確率的潜在クラスモデルに基づく初期購買を考慮した RFM 分析モデルとその可視化に関する研究

情報数理応用研究　　5214C045-1　張倩
指導教員　後藤正幸

Analysis of Purchase History Data Using Extended RFM Analysis Based on Probabilistic Latent Class Model

ZHANG Qian

1　研究背景・目的

近年の情報技術の発達に伴い，顧客に関する多様なデータを大量に取得できるようになった．本研究で対象とする小売店舗においても，顧客の購買行動データの分析に基づくマーケティング施策の立案が求められている．一般に，企業が事業を運営する上で，既存顧客を維持するために必要なコストは新規顧客を獲得するために必要となるコストの約 5 分の 1 程度であると言われている．加えて，既存顧客の中には「優良顧客」が存在し，これらの顧客が利益の大半を創出している傾向にある．従って，これらの優良顧客を効率的に抽出し，その特徴を把握できる手法が求められる．本研究では，このための手法として顧客クラスタリングに着目し，優良顧客をはじめとする特徴的な顧客群を抽出するための方法を提案する．

マーケティング目的で顧客をクラスタリングするための分析手法として，顧客の購買行動を R(最終購買日)，F(購買頻度)，M(購買金額) の 3 変数により表現し，特性を総合的に判断する RFM 分析 [1] が幅広く活用されている．他方，機械学習の分野では，Probabilistic Latent Semantic Analysis（以下 PLSA）[2，3] に関する研究が数多く展開され，応用事例も多く存在する．PLSA は，潜在的に類似した顧客群を抽出すると同時に，クラスタリングによって，顧客を潜在的に類似したいくつかのグループへと分割することを可能とする．

本研究では，特徴的な顧客を抽出すると同時に顧客のグループ分けを行うために，これら 2 手法を統合し，RFM 分析の 3 種の特徴量を用いた潜在クラスモデルの提案を行う．これにより，RFM 分析の問題点である「R，F，M のスコアリングの基準が分析者によって異なる」という問題点を回避することができる．一方，RFM 分析では，その顧客がいつから購買を始めたかという情報は，顧客の入れ替わりのある小売店では重要であるもののモデルには取り入れられていない．また，潜在クラスモデルによる分析結果を分かり易く可視化する方法が望まれる．そこで，本研究ではさらに，提案モデルを初期購買日を新たな特徴量とした潜在クラスモデルに基づく RFM モデルへと拡張する．さらに，拡張したモデルを活用し，自己組織化マップ（SOM）[5] により，分析結果を可視化する方法を示す．経営科学系研究部会連合協議会主催のデータ解析コンペティションで提供された，株式会社アイディーズの i-code データサービスにおける購買履歴データを対象とした分析を行い，提案手法の有効性を示す．

2　準備

2.1　RFM 分析

RFM 分析は，顧客の購買行動のうち R（Recency：最終購買日），F（Frequency：購買頻度），M（Monetary：購入金額）に注目して分析を行う方法である．これらの各指標の数値に重みを付与し，合算したランキングを作成することで，優良顧客を始めとする特徴的な顧客群を抽出することが可能となる．例として，R，F，M それぞれを表 1 のような基準を用い，5 段階に分け，これらの情報を基にした顧客のクラスタリングを行う．

一般の RFM 分析では表 1 のように顧客ごとの購買行動をスコアリングし，その得点によって顧客を特徴づける．例えば，合計得点が 15 の顧客は超優良顧客，13 以上が優良顧客と解釈することができる．他方，合計ランク 3 の顧客は非優良顧客と解釈することができる．

表 1: RFM 分析のスコアリングの例

	R	F	M
5	一週間以内	40 回以上	10 万円以上
4	一ヶ月以内	20 回以上	4 万円以上
3	二ヶ月以内	10 回以上	2 万円以上
2	三ヶ月以内	5 回以上	5 千円以上
1	三ヶ月より前	5 回未満	5 千円未満

ここで，通常の RFM 分析の問題点として，分析者と分析対象データにより，表 1 のような重み付与の基準が異なるため，分析結果は恣意的になるという点が存在する．

2.2　Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis（PLSA）[2，3] は，嗜好の類似したユーザの集合である潜在クラスを仮定し，ユーザとアイテムの共起関係を潜在クラスによる条件付確率分布で表したモデルである．このモデルではユーザとアイテムが唯一の潜在クラスに属するのではなく，複数の異なる潜在クラスに所属することを許容している．この仮定により，ユーザの嗜好とアイテムの被購入傾向の多様性を表現することができる．

いま，ユーザを $u_m(m \in \{1,\cdots,M\})$，アイテムを $a_n(n \in \{1,\cdots,N\})$，潜在クラスを $z_k(k \in \{1,\cdots,K\})$ とそれぞれ定義する．ここで，M，N，K はそれぞれ全ユーザ数，全アイテム数，総潜在クラス数とする．ユーザの潜在クラスに対する所属確率 $P(u_m|z_k)$ とアイテムの潜在クラスに対する所属確率 $P(a_n|z_k)$ を用いると，ユーザとアイテムの共起関係は式 (1) で表される．

$$P(u_m, a_n) = \sum_{k=1}^{K} P(z_k)P(u_m|z_k)P(a_n|z_k) \qquad (1)$$

ただし，$P(z_k)$ は $\sum_{k=1}^{K} P(z_k) = 1$ を満たすものとする．ここで各パラメータ $P(z_k)$，$P(u_m|z_k)$，$P(a_n|z_k)$ は EM アルゴリズム [4] により対数尤度を最大にするパラメータとして推定することができる．

2.3　自己組織化マップ

自己組織化マップ（SOM：Self-Organizing Map）は，T.Kohonen により提案された教師なし学習を行うニューラルネットワークモデルで，高次元の属性を持つデータを非線形写像により 2 次元平面上へデータ間の類似性に応じてプロットする手法である．

自己組織化マップは，入力層と出力層により構成され，クラスタリング，可視化と抽象化などの問題によく用いられる．SOM を用いて高次元構造のデータは 2 次元空間内に可視化され，各データはマップのユニットにプロットされる．マップでは，似ているデータが近いユニットに配置され，異なるデータが遠いユニットに配置される．

2.4 対象とする購買履歴データに関する基本分析

本研究では，某スーパーマーケットの代表的店舗における 2014 年 1 月から 2014 年 12 月の購買履歴データを用いて解析する．対象とする購買履歴データから，RFM 分析に用いる R，F，M の三つの特徴量を顧客ごとに抽出し，これらの分布を図 1-3 に示す．図 1 より，最終購買日は現時点に近ければ近いほど顧客数が多いことがわかる．図 2 より，購買頻度は多ければ多いほど顧客数が少ないことがわかる．図 3 より，購買金額は同様に高ければ高いほど顧客数が少ないことがわかる．

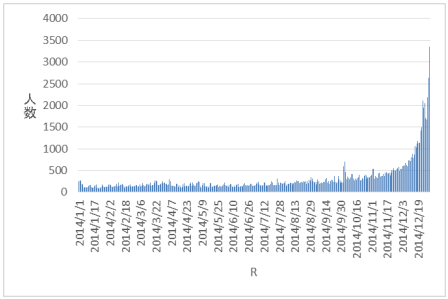

図 1: 顧客の R の分布

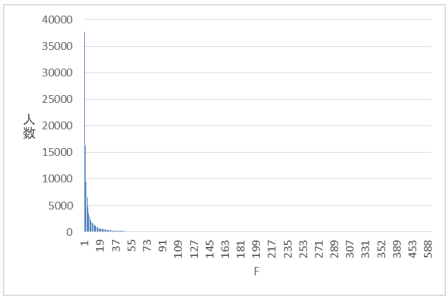

図 2: 顧客の F の分布

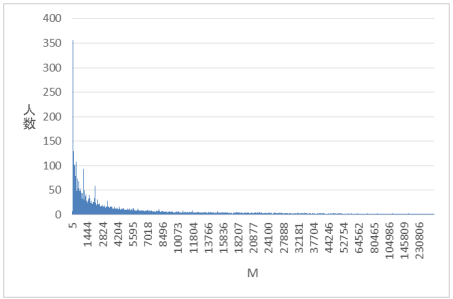

図 3: 顧客の M の分布

表 2 に全顧客の各特徴量の平均と中央値を示す．F の平均が 12.3 に対して中央値が 3.0 であるからも，購買頻度が 1，2，3 の顧客が大半を占めることがわかる．

表 2: 顧客の各特徴量の平均と中央値

	R	F	M
平均	2014/10/27	12.3	17979.2
中央値	2014/9/9	3.0	3265.0

3 潜在クラスモデルに基づく RFM 分析モデル

3.1 定式化

前述の通り，RFM 分析では特徴量 R，F，M を用いてデータをスコアリングする．しかし，分析者と購買データによりスコアリング基準が異なるため，分析結果は恣意的になる．顧客の購買情報 R，F，M を適切に活用するためには，スコアリング基準を自動的に決定し，顧客クラスタリングするための手法が必要となる．さらに，R，F，M の基準を自動で決定することにより，どのデータに対しても共通した分析を可能とし，適用することができれば，企業の経営活動の一助となると考えられる．

以下，本節では，前述の RFM 分析により抽出した購買行動の R，F，M を同時に表現する提案モデルについて述べる．一般に PLSA[2] はユーザとアイテムの間に潜在クラスの存在を仮定する．本研究では，購買行動の R，F，M という特徴量の背後に，嗜好の類似したユーザの集合である潜在クラスを仮定し，これらの共起関係を潜在クラスによる条件付確率分布で表した潜在クラスモデルを提案する．

まず，RFM のそれぞれの特徴量を $x_{ni}(i \in \{1,\cdots,L\}, n = \{1,2,3\})$，$L$ を総ユーザ数，n=1 は R，n=2 は F，n=3 は M の特徴量とする．提案する潜在クラスモデルのグラフィカルモデルを図 4 に示す．

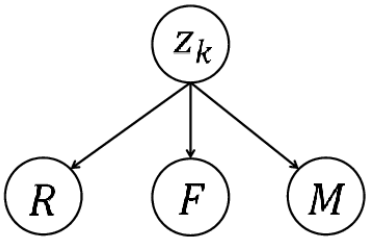

図 4: 提案手法のグラフィカルモデル

提案手法では，顧客 i が持っている R，F，M という事象 (x_{1i}, x_{2i}, x_{3i}) の集合である $\mathcal{D} = \{(x_{1i}, x_{2i}, x_{3i})\}_{i=1}^{L}$ を学習データとしてモデルの学習に用いる．このとき，潜在クラスにおける各特徴量 R，F，M の出現傾向を表す確率密度関数 $P(x_{ni}|z_k)$ を用いて，RFM に基づく顧客の購買行動を表す確率分布 $P(x_{1i}, x_{2i}, x_{3i})$ を式 (2) で定式化する．

$$P(x_{1i}, x_{2i}, x_{3i}) = \sum_{k=1}^{K} P(z_k)P(x_{1i}|z_k)P(x_{2i}|z_k)P(x_{3i}|z_k) \quad (2)$$

ただし，式 (2) における $P(x_{ni}|z_k)$ はそれぞれ独立に正規分布に従うものと仮定し，その平均を μ_{nk}，分散を σ_{nk}^2 とすれば，$P(x_{ni}|z_k)$ は式 (3) により表現することができる．

$$P(x_{ni}|z_k) = \frac{1}{\sqrt{2\pi\sigma_{nk}^2}} \exp\left\{-\frac{(x_{ni}-\mu_{nk})^2}{2\sigma_{nk}^2}\right\} \quad (3)$$

3.2 パラメータ推定法

提案モデルのパラメータを推定するため，本研究では EM アルゴリズムを適用する．学習データセットに対する対数尤度関数 LL は式 (4) のように示される．

$$LL = \sum_{i=1}^{L} \log P(x_{1i}, x_{2i}, x_{3i}) \quad (4)$$

式 (4) に対して E-Step(尤度の最大化)，M-Step(パラメータ推定) の 2 ステップを収束するまで繰り返すことにより対数尤度関数 LL を最大化するパラメータを推定するための EM アルゴリズム [4] を適用し，パラメータを求める方法を提案する．このときの EM アルゴリズムのパラメータ推定式を式 (5)-(8) に示す．

[E-Step]

$$P(z_k|x_{1i},x_{2i},x_{3i}) = \frac{P(z_k)P(x_{1i}|z_k)P(x_{2i}|z_k)P(x_{3i}|z_k)}{\sum_k P(z_k)P(x_{1i}|z_k)P(x_{2i}|z_k)P(x_{3i}|z_k)} \quad (5)$$

[M-Step]

$$P(z_k) = \frac{1}{L}\sum_i P(z_k|x_{1i},x_{2i},x_{3i}) \quad (6)$$

$$\mu_{nk} = \frac{\sum_i x_{ni}P(z_k|x_{1i},x_{2i},x_{3i})}{\sum_i P(z_k|x_{1i},x_{2i},x_{3i})} \quad (7)$$

$$\sigma_{nk}^2 = \frac{\sum_i (x_{ni}-\mu_{nk})^2 P(z_k|x_{1i},x_{2i},x_{3i})}{\sum_i P(z_k|x_{1i},x_{2i},x_{3i})} \quad (8)$$

3.3 実データの解析結果と考察

提案手法の有効性を検討するため，本研究では，前述した某スーパーマーケットの代表的店舗における 2014 年 1 月から 2014 年 12 月の購買履歴データに提案モデルを適用する．このときの顧客数は L=111,753 であり，事前分析により，潜在クラス数 K=20 と設定した．

提案モデルを用いた分析の結果を表 3 に示す．

表 3: 提案モデルによる分析結果

	R	F	M	割合
z_1	183.8	1.0	1261.5	33.62%
z_2	135.4	2.0	2583.5	14.56%
z_3	107.6	3.0	3968.8	8.47%
z_4	91.4	4.0	5370.1	5.79%
z_5	80.1	5.0	6909.7	4.21%
z_6	73.0	6.0	8371.8	3.10%
z_7	102.9	7.3	6658.9	1.53%
z_8	9.7	8.4	8766.3	1.79%
z_9	68.3	9.2	10771.2	2.32%
z_{10}	24.9	11.2	12903.6	3.05%
z_{11}	203.3	12.3	14676.9	1.54%
z_{12}	8.2	15.6	19163.1	3.70%
z_{13}	53.2	16.0	22780.8	2.50%
z_{14}	19.6	25.5	36713.3	2.78%
z_{15}	4.6	31.6	40820.5	2.89%
z_{16}	128.8	36.3	48021.0	1.21%
z_{17}	8.1	55.3	80366.0	3.21%
z_{18}	73.0	104.2	163793.0	0.46%
z_{19}	3.8	107.5	165763.0	2.37%
z_{20}	1.8	219.7	397618.0	0.89%

表 3 は提案モデルにより得られた各潜在クラスを F の値の昇順に並べたものである．表の各列はデータの一年間の最終購買日 $\mu_{1k}(k \in \{1,\cdots,20\})$(数字が小さい方が現時点に近い)，総購買頻度 μ_{2k}，総購買金額 μ_{3k}，各潜在クラスの顧客割合を表している．

表 3 より，潜在クラス z_1，z_2 では，R の値が大きく，F と M の値が小さいことから，非優良顧客のクラスタと解釈することができる．このような非優良顧客がほぼ 50%程度存在することが分かる．一方，潜在クラス z_{20} は R の値が最も小さく，F と M の値が最も大きいことから，超優良顧客と解釈可能である．加えて，潜在クラス z_{19} は z_{20} より R の値が少し大きく，F と M の値が少し小さいものの，同様の傾向を示しており，優良顧客と考えられる．また，潜在クラス z_{18} は F と M の値が z_{19} と同様に大きく，優良顧客であるとも考えられるが，R が大きな値を示しており，離反顧客と想定される．

4 初期購買を考慮した潜在クラスモデルに基づく RFM 分析モデル

4.1 定式化

前述の潜在クラスを用いた RFM 分析では，最終購買日，購買頻度，購買金額に着目していた．しかし，例えば，ある一年間の R，F，M の値が同じ顧客を考えた場合でも，最初の購買日 (初期購買日) の値により，顧客の解釈が大きく変化するものと思われる．そこで本研究では，初期購買日を新たな特徴量として導入したモデルの構築を行う．

いま，i 番目の顧客の初期購買日を x_{4i} とする．潜在クラスにおける最初購買日の出現確率 $P(x_{4i}|z_k)$ を用いて，式 (2) に対して，初期購買日を考慮したモデルを式 (9) で定式化する．

$$P(x_{1i},x_{2i},x_{3i},x_{4i}) = \sum_{k=1}^{K} P(z_k)P(x_{1i}|z_k)P(x_{2i}|z_k)P(x_{3i}|z_k)P(x_{4i}|z_k) \quad (9)$$

ただし，$P(x_{4i}|z_k)$ は，式 (2) の $P(x_{ni}|z_k)$ と同様に正規分布に従うものと仮定する．

4.2 パラメータの推定法

3.2 節と同様に EM アルゴリズムによりパラメータの推定を行う．E-step，M-step はそれぞれ以下のように表される．

[E-Step]

$$P(z_k|x_{1i},x_{2i},x_{3i},x_{4i}) = \frac{P(z_k)P(x_{1i}|z_k)P(x_{2i}|z_k)P(x_{3i}|z_k)P(x_{4i}|z_k)}{\sum_k P(z_k)P(x_{1i}|z_k)P(x_{2i}|z_k)P(x_{3i}|z_k)P(x_{4i}|z_k)} \quad (10)$$

[M-Step]

$$P(z_k) = \frac{1}{L}\sum_i P(z_k|x_{1i},x_{2i},x_{3i},x_{4i}) \quad (11)$$

$$\mu_{nk} = \frac{\sum_i x_{ni}P(z_k|x_{1i},x_{2i},x_{3i},x_{4i})}{\sum_i P(z_k|x_{1i},x_{2i},x_{3i},x_{4i})} \quad (12)$$

$$\sigma_{nk}^2 = \frac{\sum_i (x_{ni}-\mu_{nk})^2 P(z_k|x_{1i},x_{2i},x_{3i},x_{4i})}{\sum_i P(z_k|x_{1i},x_{2i},x_{3i},x_{4i})} \quad (13)$$

4.3 実データの解析結果と考察

初期購買を考慮した拡張モデルの有効性を示すために 4 節の実験と同様に実購買履歴データを用いて解析する．実験条件は 4 節の実験と同様に顧客数は L=111,753 で，事前分析により潜在クラス数 K=20 と設定した．

初期購買を考慮した拡張モデルを用いた分析の結果を表 4 に示す．

表 4: 初期購買を考慮したモデルによる分析結果

	初期購買日	R	F	M	割合
z_1	183.7	183.7	1.0	1267.0	33.64%
z_2	230.1	135.4	2.0	2584.9	14.57%
z_3	254.5	107.6	3.0	3969.8	8.47%
z_4	215.5	44.7	4.0	5927.0	2.91%
z_5	323.2	138.2	4.0	4854.8	2.89%
z_6	282.6	80.1	5.0	6909.6	4.21%
z_7	280.3	52.8	6.5	6511.8	2.47%
z_8	339.4	148.7	7.6	8246.3	2.40%
z_9	226.3	13.9	8.8	10423.7	3.09%
z_{10}	349.0	36.3	10.1	11674.2	3.28%
z_{11}	280.1	77.1	12.5	18488.0	2.33%
z_{12}	340.7	16.0	16.0	21301.7	3.91%
z_{13}	357.3	136.7	17.8	22693.6	2.17%
z_{14}	222.2	8.2	21.3	28605.4	1.91%
z_{15}	354.8	11.2	31.9	44271.8	4.29%
z_{16}	360.9	102.4	52.2	71734.4	0.90%
z_{17}	278.2	4.8	57.2	82986.8	1.06%
z_{18}	359.7	5.6	74.2	108172.0	3.31%
z_{19}	313.2	51.9	96.6	159470.0	0.62%
z_{20}	362.5	2.4	182.1	314337.0	1.55%

表 4 より，3.3 節の実験と同様に，各潜在クラスについて解釈ができる．例えば，表 4 の潜在クラス z_1，z_2 では，R の値が大きく，F と M の値が小さいことから，非優良顧客のクラスタと考えられる．一方，潜在クラス z_{20} は，R の値が最も小さく，F と M の値が最も大きいことから，超優良顧客と解釈可能である．また，表 4 の z_4，z_5 に着目すると，F の値が 4.0 で等しくなっているのに対して，それ以外の指標の値は大きく異なっている (z_4 の方が初期購買日と R が小さく，M が大きい) ことから，z_4 は通常顧客，z_5 は離反顧客と解釈することができる．ここで，4 節の実験結果 (表 3) の潜在クラス z_4 に着目すると，表 4 における z_4 と z_5 の中間的な値をとっている．初期購買日の指標を追加したことにより，離反顧客のクラスタを作ることができたものと思われる．

5 自己組織化マップによる可視化に関する検討

本研究では，初期購買日，R，F，M の 4 次元の特徴量を用い，SOM により 2 次元平面上へ顧客をプロットし，顧客の類似性を可視化する．初期購買を考慮したモデルの結果を用い，各顧客の各潜在クラスへの所属確率に着目し，その値が最も大きいクラスに顧客を割り当てる．さらに，クラスタごとの顧客の初期購買日，R，F，M のデータを SOM に当てはめて顧客の関係性を可視化する．

図 5 はユニット数を 30 と設定して全データを SOM で分析し，各ユニットの特徴を可視化したものである．各ユニットの星グラフの 4 つの軸は反時計回り (右上:初期購買日，左上:R，左下:F，右下:M) に顧客の特徴量を表す．

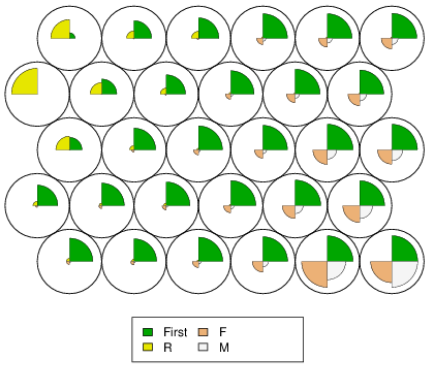

図 5: SOM による全データの図示

次に，提案モデルによって求めた潜在クラスのうち，潜在クラス z_1 と潜在クラス z_{18} を例とし，潜在クラスの可視化を行う．図 6 は，非優良顧客を表している潜在クラス z_1 に属するデータを黒字で，優良顧客を表している潜在クラス z_{18} に属するデータを灰色で，図 5 で得られた SOM 上にプロットしたものである．

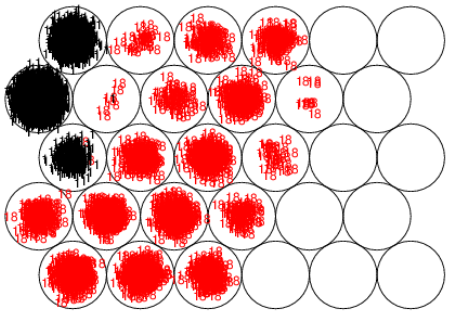

図 6: SOM による z_1,z_{18} に属するデータの図示

図 5 と図 6 を対照させることでクラスタの特徴を図で可視化することができる．非優良顧客は図 6 の左上にプロットされ，初期購買日 (First) が全体的に小さく，R が全体的に大きく，F と M が小さいという特徴をもつ．また，優良顧客は図 6 の中央左側にプロットされ，特徴としては初期購買日が全体的に大きく，R が小さく，F が相対的に大きい．このように，潜在クラスごとの特徴を SOM によって可視化することができる．

6 まとめと今後の課題

本研究では，RFM 分析をベースにし，正規分布を用いた潜在クラスモデルに基づく顧客クラスタリング手法を提案した．また，小売店には重要な情報として，顧客の初期購買日を考慮し，提案モデルを拡張した．実際の顧客の購買行動履歴データを用いて提案手法の有効性を示した．さらに，自己組織化マップにより，潜在クラスごとの特徴を可視化する方法を示した．

今後の課題として，新規顧客の抽出方法が挙げられる．また，潜在クラスごとの購買傾向（購買アイテム）を SOM により可視化する方法の構築も今後の課題である．

参考文献

[1] C. H. Cheng, Y. S. Chen, “Classifying the segmentation of customer value via RFM model and RS theory,” *Expert Systems with Applications*, vol. 36, pp. 4176–4184, 2009.

[2] T. Hofmann, “Probabilistic Latent Semantic Indexing,” *Proc. of 22nd Ann. International ACM SIGIR Conf.*, pp.50-57, 1999.

[3] T. Hofmann, J. Puzicha, “Latent Class Models for Collaborative Filtering,” *Proc. 16th International joint Conference on Artificial Intelligence*, pp. 688–693, 1999.

[4] A. P. Dempster, N. M. Laird, D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Royal Statistical Society. Series B*, vol. 39, no. 1, pp. 1–38,1977.

[5] P. Hanafizadeh, M. Mirzazadeh, “Visualizing market segmentation using self-organizing maps and fuzzy Delphi method - ADSL market of a telecommunication company,” *Expert Systems with Applications*, vol. 38, pp. 198–205, 2011.