Breast cancer diagnostic typologies by grade-of-membership fuzzy modeling

José G. Dias
ISCTE-Instituto Universitário de Lisboa

Breast cancer diagnostic typologies by grade-of-membership fuzzy modeling

JOSÉ G. DIAS
ISCTE – University Institute of Lisbon
Department of Quantitative Methods
Edifício ISCTE, Av. das Forças Armadas, 1649–026 Lisboa
PORTUGAL
jose.dias@iscte.pt

Abstract: This paper proposes de definition of breast cancer diagnostic typologies by the grade-of-membership approach. This fuzzy clustering model is described theoretically, and a fixed point algorithm is used in its estimation. An application to breast cancer diagnostic classification shows the existence of two distinct patterns. The graphical representation of the grade-of-membership estimates confirms the good fuzzy properties of the two-cluster solution.

Key–Words: Grade-of-membership model, clustering, fuzzy partition, Breast cancer diagnostics

1 Introduction

Breast cancer is a lending cause of death for women worldwide. Indeed, it is the most common cause of cancer mortality, accounting for 16% of cancer deaths in adult women [11]. There is strong evidence that early detection through mammography screening and adequate treatment of women with a positive result could significantly reduce mortality from this life-threatening disease. This has propelled a considerable amount of research on breast cancer.

After diagnosis the main types of breast cancer remains surgery followed by radiotherapy with hormonal and chemotherapeutic agents often used to treat presumed micro-metastatic disease. Surgery removes any local tumor and allows that a sample can be taken to analysis the nodes to describe the disease. The examination allows the characterization of the disease that makes local recurrence and death more likely using characteristics such as grade of the tumor (degree of abnormality displayed by the cells), the size of the tumor (maximum diameter, in mm) and the number of involved nodes [9].

Clustering is the search of homogeneous subsets in a data set. The application of clustering algorithms has been extensive in the context of Statistics (e.g., [3]) and Fuzzy Set Theory (e.g., [1]). For example, in marketing, market segmentation means the identification of groups of customers with similar behavior given a large database of customer data. On the other hand, in oncological studies, a database on breast cancer characteristics may allow the identification of different stages and patterns of the development of the disease.

An important aspect in therapy planning is to anticipate the risk of further disease in such a way that the treatment can be adjusted. Thus those women at higher risk should be treated with stronger and more invasive treatment. This type of strategy reduces side effects of unneeded invasive treatments and saves resources. Here we assume that there is two groups of women with heterogeneous needs. The goal is to identify the existing clustering structure in the data and classify each woman into each group provided her profile. We apply the grade-of-membership model to identifying the fuzzy clustering structure. This model has become very popular em health and related fields (see e.g. [8, 10]).

The remainder part of this paper is organized as follows: Section 2 defines the conceptual fuzzy clustering framework; Section 3 provides a description of the data and its empirical analysis. The paper concludes with a summary of main findings, implications, and suggestions for further research.

2 The Grade-of-Membership model

Let us have a data set of n objects to be clustered. An object is denoted by i ($i = 1, \ldots, n$). Each object is characterized by J categorical variables $Y_j, j =$

$1, ..., J$ with L_j categories ($L_j \geq 2$). Thus, y_{ij} indicates the category of Y_j for individual i, with $1 \leq y_{ij} \leq L_j$. Associated with each individual response there are L_j binary variables Y_{ijl} ($i = 1, ..., n$; $j = 1, ..., J$; $l = 1, ..., L_j$), where $y_{ijl} = I(y_{ij} = l)$, *i.e.*

$$y_{ijl} = \begin{cases} 1, & y_{ij} = l \\ 0, & y_{ij} \neq l \end{cases},$$

with $\sum_{l=1}^{Lj} y_{ijl} = 1$ and $\sum_{j=1}^{J} \sum_{l=1}^{Lj} y_{ijl} = J$.

The fuzzy clustering Grade-of-Membership (GoM) [12, 5] model is a fuzzy-set classification approach that identifies the number, say K, of profiles or pure types that best describe the pattern of association between the categories of variables. The cluster or group of subject i is indicated by $C_i \in \{1, ..., K\}$, $i = 1, ..., n$.

The Grade-of-Membership (GoM) model [5] is defined by two sets of parameters:

1. the first one relates the K pure types and the J variables
$$\lambda_{kjl} = \Pr(Y_{ijl} = 1 \mid C_i = k)$$
i.e. it is the probability that individual i in pure type k has the response l to the variable j;

2. the second set of coefficients (g_{ik}) relates each observation to the K pure types, *i.e.,* they describe how close the individual profile i is to each of K pure types with $g_{ik} \geq 0$ and $\sum_{k=1}^{K} g_{ik} = 1$.

Thus, the GoM model parameterizes $\pi_{ijl} = \Pr(Y_{ijl} = 1)$ as

$$\begin{aligned} \pi_{ijl} &= \sum_{k=1}^{K} g_{ik} \Pr(Y_{ijl} = 1 \mid C_i = k) \\ &= \sum_{k=1}^{K} g_{ik} \lambda_{kjl}, \end{aligned} \qquad (1)$$

where π_{ijl} represents the (unconditional) probability that individual i has level l on variable j, and coefficients g_{ik} are mixed weights indicating the degree to which a given individual is represented by each of K classes. These parameters and their constraints determine the geometry of the space, in which each observation is represented as a convex combination of K coordinates defining the extreme points in the space and being referred to as "pure types" or "extreme points". In a clustering approach, they can also be interpreted as cluster or segment centroids [4].

The GoM model assumes local independence, i.e., the variables are independent within each cluster [6], hence

$$f(\mathbf{y}_i; \boldsymbol{\lambda}, \mathbf{g}_i) = \prod_{j=1}^{J} f_j(\mathbf{y}_{ij}; \boldsymbol{\lambda}_j, \mathbf{g}_i),$$

where $\boldsymbol{\lambda}_j = \{\lambda_{1jL_j}, ..., \lambda_{KjL_j}\}$, $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_J\}$, and $\mathbf{g}_i = \{g_{i1}, ..., g_{iK}\}$.

Assuming that $\mathbf{y}_{ij}$ follows a multinomial distribution

$$\mathbf{y}_{ij} \sim Multi_{L_j}(1, \pi_{ij1}, ..., \pi_{ijL_j})$$

with density $f_{ij}(\mathbf{y}_{ij}; \boldsymbol{\lambda}_j, \mathbf{g}_i) = \prod_{l=1}^{L_j} (\pi_{ijl})^{y_{ijl}}$, we have

$$f_i(\mathbf{y}_i; \boldsymbol{\lambda}, \mathbf{g}_i) = \prod_{j=1}^{J} \prod_{l=1}^{L_j} (\pi_{ijl})^{y_{ijl}}.$$

From equation (1), the definition of the Grade-of-Membership model for a given observation i is

$$f_i(\mathbf{y}_i; \boldsymbol{\lambda}, \mathbf{g}_i) = \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} g_{ik} \lambda_{kjl} \right)^{y_{ijl}}.$$

There are $nK + K \sum_{j=1}^{J} L_j$ parameters to estimate; the corresponding free parameters are $p_K = n(K-1) + K \sum_{j=1}^{J} (L_j - 1)$ due to the constraints: $\lambda_{kjL_j} = 1 - \sum_{l=1}^{L_j - 1} \lambda_{kjl}$ and $g_{iK} = 1 - \sum_{k=1}^{K-1} g_{ik}$.

Under the assumption that $\mathbf{y}_1, ..., \mathbf{y}_n$ are independent realizations of the feature vector $\mathbf{y}$, the likelihood function for $\boldsymbol{\varphi}$ is given by $L(\boldsymbol{\varphi}; \mathbf{y}) = \prod_{i=1}^{n} f_i(\mathbf{y}_i; \boldsymbol{\lambda}, \mathbf{g}_i)$, and $\boldsymbol{\varphi} = \{\boldsymbol{\lambda}, \mathbf{g}\}$ represents the GoM parameters, with $\mathbf{g} = \{\mathbf{g}_1, ..., \mathbf{g}_n\}$. Thus, the likelihood function of the GoM model is

$$L(\boldsymbol{\varphi}; \mathbf{y}) = \prod_{i=1}^{n} \prod_{j=1}^{J} \prod_{l=1}^{L_j} \left(\sum_{k=1}^{K} g_{ik} \lambda_{kjl} \right)^{y_{ijl}},$$

and the log-likelihood function for K pure types ($\ell_K(\boldsymbol{\varphi}; \mathbf{y}) = \log L(\boldsymbol{\varphi}; \mathbf{y})$) is

$$\ell_K(\boldsymbol{\varphi}; \mathbf{y}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{l=1}^{L_j} y_{ijl} \log \left(\sum_{k=1}^{K} g_{ik} \lambda_{kjl} \right). \quad (2)$$

The maximum likelihood estimate of $\boldsymbol{\varphi}$ is provided by the score equation

$$\frac{\partial \ell_K(\boldsymbol{\varphi}; \mathbf{y})}{\partial \boldsymbol{\varphi}} = \mathbf{0}.$$

The GOM model is estimated by the fixed point algorithm defined as follows:

1. Set K and the tolerance level (ε), the number of pure types and stop tolerance level, respectively; set $m \leftarrow 1$; initialize $g_{ik}^{(0)}$ from a random sample:

$$\begin{aligned} g_{ik}^{(0)} &\sim Uniform[0,1] \\ g_{ik}^{(0)} &\leftarrow \frac{g_{ik}^{(0)}}{\sum_{h=1}^{K} g_{ih}^{(0)}} \end{aligned}$$

   and

$$\lambda_{kjl}^{(0)} = \frac{1}{L_j};$$

2. Compute

$$g_{ik}^{(m)} = \frac{1}{J} \sum_{j=1}^{J} \sum_{l=1}^{L_j} y_{ijl} \frac{g_{ik}^{(m-1)} \lambda_{kjl}^{(m-1)}}{\sum_{h=1}^{K} g_{ih}^{(m-1)} \lambda_{hjl}^{(m-1)}}$$

3. Compute

$$\begin{aligned} \lambda_{kjl}^{(m)} &= \frac{1}{\sum_{i=1}^{n} \sum_{l=1}^{L_j} y_{ijl} \frac{g_{ik}^{(m)} \lambda_{kjl}^{(m-1)}}{\sum_{h=1}^{K} g_{ih}^{(m)} \lambda_{hjl}^{(m-1)}}} \\ &\quad \times \sum_{i=1}^{n} y_{ijl} \frac{g_{ik}^{(m)} \lambda_{kjl}^{(m-1)}}{\sum_{h=1}^{K} g_{ih}^{(m)} \lambda_{hjl}^{(m-1)}} \end{aligned}$$

4. Compute

$$\ell(\boldsymbol{\varphi}^{(m)}; \mathbf{y}) = \sum_{i=1}^{n} \sum_{j=1}^{J} \sum_{l=1}^{L_j} y_{ijl} \log \left(\sum_{k=1}^{K} g_{ik}^{(m)} \lambda_{kjl}^{(m)} \right)$$

   If $\left| \ell_K(\boldsymbol{\varphi}^{(m)}; \mathbf{y}) - \ell_K(\boldsymbol{\varphi}^{(m-1)}; \mathbf{y}) \right| < \varepsilon$, stop; otherwise $m \leftarrow m + 1$ and go to step 2.

Since this algorithm (any iterative algorithm) can not ensure the convergence to the global maximum for a given value of K, the algorithm should be repeated using different starting values. This algorithm is implemented in MATLAB 7 [7]. We run the algorithm 100 times with initial random solutions. As stopping rule we set $\varepsilon = 10^{-7}$.

3 Application

We apply the GoM model to the Ljubljana Breast Cancer Data Set with 277 instances of real patient data. The data set is available from UCI repository (http://archive.ics.uci.edu/ml/datasets/Breast+Cancer) and contains the following variables (see categories in Tables 1):

1. Age: age (in years at last birthday) of the patient at the time of diagnosis;

2. Menopause: whether the patient is pre- or post-menopausal at time of diagnosis;

3. Tumor size: the greatest diameter (in mm) of the excised tumor;

4. Inv-nodes: the number (range 0 - 39) of axillary lymph nodes that contain metastatic breast cancer visible on histological examination;

5. Node caps: if the cancer does metastasise to a lymph node, although outside the original site of the tumor it may remain "contained" by the capsule of the lymph node. However, over time, and with more aggressive disease, the tumor may replace the lymph node and then penetrate the capsule, allowing it to invade the surrounding tissues;

6. Degree of malignancy: the histological grade (range 1-3) of the tumor. Tumors that are grade 1 predominantly consist of cells that, while neoplastic, retain many of their usual characteristics. Grade 3 tumors predominately consist of cells that are highly abnormal;

7. Breast: breast cancer may obviously occur in either breast;

8. Breast quadrant: the breast may be divided into four quadrants, using the nipple as a central point;

9. Irradiation: radiation therapy is a treatment that uses high-energy x-rays to destroy cancer cells.

Table 1 provides the estimates of λ_{kjl} for $K = 2$ and the observed sample frequencies for each category. In fact, the GoM model provides a clear split or clustering of the sample into two groups with different profile: class 1 - *Early cancer stage* and class 2 - *Advanced cancer stage*. The estimates of λ_{kjl} suggest that the pure type I contains older women aged 50 and above, in opposition to pure type 2 characterized by women aged up to 49. This result shows that younger women are less aware of the disease and most of the time they diagnose the problem in an advanced stage of its development, in opposition to the older women. This result is in agreement with the second variable: given the women are pure type II, the probability of being premenopausal is 1.00.

In the early stage (cluster 1) the tumor size tends to be smaller, in contrast with cluster 2 in which tumor

Table 1: GoM model estimates ($\hat{\lambda}_{kjl}$)

Variables	GoM Model		Frequency
	Pure type 1	Pure type 2	
Age			
10-19	0.00	0.00	0.00
20-29	0.00	0.01	0.36
30-39	0.00	0.29	13.00
40-49	0.00	0.71	32.13
50-59	0.60	0.00	32.85
60-69	0.36	0.00	19.86
70-79	0.03	0.00	1.81
Menopause			
lt40	0.04	0.00	1.81
ge40	0.96	0.00	44.40
premeno	0.00	1.00	53.79
Tumor-size			
0-4	0.06	0.00	2.89
5-9	0.03	0.00	1.44
10-14	0.20	0.00	10.11
15-19	0.17	0.04	10.47
20-24	0.15	0.20	17.33
25-29	0.13	0.24	18.41
30-34	0.15	0.26	20.58
35-39	0.00	0.14	6.86
40-44	0.09	0.07	7.94
45-49	0.00	0.02	1.08
50-54	0.03	0.03	2.89
Inv-nodes			
0-2	1.00	0.52	75.45
3-5	0.00	0.24	12.27
6-8	0.00	0.12	6.14
9-11	0.00	0.05	2.53
12-14	0.00	0.02	1.08
15-17	0.00	0.04	2.17
18-20	0.00	0.00	0.00
21-23	0.00	0.00	0.00
24-26	0.00	0.01	0.36
Node-caps			
No	1.00	0.60	79.78
Yes	0.00	0.40	20.22
Deg-malig			
Low	0.46	0.00	23.83
Medium	0.33	0.61	46.57
High	0.21	0.39	29.60
Breast			
Left	0.56	0.49	52.35
Right	0.44	0.51	47.65
Breast-quad			
Left-up	0.36	0.32	33.94
Left-low	0.37	0.39	38.27
Right-up	0.08	0.16	11.91
Right-low	0.04	0.13	8.30
Central	0.15	0.00	7.58
Irradiation			
No	1.00	0.56	77.62
Yes	0.00	0.44	22.38

size is larger than 15 mm. This confirms that the number of axillary lymph nodes that contains metastatic breast cancer visible on histological examination (Inv-nodes) are in cluster 1 in maximum 2, in opposition to cluster 2. Node Caps indicates whether the cancer does metastasise to a lymph node. One concludes that in pure type I the answer is *no*, because of its early stage. Pure type II is more heterogeneous as it contains all the cases *yes*, and still some answers *no*.

The degree of malignancy in pure type I tends to be low, in opposition to pure type II in which it is at least medium degree. Note that the propensity for levels medium and high is two times more likely in pure type II than in pure type I.

For the variables associated with the location of the cells with cancer – Breast and Breast-quad –, the results tend to be heterogeneous, in each breast and in each quadrant of each breast. Thus, these variables cannot discriminate the two pure types.

Finally, irradiation shows that women in cluster I were not treated with high-energy x-rays. The second pure type a subgroup was treat to irradiation to destroy cancer cells, and probably for 56% the tumor was already widespread that did not justify the treatment given the cost-benefit trade-off.

Now we focus our analysis on the grade-of-membership parameter estimates ($\hat{g}_{ik}$). Figure 1 depicts the grade-of-memberships for the 277 women on a scatter plot. For $K = 2$, one has $g_{i1} + g_{i2} = 1$, and therefore the observations fall on the line $(g_{i1}, 1-g_{i1})$. The Figure gives as well the frequency in each point. Overall, a very small proportion of observations occupies the interior of the interval. Indeed most of the grades-of-membership values are extreme – $(0, 1)$ or $(1, 0)$ –, which shows a good level of separation of the two clusters.

To quantify the level of separation of the two groups we use the entropy or fuzzyness in the parameters g_{ik}. This approach is rather popular in the context of mixture models (see e.g. [2]). For GoM models, the relative entropy is given by

$$E_K = 1 + \frac{1}{n \log K} \sum_{i=1}^{n} \sum_{k=1}^{K} g_{ik} \log g_{ik}.$$

E_K varies between 0 and 1. Assuming $0log0 = 0$, in a hard partition with $g_{ik} \in \{0, 1\}$, $g_{ik} \log g_{ik} = 0$, and consequently $E_K = 1$. Complete fuzzyness means $g_{ik} = 1/K$, resulting in $E_K = 1$. In our case, $E_K = 0.59$ and consequently this two-cluster typology provides a good level of separation.

4 Conclusion

In this paper we focused on the Grade-of-Membership model as a breast cancer diagnostic tool. This type of computer-based system for identification of the breast

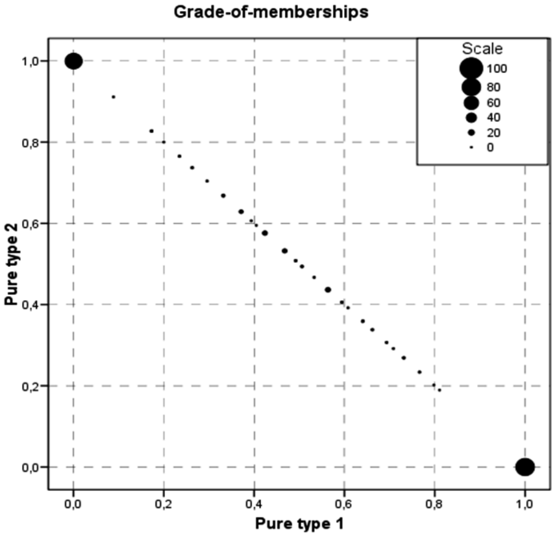

Figure 1: Scatter plot representation of the grade-of-membership estimates

cancer patterns can be very useful in diagnosis and management of the disease progression. After providing a short introduction on the importance of these tools in cancer research, we gave an overview of the Grade-of-Membership model and its estimation. We illustrate its performance in the understanding of the relation between variables collected in a cancer tumor diagnostic study. The two-cluster solution provides a good separation into the two groups of women set a priori. The pure types were well described by the variables that we would expect a priori (in Table 1). Future research can extend the model for predicting recurrence in breast cancer from the problem in breast cancer prognosis into two categories – no-recurrence-events and recurrence-events – from this clustering setting.

References

[1] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algoritms*, Plenum Press, New York 1981

[2] J.G. Dias and J.K. Vermunt, Bootstrap methods for measuring classification uncertainty in latent class models. In A. Rizzi and M. Vichi (eds.), *COMPSTAT2006. Proceedings in Computational Statistics*, pp. 31-41, Heidelberg, Physica/Springer–Verlag 2006

[3] J.G. Dias and M.J. Cortinhal, The SKM algorithm: A K-means algorithm for clustering sequential data, *Advances in Artificial Intelligence IBERAMIA 2008, Lecture Notes in Artificial Intelligence*, pp. 173-182, Springer–Verlag, Berlin 2008

[4] A. Maetzel, S.H. Johnson, M.A. Woodbury, C. Bombardier, Use of grade of membership analysis to profile the practice styles of individual physicians in the management of acute low back pain, *Journal of Clinical Epidemiology*, 53, 2000, pp. 195-205.

[5] K.G. Manton and M.A. Woodbury, A new procedure for analysis of medical classification, *Methods of Information in Medicine*, 21, 1982, pp. 210-220.

[6] K.G. Manton, M.A. Woodbury, E. Stallard, L.S. Corder, The use of grade-of-membership techniques to estimate regression relationships, *Sociological Methodology*, 22, 1992, pp. 321-381.

[7] MathWorks, *MATLAB 7.0*, The MathWorks, Natick, MA 2004

[8] P. McNamee, A comparison of the grade of membership measure with alternative health indicators in explaining costs for older people, *Health Economics*, 13(4), 2004, pp. 379 - 395.

[9] M.A. Richards, I.E. Smith, and J.M. Dixon, Role of systemic treatment for primary operable breast cancer, *BMJ*, 309, 1994, pp. 1263–1366.

[10] S. Szadoczky, S. Rozsa, S. Patten, M. Arato, and J. Furedi, Lifetime patterns of depressive symptoms in the community and among primary care attenders: an application of grade of membership analysis, *Journal of Affective Disorders*, 77, 2003, pp. 31-39.

[11] WHO, *World Health Statistics 2008*, World Health Organization, Geneva, Switzerland 2008

[12] M.A. Woodbury, J. Clive, Clinical pure types as a fuzzy partition, *Journal of Cybernetics*, 4, 1974, pp. 111-121.