Model-based multivariate discretization for logistic regression

Adrien Ehrhardt[1,4], Christophe Biernacki[1,2], Vincent Vandewalle[1,3], Philippe Heinrich[2]

[1]Inria Lille Nord-Europe, Modal team ; [2]Université de Lille 1, Laboratoire Paul Painlevé ; [3]Université de Lille 2, EA 2694 ; [4]Crédit Agricole Consumer Finance

Motivation

Credit Scoring: $\underbrace{\text{estimating the probability of an applicant to a loan to default.}}_{\textbf{logistic regression of parameter } \theta}$

Modelers traditionally **manually** perform two **pre-processing tasks**:

- **Discretization** of continuous attributes,
- **Grouping** of values of qualitative attributes.

BUT WHY?

- Resulting model **more understandable**, allows to address subgroups,
- **Increased predictive power**.

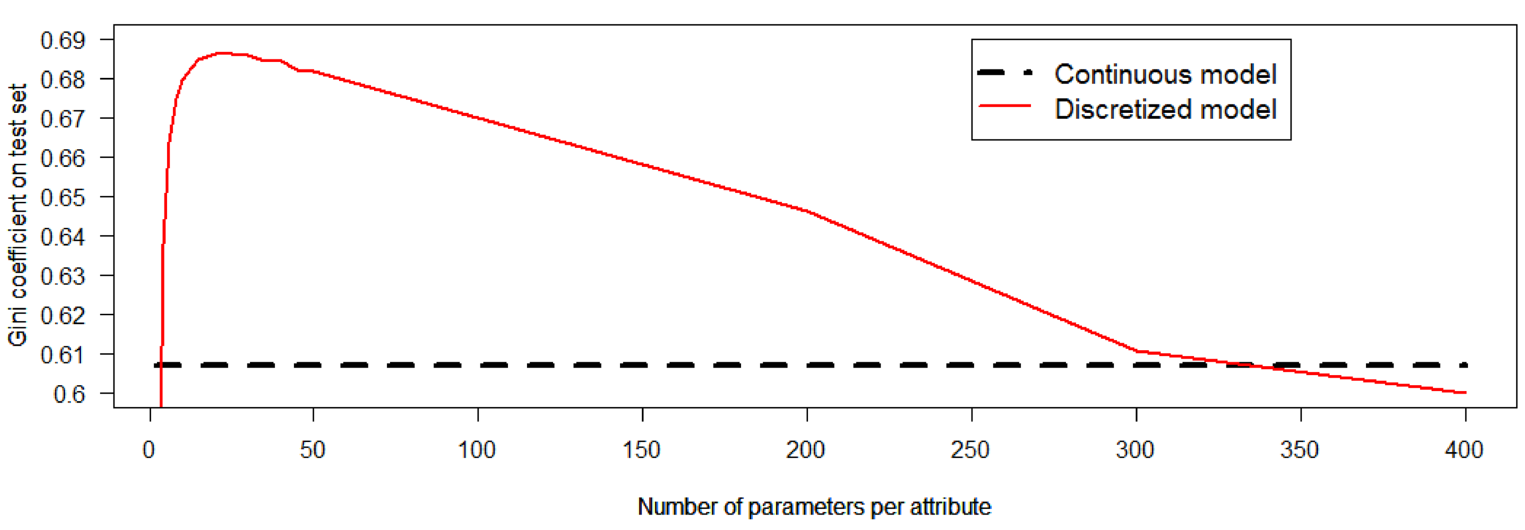

Figure 1: *equal-freq* discretization (same number of observations in each bin) with varying number of bins

Notations

Target variable: Y in $\{0;1\}$ (good/bad clients).

Predictive attributes: $X = (X^j)_1^d$ where X^j is continuous or qualitative.

Discretized attributes: $E = (E^j)_1^d$ where $E \in \mathcal{E}$ and $E^j \in \{1, \ldots, \underbrace{m_j}_{\text{unknown}}\}$.

Continuous: $E^j = 1$ $E^j = 2$ $E^j = 3$ X^j

Some intuition

Hint: The set $\mathcal{E}$ of all possible discretizations is huge!

Implicit discretization hypothesis: E "squeezes" the info in X about Y:

$$\boxed{\forall\, y, x, e,\; p(y|x,e) = p(y|e).}$$

Using this hypothesis we have:

$$\boxed{p(y|x) = \sum_{e \in \mathcal{E}} p(y|x,e)p(e|x) = \sum_{e \in \mathcal{E}} \underbrace{p(y|e)}_{\text{logistic}} \underbrace{p(e|x)}_{\text{to be defined}}.}$$

As E is unknown, this problem is **too hard** for an EM-algorithm.

All discretization methods add hypotheses to simplify the problem.

Optimized criterion

According to Figure 1 there is an optimal discretization; so we seek:

$$\boxed{\boldsymbol{e}^\star = \underset{e \in \mathcal{E}}{\operatorname{argmax}} \operatorname{AIC}(\mathrm{m}_e),}$$

Lots of candidates $\boldsymbol{e}$: it is untractable to optimize this criterion on $\mathcal{E}$.

Idea: Generate "good" candidates and choose $\boldsymbol{e}^\star$ among few candidates.

Statistical modeling

Hypothesis 1: conditionally to X, each r.v. E^j is assumed independent:

$$\boxed{\forall\, j \neq k,\; E^j|x^j \perp E^k|x^k.}$$

Hypothesis 2: each E^j is linked to X^j via multinomial logistic regression:

$$\boxed{\forall\, j, e, x,\; p(e^j|x^j) = p(e^j|x^j;\alpha_j).}$$

$\mathcal{E}$ is thus "reduced" to the multinomial logit family.

Problem: $(\alpha_j)_1^j$ cannot be estimated as $(E^j)_1^d$ are **latent variables**.

SEM-Gibbs estimation

Idea: use an **SEM-algorithm** as $p(y,e|x) = p(y|e)\prod_{j=1}^d p(e^j|x^j)$.

Trick: Gibbs-sampling from a multinomial model with parameters:

$$\boxed{p(e^j|x,y,e^{\{-j\}}) \propto p(y|e;\theta)p(e^j|x^j;\alpha_j)}$$

1. Initialize e^j randomly in $\{1,\ldots,m_j^{(0)}\}$ ($m_j^{(0)}$: user-def. max. number of intervals).
2. Repeat until $i \leq$ max_iter (user-defined) and $\exists\, j$ s.t. $m_j^{(i)} > 1$:
   1. Adjust logistic regression $p(y|e;\theta) = \text{logit}^{-1}(\theta_0 + \sum_{j=1}^d \sum_{m=1}^{m_j} \theta_m^j 1_{\{e^j=m\}})$.
   2. For all continuous attributes j, adjust multinomial logistic regressions $p(e^j|x^j;\alpha_j)$.
   3. For all qualitative attributes j, calculate $p(e^j|x^j;\alpha_j)$ through the contingency table.
   4. Use the expression above to draw $e^{(i)}$.
   5. Calculate the new candidate discretization $e_{\text{MAP}}^{(i)} = (\text{argmax}_k\, p(E^j = k|x^j;\alpha_j)))_1^d$.

Estimation performance on simulated data

More than 200 existing algorithms [1], among which *ChiMerge* [2] and *MDLP* [3].

1. Estimation precision of cut-off values knowing m_j, Table 1a.
2. Estimation precision of m_j, Table 1b.
3. Performance in presence of (hidden) interaction attributes, Table 1c.

$n=800$	$S_1 = \frac{1}{3}$	$S_2 = \frac{2}{3}$
E^1	[0.331 ; 0.335]	[0.669 ; 0.671]
E^2	[0.332 ; 0.362]	[0.662 ; 0.667]

(a) 95% CI of estimated cut-off (m_j known)

$n=1000$	Mode
$m_1 = 3$	4
$m_2 = 3$	4

(b) Mode of estimated m_j

$n=1000$	Our approach	ChiMerge	MDLP
Gini	[80 ; 81.2]	[48.5 ; 51.6]	[76.2 ; 77.9]

(c) 95% CI on test set Gini (all models misspecified)

Table 1: Different performance estimations using simulated data

Predictive performance on real data

3 portfolios: 3 different populations, products, …3 different scorecards!

Total time spent on developing a scorecard: approx. 6 months, among which approx. 3 on attribute selection, discretization, grouping and modeling.

	Portfolio 1	Portfolio 2	Portfolio 3
Current performance	57,5	27	70
Our approach	58	30	71.3
ChiMerge	16,5	26,7	0 ($\lvert\theta\rvert = 2000$)
MDLP	58	29,2	71.3

Table 2: Gini on test set of different discretized models on 3 portfolios

Conclusion

1. Our approach is a **generic way** to discretize,
2. It shows **good performance** in the simulated misspecified model case,
3. It shows **comparable** results on real data, but it is faster and automatic,
4. Perspectives:
   - Automatic creation of **interaction terms**,
   - Extension to **other model types** $p(e^j|x^j;\alpha_j)$.
5. **Implementation available** in R, Python and soon in PySpark!

References

[1] Sergio Ramírez-Gallego, Salvador García, Héctor Mouriño-Talín, David Martínez-Rego, Verónica Bolón-Canedo, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera.
Data discretization: taxonomy and big data challenge.
Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6(1):5–21, 2016.

[2] Randy Kerber.
Chimerge: Discretization of numeric attributes.
In *Proceedings of the tenth national conference on Artificial intelligence*, pages 123–128. Aaai Press, 1992.

[3] Usama M. Fayyad and Keki B. Irani.
Multi-interval discretization of continuous-valued attributes for classification learning.
In *IJCAI*, pages 1022–1029, 1993.

Try it out!

- R: github.com/adimajo/scoring
- Python: ~/poster_discretization

Contact Information

- Web: adimajo.github.io
- Email: adrien.ehrhardt@inria.fr

ID049

CRÉDIT AGRICOLE CONSUMER FINANCE

Inria INVENTEURS DU MONDE NUMÉRIQUE