Product Analytics Based On Demographic Democratization

Harry Katzan, Jr., Savannah State University, USA
William A. Dowling, Savannah State University, USA
F. Ruth Smith, Savannah State University, USA
Paul D. Torres, Savannah State University, USA

ABSTRACT

Product analytics is a blend of computational methods with the express purpose of facilitating the multifaceted process of decision-making based on demographic and consumer preferences. This complex subject is derived from consensus theory and includes structured analytics, categories, and the combination of evidence. The methodology is applicable to a wide range of business, economic, social, political, and strategic decisions. The paper describes a product allocation application to demonstrate the concepts.

Keywords: Categorical analysis, democratization, consensus theory, uncertainty, evidence, Dempster's rule, analytics, categories, Dempster-Shafer theory

INTRODUCTION

The efficacy of decisions made under uncertainty on product allocation and selection is dependent upon two important concepts: the representation of the problem domain and the completeness of the solution space. A *category* is a means of representing the problem domain so that relevant structural information, based on demographics and consumer preferences, can be determined, resulting in improved decision-making. A *frame of discernment* (Shafer 1976; Katzan 1992, 2006, 2008) is a set of mutually exclusive and collectively exhaustive possibilities for the solution space. We are going to apply analytics to the task of product determination as a means of reducing the risk inherent in conventional statistical methods.

Demographics and Preferences

The selection of a product mix based on client preferences is an exceedingly complex task, because of the combinatorics of the independent variables. In an automobile selection process, for example, the number of consumer categories, such as gender, age, and education, is reasonably large, and the structural elements within each category are numerous enough to yield a large number of combinations. With the following categories, for example:

C = {gender, age, education}
gender = {male, female}
age = {<=25, 26-55, >=56}
education = {none, attended, grad}

the number of combinations of structural elements is 2×3×3 or 18. In a typical set of eight demographic categories, the number of combinations is considerably greater than 2^8.

We propose a methodology based on categorical analytics and the use of consensus theory (Katzan 2006) for combining information. The steps that comprise product analytics are:

1. Compute analytic alternatives for each category based on historical data. With products A and B for the gender category the process could yield the following data snippet:

Category	Structural Element	Structural Probability	Preference A	Preference B
gender	male	0.4	0.7	0.3
gender	female	0.6	0.5	0.5

2. Compute composite categorical probabilities by combining the analytic alternatives with preferences, as follows:

Category	Probability A	Probability B
gender	0.58	0.42

where the value 0.58 is computed as 0.4×0.7+0.6×0.5, and the value 0.42 is computed as 0.4×0.3+0.6×0.5.

3. Combine the resulting set of composite categorical probabilities, such as

Category	Probability A	Probability B
gender	0.58	0.42
age	0.44	0.56
education	0.46	0.54

using consensus theory yielding a set of probabilities for the solution set from which a decision maker can establish a product mix based on posterior probabilities concerning the client is most likely to purchase.

We call this process the **democratization** of product offerings based on preference data, because previous customers are essentially voting on what products to offer.

Category

A category is a means of structuring a problem domain with the objective of engaging in a predictive modality in which one or more future states may be identified and analyzed. Let $\mathbf{C}_i$ be one of the categories used to stratify the problem domain such that the collection

$$\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots , \mathbf{C}_n\}$$

represents a complete conceptualization of the dynamics under investigation and n is the number of categories.

Associated with each category is a set of probabilities representing an assessment of a future outcome based on its underlying categorical imperative. Thus, a category is a mechanism for isolating a single view of the problem under consideration. The ontological definition of a *category*, as a conceptual entity with no attributes in common with other categories, is adopted in this paper. The mutually exclusive set of possibilities under investigation is known as the frame of discernment. It is covered next, followed by a presentation of an approach to the structural analysis of categories. Finally, a product selection application is used as a demonstrative example that gives some insight into how the methods can be applied to other problems.

Frame of Discernment

A frame of discernment is a means of representing the possibilities under consideration, as in the following examples:

P = {sedan, wagon, roadster}
E = {stocks, bonds, gold}

Clearly, the elements in a frame of discernment are, in fact, propositions that can be interpreted as events or states. Thus, if component s_i of system S over domain V were associated with the symbol “sedan,” then that state is equivalent to the proposition, “The true value of *V* for component s_i is sedan,” or in ordinary language, “s_i prefers sedan.”

Accordingly, the set S of propositions S_i,

$$S = \{S_1, S_2, \ldots, S_n\}$$

represents the collection of states of a system under analysis. Clearly, at an agreed upon point in time, one proposition is true and the others are false.

Uncertainty

Prior to the agreed point in time (τ), we obviously do not know the state of the system under analysis or its components with any degree of certainty. The expectation that a part of the system will be in a particular state at time τ is denoted by a real number $p(S_i)$ associated with each of the propositions in the frame S = $\{S_i\}$, i=1,2,…,*n*, such that

$$0 \le p(S_i) \le 1$$

and

$$\sum_{i=1}^{n} p(S_i) = 1$$

This is simply the addition rule for mutually exclusive events.

CONSENSUS THEORY

Consensus theory is a methodology for combining evidence based on Dempster-Shafer theory (Shafer 1976; Katzan 1992, 2006) and the mathematical combination of evidence (Dempster 1967). Dempster-Shafer theory has commanded a considerable amount of attention in the scientific and business communities, because it allows a knowledge source to assign a numerical measure to a proposition from a problem space and provides a means for the measures accorded to independent knowledge sources to be combined. Dempster-Shafer theory is attractive because conflicting, as well as confirmatory, evidence from multiple sources may be combined.

The basis of Dempster-Shafer theory is the frame of discernment (Θ), introduced previously. Accordingly, a knowledge source may assign a numerical measure to a distinct element of Θ, which is equivalent to assigning a measure of belief to the corresponding proposition. In most cases, the numerical measure will be a basic probability assignment. A measure of belief may also be assigned to a subset of Θ or to Θ itself.

Support Functions

Consider a frame of discernment Θ and its power set denoted by 2^{Θ}. For example, given the frame:

$$\Theta = \{a, b, c\}$$

The power set is delineated as follows:

$$2^{\Theta} = \{\{a, b, c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a\}, \{b\}, \{c\}\}$$

In Dempster-Shafer theory, a knowledge source apportions a unit of belief to an element of 2^{Θ}. This belief can be regarded as a mass committed to a proposition and represents a judgment as to the strength of the evidence supporting that proposition. When viewed in this manner, evidence focuses on the set corresponding to a proposition; this set is called a *focal set*.

The support for a focal set is a function m that maps an element of 2^{Θ}, denoted by A, onto the interval [0,1]. Given a frame of discernment Θ and function $m: 2^{\Theta} \rightarrow [0,1]$, a support function is defined as follows:

$m(\phi) = 0$, where ϕ is the null set
$0 \leq m(A) \leq 1$, and

$$\sum_{A \subset 2^{\Theta}} m(A) = 1$$

The support function m is called a *basic probability assignment*, which is assigned by the knowledge engineer or domain specialist.

A support function is called a *simple support function* if it reflects, at most, one focal set not equal to Θ. A simple support function assigns a measure of belief to the focal set A, as follows:

$m(A)>0$
$m(\Theta)=1-m(A)$
$m(B)=0$, for all $B \subset 2^{\Theta}$ and $B \neq A$

The simple support function for a focal set A assigns a portion of the total belief exactly to A and not to its subsets or supersets. The remainder of the belief is assigned to Θ. Because certainty function must add up to 1, $m(\Theta)=1-m(A)$.

It is possible that a body of knowledge or evidence supports more than one proposition, as in the following case. If

$\Theta = \{a, b, c, d\}$
$A = \{a, b\}$

and

$B = \{a, c, d\}$

then the evidence supports two focal sets, which in the example, are A and B. If $m(A)=0.5$ and $m(B)=0.3$, then $m(\Theta)=0.2$. A support function with more than one focal set is called a *separable support function*. Separable support functions are normally generated when simple support functions are combined.

The notion of combining simple support functions is a practical approach to the assessment of evidence. An analyst obtains information from a knowledge source, and it leads to an immediate conclusion – not with certainty, but with a certain level of belief. This is a straightforward means of handling human affairs and is precisely what people do. When additional information comes in, the various pieces of evidence are combined to obtain a composite picture of the situation.

Combination of Evidence

A method of combining evidence is known as Dempster's rule of combination (Dempster 1967). Evidence would normally be combined when it is obtained from two different observations, each over the same frame of discernment. The combination rule computes a new support function reflecting the consensus of the combined evidence.

If m_1 and m_2 denote two support functions, then their combination is denoted by $m_1 \oplus m_2$ and is called their *orthogonal sum*. The combination $m_1 \oplus m_2$ is computed from m_1 and m_2 by considering all products of the form $m_1(X)\bullet m_2(Y)$, where X and Y range over the elements of Θ; $m_1(X)\bullet m_2(Y)$ is the set intersection of X and Y combined with the product of the corresponding probabilities.

For example, consider the frame of discernment

Θ = {healthy, tests, sick}

and views A and B, based on two different observation over the same frame:

X = {{healthy},0.6},{{tests},0.3},{{sick},0.1}}
Y = {{healthy},0.4},{{tests},0.4},{{sick},0.2}}

The entries are combined as follows using Dempster's rule of combination:

$m_1 \oplus m_2(\{\text{healthy}\}) = 0.24$
$m_1 \oplus m_2(\{\text{tests}\}) = 0.12$
$m_1 \oplus m_2(\{\text{sick}\}) = 0.02$
$m_1 \oplus m_2(\{\varnothing\}) = 0.62$

Thus, for $A_i \cap B_j = \boldsymbol{A}$ and $m_1 \oplus m_2 = \boldsymbol{m}$, the combination rule is defined mathematically as:

$$\boldsymbol{m}(\boldsymbol{A}) = \sum_{A_i \cap B_j = \boldsymbol{A}} m_1(A_i)\bullet m_2(B_j) / (1 - \sum_{A_i \cap B_j = \varnothing} m_1(A_i)\bullet m_2(B_j))$$

The denominator reflects a normalization process to insure that the pooled values sum to 1. So, in this instance, the normalization process yields the combination

X⊕Y = {{healthy},0.63},{{tests},0.32},{{sick},0.05}}

after normalization by dividing the combined assessment by (1-0.62) or 0.38. Because the problem is well-structured, the representation can be simplified as

X⊕Y = {0.63,0.32,0.05}

For views A = $\{A_1, A_2, \ldots, A_n\}$ and B=$\{B_1, B_2, \ldots, B_n\}$, the combination rule can be simplified as

$$\mathrm{A} \oplus \mathrm{B} = \{A_1 \times B_1/k, A_2 \times B_2/k, \ldots, A_n \times B_n/k\} \qquad \textbf{[1]}$$

N

where $k = \sum_{i=1} A_i \times B_i$

We will refer to equation **[1]** as the *simplification rule*. An example of the preceding concepts is demonstrated through the elicitation of expert opinion.

Elicitation of Expert Opinion

Typically, experts do not agree, especially when system failure is concerned. A typical example might be the crash of an expensive fighter aircraft or the collapse of a building. Consider a situation where the frame of discernment is $\{A,B,C\}$ denoting that the failure could be caused by Component A, Component B, or Component C.

Expert #1 believes the failure is due to Component A with probability 0.75, Component B with probability 0.15, or Component C with probability 0.10. Expert #2 believes the failure is due to Component A with probability 0.30, Component B with probability 0.20, or Component C with probability 0.50. The support functions are:

Expert #1 = {{{A},0.75}, {{B},0.15}, {{C},0.10}} = {0.75, 0.15, 0.10}
Expert #2 = {{{A},0.30}, {{B},0.20}, {{C},0.50}} = {0.30, 0.20, 0.50}

Table 1 summarizes the application of the simplification rule to this problem. The opinion of the experts is summarized and reflects the differing opinions.

Table 1: Elicitation of Expert Opinion

Support Function	Probability Assignment	Entropy
Expert #1 (=X)	{0.75, 0.15, 0.10}	1.05
Expert #2 (=Y)	{0.30, 0.20, 0.50}	1.49
X×Y	{0.738, 0.098, 0.164}	1.08

The strong opinion of Expert #1 in favor of Component A, reflected in the low entropy (Theil 1967), has a major influence on the consensus.

STRUCTURAL ANALYTICS

A problem domain is composed of categories, each of which is defined by a set of alternate structures. In a product example, for example, the category gender could be defined as

gender = {male, female}

based on a structural assessment, such as demographics. In this instance, the category gender is one of many viewpoints of an underlying decision situation, which could be a vote in an election or a position on an important issue. We are going to argue that in many unstructured decision-making problems, the probabilistic outcome can be based on structural, rather than, preferential elements. **What makes an unstructured decision so complex is that there are usually several categories "tugging at the decision maker."** We are going to show how categorical assessments can be combined to form a composite assessment of a decision under consideration. Through the technique known as structural analysis, we are going to assign probabilities to the elements of the frame of discernment from a given category, and then use consensus theory to combine the probabilities from the various categories. For example, a choice based on gender could go one way and a choice based on age could go another way. A realistic assessment would involve the combination of the two factors.

Structural Elements

Each category $\mathbf{C}_\mathrm{i}$ is comprised of a set of structural elements $\mathbf{S}_\mathrm{i} = \{S_{\mathrm{i}1}, S_{\mathrm{i}2}, S_{\mathrm{i}3}, \ldots, S_{\mathrm{i}k}\}$, where k is the number of structural elements in category $\mathbf{C}_\mathrm{i}$. Consider the previously given universe defined as:

$\mathrm{C} = \{\mathrm{C}_1, \mathrm{C}_{2,}\,\mathrm{C}_3\}$ = {gender, age, education}, where
$\mathbf{S}_1$ = gender = {male, female}, and
$\mathbf{S}_2$ = age = {<=25, 26-55, >=56}
$\mathbf{S}_3$ = education = {none, attended, grad}

where $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_3$ are defined respectively as

S_{11} = male	S_{21} = <=25	S_{31} = none
S_{12} = female	S_{22} = 26-55	S_{32} = attended

and k = 2	S_{23} = >=56	S_{33} = grad
	and k = 3	and k = 3

Each problem domain is represented by a set of categories, each of which is a special lens into the underlying problem. Each category is defined as a set of structural components that define it. The categorical demographics in an election, for example, could be party, gender, age, and so forth. In the immediate example, the categories are gender, age, and education.

Structural Probabilities

Each structural element has a demographic probability $p(S_{ij})$

where $\sum_{j=1}^{k} p(S_{ij}) = 1$ for category i and structural element j in category i

and k is the cardinality of $\mathbf{S}_{i.}$ Accordingly, for category $\mathbf{C}_i$ and its structure $\mathbf{S}_i$, the probability set would be expressed as:

$\mathbf{P}_i = \{p(S_{i1}), p(S_{i2}), \ldots, p(S_{ik})\}$

For example, the probability set for category #1 (gender), could be

$\mathbf{P}_1 = \{0.4, 0.6\}$

representing male and female, respectively. Each $\mathbf{P}_i$ represents the "probability of occurrence" in the universe of study of the structural elements of category *i*. This is demographic information.

Analytic Alternatives

In this form of analysis, each category $\mathbf{C}_i$ has an associated probability set $\mathbf{P}_i$. Each structural element has a corresponding probability $p(S_{ij})$ in $\mathbf{P}_i$. That probability represents the likelihood that an object selected at random from category $\mathbf{C}_i$ would be S_{ij}. Another interpretation is that a value in $\mathbf{P}_i$ gives the proportion of the corresponding structural element in $\mathbf{C}_i$. Table 2 gives another example:

Table 2: Analytic Alternatives

Category	Structural Element	Probability
age	<25	0.4
age	26-55	0.4
age	>=56	0.2

The structural probabilities, alternately regarded as structural proportions, give a means of describing the environment in which a decision is to take place. In a product analysis, the environment would be the consumer demographics.

Preference Set

Each structural element is assigned a *preference set* over the frame of discernment from a knowledge source, such as a poll, survey, or statistical data. The probabilities in the preference set are the decision variables. For example, we might know that male person prefers product A with probability 0.7 and product B with probability 0.3. The set {0.7, 0.3} is known as the *preference set.*

Thus, for each structural element S_{ij} for all categories, there exists a preference set

$ps(S_{ij}) = \{p_{ij}(\Theta^1), p_{ij}(\Theta^2), \ldots, p_{ij}(\Theta^t)\}$, where t is the cardinality of the frame of discernment,

and

$\boldsymbol{\Theta}^{\mathbf{k}} = \{\Theta^1, \Theta^2, \ldots, \Theta^t\}$

Clearly, $\sum_{k=1}^{t} p_{ij}(\Theta^k) = 1$ for all i and j.

Composite Probabilities

Composite categorical probabilities for each element in the frame of discernment are computed by combining the structural probabilities and corresponding preference set as follows:

$$P(\Theta_{it}) = \sum_{j=1}^{\mathbf{t}} ((p(S_{\mathbf{ij}}) \cdot ps(S_{ij})) \qquad \textbf{[2]}$$

where the index i runs through the categories and the index t runs through the alternatives in the frame of discernment.

Categorical Probabilities

The composite probabilities represent a summation of the preference for each element of the frame of discernment for each category. The result is a set of independent categorical assessments of the problem domain from different viewpoints represented as probabilities, as follows:

$\mathbf{C}_i = \{P(\Theta_{i1}), P(\Theta_{i2}), \ldots, P(\Theta_{it})\}$

where t is the cardinality of the frame of discernment, as defined previously. Using the simplification rule **[1]**, we derive a combined assessment of categories $\mathbf{C}_i$ and $\mathbf{C}_j$ of the form

$\mathbf{C}_i \oplus \mathbf{C}_j$

So that if

$\mathbf{C}_1 = \{0.54, 0.46\}$ and $\mathbf{C}_2 = \{0.58, 0.42\}$

Then

$\mathbf{K} = \mathbf{C}_i \oplus \mathbf{C}_j = \{0.62, 0.38\}$

The evidence is complementary, and that fact is reflected in the combined assessment.

PRODUCT ALLOCATION APPLICATION

One of the most familiar unstructured decision applications is the prior assessment of the products that customers will purchase. The major determinants of what products people will purchase can be combined into four well-known categories: gender, age, education, and race. The structural elements for each of the categories are given in Sheet 1, along with the respective structural probabilities. The columns are titled “Demographics.” For the category gender, the structural element male has a probability (or proportion), for example, of 0.4. Associated with each structural element is a preference set for that element over the frame of discernment, which is {A, B}. In this case, a person in gender/male, would choose product A with probability 0.7 and B with probability 0.3.

Categorical probabilities are calculated as a set of composite probabilities using equation **[2]**, as shown in Sheet 2, which gives spreadsheet functions that compute the respective probabilistic elements in the category probability set. Sheet 1 gives the computed probabilities for this example in the "Categorical Probabilities" section.

Finally, the consensus probabilities are computed using the simplification rule (equation **[1]**) in the "Consensus" section of Sheet 2. The results of the actual calculations are given in the "Consensus" section of Sheet 1. The probabilities are combined from top down, starting with the gender category and ending with race.

The results are more sensitive to demographics than they are to the preferences, as evidenced through experimentation with the spreadsheet recalculation facility.

SUMMARY

An admixture of methods has been given to structure a problem domain into categories and to compute categorical probabilities from structure elements and preference sets. The categorical probabilities are then combined using Dempster's rule of combination to obtain a composite assessment of the decision landscape. A demonstrative product allocation application is given. A sales organization must make an assessment of the products to stock, and product analytics provides a methodology of formalizing the selection process. The methods can be applied to product features, as well as to products.

AUTHOR BIOGRAPHIES

Drs. Katzan, Dowling, Smith, and **Torres** are on the faculty of the College of Business Administration at Savannah State University, the cornerstone of academic research in the southeastern United States. Their disciplines are information systems, finance, marketing, and accounting, respectively. Information on their bodies of work can be found on the Web using a major search engine.

REFERENCES

1. Dempster, A.P. 1967, "Upper and Lower Probabilities Induced by a Multivalued Mapping," *The Annals of Statistics* 28:325-339.
2. Katzan, H. 1992, *Managing Uncertainty: A Pragmatic Approach*, New York: Van Nostrand Reinhold Co.
3. Katzan, H. 2006, "Consensus," Proceedings of the Decision Science Institute National Conference, San Antonio TX. (November 2006).
4. Katzan, H. 2008, "Categorical Analytics Based on Consensus Theory," *Journal of Business and Economics Research*, 6(8): 89-102.
5. Shafer, G. 1976, *A Mathematical Theory of Evidence*, Princeton, NJ: Princeton University Press.
6. Theil, H. 1967, *Economics and Information Theory*, New York: American Elsevier Publishing Company, Inc.

Sheet 1: Spreadsheet for the Product Allocation Application showing Demographics, Preferences, Categorical Probabilities, and the Product Consensus for products A and B

	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**Product Consensus (A/B)**			**Categorical Probabilities (A/B)**				**Demographics**				**Product Preferences (A/B)**	
2	0.58	0.42		Gender	0.58	0.42		Gender	Male	0.4		0.7	0.3
3									Female	0.6		0.5	0.5
4	0.520391517	0.479608483		Age	0.44	0.56		Age	<=25	0.4		0.2	0.8
5									26-55	0.4		0.5	0.5
6									>=56	0.2		0.8	0.2
7	0.480327332	0.519672668		Education	0.46	0.54		Education	None	0.4		0.2	0.8
8									Attended	0.2		0.5	0.5
9									Grad	0.4		0.7	0.3
10	0.530447127	0.469552873		Race	0.55	0.45		Race	B/C	0.1		0.2	0.8
11									Cauc	0.8		0.6	0.4
12									other	0.1		0.5	0.5

Sheet 2: Spreadsheet for the Product Allocation Application Giving Functions for the Calculations in Sheet 1

	B	C	D	E	F	G
1	**Product Consensus (A/B)**			**Categorical Probabilities (A/B)**		
2	=F2	=G2		Gender	=K2*M2+K3*M3	=K2*N2+K3*N3
3						
4	=(B2*F4)/(B2*F4+C2*G4)	=(C2*G4)/(B2*F4+C2*G4)		Age	=K4*M4+K5*M5+K6*M6	=K4*N4+K5*N5+K6*N6
5						
6						
7	=(B4*F7)/(B4*F7+C4*G7)	=(C4*G7)/(B4*F7+C4*G7)		Education	=K7*M7+K8*M8+K9*M9	=K7*N7+K8*N8+K9*N9
8						
9						
10	=(B7*F10)/(B7*F10+C7*G10)	=(C7*G10)/(B7*F10+C7*G10)		Race	=K10*M10+K11*M11+K12*M12	=K10*N10+K11*N11+K12*N12