Multivariate Rank Based Methods For Classification

Biman Chakraborty

School of Mathematics, University of Birmingham,
Edgbaston, Birmingham B15 2TT, United Kingdom

Multi-Disciplinary Research in Data Mining - BTG Away Day
20-21 September, 2010

Univariate Sign Function

$$sign(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ +1 & \text{if } x > 0 \end{cases}$$

$$= \begin{cases} \frac{x}{|x|} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

$$= \begin{cases} \frac{d}{dx}|x| & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Univariate Ranks

- Centered Rank:

$$Rank(x) = \frac{1}{n}\sum_{i=1}^{n} sign(x - X_i).$$

- Note that: $E[Rank(x)] = 2F(x) - 1$.
- $-1 \leq Rank(x) \leq 1$
- $Rank(x) = 0 \Longrightarrow x$ is sample median.

Multivariate Signs:

Consider $\boldsymbol{x} \in \mathbb{R}^d$

$$
\begin{aligned}
Sign(\boldsymbol{x}) &= \begin{cases} \frac{\partial}{\partial \boldsymbol{x}} \|\boldsymbol{x}\| & \text{if } \boldsymbol{x} \neq \boldsymbol{0} \\ \boldsymbol{0} & \text{if } \boldsymbol{x} = \boldsymbol{0} \end{cases} \\
&= \begin{cases} \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} & \text{if } \boldsymbol{x} \neq \boldsymbol{0} \\ \boldsymbol{0} & \text{if } \boldsymbol{x} = \boldsymbol{0} \end{cases}
\end{aligned}
$$

where $\|\boldsymbol{x}\| = \{|x_1|^2 + \cdots + |x_d|^2\}^{1/2}$.

Multivariate Ranks:

- Data: $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n \in \mathbb{R}^d$.
- $$Rank(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} Sign(\boldsymbol{x} - \boldsymbol{X}_i).$$
- $\|Rank(\boldsymbol{x})\| \leq 1$
- $Rank(\boldsymbol{x}) = \boldsymbol{0} \Longrightarrow \boldsymbol{x}$ is the spatial median.
- Population Version:

$$R_F(\boldsymbol{x}) = E_F[Sign(\boldsymbol{x} - \boldsymbol{X})], \quad \boldsymbol{X} \sim F.$$

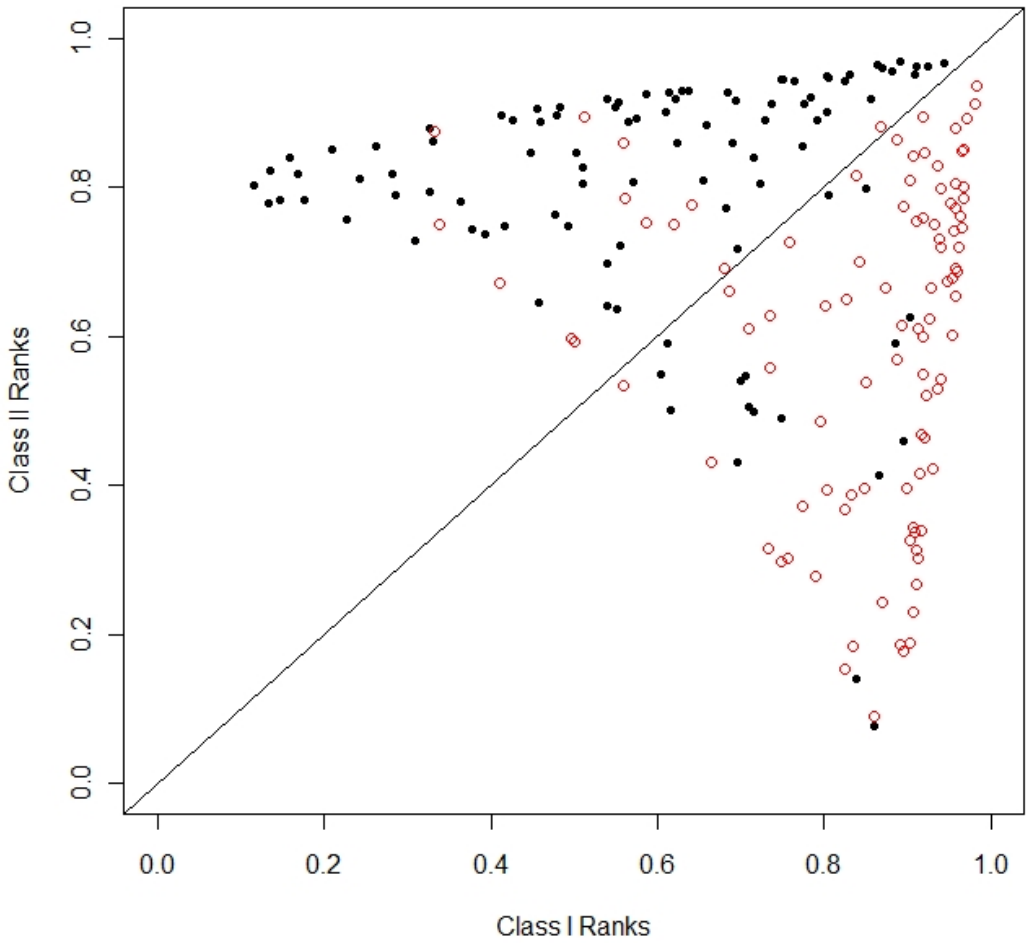
Relation With Quantiles:

- Univariate Case:
  - $Rank(x) = u, \quad -1 \leq u \leq 1$ then x is the $(u+1)/2$-th quantile.
- Multivariate Case: (Chakraborty, 2001)
  - $Rank(\boldsymbol{x}) = \boldsymbol{u}, \quad \|\boldsymbol{u}\| < 1$ then $\boldsymbol{x}$ is the $\boldsymbol{u}$-th geometric quantile.
- A Related Depth Function:
  - $D(\boldsymbol{x}) = 1 - \|Rank(\boldsymbol{x})\|$.
  - Depth of the median $= 1$.

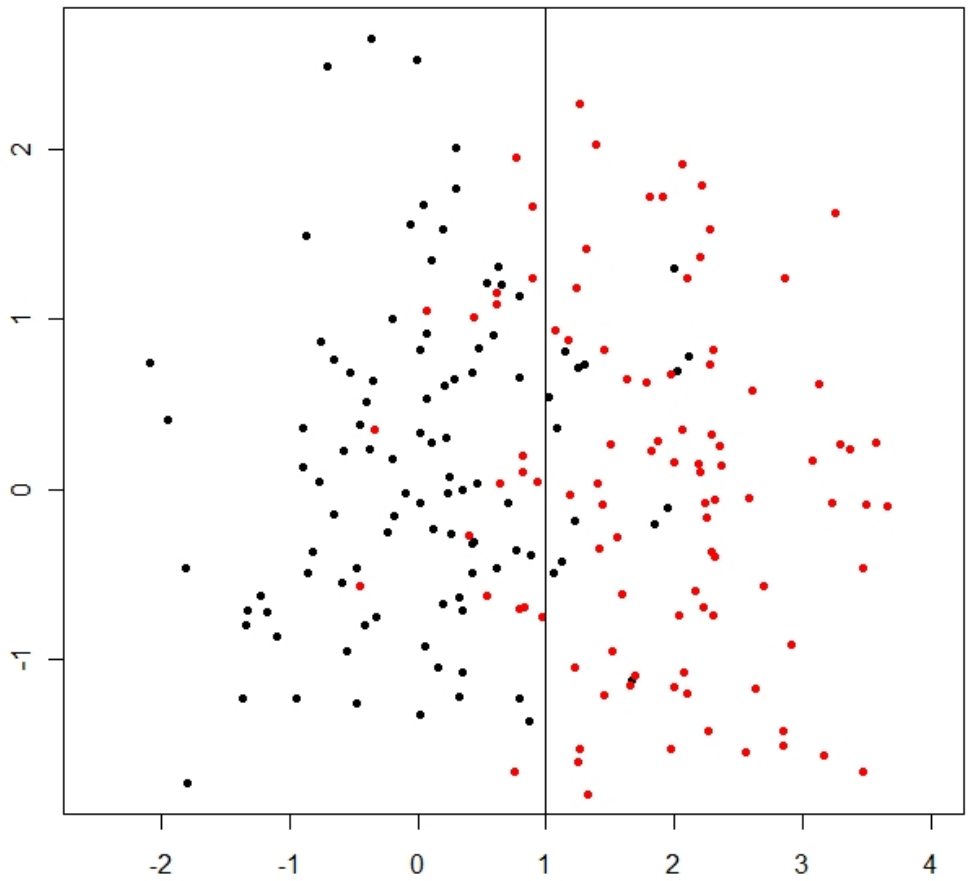
A Simple Classification Rule

- Two-class Problem:
  - Let C_1 and C_2 be the two classes of population with the d-dimensional feature vector $\boldsymbol{X}$ having distributions F and G, respectively.
  - Rule: Classify $\boldsymbol{x} \in \mathbb{R}^d$ to C_1 if $\|R_F(\boldsymbol{x})\| < \|R_G(\boldsymbol{x})\|$ and classify $\boldsymbol{x}$ to C_2 otherwise.
- K-class Problem:
  - Let the classes $C_1, \ldots, C_K$ have distributions $F_1, \ldots, F_K$.
  - Rule: Classify $\boldsymbol{x}$ to C_i if $\|R_{F_i}(\boldsymbol{x})\| = \min_{1 \le j \le K} \|R_{F_j}(\boldsymbol{x})\|$.

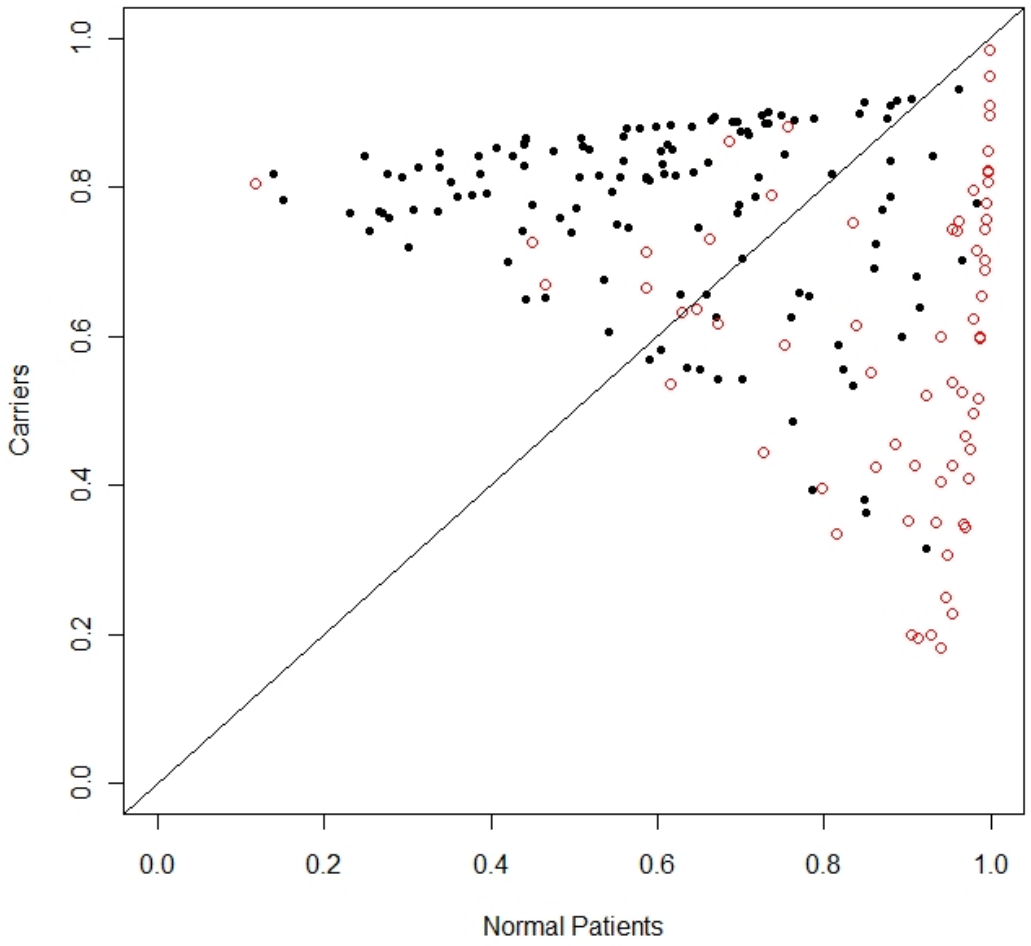
A Simulated Example: Rank-Rank Plot



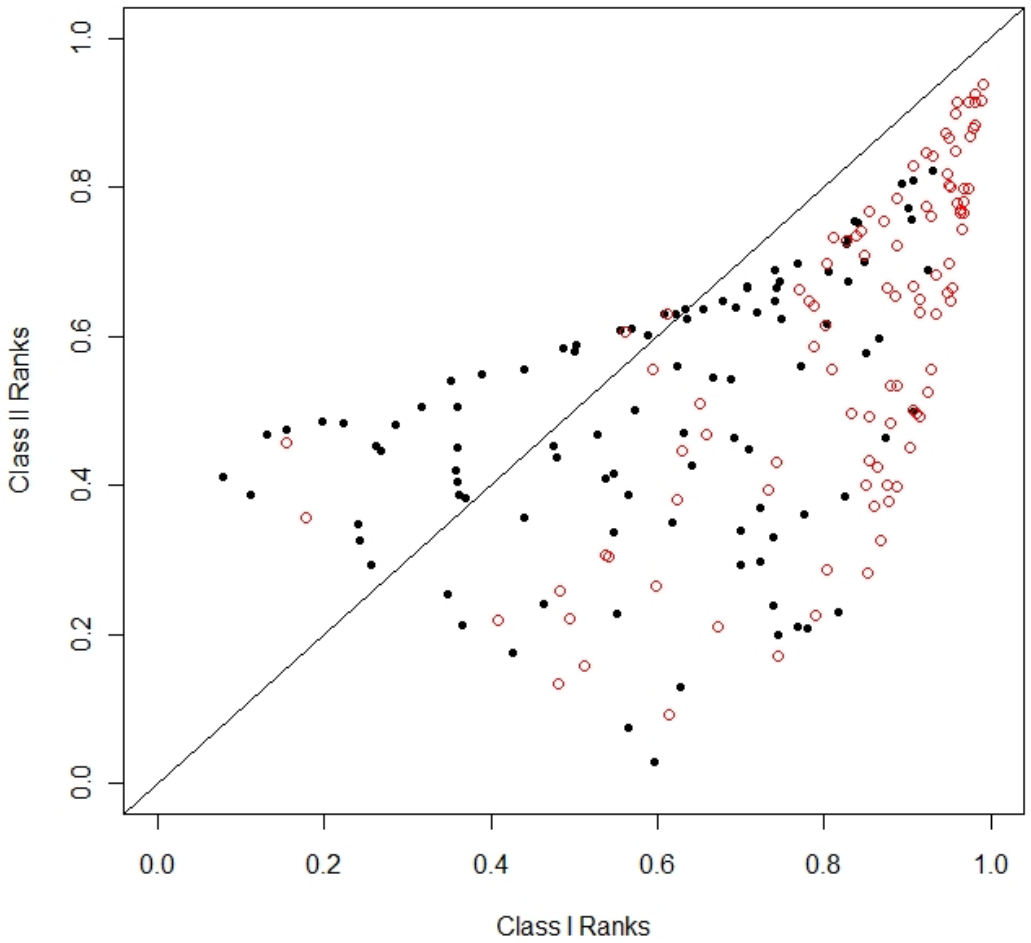
The Example: In The Feature Space



Blood Measurements Data



A Difference in Scale Example



Comments

- The previous rule is Bayes optimal for elliptically symmetric distributions with a location shift.
- It does not perform well for difference in scales.
- We may consider rules like: Classify to C_1 if $\|R_F(\boldsymbol{x})\| < a\|R_G(\boldsymbol{x})\| + c$ for some optimal choice of a and c.
- In general, one can consider: Classify to C_1 if $\|R_F(\boldsymbol{x})\| < h(\|R_G(\boldsymbol{x})\|)$ for some function h.

Comments

- Similar ideas can be implemented for clustering.
- We can also consider some local version of multivariate ranks.
- Similar ideas can be explored with other notions of data depths.
- Problem: Really large and high dimensional dataset.
- Problem: Feature vector containing discrete variables.