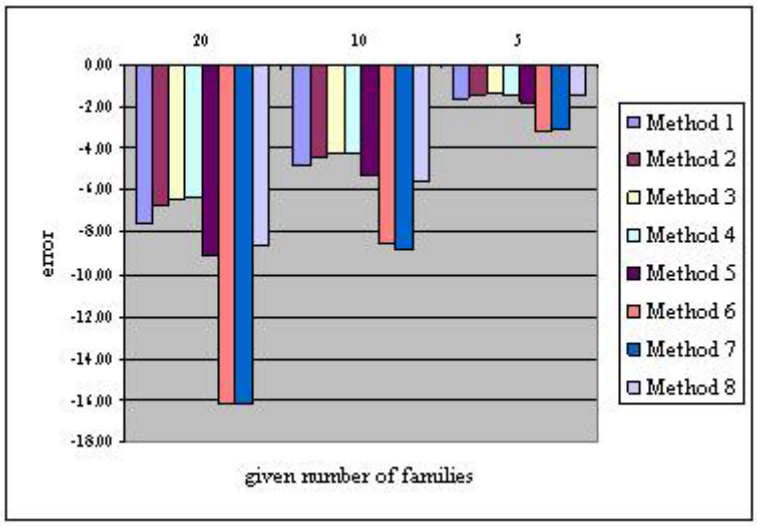


Fig. 11. The error incurred by classification

Table 7. The error incurred by classification

Given number of families	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8
5	-1.59 (2.09)	-1.46 (1.81)	-1.35 (1.73)	-1.38 (1.73)	-1.77 (2.01)	-3.20 (2.44)	-3.16 (2.38)	-1.43 (1.82)
10	-4.80 (4.47)	-4.43 (3.79)	-4.26 (3.66)	-4.25 (3.61)	-5.29 (4.30)	-8.49 (5.29)	-8.79 (5.13)	-5.55 (4.31)
20	-7.61 (7.59)	-6.70 (5.99)	-6.46 (5.78)	-6.36 (5.64)	-9.05 (7.72)	-16.22 (9.18)	-16.22 (8.96)	-8.62 (7.14)

7. CONCLUSIONS

In this paper we discussed a data aggregation methodology for capacity management. The experimental results illustrated the methodology have potential although the experimental result of the classification was poor. The data reduction of products (DRP) was shown to be efficient and the solutions obtained were close to the given number of families. In variable selection the relevant attributes were found accurately as well, but in product classification (PCP) the result was poor. For the better classification it is asked to develop the other methods. The regression model approach is one of suggested approaches, which build a model to predict u using the selected attribute subset and classifies the new product according to the predicted value of u obtained by the regression model. It would be better than the proposed model.

Future directions for research are the studies of developing more accurate classification method, applying the proposed methodology to more realistic design environment and the extension of the methodology to other area's problems, quality control, marketing and so on. The applicable situation is; if performance is represented by primitive variables, the methodology could be applied to a situation that need to categorize performance, select significant variables to variance to performance, and expect how a change of the primitive variables influences on the performance.

REFERENCES

Axsäter, S. (1980). Aggregation of product data for hierarchical production planning. *Operations Research*, *29,* 744 – 756.

Axsäter, S. & Jönsson, H. (1984). Aggregation and disaggregation in hierarchical production. *European Journal of Operational Research*, *12*, 338 - 350.

Bitran, G.R., Hass, E.A., & Hax, A.C. (1981). Hierarchical production planning: Single stage system. *Operations Research*, *29*, 717 – 743.

Dempster, A.P. (1969). *Elements of continuous multivariate analysis*. Addison Wesley, Reading, MA.

Dillon, W.R. & Goldstein, M. (1984). *Multivariate analysis-method and application.* John Wiley & Sons, New York, NY.

Dowlatshahi, S. & Nagaraj, M. (1998). Application of group technology for design data management. *Computers Ind. Engng*, *34*, 235-255.

Elmasri, R. & Navathe, S.B. (2000). *Applied fundamentals of database systems*. 3rd ed. Addison-Wesley, Reading, MA.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of ACM*, *39*, 27-34.

Fisher, R.A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics*, *8*, 376-386.

Geoffrion, A.M. (1977). *A priori* error bounds for procurement commodity aggregation in logistics planning models. *Naval Research Logistics Quarterly*, *24*, 201-212.

Goldstein, R.C. & Storey, V.C. (1999). Data abstractions: why and how. *Data & Knowledge Engineering*. *29*, 293-311.

Graefe, G. (1993). Query evaluation techniques for large database. *ACM Computing Surveys*, *25*, 73-170.

Hu, X. & Cercone, N. (1994). Discovery of decision rules in relational data bases: a rough set approach. *Proceedings of the 3rd International Conference on Information and Knowledge Management,* 392-400.

Hull, R. & King, R. (1987). Semantic database modeling: survey, applications, and research issues. *ACM Computing Surveys*, *19*, 201-260.

Johnson, R.A. & Wichern, D.W. (1998). *Applied multivariate statistical analysis.* 4th ed. Prentice Hall, Englewood Cliffs, NJ.

Lelewer, D.A. & Hirschberg, D.S. (1987). Data compression. *ACM Computing Surveys*, *19*, 262-296.

Montgomery, D.C. (1997). *Design and analysis of experiments*. 4th ed. John Wiley & Sons, New York, NY.

Murray, G.D. (1977). Cautionary note on selection of variables in discriminant analysis. *Applied Statistics*, *26*, 246-250.

Nickerson, J.A. & Sloan, T.W. (1999). Data reduction techniques and hypothesis testing for analysis of benchmarking data. *IJPR*, *37*, 1717-1741.

Pawlak, Z. (1982). Rough sets. *Int.J. Comput. Inf. Sci. 11*, 341-356.

Pieńkosz, K. & Toczylowski, E. (1991). On aggregation of single-stage production systems with limited inventory levels. *Operations Research*, *41*, 419 – 426.

Shapiro, J.F. (2001). *Modeling the supply chain*. Duxbury Thomson Learning, Pacific Grove, CA.

Smith, J.M. & Smith, C.P. (1977). Database abstractions: aggregation and generalization. *ACM Transactions on Database Systems*, *2*, 105-133.

VITA

Yong Woo Lee was born on January 30, 1970 in Seoul, Korea, son of Dukjoo Lee and Sunja Kim. He earned a B.E. degree in industrial automation engineering from Inha University, Korea, in 1995. After graduation, he worked for the Automobile Parts Division, Samsung Electromechanics Co., Ltd. as a production engineer for four years and Baikdoo Electronics, Co., Ltd. as a system developer for a year. In 1999 he received an award from Samsung for the authentification of the QS9000 quality system.

In August 2000, he began his M.S. in industrial engineering at Texas A&M University. He will start his Ph.D. program in the same department in the fall of 2003. He is married to Ahyoung Han, daughter of Dongil Han and Hangsuk Seo and has a daughter, Diane Lee.

His permanent mailing address is:

935 Willow Pond Street

College Station TX 77845, U.S.A.