ELSEVIER Computational Statistics & Data Analysis 41 (2003) 531–537

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Latent class models for classification

Jeroen K. Vermunt[a,*], Jay Magidson[b]

[a] *Department of Methodology and Statistics, Tilburg University, 5000 LE Tilburg, Netherlands*
[b] *Statistical Innovations Inc., Belmont, MA 02478, USA*

Received 1 February 2002; received in revised form 1 March 2002

Abstract

An overview is provided of recent developments in the use of latent class (LC) and other types of finite mixture models for classification purposes. Several extensions of existing models are presented. Two basic types of LC models for classification are defined: supervised and unsupervised structures. Their most important special cases are presented and illustrated with an empirical example.

Keywords: Classification; Feed-forward neural networks; Bayesian networks; Latent class models; Finite mixture models; Latent class regression models; Latent class factor models; Mixtures of experts; Supervised learning

1. Introduction

Let y denote a discrete dependent, outcome, target, or output variable, and $\mathbf{z}$ a vector of independent, input, predictor, or attribute variables.[1] Classification involves predicting the discrete outcome variable y as accurate as possible using the information on the z variables. Recently, latent class (LC), or finite mixture (FM), models have been proposed as classification tools in the field of neural networks (Jacobs et al., 1991; Bishop, 1995, pp. 212–220), as well as in the field of Bayesian (or belief) networks (Kontkanen et al., 1996; Monti and Cooper, 1999; Meilă and Jordan, 2000). This paper gives an overview of these developments and presents several extensions of the proposed models.

* Corresponding author. Tel.: +31-(0)134662748; fax: +31-(0)134663002.
E-mail address: j.k.vermunt@kub.nl (J.K. Vermunt).

[1] The discrete y variable is often denoted as the class variable and its categories as classes. Here, we will not use the term class(es) in order to prevent confusion with the term latent class(es).

0167-9473/03/$ - see front matter
PII: S0167-9473(02)00179-2

Classification using a statistical model involves specifying either a model for $P(y|\mathbf{z})$, as in regression analysis, or a model for $P(\mathbf{z}|y)$, as in discriminant analysis. In the next two sections, we present two basic types of LC models for classification: they involve specifying a model for $P(y|\mathbf{z})$ and $P(\mathbf{z}|y)$, respectively. Subsequently, we illustrate the most important special cases of these two basic types with an empirical example. The paper ends with a short discussion.

2. Supervised classification structures

The first basic type of LC model for classification involves specifying a model for the conditional distribution of y given $\mathbf{z}$, where a discrete hidden variable x serves as intervening variable. More precisely, the assumed probability structure for $P(y,\mathbf{z})$ is

$$P(y,\mathbf{z}) = P(\mathbf{z})P(y|\mathbf{z}) = P(\mathbf{z})\sum_{x} P(x|\mathbf{z})P(y|\mathbf{z},x), \tag{1}$$

where $P(\mathbf{z})$ is treated as fixed. Besides the above probability structure, regression-type constraints are imposed on the model probabilities. Since both the latent variable and the outcome variable are assumed to be discrete, it is most natural to restrict $P(x|\mathbf{z})$ and $P(y|\mathbf{z},x)$ using (multinomial) logit models. Note that the model is similar to the concomitant-variable LC model proposed by Dayton and Macready (1988). When the z variables are qualitative, the model is also a special case of the log–linear models with latent variables proposed by Hagenaars (1990, Chapter 3) and extended by Vermunt (1997, Chapter 3).

Maximum-likelihood or maximum posterior estimation is based on a likelihood function containing only the term $P(y|\mathbf{z})$, which implies that there is a direct relationship between model fit and classification performance. These LC models belong, therefore, to the family of supervised classification or supervised learning methods.

In the neural networks field, the model described in Eq. (1) is known as the *mixture-of-experts* model (Jacobs et al., 1991; Bishop, 1995, pp. 212–220). The original motivation was to provide a mechanism for partitioning the prediction problem between several neural networks. The separate neural networks can be much simpler than the model that would be needed without such a partitioning. In other words, rather than a complicated non-linear regression model describing the relationship between $\mathbf{z}$ and y, there is a standard (logistic) regression model for each LC or mixture component. As can be seen, the mixing proportions depend on the input variables (predictors), which means that the input space is divided into regions that differ with respect to the nature of the relationships between the input variables and the output variable. In the interpretation of the model parameters we have to take into account the double role of the input variables: component-specific regression coefficients defining $P(y|\mathbf{z},x)$ hold for certain combinations of z variables.

A special case of the mixture-of-experts model is the *LC regression* or *mixture regression* model, a model that is popular in the field of marketing research (Wedel and DeSarbo, 1994). In this model, the mixing distributions are assumed to be independent of $\mathbf{z}$; that is, $P(x|\mathbf{z}) = P(x)$. Interpretation of the model parameters is straightforward:

latent classes differ with respect to the size of regression coefficients. The predicted value of y is obtained by a weighted average of the component specific predictions, where the mixing proportions serve as weights.

Another special case is obtained by assuming that $P(y|\mathbf{z},x) = P(y|x)$; that is, when the effects of the z variables on the y go completely through x. This yields a structure that is similar to a feed-forward neural network with a single hidden layer. More precisely, a model with K latent classes is comparable to a neural network with $K-1$ nodes in the hidden layer. We label this model the *LC feed-forward* model. Siciliano and Mooijaart (2000) proposed a similar structure for a single discrete input variable, labeled latent budget model. In the literature, we have, however, not seen the model in its general form with a logit parameterization of $P(x|\mathbf{z})$ used for classification purposes. Feed-forward neural networks are often criticized because they yield results that are difficult to interpret. In contrast, the interpretation of the parameters of the proposed LC model is extremely easy. It is assumed that there are K basic output profiles defined by $P(y|x)$ and that the probability of having one of these output profiles depends on the input variables. Predicting y consists of taking a weighted average of the basic output profiles, where the weights depend on the input variables.

Rather than having a hidden layer consisting of single nominal latent variable, it is also possible to have a hidden layer with several dichotomous latent variables. Using the terminology introduced by Magidson and Vermunt (2001), this amounts to working with a LC factor rather than a LC cluster structure. A model with J (dichotomous) latent variables (labeled factors) that are mutually independent given $\mathbf{z}$ can be defined as

$$P(y|\mathbf{z}) = \sum_x \left[\prod_j P(x_j|\mathbf{z})\right] P(y|\mathbf{x}),$$

where $P(x_j|\mathbf{z})$ and $P(y|\mathbf{x})$ are again restricted using logit models. Conceptually, this factor variant of the LC feed-forward model is even more similar to a feed-forward neural network: each factor plays the role of a node in the hidden layer. A nice feature of this model is that the effects of the input variables on the output variable are explicitly split up into J independent dimensions. Similarly, factor variants of the mixture-of-experts and the LC regression model can be defined.

3. Unsupervised classification structures

In the second basic type of LC model for classification, one models the conditional distribution of the z variables given y, $P(\mathbf{z}|y)$. The decomposing of $P(y,\mathbf{z})$ is now

$$P(y,\mathbf{z}) = P(y)P(\mathbf{z}|y) = P(y)\sum_x P(x|y)P(\mathbf{z}|y,x). \tag{2}$$

Since the likelihood function used in the estimation is based on $P(\mathbf{z}|y)$ or $P(y,\mathbf{z})$, there is no direct relationship between model fit and classification performance. These methods belong, therefore, to the family of unsupervised classification or unsupervised

learning methods. The predictive distribution of y given $\mathbf{z}$, $P(y|\mathbf{z})$, that is needed to perform the classification task can be obtained by the well-known Bayes' theorem

$$P(y|\mathbf{z}) = \frac{P(y)P(\mathbf{z}|y)}{\sum_y P(y)P(\mathbf{z}|y)}.$$

With a single mixture component, several well-known classifiers arise depending on the form of $P(\mathbf{z}|y)$. The Naive Bayes (NB) classifier, for example, assumes mutual independence of the z variables within levels of y, $P(\mathbf{z}|y) = \prod_\ell P(z_\ell|y)$. Of course, the exact form of the conditional density $P(z_\ell|y)$ depends on the scale type of z_ℓ. Less restricted forms for $P(\mathbf{z}|y)$ are used in Bayesian tree classifiers and in discriminant analysis.

Kontkanen et al. (1996) proposed using a *standard LC*, or *standard FM*, model as classifier. This is the special case of the model defined in Eq. (2) obtained when $P(\mathbf{z}|y,x) = \prod_\ell P(z_\ell|x)$. Typical for this classifier is that the outcome variable has no differential status: all variables, including y, are assumed to be independent of one another within latent classes.

Monti and Cooper (1999) proposed combining elements of the standard LC model and the NB classifier. Their finite-mixture augmented Naive-Bayes (FAN) classifier has the form

$$P(y,\mathbf{z}) = P(y)\sum_x P(x)\prod_\ell P(z_\ell|y,x).$$

As in the NB model, the distribution of the z_ℓ variables depends on the outcome variable y. However, now the latent variable x is included to relax the assumption that the z_ℓ variables are conditionally independent given the outcome variable y. Compared to the LC model, y has back its differential status, which should yield a better classifier.

Note that the FAN model is not a full mixture of NB models since x is assumed to be independent of y. A natural extension of the FAN is obtained by including such a dependence; that is, by replacing $P(x)$ with $P(x|y)$. We label this model *mixture-of-NB* classifier.[2] Another extension of the FAN model is the mixture-of-trees model proposed by Meilã and Jordan (2000).

As discussed in the previous section, also with unsupervised structures it is possible to replace the single nominal latent variable by several dichotomous latent variables, which yields what we labeled LC factor models (Magidson and Vermunt, 2001). A factor variant of the mixture-of-NB model is

$$P(y,\mathbf{z}) = P(y)\sum_{\mathbf{x}}\left[\prod_j P(x_j|y)\right]\sum_x\prod_\ell P(z_\ell|y,\mathbf{x}).$$

Here, $P(z_\ell|y,\mathbf{x})$ is further restricted in order to exclude interaction effects between the factors. By setting $P(z_\ell|y,\mathbf{x}) = P(z_\ell|\mathbf{x})$ we obtain a factor variant of the standard LC model, and with $P(x_j|y) = P(x_j)$ we get a factor variant of Monti and Cooper's FAN.

[2] When all x variables are qualitative, the mixture-of-NB model is equivalent to the multiple-group LC model proposed by Clogg and Goodman (1984), where y serves as grouping variable.

Table 1
Error rates and number of parameters for the estimated LC models

Model type	Number of latent classes				
	1	2	3	4	2-by-2
LC feed-forward	0.437 (1)[a]	0.215 (13)	0.153 (25)	0.151 (37)	0.153 (25)
LC regression	0.260 (11)[b]	0.213 (23)	0.164 (35)	0.169 (47)	0.169 (35)
Mixture-of-experts	0.260 (11)[b]	0.163 (33)	0.151 (55)	0.151 (77)	0.151 (55)
Standard LC	0.437 (10)	0.278 (22)	0.241 (34)	0.209 (46)	0.219 (34)
FM-augmented NB	0.282 (20)[c]	0.203 (41)	0.171 (62)	0.154 (83)	0.185 (62)
Mixture-of-NB	0.282 (20)[c]	0.197 (42)	0.174 (64)	0.158 (86)	0.156 (64)

[a]Intercept-only model.
[b]Standard logit model.
[c]Naive Bayes model.

4. An application

We applied the various LC models for classification to data of 9949 employees of a large national (American) corporation who where asked about their job satisfaction (see Table 5.10 in Agresti, 1990). The outcome variable (job satisfaction) has two levels: satisfied and not satisfied. The predictors are race, gender, age (three age groups) and regional location (seven regions). The data set was randomly split into a training and a validation sample, consisting of 5007 and 4942 cases, respectively.

This example was selected because application of a standard logit model revealed that, besides the main effects, almost all first- and higher-order interactions are needed in order to get a model that fits the data and that classifies as well as possible. We wanted to know how well the various LC models performed in such a situation.

Table 1 reports the misclassification rates and the number of parameters for six types of LC models. For each type, we estimated models with 1–4 components, as well as a model with two dichotomous latent variables. Classification is based on $\max[P(y|\mathbf{z})]$. In models with multiple maxima, we report the error rate of the solution with the largest log-likelihood value.

The intercept-only model yields the upper bound for the error rate, in this case 0.437. The other two models with a single component, the standard logit model and the NB model, have error rates of 0.260, and 0.282, respectively. We are interested in determining whether the LC classifiers improve upon these standard classifiers.

It can be seen that with a small number of latent classes, the classification performance of the supervised methods is better than that of the unsupervised methods. Among the supervised methods, the mixture-of-experts and the LC feed-forward models yield the lowest error rates. Taking into account parsimony and easiness of interpretation, either the 3-class or 2-factor feed-forward model would be our choice for this data set. The 3-class model yields classes with probabilities of 0.114, 0.909, and 0.902 of being satisfied. The logit parameters in the model for x indicate which subgroups are most satisfied, or have a higher probability of belonging to classes 2 or 3.

Table 2
L^2 values and df for the estimated models

Model type	Number of latent classes				
	1	2	3	4	2-by-2
LC feed-foward	3124 (83)[a]	1408 (71)	376 (59)	262 (47)	376 (59)
LC regression	1739 (73)[b]	844 (61)	594 (49)	467 (37)	594 (49)
Mixture-of-experts	1739 (73)[b]	333 (51)	66 (29)	13 (7)	53 (29)
Standard LC	5602 (156)	3969 (144)	2864 (132)	2286 (120)	2811 (132)
FM-augmented NB	4019 (146)[c]	2316 (125)	1425 (104)	834 (83)	1337 (104)
Mixture-of-NB	4019 (146)[c]	2305 (124)	1421 (102)	806 (80)	1264 (102)

[a]Intercept-only model.
[b]Standard logit model.
[c]Naive Bayes model.

The extended NB classifiers perform better than the standard LC model, which is not surprising given the quite restrictive assumptions of the latter. This is, however, at the cost of a much larger number of parameters in the former. The error rate of the standard LC model can, however, substantially be decreased by increasing the number of latent classes: a model with 6 classes has an error rate of 0.156.

We first concentrated on the classification performance of the various LC models because that is our main interest. It is, however, also informative to compare their goodness-of-fit. Table 2 reports the value of the likelihood-ratio statistic (L^2) and the number of degrees of freedom (df) for our six types of LC models.

The goodness-of-fit information shows that the supervised methods fit the data much better than the unsupervised methods. Actually, with the more restricted unsupervised structures we should considerably increase the number of latent classes to obtain a reasonable fit. It can also be seen that the standard logit model fits badly, which shows that there are important higher-order interactions. The various supervised methods capture these higher-order interactions quite well. If we would base model selection on strict goodness-of-fit tests, the 4-class mixture-of-experts model should be the final model because it is the only one that passes the test at a 5% significance level. In our point of view, however, the more parsimonious models that classify equally well should be preferred.

5. Discussion

We described two basic types of LC models for classification. Advantages of the unsupervised methods are that their estimation is much faster, that they are less prone to local maxima, and that they can easily deal with missing data in the predictor variables. The most important advantage of the supervised methods is their better classification performance.

Among the unsupervised methods, the standard LC model (including factor variant) yields results that are most easy to interpret. Classification may, however, be poor with

a small number of components. In such cases, FAN and mixture-of-NB models yield much lower misclassification rates.

Among the supervised methods, the new LC feed-forward structure seems to be most attractive. Its computation time is lower and interpretation easier than of the mixture-of-experts model, and prediction is usually better than with the LC regression model. Like feed-forward neural networks, it is able to describe complicated interaction effects. Unlike feed-forward neural networks, the parameters of LC feed-forward models are easy to interpret.

References

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarence Press, Oxford.

Clogg, C.C., Goodman, L.A., 1984. Latent structure analysis of a set of multi-dimensional contingency tables. J. Amer. Statist. Assoc. 79, 762–771.

Dayton, C.M., Macready, G.B., 1988. Concomitant-variable latent-class models. J. Amer. Statist. Assoc. 83, 173–178.

Hagenaars, J.A., 1990. Categorical Longitudinal Data—Loglinear Analysis of Panel, Trend and Cohort Data. Sage, Newbury Park.

Jacobs, R.A., Jordan, M.I, Nowlan, S.J., Hinton, G.E., 1991. Adaptive mixtures of local experts. Neural Comput. 3, 79–87.

Kontkanen, P., Myllymäki, P., Tirri, H., 1996. Predictive data mining with finite mixtures. In: Simoudis, E., Han, J., Fayyad, U. (Eds.), Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, pp. 176–182.

Magidson, J., Vermunt, J.K., 2001. Latent class factor and cluster models, biplots and related graphical displays. Sociol. Methodol. 31, 223–264.

Meilã, M., Jordan, M.I., 2000. Learning with mixtures of trees. J. Mach. Learning Res. 1, 1–48.

Monti, S., Cooper, G.F., 1999. A Bayesian network classifier that combines a finite mixture model and a naïve Bayes model. In: Blackmond Laskey, K., Prade, H. (Eds.), Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, pp. 447–456.

Siciliano, R., Mooijaart, A., 2000. Unconditional latent budget analysis: a neural network approach. In: Borra, S., Rocci, R., Vichi, M., Schader, M. (Eds.), Advances in Classification and Data Analysis. Springer, Berlin, pp. 127–136.

Vermunt, J.K., 1997. Log-linear Models for Event Histories. Sage Publications, Thousand Oakes.

Wedel, M., DeSarbo, W.S., 1994. A review of recent developments in latent class regression models. In: Bagozzi, R.P. (Ed.), Advanced Methods of Marketing Research. Blackwell Publishers, Cambridge, pp. 352–388.