ELSEVIER

European Journal of Operational Research 123 (2000) 640–651

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/dsw

Theory and Methodology

Clustering of objects and attributes for manufacturing and marketing applications ☆

Asoo J. Vakharia [a,*], Jayashree Mahajan [b]

[a] Department of Decision and Information Sciences, Warrington College of Business Administration, University of Florida, PO Box 117169, Gainesville, FL 32611-7169, USA

[b] Department of Marketing, Warrington College of Business Administration, University of Florida, Gainesville, FL 32611, USA

Received 1 December 1997; accepted 1 February 1999

Abstract

The problem of optimally grouping a set of '*n*' objects into a set of '*k*' clusters is one of the classic unresolved problems in the clustering literature. Thus, prior work in clustering has primarily focused on developing computationally efficient heuristics for grouping objects into clusters. However, there are a variety of applications where both the object and associated attribute set needs to be simultaneously grouped. In this paper, we propose such an alternative view of the clustering problem. Essentially, we adopt the perspective that not only do we want to group the set of '*n*' objects into clusters but that the attribute set '*q*' (associated with the object set '*n*') is also to be simultaneously grouped into clusters. Given the combinatorial nature of this problem, we develop a two-phase sequential algorithm. At the first phase, we identify the optimal set of '*k*' object clusters, this is followed by a polynomially bounded procedure at the second phase for optimally assigning the attribute set to each object cluster identified in the first phase. We illustrate our approach by applying it to two decision problems in manufacturing and retailing that require a determination of object and attribute clusters. © 2000 Elsevier Science B.V. All rights reserved.

1. Introduction

In general, the clustering problem is concerned with identifying 'natural groups of objects such that the degree of natural association is *high* between objects within each group and *low* between objects in different groups'. The combinatorial nature of this problem stems from the fact that the number of alternative ways of clustering n objects into m groups is given by $S_n^{(m)}$ (a Sterling number of the second kind) where

[*] Corresponding author. Tel.: +1 352 392 8571; fax: +1 352 392 5438.

E-mail address: vakharaj@dale.cba.ufl.edu (A.J. Vakharia).

☆ This paper is based upon work supported in part by the National Science Foundation under Grant No. DDM-92-15432. We would also like to acknowledge the support of the Center for Retailing Education and Research, Warrington College of Business Administration at the University of Florida for providing data used in the paper.

PII: S0377-2217(99)00103-4

$$S_n^{(m)} = \frac{1}{m!}\sum_{k=0}^{m}\left[(-1)^{m-k}\binom{m}{k}k^n\right]. \quad (1)$$

For $n = 25$, $m = 5$, $S_{25}^{(5)} = 2,436,684,974,110,751$. Further, note that the computational complexity of the problem increases even more if the object and attribute sets have to be simultaneously grouped. However, often there are business situations where there is a close link between objects and attributes so that a manager may need insights on the object groups along with the attribute groups. Consider the following two applications in manufacturing and retailing.

In cellular manufacturing, one of first design issues deals with the identification of product families and production cells. Typically, these product families contain individual products which are processed on a set of similar equipment types while production cells contain distinct equipment types which process similar products. A multitude of approaches have been proposed to solve the cell formation problem (i.e., the problem of identifying product families and production cells) and a large subset of these methods uses clustering techniques. The reader is referred to Selim et al. (1998) for a recent review of alternative cell formation approaches. When using clustering for cell formation, researchers have either identified product families first and then based on equipment types requirements and availabilities, created corresponding production cells. Alternatively, other researchers have identified production cells first, and then allocated products for processing to one or more cells. Regardless of the approach taken, there are two major shortcomings with this prior work. First, none of these studies have relied upon an optimization model for cluster identification (of families or cells). Given that clustering results are known to be technique specific, the absence of an optimization model for clustering products or equipment types makes it very difficult, if not impossible, to identify the preferred clustering method. Second, all clustering methods for cell formation require the user to prespecify the number of clusters to be identified. Although this is not a serious problem if analyzing a single data set, it becomes more of an issue for providing general guidelines for users of a clustering based methodology. Additionally by decomposing the problem and relying on heuristic based approaches results in a less than desirable solution. In sum, these shortcomings of current research in cell formation raises the need for a clustering methodology which uses an optimization model, helps to detect the number of clusters in any data set, and can be used to identify product families and production cells.

In retailing there is an analog to the above situation that we refer to as the market segmentation/product assortment problem. A critical decision faced by retailers is to determine the optimal product assortment for each retail location (Levy et al., 1998). However, this assortment could be dependent on the market segment served by that retail location. Typically retailers identify market segments a priori and subsequently, develop a generic assortment plan across locations. The assortment of products varies across store locations only due to the physical capacity of the store. In this context, cluster analysis is frequently applied to identify market segments (Hoek et al., 1996). Clearly, a modelling approach that develops product assortment groupings based on specific customer segments would be beneficial. Our clustering methodology incorporates a formal algorithm which can be used for making both the market segmentation and product assortment decisions.

Hence, what is relevant to both applications described above is that both product families or market segments and production cells or product assortments need to be identified. The reason for this is that in both applications, the objects and attributes are very closely related so that identifying object clusters is dependent on the attribute set while identifying the attribute clusters is a function of the object set. The focus of this paper is to propose an integrated model for creating object *and* attribute sets. Given the computational complexity of this integrated model, we decompose the problem into two subproblems. At the first phase, we use a modified version of the p-median model which identifies the optimal number of object clusters and groups individual objects into non-overlapping clusters. Based on this solution, we propose a polynomially bounded algorithm for optimal individual attribute assignment to each given object cluster.

The remainder of this paper is structured as follows. In the next section, we briefly review the relevant clustering literature. This is followed by a description of the general model for simultaneously clustering objects and attributes in Section 3. In Section 4, we describe the two-phase methodology for solving this model. An illustration of our approach to published data sets for cell formation and to data obtained from a retailer on the market segmentation/product assortment problem is discussed in Section 5. Finally, in Section 6, we describe the implications and conclusions of this paper.

2. Relevant literature

A review of clustering reveals that due to the combinatorial nature of this problem, research has primarily focused on developing heuristic algorithms for obtaining good solutions in a reasonable period of time. The more common heuristic clustering approaches are briefly described below (Anderberg, 1973).

Hierarchical agglomerative methods: All these procedures start by assuming that each object is a separate cluster by itself and in a hierarchical manner, objects are combined to form larger clusters. If no stopping criterion is specified, then the procedures terminate when all objects are in a single cluster. Examples are the Linkage methods (Single, Complete and Average), WARD's method, Centroid method, Median method and the Set Merging method.

Hierarchical divisive methods: These procedures start by assuming that all objects belong to a single cluster and hierarchically, we divide the clusters. Again, if no stopping criterion is specified, then the procedures terminate when each object represents an individual cluster. These methods can be categorized as Monothetic – based on possession/non-possession of a single attribute (examples are Association Analysis and the Automatic Identification Detector method); and Polythetic – based on values taken by all attributes (an example is the Mc-Naughton-Smith et al. procedure).

Iterative partitioning (non-hierarchical divisive) methods: These start by assuming an initial partitioning of objects into a predetermined number of clusters with each cluster identified with a seed point (or nucleus). Iteratively, objects are reallocated between clusters to improve an objective and for each reallocation, the seed (nucleus) for the affected clusters is recomputed. An example is the *K*-means method which is commonly used in social science applications.

Density search methods: These methods attempt to identify regions of high 'density'. Some examples are the TAXMAP method, CARTETT COUNT method, Mode Analysis and the Method of Mixtures.

Factor analytic methods: These methods attempt to identify underlying dimensions of objects to create clusters consisting of objects with similar 'loadings' on these dimensions. The most common examples are Principal Components Analysis and Factor Analysis.

Clumping methods: This is the only type of clustering procedure which identifies overlapping clusters. Typically partitions are based on minimizing a cohesion function between groups.

Graph theoretic methods: Using these methods, a graph with nodes representing objects and edges between nodes weighted using some criterion to be optimized is created. A partitioning scheme used to identify a predetermined number of clusters. An example of the partitioning schemes which have been used are the Minimum Spanning Tree (MST) and Lin–Kernighan Exchange procedures.

As noted earlier, all the methods described earlier are heuristic in nature primarily due to the computational complexity of the clustering problem. However, mathematical programming and dynamic programming models have also been developed (Vinod, 1969; Jensen, 1969; Rao, 1971). The mathematical programming object clustering formulation proposed by Vinod (1969) is similar to the *p*-median formulation developed to address the uncapacitated plant location problem. The primary differences between these two models stems from two aspects: (i) in identifying clusters using Vinod's model, we select an object around which a cluster is created while in the *p*-median model, facilities and products are treated as distinct entities; and (ii) The *p*-median model for uncapacitated facility location typically assumes positive fixed plant location costs while Vinod's clustering model assumes zero cluster creation costs. Optimization based heuristics to solve the *p*-median model with non-negative plant location costs have been proposed by several researchers. For example, Mulvey and Crowder (1979) propose the use of a sub gradient method to determine lower bounds and a simple search method for determining upper bounds while Klastorin (1985) proposed an adaptation of Erlenkotter (1978) DUALOC method for this problem. In another paper, Mulvey and Beck (1984) develop a primal heuristic augmented with local search to solve a capacitated version of the *p*-median model. In this paper, we use a modified version of Vinod's object clustering model.

One of the key concerns in evaluating the efficacy of any clustering method is how to determine the quality of a cluster solution. Although several authors have evaluated the quality of solutions obtained using one or more of these clustering techniques (e.g., Blashfield, 1976; Milligan, 1980; Milligan and Cooper, 1987), the emphasis in such comparisons seems to have been on evaluating methods to determine which ones extract predetermined and inherent clusters in data. The primary reasons for this are twofold. First, there is no comprehensively acceptable measure of clustering efficiency which has been accepted across all clustering applications. Thus, researchers have developed measures for addressing specific applications (see, for example, Chandrasekharan and Rajagopalan (1986) who have developed measures for evaluating cell formation solutions). In this paper, we propose a clustering objective that captures the essence of prior work on measurement in the sense that it attempts to maximize the association of objects within all clusters *and* minimize the association between objects not in the same cluster. Second, and more importantly, even if there was an acceptable measure of solution quality, the combinatorial nature of the clustering problem, makes it difficult (if not impossible) to evaluate the efficacy of heuristic methods since obtaining optimal solutions would require extensive computing resources. Although recent advances in computing technology have facilitated the comparison process, this comparison has not been carried out to date. In this paper, we provide such a comparison.

3. Clustering model

Since the applications that motivated this study drive the development of our clustering model, we first elaborate on these in more detail. For the cell formation problem, the data typically used is product routing data. This is usually represented in a binary format with a '1' indicating that a product needs to be processed on an equipment type while a '0' indicates the converse. We assume this standard binary data representation in developing our integrated clustering model and represent products as individual objects and equipment types as attributes. For the product assortment/market segmentation problem faced by retailers, we use prior customer purchase data which indicates whether or not an individual customer has purchased a product. As with the earlier application, this can be represented in binary format with a '1' indicating that a product was purchased by a customer and a '0' indicates the converse. Assuming this binary representation in our integrated model, we treat each customer as an object, and each product as an attribute. Obviously, in the context of both applications, we would prefer to develop a simultaneous grouping of objects and attributes. Based on this discussion, we extend the traditional clustering objective as follows:

> The problem of *simultaneously* finding natural groups of objects *and* attributes such that the degree of natural association is *high* between objects and attributes within each group and *low* between objects and attributes in different groups.

A mathematical model for formalizing this problem is as follows:

minimize Z

subjected to:

$$\sum_{k\in m} x_{ik} = 1 \quad \forall i \in n, \tag{2}$$

$$\sum_{k \in m} y_{jk} = 1 \quad \forall j \in q, \tag{3}$$

$$x_{ik} \in \{0, 1\} \quad \forall i \in n; k \in m, \tag{4}$$

$$y_{jk} \in \{0, 1\} \quad \forall j \in q; k \in m, \tag{5}$$

where n is the object set, q the attribute set, k the number of clusters to be identified, x_{ik} is 1 if object i is assigned to cluster k and 0 otherwise, and y_{jk} is 1 if attribute j is assigned to cluster k and 0 otherwise. The constraint sets (2) and (3) enforce the restrictions that each object and attribute are assigned to a single cluster, respectively and constraint sets (4) and (5) enforce the technological constraints on the decision variables. The objective function we formulate simultaneously minimizes the Between Group Association (BGA) and Within Group Non-Association (WGNA). While BGA looks at the 'closeness' between clusters, WGNA assesses the degree to which objects/attributes within each cluster are dissimilar. In order to consider BGA and WGNA simultaneously, our objective function is specified as a convex combination of both these measures:

$$Z = \beta \text{BGA} + (1 - \beta)\text{WGNA}, \tag{6}$$

where $0 \leqslant \beta \leqslant 1$.

Given that we use a binary data representation for both applications as described previously, our measure of clustering efficiency is also based on binary data and is formulated as follows. Let $a_{ij} = 1$, if object i and attribute j are associated; and 0 otherwise. There are several measures of efficiency which have been developed for evaluating a clustering of objects and attributes when binary data is used to specify the relationship (e.g., McCormick et al., 1972; Chandrasekharan and Rajagopalan, 1986; Kumar and Chandrasekharan, 1990; Miltenburg and Zhang, 1991; Ng, 1993). Our work on measure development draws upon this prior work and we define BGA and WGNA as follows:

- Between Group 'Association' (BGA) measure

$$\text{BGA} = 1 - \left[\frac{\sum_i \sum_j \sum_k a_{ij} x_{ik} y_{jk}}{\sum_i \sum_j a_{ij}}\right]. \tag{7}$$

BGA computes the percentage of total association entries (i.e., non-zero entries) which are *not* included in each object/attribute cluster (i.e., $0 \leqslant \text{BGA} \leqslant 1$). We attempt to minimize BGA since we are interested in creating clusters with *low* association between them.

- Within Group 'Non-Association' (WGNA)

$$\text{WGNA} = \frac{\sum_i \sum_j \sum_k (1 - a_{ij}) x_{ik} y_{jk}}{\sum_i \sum_j (1 - a_{ij})}. \tag{8}$$

WGNA computes the percentage of total non-association (i.e., zero-entries) which are included in each cluster (i.e., $0 \leqslant \text{WGNA} \leqslant 1$). As with BGA, we attempt to minimize WGNA since we want to create clusters with *high* association (i.e., low non-association).

Note that both BGA and WGNA attempt to capture two important features associated with clustering applications. First, the numerator in both measures is a function of the object and attribute assignments. Given that previous researchers in clustering have pointed out that solutions are technique dependent, both these measures attempt to capture the impact of a technique when identifying clusters. Second, the denominator in both measures attempts to capture a characteristic of the data being analyzed. The primary motivation for this is that data set characteristics have tended to drive clustering results and hence, we attempt to integrate this aspect in both these measures. Turning back to our clustering model, our objective function is formulated as

$$\begin{aligned} Z &= \beta \text{BGA} + (1 - \beta)\text{WGNA} \\ &= \beta \left\{ 1 - \left[\frac{\sum_i \sum_j \sum_k a_{ij} x_{ik} y_{jk}}{\sum_i \sum_j a_{ij}}\right] \right\} \\ &\quad + (1 - \beta)\left[\frac{\sum_i \sum_j \sum_k (1 - a_{ij}) x_{ik} y_{jk}}{\sum_i \sum_j (1 - a_{ij})}\right]. \end{aligned} \tag{9}$$

Since $0 \leqslant \beta \leqslant 1$, $0 \leqslant \text{BGA} \leqslant 1$ and $0 \leqslant \text{WGNA} \leqslant 1$, this implies that $0 \leqslant Z \leqslant 1$.

The complete model described in the section is obviously NP-hard since it has a non-linear (quadratic) objective function and 0–1 decision variables. Hence, we develop a two-phase hierarchical algorithm for solving this model. This is described in the next section.

4. Solution algorithm

In general terms, our solution approach focuses around decomposing the problem into two subproblems. The first subproblem focuses on identifying similar groups of objects such that the pairwise distance between objects within a group is minimized. Another feature which we incorporate in this first subproblem is that we do *not* specify the number of clusters, a priori. Once we get this grouping of objects into clusters, we develop a polynomially bounded algorithm which specifies an optimal assignment of attributes to the object clusters created. In sum, objects are clustered at the first phase, while attributes are assigned to object clusters at the second phase. Obviously, our two-phase approach does not guarantee an optimal solution to the simultaneous clustering problem formulated previously. However, it does provide one method by which each of the decomposed problems can be optimally solved. An alternative two stage approach has been developed by Ng (1996) specifically for the cell formation problem. At the first phase of his approach, he uses the MST approach combined with a subtree partition procedure to identify machine cells. This is followed by a second stage which assigns part families to machine cells. Although our procedure is similar, it differs in terms of the method/models and algorithms formulated at each phase as well as the clustering criterion. We now proceed to describe each phase of our procedure.

4.1. Phase I – clustering problem (CP)

As noted earlier, we use a modified version of the object clustering model proposed by Vinod (1969) at this phase. The modification is as follows. While the original model required the user to input the number of clusters to be identified, we do *not* impose this restriction. Hence, an output of our model is actually an optimal set of object clusters as well as the grouping of objects into clusters. The basic clustering model is as follows:

$$\text{Minimize } Z1 = \sum_{p\in n}\sum_{r\in n} d_{pr}x_{pr} \tag{10}$$

subjected to:

$$\sum_{r\in n} x_{pr} = 1 \quad \forall p \in n, \tag{11}$$

$$x_{pr} \leqslant x_{rr} \quad \forall p \in n,\ r \in n, \tag{12}$$

$$x_{pr} \geqslant 0 \quad \forall p \in n,\ r \in n,\ p \neq r, \tag{13}$$

$$x_{rr} \in \{0,1\} \quad \forall r \in n, \tag{14}$$

where n is the set of objects to be clustered and d_{pr} is the ‘distance’ between objects p and r. The decision variables are

$$x_{pr} = \begin{cases} 1 & \text{if object } p \text{ is assigned to cluster } r,\\ 0 & \text{otherwise } (p \neq r), \end{cases}$$

$$x_{rr} = \begin{cases} 1 & \text{if cluster } r \text{ is created},\\ 0 & \text{otherwise.} \end{cases}$$

The objective function (10) attempts to minimize the pairwise ‘distances’ between objects in the same cluster. Constraints in the model are to ensure that each object is part of a single cluster (11) and that an object p is only assigned to a cluster r if it is created (Eq. (12)). Finally, constraint sets (13) and (14) enforce the technological constraints on the decision variables. This model is similar to the p-median model and the uncapacitated facility location model with a zero fixed cost for locating a facility. The primary difference stems from the fact that when identifying clusters, we select an object around which a cluster is created while in the uncapacitated facility location (or p-median) problem, facilities and products are treated as distinct entities.

In defining the pairwise distance between objects, we draw upon the clustering literature for matching coefficients. The pairwise distance between objects ‘p’ and ‘r’ ($p \neq r$) is defined as follows:

$$d_{pr} = 1 - \left\{ \beta\left[\frac{\sum_{j\in q} a_{pj}a_{rj}}{\sum_{j\in q} a_{pj} + \sum_{j\in q} a_{rj}}\right] + (1-\beta)\left[\frac{\sum_{j\in q}(1-a_{pj})(1-a_{rj})}{\sum_{j\in q}(1-a_{pj}) + \sum_{j\in q}(1-a_{rj})}\right]\right\}. \tag{15}$$

On the other hand, $d_{pp} = 1\ \forall p \in n$. This measure is similar to some of the matching measures proposed in the clustering literature (Anderberg, 1973) where the first term is the ratio of (1,1) matches between two objects divided by the total number of (1) entries for both objects while the second term is the ratio of total number of (0,0) mismatches divided by the total number of (0) entries for both objects. Since this is a matching coefficient bounded above and below by 1 and 0, respectively, we convert it to a distance measure by subtracting it from 1. Thus, d_{pr} is a traditional distance measure satisfying the metric property of symmetry and is bounded above and below by 1 and 0, respectively. Although d_{pr} appears to be a surrogate measure for Eq. (9), in our opinion, the connection is tenuous. Note that the distance measure is simply capturing the matching effect for pairs of objects while our original measure in Eq. (9) is capturing the effect of a complete clustering solution.

The model stated above is NP-hard. However, our computational experience (discussed later in the next section) indicates that a linear programming relaxation with upper and lower bounds of 1 and 0, respectively, on the x_{rr} decision variables tends to identify integer solutions in all but a few cases. For those few cases, a branch-and-bound procedure was fairly fast in converging to an optimal solution. Once we solve this model for objects, the output of this phase of our algorithm is the cluster set m and the assignment of objects to each cluster 'k' ($k \in m$). Based on this assignment, the next phase of the algorithm focuses on assigning attributes to each cluster and this is described below.

4.2. Phase II – assignment problem (AP)

Let x^I_{ik} represent the assignment of object i ($i \in n$) to cluster k ($k \in m$) based on the model outlined in Phase I. Then to assign each individual attribute $j \in q$ to one of the clusters $k \in m$, we need to solve the following problem:

Minimize $Z2$

$$= \beta\left\{1 - \left[\frac{\sum_i \sum_j \sum_k a_{ij} x^I_{ik} y_{jk}}{\sum_i \sum_j a_{ij}}\right]\right\} + (1-\beta)\left[\frac{\sum_i \sum_j \sum_k (1-a_{ij}) x^I_{ik} y_{jk}}{\sum_i \sum_j (1-a_{ij})}\right], \qquad (16)$$

subjected to:

$$\sum_{k \in m} y_{jk} = 1 \quad \forall j \in q, \qquad (17)$$

$$y_{jk} \in \{0,1\} \quad \forall j \in q;\ k \in m. \qquad (18)$$

This problem is trivial to solve since the assignment of any attribute $j_1 \in q$ is independent of the assignment of any other attribute $j_2 \in q$. Thus, the optimal algorithm detailed below sequentially assigns attributes to object clusters which will minimize the value of the objective function. If there is more than one existing cluster which will minimize the value of the objective function, then the attribute is assigned to the cluster which contains the least number of objects. The rationale behind such an assignment is to create more compact clusters. The specifics of the algorithm are as follows.

Step 1: Set $j = 0$

Step 2: $j = j + 1$. Compute the following:

$$d1_{jk} = \beta\left\{1 - \left[\frac{\sum_i a_{ij} x^I_{ik}}{\sum_i \sum_j a_{ij}}\right]\right\} + (1-\beta)\left[\frac{\sum_i (1-a_{ij}) x^I_{ik}}{\sum_i \sum_j (1-a_{ij})}\right] \quad \forall k \in m. \qquad (19)$$

Step 3: Identify all $k1 \in m$ such that

$$d1_{jk1} = \min_{(k \in m)}[d1_{jk}]. \qquad (20)$$

Let this set of $k1$ be $m1$.

Step 4: If $|m1| = 1$, then set $y_{jk1} = 1$ and $y_{jk} = 0$ $\forall k \in m,\ k \neq k1$. Else set $y_{jk2} = 1$ such that $\sum_{(i \in n)} x^I_{ik2} = \min_{(k1 \in m1)} \sum_{i \in n} x^I_{ik1}$. Set $y_{jk} = 0$ $\forall k \in m$, $k \neq k2$.

Step 5: If $j = |q|$, Stop. Else goto Step 2.

By applying CP/AP to the data at hand we can develop solutions to the cell formation and market segmentation/product assortment problems discussed earlier. Before describing these illustrations, one additional feature of our approach needs to be highlighted. Although we have stated in the entire discussion of the procedure CP/AP that in the first subproblem we cluster objects and in the second subproblem, we assign attributes to each object cluster, there is no reason why we need to carry out object clustering first and subsequently assign attributes. In fact, the procedure could be reversed without loss of generality. Thus, at Phase I, we could cluster attributes (i.e., determine the optimal values of y_{jk}) and at Phase II, we could optimally assign objects to each attribute cluster (i.e., determine x_{ik}). Given the fact that we are solving a binary clustering model at Phase I, the choice of whether to cluster objects or attributes could be made based on problem size considerations so that computation times are minimized. Of course, an alternative is to separately carry out Phases I and II for both strategies and then if there are solution differences, choose the one that optimizes the value of the overall objective function in Eq. (9).

5. Illustration

The primary focus of this section is two-fold. First, we illustrate that we can obtain optimal integer solutions to the model developed in Phase I in a reasonable period of time. Given that the algorithm for Phase II is polynomially bounded, computation times are not an issue for attribute assignments. A second objective of our illustration focuses on solution quality. The cell formation solutions obtained using our procedure are compared to the 'best' solutions in the literature using an established measure of solution quality (grouping efficacy). For the market segmentation/product assortment problem, the solutions obtained using our approach are compared to those obtained using traditional heuristic approaches currently used in the marketing literature. In this case, the basis of comparison is the measure of clustering efficiency proposed in this paper (Eq. (9)).

5.1. Cell formation

To illustrate our method, we use 24 published binary data sets from the literature on cell formation in cellular manufacturing (Vakharia and Wemmerlöv, 1995). Basic details for each data set are given in Table 1. Note that the density column in Table 1 is number of total entries which are defined as 1 divided by the total number of possible entries (i.e., $|n| \times |q|$). For each data set, we choose to implement our two-phase procedure as follows. We first specified the value of β to be 0.50. Then, for each data set, we first developed a cluster of objects (products) at Phase I and assigned attributes (equipment types) at Phase II. To solve the model in Phase I, we used the CPLEX mathematical programming library on a VAX 9600 computer and recorded the CPU times in seconds for obtaining an optimal solution. Table 1 presents the results of our procedure in terms of the number of clusters, objective function value (see Eq. (9) with $\beta = 0.50$), and the CPU times for Phase I. As can be seen, the CPU times are less than 1 minute for every problem and we needed to invoke the branch and bound procedure programmed in CPLEX for only 2 of the 24 problems (data sets 23 and 24). Thus the model proposed in Phase I solved these cell formation problems fairly fast.

In order to evaluate the solution quality of our methodology, we also computed the 'grouping efficacy' (Kumar and Chandrasekharan, 1990) for each solution for the 24 data sets. We compare this value to the best value obtained by prior researchers in cellular manufacturing and this is also shown in Table 1. Table 1 indicates that our methodology provides identical efficacy values for 14 of the 24 data sets and better quality solutions for the remaining 10 data sets. Thus our procedure provides equivalent or better solutions as compared to those obtained using existing cell formation methods.

5.2. Market segmentation/product assortment

In marketing, identifying customer segments is one of the key decisions where cluster analysis has been used extensively. Recent applications have relied on heuristic procedures for identifying homogeneous groups of customers. Our approach differs substantially from this prior work in two aspects. In this application, we identify customer segments within each region based on products purchased since the retailer was interested in identifying product-group clusters which could help to make stocking decisions at each individual

Table 1
Cell formation results

Data set	Description			Results			Grouping efficacy	
	Objects (products)	Attributes (eq. types)	Density	# of clusters	Obj. func. value	CPU time (s)	Current	Prior
1	20	10	0.2000	4	0.41	0.07	0.8163	0.8163
2	23	14	0.1802	4	0.36	0.03	0.6824	0.6432
3	40	24	0.1094	6	0.31	0.85	0.6213	0.6213
4	43	14	0.1395	5	0.33	0.32	0.6667	0.6566
5	50	30	0.1027	11	0.53	5.22	0.5661	0.5632
6	18	24	0.2037	4	0.38	0.18	0.4891	0.4891
7	19	12	0.3290	4	0.55	0.37	0.5875	0.5656
8	20	8	0.3813	3	0.57	0.12	0.8525	0.8192
9	20	23	0.2457	5	0.55	0.72	0.5321	0.4936
10	22	11	0.3223	3	0.47	0.12	0.7312	0.7312
11	24	14	0.1816	4	0.36	0.28	0.6555	0.6555
12	30	16	0.2417	4	0.47	0.93	0.6783	0.6783
13	35	20	0.1943	4	0.39	0.38	0.7571	0.7514
14	40	24	0.1354	7	0.50	0.67	0.8511	0.8511
15	40	24	0.1365	7	0.50	1.40	0.7351	0.7351
16	40	24	0.1344	8	0.53	3.58	0.4909	0.4327
17	40	24	0.1354	9	0.56	3.95	0.4451	0.4451
18	40	24	0.1354	10	0.57	7.35	0.4233	0.4167
19	41	30	0.1041	7	0.37	2.63	0.5543	0.5543
20	43	14	0.1445	5	0.36	0.70	0.6434	0.6434
21	43	16	0.1831	7	0.51	7.03	0.5439	0.5439
22	46	28	0.1638	8	0.56	8.47	0.3688	0.3301
23	90	30	0.1126	15	0.60	56.32	0.3941	0.3941
24	100	40	0.1050	9	0.50	10.07	0.8392	0.8392

store. This would facilitate better inventory decisions in the distribution component of the entire supply chain. To address this issue, we conceptualized the problem as one of determining the clusters of similar customer segments based on prior product purchasing patterns. Hence, we visualize the overall problem as one of simultaneously determining customer segments and related product groups using individual customer purchase data.

The market segmentation/product assortment illustration described in this section uses data obtained from a large US apparel retailer. In a single territorial market in the US, the retailer marketed products through stores in 4 regions. In each region, stores tracked individual customer purchases of 36 product groups over a given month. By aggregating this store level data at the regional level, the retailer had information of which product group was purchased by each customer. The number of unique customers tracked in each region is shown in Table 2. Since not all 36 product groups were stocked/purchased in all regions, the number of distinct product groups purchased in each region is also shown in Table 2.

In terms of applying our solution procedure, we set the value of β to be 0.50. Then we applied the two-phase solution method CP/AP to each region data and determined the optimal number of clusters. If we use the individual customer data, the size of the problem to be solved in Phase I is prohibitively large. Hence, we reduced this problem size by grouping customers which had identical product purchases within a region and this reduced the problem size considerably. For example, for region 1, from an original of 2614 customers, we were able to collapse the data to 72 distinct customers. The resulting four problems (one for each region) for Phase I were solved in at most 17.32 min on a VAX 9600 computer with a floating point accelerator using the CPLEX mathematical programming library (the largest problem solved in this case

Table 2
Market segmentation/product assortment results

Region	Customers (objects)	Products (attributes)	# of clusters	Results and comparison			
				CP/AP	AL/AP	SL/AP	KM/AP
A	2614	32	4	0.3186	0.3464	0.5784	0.3328
B	2301	34	6	0.3079	0.3349	0.5864	0.3434
C	2540	31	5	0.4126	0.488	0.6874	0.5679
D	2105	36	3	0.2878	0.3828	0.5690	0.4143

(1) CP/AP is our solution algorithm; AL/AP is use of Average linkage combined with Phase II of our algorithm; SL/AP is use of Single linkage combined with Phase II of our algorithm; and KM/AP is the *K*-means algorithm with Phase II of our algorithm.
(2) The values given in the columns of results and comparison are the objective function values for each case computed using $\beta = 0.5$ in Eq. (9).

included 72 integer variables, 5112 continuous variables, and 5256 constraints). For each problem, we needed to invoke the branch and bound procedure programmed in CPLEX to obtain optimal integer solutions. Given that determining product assortments based on market segments is a retail store design problem, such computation times are not unacceptable. This indicates that Phase I of our procedure can obtain optimal solutions to substantially large problems in a reasonable amount of time.

In column 4 of Table 2, we show the number of market segment clusters identified using Phase I of CP/AP. In order to compare our solution to that which could be obtained at Phase I using heuristic clustering methods, we chose to do the following.

(1) We chose to compare the quality of our solutions to that obtained using two hierarchical methods (Single Linkage and Average Linkage) and one non-hierarchical method (*K*-means). Both the hierarchical methods were used in conjunction with the Jaccard similarity index. These three methods are the most widely used clustering methods in the marketing literature (see, for example, Fader and Lodish, 1990; Morowitz and Schmittlein, 1992; Krieger and Green, 1996).

(2) For each clustering method we used (i.e., Single Linkage, Average Linkage, and *K*-means), we prespecified that the number of clusters to be identified is set equal to that identified using CP/AP. The primary rationale for setting this value is that it would facilitate comparison of alternative solutions.

(3) After a customer segment clustering solution was obtained using any of the three clustering heuristics, we used Phase II (AP) of our solution algorithm to assign product groups and identify the associated product assortments. The reason for using AP at Phase II for all four procedures was two fold. First, if we know the object to cluster groupings, AP provides optimal attribute assignments and hence, is a valid approach for use regardless of which method is used in Phase I. Second, this allows us to evaluate whether the optimal clustering model used in Phase I outperforms the traditional heuristic methods in developing object clusters.

(4) Finally, all the resulting solutions (including the one obtained using CP/AP) was evaluated by using $\beta = 0.5$ is Eq. (9).

The results of our comparison are shown in the last four columns of Table 2. As can be seen, for each region, the CP/AP algorithm provides superior solutions as compared to any of the three heuristic clustering techniques. Further, the primary cause for this is the use of the optimization model used in Phase I (since we used Phase II – AP across all solutions). This leads us to conclude that the CP/AP method has the potential to identify clusters with higher within cluster association and low between cluster association.

6. Implications and conclusions

In this paper, we have proposed an integrated model for simultaneously clustering objects and attributes. Given that the basic object clustering problem is computationally complex, it is not surprising that the proposed model is NP-hard.

Thus, we develop a hierarchical two-phase method for solving the problem. At the first phase, we use an integer programming model for clustering objects and at the second phase, we use a polynomially bounded algorithm for optimally assigning individual attributes to each object cluster. There are several unique features of our approach. First, our hierarchical approach is one of the first to focus on clustering of objects and attributes. As illustrated in this paper, it can be applied to several practical problems which require such a perspective. Second, prior research in clustering has pointed out that clustering solutions obtained by any technique are highly sensitive to the number of clusters that are identified. To overcome this problem, the integer programming model used to develop object clusters also helps to identify the optimal *number* of object clusters. Finally, we have also demonstrated that the solutions obtained from using our Phase I model are superior to those obtained using existing clustering methods. Further, these solutions can be obtained in a reasonable period of time. This implies that the model proposed in Phase I can be used as an alternative to existing clustering techniques and thus, broadens the applicability of our modelling effort.

At a more pragmatic level, we demonstrate the usefulness of our model in solving specific problems in manufacturing and retailing. For cell formation, we have shown that our approach generates solutions that are superior (or at worst equivalent) to those generated using other methods. For cases where our approach identifies better solutions, a higher grouping efficacy implies the following. If for these cases, product families and related production cells are created based on our approach, it should lead to increased equipment utilization within cells as well as fewer movements of materials between cells as compared to cell formation solutions identified using existing approaches. Thus, operational efficiencies associated with creating cells should be higher for solutions with greater grouping efficacy.

For the market segmentation/product assortment problem, we were able to show that our procedure generated more compact and distinct product and customer clusters than the traditional heuristic based methods which have been used to address this problem. In essence, this points to the fact that the product assortment decision in retailing can be addressed in a more formal manner by relying on prior customer purchase patterns. In the context of market segmentation, we have shown that optimization based clustering approaches can be utilized successfully to identify market segments. Further, our procedure also identifies good or close to optimal solutions for determining the number of market segments as compared to the current trial and error approaches (e.g., the between group to within group *F*-ratio test) used in marketing.

References

Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press, New York.

Blashfield, R., 1976. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical procedure. Psychological Bulletin 83, 377–388.

Chandrasekharan, M.P., Rajagopalan, R., 1986. An ideal seed non-hierarchical clustering algorithm for cellular manufacturing. International Journal of Production Research 24, 451–464.

Erlenkotter, D., 1978. A dual-based procedure for uncapacitated facility location. Operations Research 26, 992–1009.

Fader, P.S., Lodish, L.M., 1990. A cross-category analysis of category structure and promotional activities for grocery products. Journal of Marketing 54, 52–65.

Hoek, J., Gendall, P., Esslemont, D., 1996. Market segmentation: A search for the holy grail. Journal of Marketing Practice 2 (1), 25–34.

Jensen, R.E., 1969. A dynamic programming algorithm for cluster analysis. Operations Research 12, 1034–1057.

Klastorin, T.D., 1985. The *p*-median problem for cluster analysis: A comparative test using the mixture model approach. Management Science 31, 84–95.

Krieger, A.M., Green, P.E., 1996. Modifying cluster-based segments to enhance agreement with an exogenous response variable. Journal of Marketing Research 33, 351–363.

Kumar, C.S., Chandrasekharan, M.P., 1990. Grouping Efficacy: A quantitative criterion for the goodness of block diagonal forms of binary matrices in group technology. International Journal of Production Research 28, 233–243.

Levy, M, Weitz, B.A. 1998. Retailing Management. Irwin-McGraw Hill, Boston.

McCormick, W.T., Schweitzer, P.J., White, T.W., 1972. Problem decomposition and data reorganization by a clustering technique. Operations Research 20, 993–1009.

Milligan, G.W., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45, 325–342.

Milligan, G.W., Cooper, S.C., 1987. Methodology review: Clustering methods. Applied Psychological Measurement 11, 329–354.

Miltenburg, J., Zhang, W., 1991. A comparative evaluation of nine well-known algorithms for solving the cell formation problem in group technology. Journal of Operations Management 10, 44–72.

Morowitz, V.G., Schmittlein, D., 1992. Using segmentation to improve sales forecasts based on purchase intent. Journal of Marketing Research 29, 391–405.

Mulvey, J.M., Beck, M.P., 1984. Solving capacitated clustering problems. European Journal of Operational Research 18, 339–348.

Mulvey, J.M., Crowder, H.P., 1979. Cluster analysis: An application of Lagrangian relaxation. Management Science 25, 340–392.

Ng, S.M., 1993. Worst-case analysis of an algorithm for cellular manufacturing. European Journal of Operational Research 69, 384–398.

Ng, S.M., 1996. On the characterization and measure of machine cells in group technology. Operations Research 44 (5), 735–744.

Rao, M.R., 1971. Cluster Analysis and Mathematical Programming. Journal of the American Statistical Association 66, 622–626.

Selim, H., Askin, R.G., Vakharia, A.J., 1998. Cell formation in group technology: Review, evaluation, and directions for future research. Computers and Industrial Engineering 34 (1), 3–20.

Vakharia, A.J., Wemmerlöv, U. 1995. A comparative investigation of hierarchical clustering techniques and dissimilarity measures applied to the cell formation problem. Journal of Operations Management 13, 117–138.

Vinod, H.D., 1969. Integer programming and the theory of groups. Journal of the American Statistical Association 64, 506–519.