Development of Classification Models from Imbalanced Datasets Using Mathematical Programming

John Glen

Konstantinos Falangis

University of Edinburgh Business School

Outline

- Background
- Methods for dealing with imbalanced datasets
- Mathematical programming (MP) discriminant analysis models:
  - Minimisation of sum of deviations (MSD) model
- Extending the MSD model for imbalanced datasets
- Experimental studies
- Conclusions

Imbalanced Datasets

- Ideally, classification models should be developed from datasets with approximately same number of observations in each class.
- In practice, datasets imbalanced, e.g.
  - Credit scoring − typically less than 10% bad
  - Fraud scoring − typically less than 1% fraudulent
- Classifiers developed from imbalanced datasets can be of limited value, e.g.
  - If 99% of observations in majority class, assigning all observations to this class has 99% accuracy

Dealing with Imbalanced Datasets

- Pre-process data to produce balanced dataset:
  - Over-sample the minority class
  - Under-sample the majority class
  - Generate synthetic minority class observations between adjacent minority class observations
- Focus on observations close to class boundary
  - Iterative approach (e.g. Vinciotti and Hand, 2003)
- Incorporate misclassification costs:
  - Directly in MP models for minimising misclassifications and machine intelligence methods
  - Cost based threshold in statistical methods, (e.g. Hand and Vinciotti, 2003)

MP Discriminant Analysis Models

Methods for assessing group separation:

- In terms of deviations of misclassified observations from discriminant function
  - minimisation of sum of deviations (MSD)
- Consider external and internal deviations
  - goal programming model
- By maximisation of classification accuracy (or minimisation of misclassifications)
  - requires a binary variable for each observation
  - iterative approach for larger problems (Glen, 2003)

MP Discriminant Analysis Models

- Only the two-group problem considered
- Training sample, G, contains m observations known to belong to group 1 (G_1) or group 2 (G_2)
- Each observation a set of values of n features
- X_{ij} – value of feature j, j=1,2,....,n, in observation i, i=1,2,....,m
- Group 1 observations below discriminant function
- Group 2 observations above discriminant function
- Determine discriminant function in terms of coefficient, a_j, j=1,2,....,n, of feature j, and constant term, a_0, using an appropriate objective

MSD Model

Let d_i, $d_i \geq 0$, be deviation of observation i from function, where $d_i > 0$ if observation misclassified.

$$\text{Minimise} \quad \sum_i d_i$$

$$\text{subject to} \quad \sum_j X_{ij} a_j - a_0 - d_i \leq 0 \qquad i \in G_1$$

$$\sum_j X_{ij} a_j - a_0 + d_i \geq 0 \qquad i \in G_2$$

$$a_j \text{ free}, \; j = 0,1,....,n; \; d_i \geq 0, \; i = 1,2,....,m$$

MP Discriminant Analysis Models: Normalisation

- Standard method – set a_0 to a constant.
- Limitations of standard normalisation:
  - unacceptable solutions, i.e. coefficients zero
  - model should be solved twice, with positive and negative normalisation constants
  - functions with zero constant term not permitted
  - coefficients, a_j, j=1,2,…,n, not invariant under origin shift in problem data.

Integer Programming Normalisation

Represent free variables a_j, $j=1,2,....,n$, by non-negative variables a_j^+ and a_j^-, where $a_j = a_j^+ - a_j^-$

Define binary variables δ_j and γ_j such that $\delta_j = 1 \Leftrightarrow a_j^+ \geq \varepsilon$ and $\gamma_j = 1 \Leftrightarrow a_j^- \geq \varepsilon$, where $\varepsilon, \varepsilon > 0$, is small and add constraints:

$$a_j^+ - \varepsilon\delta_j \geq 0 \qquad j=1,2,....,n$$
$$a_j^+ - \delta_j \leq 0 \qquad j=1,2,....,n$$
$$a_j^- - \varepsilon\gamma_j \geq 0 \qquad j=1,2,....,n$$
$$a_j^- - \gamma_j \leq 0 \qquad j=1,2,....,n$$
$$\delta_j + \gamma_j \leq 1 \qquad j=1,2,....,n$$
$$\sum_j (a_j^+ + a_j^-) = 1$$

In practice, more efficient to represent each variable pair a_j^+ and a_j^- as a special ordered set of type 1 (SOS1).

Maximisation of Classification Accuracy − MCA Model

Define binary variables β_i, where β_i=1 if observation i is classified correctly

$$\text{Maximise} \quad \sum_i \beta_i$$

$$\text{subject to} \quad \sum_j X_{ij} a_j - a_0 + U\beta_i \le U \qquad i \in G_1$$

$$\sum_j X_{ij} a_j - a_0 - U\beta_i \ge -U \qquad i \in G_2$$

$$a_j \text{ free}, \; j = 0,1,....,n; \; \beta_i = 0,1, \; i = 1,2,....,m$$

where U, U>0, is large.

Extending the MSD Model: 1

- Modify the objective function to achieve balance across both groups (Glover, 1990)
- Assuming m_1 observations in G_1 and m_2 in G_2, objective function becomes:

$$\text{Minimise} \quad m_2 \sum_{i \in G_1} d_i + m_1 \sum_{i \in G_2} d_i$$

Extending the MSD Model: 2

- Impose constraints on proportion of observations misclassified in each group (Koehler, 1990)
- In MIP model for minimising misclassifications,
  - assume m_k, k=1,2, observations in G_k and
  - let z_k, k=1,2, be number of misclassified observations in G_k

  then for γ>0 and small, add constraint:

$$-m_1 m_2 \gamma \leq m_2 z_1 - m_1 z_2 \leq m_1 m_2 \gamma$$

Extending the MSD Model: 2

- For MSD model, impose constraints on mean deviation in each group
- In MSD model,
  - assume m_k, k=1,2, observations in G_k and
  - let d_i, i=1,2,….,m, be deviation of observation i

  then for δ>0 and small, add constraint:

$$-\delta \le \frac{1}{m_1}\sum_{i\in G_1} d_i - \frac{1}{m_2}\sum_{i\in G_2} d_i \le \delta$$

Extending the MSD Model: 3

- Glover and Better (2007) suggested that violations should be balanced in each group by constraint:

$$m_1 \sum_{i \in G_2} d_i = m_2 \sum_{i \in G_1} d_i$$

- In practice the main difficulty with imbalanced datasets is that observations tend to be assigned to the majority class. Hence constrain so that mean deviation in minority class (G_1) does not exceed mean deviation in majority class (G_2):

$$\frac{1}{m_1} \sum_{i \in G_1} d_i \leq \frac{1}{m_2} \sum_{i \in G_2} d_i$$

Test Datasets

- Dataset 1:
  13516 observations, 184 (1.4%) bad
  - 12 variables transformed to 32 features by WOE
- Dataset 2:
  15050 observations, 218 (1.4%) bad
  - 8 variables transformed to 18 features by WOE
- Dataset 3:
  29389 observations, 1006 (3.4%) bad
  - 11 variables transformed to 25 features by WOE
- Dataset 4:
  10375 observations, 375 (1.4%) bad
  - 21 variables transformed to 37 features by WOE

Method	Accuracy (%)		
	Total	Goods	Bads
Logistic Regression	99	99	1
MSD – Basic Model	99	100	0
MSD – Balancing Objective	75	76	61
MSD – Range Constraints: δ=0.0009	79	80	57
MSD – Range Constraints: δ=0.0001	76	76	60
MSD – Balancing Constraint	75	76	61

Table 1: Holdout Sample Performance – Dataset 1

Method	Accuracy (%)		
	Total	Goods	Bads
Logistic Regression	98	99	1
MSD – Basic Model	99	99	3
MSD – Balancing Objective	74	74	57
MSD – Range Constraints: δ=0.0009	71	71	60
MSD – Range Constraints: δ=0.0001	69	69	63
MSD – Balancing Constraint	69	70	63

Table 2: Holdout Sample Performance – Dataset 2

Method	Accuracy (%)		
	Total	Goods	Bads
Logistic Regression	96	100	0
MSD – Basic Model	98	98	3
MSD – Balancing Objective	70	70	68
MSD – Range Constraints: δ=0.0009	74	74	66
MSD – Range Constraints: δ=0.0001	70	70	68
MSD – Balancing Constraint	70	71	68

Table 3: Holdout Sample Performance – Dataset 3

Method	Accuracy (%)		
	Total	Goods	Bads
Logistic Regression	96	100	0
MSD – Basic Model	99	99	7
MSD – Balancing Objective	85	85	76
MSD – Range Constraints: δ=0.0009	88	89	70
MSD – Range Constraints: δ=0.0001	85	85	75
MSD – Balancing Constraint	85	85	76

Table 4: Holdout Sample Performance – Dataset 4

Conclusions

- In practice, datasets are often imbalanced.
- Limitations of established methods for dealing with imbalanced datasets:
  - over-sampling minority class may result in overfitting
  - under-sampling majority class may ignore some data areas
  - difficulties in assessing costs in cost based methods
- Other approaches can be used with MP methods:
  - additional constraints can be incorporated
  - objective function can be modified