RC24038 (W0608-091) August 28, 2006
Computer Science

IBM Research Report

Identifying Bundles of Product Options Using Mutual Information Clustering

Claudia Perlich, Saharon Rosset
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Identifying Bundles of Product Options using Mutual Information Clustering

Claudia Perlich, Saharon Rosset
{perlich,srosset}@us.ibm.com

IBM T.J. Watson Research Center

Abstract. Mass-produced goods tend to be highly standardized in order to maximize manufacturing efficiencies. Some high-value goods with limited production quantities remain much less standardized and each item can be configured to meet the specific requirements of the customer. In this work we suggest a novel methodology to reduce the number of options for complex product configurations by identifying meaningful sets of options that exhibit strong empirical dependencies in previous customer orders. Our approach explores different measures from statistics and information theory to captures the degree of interdependence between the choices for any pair of product components. We use hierarchical clustering to identify meaningful sets of components that can be combined to decrease the number of unique product specifications and increase production standardization. The focus of our analysis is on different similarity measure including chi-squared statistics and versions of mutual information on the ability of the clustering to find meaningful clusters.

1 Introduction and Motivating Example

While bundling of products has received significant attention in the economic literature (e.g., [1–3]), the bundling of product options is typically limited to considerations of production efficiency and engineering. In order to optimize the tradeoffs between maximizing production efficiencies and making products that meet the individualized requirements of particular customers, manufacturers have developed techniques of combining options into bundles so that batches of similarly customized products may be made together, rather than making each customized product individually. Existing approaches to developing such bundles, however, have been driven by the choices of product designers and have not afforded a systematic way of incorporating customer preference data. Examples of products that offer option bundles can be observed in the car industry. Toyota for instance offers an 'All Weather Guard Package' that includes an Intermittent Rear Window Wiper, Windshield Molding, Heavy-Duty Heater and Rear-Seat Heater Ducts. All of the above components appear related to the requirements of driving under harsh weather. However, Toyota also offers Option Combination B for a Toyota Matrix that contains 50 State Emissions, Cruise Control, and Power Package. For this combination it is much less clear whether there is

any relationship between the components. In this work we explore the task of finding good sets of components for bundling on the example of truck configurations. Even considering all the different options, passenger cars remain a highly standardized product class. Trucks on the other hand are ordered for a specific use and the customer can specify all major components separately. A truck will only be produced after the customer has made his choices. However, one would suspect a limited set of usage categories and certain recurring patterns in the customer orders. A customer that typically uses a truck for long-haul purposes is likely to require optional sleeper cab facilities, while a customer using the same type of truck for local hauling would not be willing to pay for. In addition, requirements for safety and handling options relevant to various types of road and weather conditions may differ from customer to customer. So the objective of our bundling task is to find component combinations that empirically exhibit strong customer-choice interdependencies and will appeal to future customers of new orders. To address this objective we need to quantify dependencies between components. In order to achieve this goal we will explore potential measures of dependencies between nominal variables in Section 3 and discuss properties of such similarity measures. Given that such measures can capture only pair wise dependencies, we propose in Section 4 the use of hierarchical clustering to find larger sets of components that all exhibit large pair wise dependencies. We will illustrate the issues and results on the example of the truck configuration domain.

While the work in this paper is applied, but we think that the business problem is of general interest and we are not aware of prior work on this formalization as hierarchical clustering under an appropriate similarity measure. Other contributions are the identification of desirable properties in the given context of nominal variables with difference in the number of choices and the skew of the probabilities. While there has been substantial work on clustering using chi-square based similarities as well as clustering with mutual information (e.g., [4–6]), we are not aware of the proposed combined methods that incorporates both, the information content and the statistical reliability.

2 Notation and Formalization

Formally, a complex product consists of n components $C_1, ..., C_n$. For every component C_j, there is a limited set of k_j possible choices $\{c_{j1}, ..., c_{jk_j}\}$ where the number of choices k_j differs across components. We also assume that we have N past observations that indicate for each order the particular choices as a vector $o_1, ..., o_n$ with $o_j \in \{c_{j1}, ..., c_{jk_j}\}$. Note that this setup differs considerably from the typical basket analysis of customer choices that motivated the work on large itemsets and mining of association rules [7]. The notion of components imposes additional constraints:

- all customers have the identical number of n components and
- for each component only one choice is permissible.

While frequent itemsets may be indicators of semantic interdependencies between choices, they do not measure the interdependence of components. Each itemset considers only one particular choice $c_{jg} \in \{c_{j1}, ..., c_{jk_j}\}$ and how often it appears with another choice for another component, but not how much each possible choice $c_{j1}, ..., c_{jk_j}$ for component c_j correlates with the choices for the other component. Another problem with the notion of frequent itemsets is its dependence on the prior probability of a particular choice. In particular, a frequent itemset analysis identifies typically combinations of default values for components with one very common default value and a small set of much less common values. That does not mean that there is any deeper semantic dependency between the components. It is just an artifact of the high skew of the probabilities. While there are measures of the 'unexpectedness' of an itemset, these measures are typically a function of the size of the set and with larger sets exhibiting much more unexpected behavior.

To address our specific bundling objective we need to quantify dependencies between sets of components, not sets of choices. In order to achieve this goal we will explore potential measures of dependencies between nominal variables in the following Section and discuss properties of such similarity measures.

3 Measuring Dependence

The objective in our bundling task is to find sets of components where past customer choices exhibit some form of dependence. So far we have used the term dependence rather loosely in a non-technical sense of some form of a semantic connection. While it is difficult to formalize dependence without a clear prior notion of how things depend on each other, there is a clear statistical notion of the opposite: independence between random variables. We can formalize the observation of a customer choice o_i for a particular component C_i as the outcome of a random experiment over the sample space $\Omega_i = \{c_{i1}, ..., c_{ik}\}$. Formally, two random variables are independent if their joint distribution is equal to the product of their individual distribution functions

$$P(o_j = c_{hp}, o_l = c_{lm}) = P(o_h = c_{hp}) * P(o_l = c_{lm}) \tag{1}$$

for all elements of the Cartesian product of the two sample spaces $\Omega_i \times \Omega_j$ (all possible choice pairs for the two components). Independence is defined generally over an arbitrary number of variables and we could attempt to devise a measure the interdependence within entire sets of components. However, such a strategy will not lead to non-overlapping bundles as desired in our case. In addition, given the somewhat vague business objective, the final choice of bundles is potentially subject to many additional production constraints and considerations. We will therefore restrict our work to pairs of components and employ hierarchical clustering to suggest potential non-overlapping bundles.

We can now measure dependence in terms of the degree of violation of this equality 1 over all pairs of choices c_{hp}, c_{lm}) for a pair of components (C_h, C_l).

This requires initially the estimation of the distribution for all possible components and their choices $P(o_l = c_{lm})$ and choice pairs $p(c_{hp}, c_{lm})$. We will simplify the notation and use $p(c_{hp})$ to denote $P(o_h = c_{hp})$ and $p(c_{hp}, c_{lm})$ for $P(o_j = c_{hp}, o_l = c_{lm})$ respectively. Note that for the posed business problem, we do not have a clear evaluation metrics. Otherwise we could hope to derive (either implicit or explicitly) an appropriate similarity measure subject to optimal bundling performance. Our results will depend very much on the particular choice of similarity. We will therefore discuss in more detail some desirable and useful properties and frame existing measure with respect to this properties.

While there are many possible choices of a similarity measure $D([0,1]^s, [0,1]^s) \rightarrow \mathbb{R}$ (where $s = k_h * k_l$ is the number of choice pairs), reasonable candidates can be constructed from i) an 'atomic' measure of similarity $D_0([0,1],[0,1]) \rightarrow \mathbb{R}$ of the elements (c_{hp}, c_{lm}) of the Cartesian product over the sample spaces and ii) an aggregation function $A(\mathbb{R}^n) \rightarrow \mathbb{R}$ over all the atomic similarities.

In order to be suitable for our bundling task, we would like the similarity to exhibit three other desirable properties:

- It has to be **symmetric** with $D(C_h, C_l) = D(C_l, C_h)$, since there is no special order on the components;
- It should to be **comparable** across component pairs. In particular, it should be rather insensitive to the specific size of the Cartesian product of the sample spaces;
- It should be **robust** towards estimation errors of the distribution. Given a limited sample of prior customer orders and a large sample space for some components with many possible values, the estimation quality of both the single probabilities and even more so the probabilities of choice pairs will be limited. This problem is particularly dominant for rare choices.

The issue of assessing independence has been considered in different fields and contexts including the analysis of contingency tables in statistics and information theoretical work on the information content of signals.

3.1 Chi-Square Based Similarities

Measures of association have a long history in the context of the analysis of contingency tables. For an extensive overview consider [8]. However, the majority is not very suitable for our task for various reasons including a lack of symmetry, and focus on the conditional mode of the distribution while ignoring less common choices. One standard approach to evaluate the significance of statistical dependencies of two nominal random variables (C_h and C_l) is based on a Chi-square test.

$$\chi^2(C_h, C_l) = N \sum_{i=1}^{k_h} \sum_{j=1}^{k_l} \frac{(p(c_{hp}, c_{lm}) - p(c_{hp})p(c_{lm}))^2}{p(c_{hp})p(c_{lm})} \tag{2}$$

Note that this formulation uses an 'atomic' Euclidean similarity and a weighted sum as aggregation function where the weight reflects the expected probability of observing a pair under the null-hypothesis of independence. Let us make a few observations that contradict our two desirable properties for the bundling task:

- The measure from Equation 2 follows (under certain assumptions) approximately a Chi-square distribution with $(k_h - 1)(k_l - 1)$. This means that its expected value is a function of the sizes of the sample spaces and renders a comparison across component pairs impossible.
- The Qui-square statistic is known to be sensitive to small number of expected observations in the nominator. The Fisher exact test is correcting for this problem but becomes computationally infeasible already for moderate sample spaces (e.g., size of 4).

One can consider ad-hoc solutions for both issues. To address the dependence on the degrees of freedom, we can either convert the statistic into the corresponding p-value or correct it based on the Normal approximation. The p-value is derived from the cumulative distribution with the appropriate degrees of freedom and reflects the probability of such a Chi-square occurring by chance. However, as we will see in the experiments, this correction eliminates most of the information. Given the comparably large size of our dataset, almost all observed values are significant with high probability and most of the p-values are indistinguishable from 0. The second correction takes advantage of the fact that a Chi-square with large number of degrees of freedom d is approximately normally distributed with a mean equal to d and a variance equal to $2*d$. We can therefore use the following correction:

$$N_{\chi^2}(C_h, C_l) = \frac{\chi^2(C_h, C_l) - d}{\sqrt{2d}} \tag{3}$$

To address the issue of small expectations, we combine multiple rare component choices into a new value 'others'. Note that a replacement with 'other' can artificially create dependencies and should be taken with a grain of salt: the fact that for two components some cases have both the value 'other' is likely to indicate that the customer is picky and always wants something special, not that this choice of one component affects the other.

3.2 Mutual Information

Intuitively, mutual information [9] measures the information about one component that is shared by another. If the components are independent, then one contains no information about the other vice versa, so their mutual information is zero. Formally, the mutual information MI of two random variables for components C_h and C_l is defined as:

$$MI(C_h, C_l) = \sum_{i=1}^{k_h} \sum_{j=1}^{k_l} p(c_{hp}, c_{lm}) \log \frac{p(c_{hp}, c_{lm})}{p(c_{hp}) * p(c_{lm})} \tag{4}$$

In the case of mutual information the aggregation function is again a weighted sum and the 'atomic' similarity is the log of the ratio of the expected and observed probability. While this measure both symmetric and robust to small expectations due to the log transformation, it is not comparable across pairs of components. If the sample space of the two variables is identical, the maximum mutual information under complete dependence is equal to the entropy. Entropy however is a function of the sample size. In particular, a tight upper bound on the mutual information is given by

$$MI(C_h, C_l) \leq \frac{H(C_h) + H(C_l)}{2} \tag{5}$$

where $H(C_h)$ is the entropy [10] of component C_h defined as

$$P(C_h) = \sum_{i=1}^{k_h} p(c_{hi}) \log(\frac{1}{p(c_{hi})}) \tag{6}$$

We therefore define a normalized mutual information as suggested by [11] as

$$NMI(C_h, C_l) = \frac{2MI(C_h, C_l)}{H(C_h) + H(C_l)}. \tag{7}$$

3.3 Combining Mutual Information and Significance

While both measures work on the same underlying information, the objective for which they were developed is very different. The goal of the Chi-square measure is to assess significance relative to the null-hypothesis of independence. This means in particular, that it matters how many observations are provided. The power of a test is a function of the provided number of observations and as the sample becomes very large, almost every small deviation becomes significant. We can see the relevance of the sample size N in Equation 2.
Information theory ([10, 9]) on the other hand takes a different perspective. Mutual information is completely independent of the sample size N and in does not assess whether the observed amount of information could have been observed by random chance. So mutual information is a closer measure of the quantity we are interested in, the degree of dependence, but does not take randomness into account and whether the observed quantities are significant.
We therefore propose a similarity measure that incorporates both, statistical considerations of significance and the amount of information

$$SIM(C_h, C_l) = NMI(C_h, C_l) cdf(\chi^2(C_h, C_l), (k_h - 1)(k_l - 1)) \tag{8}$$

where $cdf(\chi^2(C_h, C_l), (k_h - 1)(k_l - 1))$ is the value of the cumulative density function for the Chi-square statistic $\chi^2(C_h, C_l)$ with $(k_h - 1)(k_l - 1)$ degrees of freedom. This similarity measure weights the observed amount of shared information by the probability of it not being random.

4 Hierarchical Clustering

Clustering and cluster analysis (e.g., [12, 13]) encompasses a number of different algorithms and methods for grouping objects of similar kind into respective groups. Rather than finding a fixed number of clusters in the data, hierarchical clustering as proposed by Johnson [12] provides a proceeds iteratively by combining existing clusters may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis. Examples of such a dendrograms are given Figure 1. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by a clustering algorithm (see cluster analysis). The similarities between the nodes reflect the relative similarities of the clusters. Given a set of n items to be clustered, and an $n * n$ similarity matrix, the basic process of hierarchical clustering [12] is this:

1. Start by assigning each of the n component to its own cluster. Let the similarities between the clusters the same as the similarities between the items they contain.
2. Find the closest pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute similarities between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size n.

Aside from the similarity measure, the criterion to define 'closest' in step 2 is one of the major components of the clustering algorithm and can affect the results significantly. Different criteria include:

- **Minimum**: Similarity between clusters is the smallest similarity from any member of one cluster to any member of the other cluster.
- **Average**: Similarity is the average over the similarities from any member of one cluster to any member of the other cluster. Alternatively, one can consider the median, which is more robust to similarity outliers.
- **Maximum**: Similarity between clusters is the largest similarity from any member of one cluster to any member of the other cluster.
- **Ward**: Similarity is phrased in terms of the increase in diversity in the cluster (originally considered as a total prediction error [14]). The main difference between this and the previous criteria is the consideration of the size of the cluster. While doubling the number of elements in each cluster has no effect on the previous criteria, it will increase to total diversity by a factor of two and makes it less likely to combine larger clusters.

5 Dataset and Empirical Results

Our experiments are based on non-public transaction records of a truck manufacture. To preserve the privacy of the client we have replaced component names

Code	Component	Size	Mode	Code	Component	Size	Mode
001	Model	8	0.80	420	Axle Rear Drive	57	0.21
016	Exhaust Package	15	0.64	421	Axle Ration	49	0.15
018	Brake Package	4	0.92	423	Brake Rear	13	0.12
035	Dead Axle Package	11	0.98	545	Wheelbase	164	0.19
101	Engine	54	0.14	546	Frame Rail	10	0.27
128	Retarder Driveline	12	0.65	552	Frame Overhang	115	0.23
180	Clutch	717	0.51	578	Fifthwheel	33	0.91
204	LH Fuel Tank	15	0.45	620	Suspension Front	12	0.40
206	RH Fuel Tank	17	0.45	622	Suspension Rear	73	0.11
266	Radiator	8	0.40	682	SleeperCab	2	0.99
290	BatteryBox	7	0.98	829	Cab Size	7	0.74
342	Transmission	82	0.11	A84	Business Segment	30	0.33
360	PTO Engine Front	4	0.96	A85	Vehicle Service	14	0.67
362	PTO Transmission	22	0.80	AA2	Trailer Type	12	0.86
400	Axle Front	21	0.20	AA3	Body Type	30	0.44
402	Brake Front	9	0.48				

Table 1. Component Codes and Definitions for the Example Domain. The size column represents the number of possible choices for the component (size of the sample space) and the last column presents the probability of the most common choice (Mode) as an indicator of the skew in the probabilities.

while keeping all statistical aspect identical. We selected (based on the recommendation of the manufacturer) a small subset of 30 important components for the illustration of this work and included in the analysis a total of 3500 recent orders. An overview of the components is provided in Table 1. The table also provides some information about the statistical properties including the size of the sample space for each component (Size) and the distribution of the mode for each component (Mode).

5.1 Similarity Measures

Following the discussion in Section 3 we have 7 different similarity measure to our disposal:

- N_{χ^2}: Chi-sqare corrected for degrees of freedom by Normal approximation as defined in Equation 3
- $N_{\chi^2_r}$: Chi-sqare without rare options (occurrence below 20) corrected for degrees of freedom by Normal approximation
- $p(\chi^2)$: p-values of Chi-square
- $p(\chi^2_r)$: p-values of Chi-square without rare options
- MI: Mutual information as defined in Equation 4
- NMI: Normalized mutual information as defined in Equation 7
- SIM: Combined mutual information and p-value as defined in Equation 8

Table 2 shows the correlation (which implicitly assumes a linear relationship)

	N_{χ^2}	$N_{\chi^2_r}$	$p(\chi^2)$	$p(\chi^2_r)$	MI	NMI	SIM
N_{χ^2}	1.00	0.80	0.20	0.16	0.41	0.63	0.63
$N_{\chi^2_r}$	0.80	1.00	0.18	0.16	0.59	0.84	0.84
$p(\chi^2)$	0.20	0.18	1.00	0.58	0.18	0.22	0.24
$p(\chi^2_r)$	0.16	0.16	0.58	1.00	0.16	0.20	0.20
MI	0.41	0.59	0.18	0.16	1.00	0.88	0.88
NMI	0.63	0.84	0.22	0.20	0.88	1.00	0.99
SIM	0.63	0.84	0.24	0.20	0.88	0.99	1.00

Table 2. Correlation of the different similarity measures.

between the measures. We can clearly identify three groups: measures based on mutual information (MI, NMI and SIM), measure based on the p-values and the two Chi-square values. The fact that the p-values are only very vaguely correlated with the Chi-square measures is due to the inherent non-linearity of the cumulative density function. Replacing rare values has a moderate effect both in the case of p-values and the Chi-square measures. The normalization of the mutual information has clearly an effect, much more so than the weighting by the p-value. The only exception to the nice separation of the measures into 3 groups is the high correlation between the Chi-square adjusted for rare values and the two normalized mutual information measures of 0.84.

As pointed out earlier, the measures using a p-value only reflect whether the observed degree of dependence could be random. We have a fairly large dataset and both measures assign a value of 1 to 93% of all pair wise distances. This renders it unusable as a similarity measure for the clustering objective. The only pairs that show values below 1 involve typically components with a very high probability for the mode (e.g., components 035, 682, and 360).

5.2 Clustering Results

We used the Pajek [15] implementation to perform the hierarchical clustering using the Ward and average criterion and the visualization of the corresponding dendrograms. Given our earlier analysis of the similarity measures we consider for clustering only SIM, $N_{\chi^2_r}$.

Figure 1 shows examples of dendrograms for the two main similarity measures and different clustering criteria. Unfortunately, the effect of the clustering criterion is at least as relevant as the similarity measure. In addition, the scale and skew of the similarities affect the results severely. The SIM measure is limited between 0 and 1 and is typically close to 1. The Chi-square measure with the normal correction for the degrees of freedom on the other hand ranges from -10 to 500 with a much more uniform distribution.

We can nevertheless make some observations about the results that could be used by a domain expert to identify bundles. We can find a number of groups of components that are placed together by most reasonable clusterings. This includes for instance the set {A85,AA3,A84}, {204,206}, {620,400}, and {829,001}.

They are clearly good candidates for component bundles. The descriptions in Table 1 suggest that indeed these sets are meaningful.

6 Discussion and Conclusion

We presented an analytical approach that can guide the design of appropriate bundles of components for complex products such as trucks. While the task is very relevant in practice, there is no clear measure of performance and the validity of the results can only be assessed based on domain specific information or by an domain expert. We suggest the use of mutual information, adjusting for the number of options and combining it with statistical significance, as a measure of dependence between customer choices. Our approach could identify meaningful candidate sets of components. An important observation of our analysis is the relevance of the clustering criteria and its potential interaction with properties of the similarity measure. We are not aware of studies that investigate issues of similarity scaling and distribution in the context of different clustering criteria and hope to address this topic in future work.

Acknowledgments

We are grateful to Tomasz Nowicki for suggestions and discussions about dependency measures.

References

1. Eppen, G.D., Hanson, W.A.: Bundling-new products, new markets, low risk. Sloan Management Review **32**(4) (1991) 7–14
2. Adams, W.J., Yellen, J.L.: Commodity bundling and the burden of monomoly. Quaterly Journal of Economics **90** (1976) 475–498
3. Bakos, Y., Brynjolfsson, E.: Bundling information goods: Pricing, profits and efficiency. Management Science **45**(12) (1999)
4. Strehl, A., Ghosh, J., Mooney, R.: Impact of similarity measures on web-page clustering. In: Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI 2000), 30-31 July 2000, Austin, Texas, USA, AAAI (2000) 58–64
5. Kraskov, A., Stogbauer, H., Andrzejak, R.G., Grassberger, P.: Hierarchical clustering based on mutual information. Europhysics Letters **70**(2) (2005) 278–284
6. Slonim, N., Atwal, G.S., Tkacik, G., Bialek, W.: Information based clustering: Supplementary material. Proceedings of the National Academie of Sciences **102**(51) (2005) 18297–18302
7. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: International Conference of Very Large Data Bases (VLDB). (1994) 487–499
8. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. Springer-Verlag,New York (1979)
9. Kullback, S.: Information Theory and Statistics. Dover, New York (1968)

10. Shannon, C.: A mathematical theory of communication. The Bell system technical journal **27** (1948) 379–423
11. Strehl, A.: Relationship-based clustering and cluster ensembles for high-dimensional data mining, phd thesis, the university of texas at austin (2002)
12. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika **2** (1967) 241–254
13. Tyron, R., Bailey, D.: Cluster analysis. McGraw-Hill (1970)
14. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of American Statistical Associatoin **58** (1963) 236–244
15. Batagelj, V., Mrvar, A.: Pajek - program for large network analysis. Connections **21**(2) (1998) 47–57

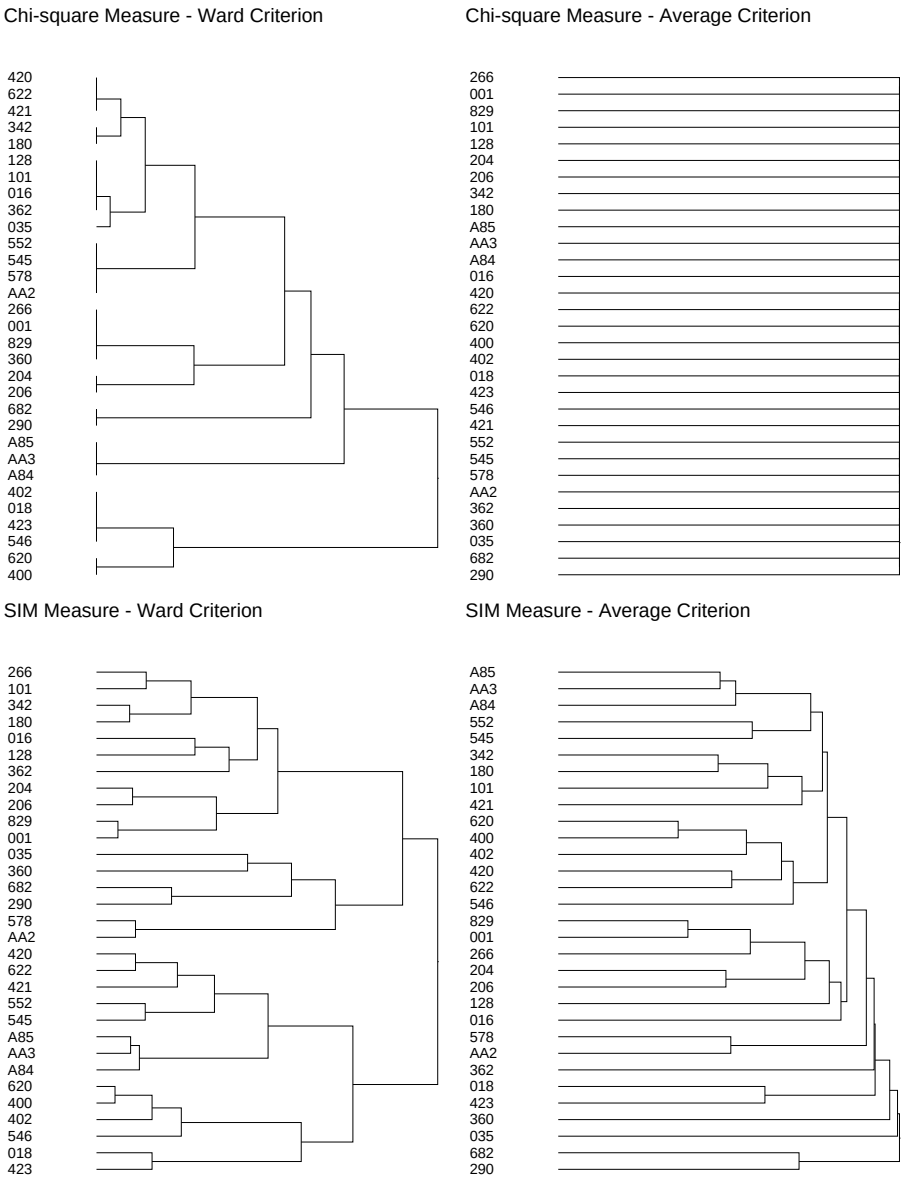

Fig. 1. Dendrograms for hierarchical clustering using the Ward and average criteria.