Risk-Based Comparison of Classification Systems

2d Lt Seth Wagenman

Air Force Institute of Technology
Department of Mathematics and Statistics

Thursday, 13 March 2007
Master's Thesis Presentation

Hypothetical Situation

2d Lt Seth Wagenman

A broker classifies stocks as “Buy”, “Hold”, or “Sell”:

- Many small factors have begun to adversely affect stocks in a particular client’s portfolio
- Companies whose stocks may fail are unaware
- Broker applies classification algorithm to gathered data, classifying 85% of portfolio companies as “in severe danger of failing” (usually < 10 %)
- Broker doesn’t know if system is “best available”
- Instead of selling stocks immediately, she questions accuracy of data inputs and calls algorithm vendor to find out if something is wrong with “the system” she is using to classify stocks
- Brokerage loses time, portfolio value deteriorates

Outline

1. Background
   - Background of ROC Analysis in Binary Classification
   - Definitions and Assumptions
   - Review of ROC Analysis Methods
2. ROC Risk Functional (RRF)
   - Development & Definition
   - Completely Unknown Class Prevalences
   - Partially Known Class Prevalences
3. Example Application
   - Uniformly Distributed Class Prevalences
   - Beta-Distributed Class Prevalences
   - Discussion of Application Results

Confusion Matrices

Binary (Two-Class) Classification System

- Results of two-class (i.e., Class 1 *Positive*, Class 2 *Negative*) experiment or observational study applied to a set of validation data:

Contingency Matrix

	Actual Class: 1	Actual Class: 2
Labeled Class: 1	TP	FP
Labeled Class: 2	FN	TN

- Calculate by-class decision frequencies using numbers of objects in Classes 1 and 2:

Confusion Matrix

	Actual Class: 1	Actual Class: 2
Labeled Class: 1	$\frac{TP}{M_1}$	$\frac{FN}{M_2}$
Labeled Class: 2	$\frac{FN}{M_1}$	$\frac{TN}{M_2}$

Transpose Stochastic Confusion Matrix

- Given the confusion matrix shown previously:

Confusion Matrix	Actual Class: 1	Actual Class: 2
Labeled Class: 1	$\frac{TP}{M_1}$	$\frac{FP}{M_2}$
Labeled Class: 2	$\frac{FN}{M_1}$	$\frac{TN}{M_2}$

- One entry per column represents all *information*:
- $TP + FN = N_1 \implies \frac{TP}{M_1} + \frac{FN}{M_1} = 1$
- $FP + TN = M_2 \implies \frac{FP}{M_2} + \frac{TN}{M_2} = 1$

Transpose Stochastic Confusion Matrix

- Given the confusion matrix shown previously:

Confusion Matrix	Actual Class: 1	Actual Class: 2
Labeled Class: 1	$\frac{TP}{M_1}$	$\frac{FP}{M_2}$
Labeled Class: 2	$\frac{FN}{M_1}$	$\frac{TN}{M_2}$

- One entry per column represents all *information*:
- $TP + FN = N_1 \implies \frac{TP}{M_1} + \frac{FN}{M_1} = 1$
- $FP + TN = M_2 \implies \frac{FP}{M_2} + \frac{TN}{M_2} = 1$

Transpose Stochastic Confusion Matrix

- Given the confusion matrix shown previously:

Confusion Matrix	Actual Class: 1	Actual Class: 2
Labeled Class: 1	$\frac{TP}{M_1}$	$\frac{FP}{M_2}$
Labeled Class: 2	$\frac{FN}{M_1}$	$\frac{TN}{M_2}$

- One entry per column represents all *information*:
- $TP + FN = N_1 \implies \frac{TP}{M_1} + \frac{FN}{M_1} = 1$
- $FP + TN = M_2 \implies \frac{FP}{M_2} + \frac{TN}{M_2} = 1$

Transpose Stochastic Confusion Matrix

- Given the confusion matrix shown previously:

Confusion Matrix	Actual Class: 1	Actual Class: 2
Labeled Class: 1	$\frac{TP}{M_1}$	$\frac{FP}{M_2}$
Labeled Class: 2	$\frac{FN}{M_1}$	$\frac{TN}{M_2}$

- One entry per column represents all *information*:
- $TP + FN = N_1 \implies \frac{TP}{M_1} + \frac{FN}{M_1} = 1$
- $FP + TN = M_2 \implies \frac{FP}{M_2} + \frac{TN}{M_2} = 1$

Transpose Stochastic Confusion Matrix

- Given the confusion matrix shown previously:

Confusion Matrix	Actual Class: 1	Actual Class: 2
Labeled Class: 1	$\frac{TP}{M_1}$	$\frac{FP}{M_2}$
Labeled Class: 2	$\frac{FN}{M_1}$	$\frac{TN}{M_2}$

- One entry per column represents all *information*:
- $TP + FN = N_1 \implies \frac{TP}{M_1} + \frac{FN}{M_1} = 1$
- $FP + TN = M_2 \implies \frac{FP}{M_2} + \frac{TN}{M_2} = 1$
- Receiver Operating Characteristic (ROC) is the *information* estimated by a confusion matrix

Mathematical Framework

Multi-Class Classification System, Probability Space

- A set E of possible experimental outcomes, or elementary events $e \in \mathrm{E}$, a sample space
- A finite set $\mathrm{L} = \{\ell_1, \ell_2, \ell_3, \ldots, \ell_\mathrm{n}\}$ of distinct labels
- Intermediate mappings between E and L:
  - A sensor $s : \mathrm{E} \longrightarrow \mathrm{D}$ with data set D as range
  - A processor $p : \mathrm{D}' \longrightarrow \mathrm{F}$ mapping from D to a set F of features (usually real variables)
  - A classifier $c_\theta : \mathrm{F} \longrightarrow \mathrm{L}$ with threshold parameter $\theta \in \Theta$, a *threshold set* of possible parameter settings
- The composition $A_\theta \equiv c_\theta \circ p \circ s : \mathrm{E} \longrightarrow \mathrm{L}$ is the classification system with threshold parameter θ:

$$\mathrm{E} \xrightarrow{A_\theta} \mathrm{L} \;\equiv\; \mathrm{E} \xrightarrow{s} \mathrm{D} \xrightarrow{p} \mathrm{F} \xrightarrow{c_\theta} \mathrm{L}$$

- A probability space $(\mathrm{E}, \mathscr{E}, P)$ defined on σ-field $\mathscr{E}$

Events of Interest

Event Set Class Partition, Classification System Pre-Image

- Label set $\mathrm{L} = \{\ell_1, \ell_2, \ell_3, \ldots, \ell_n\}$ induces a *partition* $\mathrm{E} = \bigcup_{j=1}^{n} \mathcal{E}_n$ of *disjoint* classes of event set E
- Property of class prevalences due to partition:

$$\sum_{j=1}^{n} P(\mathcal{E}_j) \equiv \sum_{j=1}^{n} p_j = 1$$

- Pre-image set function $A_\theta^\natural : \mathscr{L} \longrightarrow \mathscr{E}$ given by:

$$A_\theta^\natural\left(\{\ell_i\}\right) \equiv \{e \in \mathrm{E} : A_\theta(e) = \ell_i\} \subset \mathrm{E}$$

where $\mathscr{E}$, $\mathscr{L}$ are σ-fields over E, L, respectively (using $\natural$ instead of $^{-1}$ to avoid confusion)

Class-Conditional Probabilities

- $q_{i|j}(A_\theta) \equiv$ conditional probability that A_θ assigns label ℓ_i to an outcome e, *given that* e belongs to *class* $\mathcal{E}_j$, defined for $i, j = 1, 2, 3, \ldots, n$
- For $\mathcal{E}_j$ with $P(\mathcal{E}_j) = 0$ (i.e., *a priori* probability zero):

$$q_{i|j}(A_\theta) \equiv 0, \quad \forall\ i = 1, 2, 3, \ldots, n$$

- For $\mathcal{E}_j$ with $p_j > 0$ (i.e., positive *class prevalence*):

$$q_{i|j}(A_\theta) \equiv \frac{P\left(A_\theta^\natural(\{\ell_i\}) \bigcap \mathcal{E}_j\right)}{P(\mathcal{E}_j)}, \quad \forall\ i = 1, 2, 3, \ldots, n$$

Important Matrix Notation

- Conditional Probability Matrix $\mathbf{Q}_{A,\theta} = \left[\mathrm{q}_{\mathrm{i}|\mathrm{j}}(A_\theta)\right]_{\mathrm{i\,j}}$ is the $\mathrm{n} \times \mathrm{n}$ matrix of class-conditional probabilities for a classification system A_θ

Important Matrix Notation

- Conditional Probability Matrix $\mathbf{Q}_{A,\theta} = \left[q_{i|j}(A_\theta) \right]_{ij}$ is the $n \times n$ matrix of class-conditional probabilities for a classification system A_θ
- Cost Matrix $\mathbf{C}_A = \left[c_{i|j}(A) \right]_{ij}$ is the $n \times n$ matrix of costs or *losses* of occurrence of the events whose probabilities are $q_{i|j}(A_\theta)$, invariant for all $\theta \in \Theta$

Important Matrix Notation

- Conditional Probability Matrix $\mathbf{Q}_{A,\theta} = \left[\mathrm{q}_{\mathrm{i}|\mathrm{j}}(A_\theta) \right]_{\mathrm{ij}}$ is the $\mathrm{n} \times \mathrm{n}$ matrix of class-conditional probabilities for a classification system A_θ
- Cost Matrix $\mathbf{C}_A = \left[\mathrm{c}_{\mathrm{i}|\mathrm{j}}(A) \right]_{\mathrm{ij}}$ is the $\mathrm{n} \times \mathrm{n}$ matrix of costs or *losses* of occurrence of the events whose probabilities are $\mathrm{q}_{\mathrm{i}|\mathrm{j}}(A_\theta)$, invariant for all $\theta \in \Theta$
- Prevalence Matrix $\mathbf{P}$ is an $\mathrm{n} \times \mathrm{n}$ *stochastic* matrix with each row the *same* ordered n-tuple $\mathbf{p}^T$ consisting of the class prevalences $\{\mathrm{p}_\mathrm{j}\}_{\mathrm{j}=1}^{\mathrm{n}}$:

$$\mathbf{p} \equiv \begin{bmatrix} \mathrm{p}_1 \\ \vdots \\ \mathrm{p}_\mathrm{n} \end{bmatrix} \Longrightarrow \mathbf{P} = \begin{bmatrix} \mathbf{p}^T \\ \vdots \\ \mathbf{p}^T \end{bmatrix} = \begin{bmatrix} \mathrm{p}_1 & \cdots & \mathrm{p}_\mathrm{n} \\ & \vdots & \\ \mathrm{p}_1 & \cdots & \mathrm{p}_\mathrm{n} \end{bmatrix}_{\mathrm{n}\times\mathrm{n}}$$

Special Matrix Operators

- Given two matrices **U** and **V** of size $s \times r$:

$$[\mathbf{U} \odot \mathbf{V}]_{ij} \equiv u_{ij} v_{ij}$$

is the (i, j)-element of the Hadamard Product, and the sum of all these is the Frobenius Dot Product:

$$\langle \mathbf{U}, \mathbf{V} \rangle_F \equiv \sum_{i=1}^{s} \left[\sum_{j=1}^{r} u_{ij} v_{ij} \right]$$

Special Matrix Operators

- Given two matrices **U** and **V** of size $s \times r$:

$$[\mathbf{U} \odot \mathbf{V}]_{ij} \equiv u_{ij} v_{ij}$$

is the (i,j)-element of the Hadamard Product, and the sum of all these is the Frobenius Dot Product:

$$\langle \mathbf{U}, \mathbf{V} \rangle_F \equiv \sum_{i=1}^{s} \left[\sum_{j=1}^{r} u_{ij} v_{ij} \right]$$

- Given a matrix **X** of size $s \times r$: $\int [\cdot]\, \mathbf{dX} \equiv \int \ldots \ldots \int \ldots \int \int [\cdot]\, dx_{11}\, dx_{12} \ldots dx_{1r} \ldots \ldots dx_{sr}$ is integration with respect to the vector **dX** of all differential elements dx_{ij} of variables in **X**

Assumptions

Given a classification system A_θ:

Given a classification system A_θ and $\mathrm{i,j} = 1, \ldots, \mathrm{n}$:

- All *non-zero* class-conditional probabilities $q_{\mathrm{i|j}}(A_\theta)$ are independent of all class prevalences $\mathrm{p_j}$

Assumptions

Given a classification system A_θ:

Given a classification system A_θ and $\mathrm{i}, \mathrm{j} = 1, \ldots, \mathrm{n}$:

- All *non-zero* class-conditional probabilities $q_{\mathrm{i}|\mathrm{j}}(A_\theta)$ are independent of all class prevalences p_j
- All costs $c_{\mathrm{i}|\mathrm{j}}(A)$ are subjectively fixed and independent of all class prior probabilities $P(\mathcal{E}_\mathrm{j})$

Assumptions

Given a classification system A_θ :

Given a classification system A_θ and $\mathrm{i}, \mathrm{j} = 1, \ldots, \mathrm{n}$:

- All *non-zero* class-conditional probabilities $q_{\mathrm{i}|\mathrm{j}}(A_\theta)$ are independent of all class prevalences p_j
- All costs $c_{\mathrm{i}|\mathrm{j}}(A)$ are subjectively fixed and independent of all class prior probabilities $P(\mathcal{E}_\mathrm{j})$
- All class-conditional probabilities $\mathrm{q}_{\mathrm{i}|\mathrm{j}}(A_\theta)$ are independent of the fixed costs $\mathrm{c}_{\mathrm{i}|\mathrm{j}}(A)$

Assumptions

Given a classification system A_θ :

Given a classification system A_θ and $\mathrm{i}, \mathrm{j} = 1, \ldots, \mathrm{n}$:

- All *non-zero* class-conditional probabilities $q_{\mathrm{i|j}}(A_\theta)$ are independent of all class prevalences $\mathrm{p_j}$
- All costs $c_{\mathrm{i|j}}(A)$ are subjectively fixed and independent of all class prior probabilities $P(\mathcal{E}_\mathrm{j})$
- All class-conditional probabilities $\mathrm{q_{i|j}}(A_\theta)$ are independent of the fixed costs $\mathrm{c_{i|j}}(A)$
- $\mathbf{Q}_{A,\theta}$, $\mathbf{C}_A$, and $\mathbf{P}$ are *continuous random matrices* with marginal probability density functions $W_\mathbf{Q}(\mathbf{Q}_{A,\theta})$, $W_\mathbf{C}(\mathbf{C}_A)$, and $W_\mathbf{P}(\mathbf{P})$ satisfying $W_{\mathbf{Q,C,P}}(\mathbf{Q}_{A,\theta}, \mathbf{C}_A, \mathbf{P}) = W_\mathbf{Q}(\mathbf{Q}_{A,\theta})\, W_\mathbf{C}(\mathbf{C}_A)\, W_\mathbf{P}(\mathbf{P})$

Assumptions

Given a classification system A_θ :

Given a classification system A_θ and $\mathrm{i,j} = 1, \ldots, \mathrm{n}$:

- All *non-zero* class-conditional probabilities $q_{\mathrm{i|j}}(A_\theta)$ are independent of all class prevalences $\mathrm{p_j}$
- All costs $c_{\mathrm{i|j}}(A)$ are subjectively fixed and independent of all class prior probabilities $P(\mathcal{E}_\mathrm{j})$
- All class-conditional probabilities $\mathrm{q_{i|j}}(A_\theta)$ are independent of the fixed costs $\mathrm{c_{i|j}}(A)$
- $\mathbf{Q}_{A,\theta}$, $\mathbf{C}_A$, and $\mathbf{P}$ are *continuous random matrices* with marginal probability density functions $W_\mathbf{Q}(\mathbf{Q}_{A,\theta})$, $W_\mathbf{C}(\mathbf{C}_A)$, and $W_\mathbf{P}(\mathbf{P})$ satisfying $W_{\mathbf{Q,C,P}}(\mathbf{Q}_{A,\theta}, \mathbf{C}_A, \mathbf{P}) = W_\mathbf{Q}(\mathbf{Q}_{A,\theta})\, W_\mathbf{C}(\mathbf{C}_A)\, W_\mathbf{P}(\mathbf{P})$
- Any estimate $\widehat{\mathbf{Q}_{A,\theta}}$ of the conditional probability matrix $\mathbf{Q}_{A,\theta}$ is acceptable (i.e., substituting $\widehat{\mathbf{Q}_{A,\theta}} \approx \mathbf{Q}_{A,\theta}$ results in no appreciable error)

Family of Classification Systems, Risk

- Given event set E, label set L, and threshold set Θ, a family of classification systems of form $A_\theta : \mathrm{E} \longrightarrow \mathrm{L}$ is denoted by $\mathbb{A} = \{A_\theta : \theta \in \Theta\}$

Family of Classification Systems, Risk

- Given event set E, label set L, and threshold set Θ, a family of classification systems of form $A_\theta : \mathrm{E} \longrightarrow \mathrm{L}$ is denoted by $\mathbb{A} = \{A_\theta : \theta \in \Theta\}$
- Given conditional probability matrix $\mathbf{Q}_{A,\theta}$, cost matrix $\mathbf{C}_A$, and prevalence matrix $\mathbf{P}$, risk $R_{A,\theta}$ is:

$$
\begin{aligned}
R_{A,\theta} &= \left\langle \mathbf{Q}_{A,\theta}, (\mathbf{C}_A \odot \mathbf{P}) \right\rangle_F \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{q}_{\mathrm{i}|\mathrm{j}}(A_\theta)\, \mathrm{c}_{\mathrm{i}|\mathrm{j}}(A)\, \mathrm{p}_{\mathrm{i}\,\mathrm{j}} \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{q}_{\mathrm{i}|\mathrm{j}}(A_\theta)\, \mathrm{c}_{\mathrm{i}|\mathrm{j}}(A)\, \mathrm{p}_{\mathrm{j}}
\end{aligned}
$$

Outline

Two-Class ROC Analysis

- Recall confusion matrix as acceptable estimate of conditional probability matrix $\mathbf{Q}_{A,\theta}$:

$$\begin{bmatrix} tpr & fpr \\ fnr & tnr \end{bmatrix} \equiv \begin{bmatrix} \frac{TP}{M_1} & \frac{FP}{M_2} \\ \frac{FN}{M_1} & \frac{TN}{M_2} \end{bmatrix} \approx \begin{bmatrix} q_{1|1}(A_\theta) & q_{1|2}(A_\theta) \\ q_{2|1}(A_\theta) & q_{2|2}(A_\theta) \end{bmatrix}$$

Two-Class ROC Analysis

- Recall confusion matrix as acceptable estimate of conditional probability matrix $\mathbf{Q}_{A,\theta}$:

$$\begin{bmatrix} tpr & fpr \\ fnr & tnr \end{bmatrix} \equiv \begin{bmatrix} \frac{TP}{M_1} & \frac{FP}{M_2} \\ \frac{FN}{M_1} & \frac{TN}{M_2} \end{bmatrix}$$

$$\approx \begin{bmatrix} \mathrm{q}_{1|1}(A_\theta) & \mathrm{q}_{1|2}(A_\theta) \\ \mathrm{q}_{2|1}(A_\theta) & \mathrm{q}_{2|2}(A_\theta) \end{bmatrix}$$

- If we vary θ over some range and plot points $\left(\mathrm{q}_{1|2}(A_\theta), \mathrm{q}_{1|1}(A_\theta)\right)$, we represent a ROC curve

Two-Class ROC Analysis

- Recall confusion matrix as acceptable estimate of conditional probability matrix $\mathbf{Q}_{A,\theta}$:

$$\begin{bmatrix} tpr & fpr \\ fnr & tnr \end{bmatrix} \equiv \begin{bmatrix} \frac{TP}{M_1} & \frac{FP}{M_2} \\ \frac{FN}{M_1} & \frac{TN}{M_2} \end{bmatrix} \approx \begin{bmatrix} \mathrm{q}_{1|1}(A_\theta) & \mathrm{q}_{1|2}(A_\theta) \\ \mathrm{q}_{2|1}(A_\theta) & \mathrm{q}_{2|2}(A_\theta) \end{bmatrix}$$

- If we vary θ over some range and plot points $\left(\mathrm{q}_{1|2}(A_\theta), \mathrm{q}_{1|1}(A_\theta)\right)$, we represent a ROC curve
- If threshold set Θ is continuous, a large sample of ROC vectors $\left(\mathrm{q}_{1|2}(A_\theta), \mathrm{q}_{1|1}(A_\theta)\right)$ approaches a smooth curve on the unit square

AUC: Area Under the ROC Curve

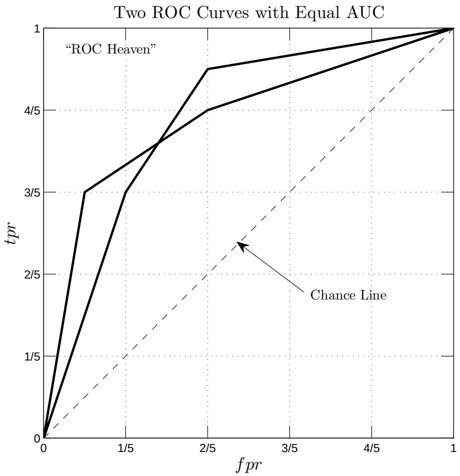

Figure: Two ROC Curves with Equal AUC

AUC for One ROC Point

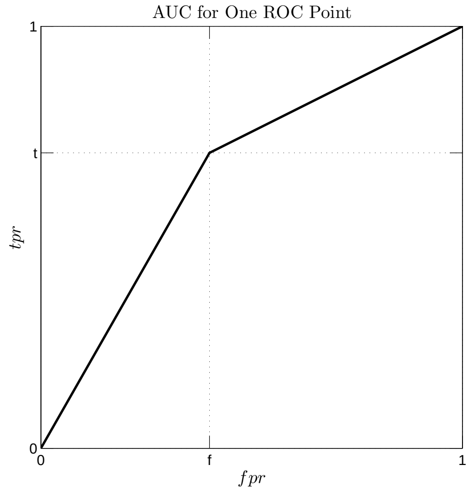

Figure: AUC for a Single ROC Point

Other Binary ROC Analysis Methods

- AUC $= 1 - \frac{tpr + tnr}{2}$, or $1 -$ "skew-insensitive" accuracy under assumptions of equal class prevalences and so-called "zero-one" cost matrix (zeroes along diagonal)

Other Binary ROC Analysis Methods

- AUC $= 1 - \frac{tpr + tnr}{2}$, or $1 -$ "skew-insensitive" accuracy under assumptions of equal class prevalences and so-called "zero-one" cost matrix (zeroes along diagonal)
- Gini Coefficient—related to accuracy

Other Binary ROC Analysis Methods

- AUC $= 1 - \frac{tpr + tnr}{2}$, or $1 -$ "skew-insensitive" accuracy under assumptions of equal class prevalences and so-called "zero-one" cost matrix (zeroes along diagonal)
- Gini Coefficient—related to accuracy
- Matthews Correlation Coefficient–also related to accuracy

Other Binary ROC Analysis Methods

- AUC $= 1 - \frac{tpr + tnr}{2}$, or $1 -$ "skew-insensitive" accuracy under assumptions of equal class prevalences and so-called "zero-one" cost matrix (zeroes along diagonal)
- Gini Coefficient—related to accuracy
- Matthews Correlation Coefficient–also related to accuracy
- Specificity, Sensitivity, Positive/Negative Predictive Values, False Discovery Rate, etc.

Other Binary ROC Analysis Methods

- AUC $= 1 - \frac{tpr + tnr}{2}$, or $1-$ "skew-insensitive" accuracy under assumptions of equal class prevalences and so-called "zero-one" cost matrix (zeroes along diagonal)
- Gini Coefficient—related to accuracy
- Matthews Correlation Coefficient–also related to accuracy
- Specificity, Sensitivity, Positive/Negative Predictive Values, False Discovery Rate, etc.
- In general, these are not robust to simultaneous changes in class prevalence and cost

Multi-Class ROC Analysis Methods

- In extending to 3 dimensions, ROC vector has 6 points and cannot be visualized

Multi-Class ROC Analysis Methods

- In extending to 3 dimensions, ROC vector has 6 points and cannot be visualized
- Some have constructed a 3-dimensional surface from the diagonals of the conditional probability matrix, ignoring off-diagonals to calculate Volume Under the ROC Surface (VUS)

Multi-Class ROC Analysis Methods

- In extending to 3 dimensions, ROC vector has 6 points and cannot be visualized
- Some have constructed a 3-dimensional surface from the diagonals of the conditional probability matrix, ignoring off-diagonals to calculate Volume Under the ROC Surface (VUS)
- Some treat *either* cost or class prevalence as threshold parameters, but not *both*

Multi-Class ROC Analysis Methods

- In extending to 3 dimensions, ROC vector has 6 points and cannot be visualized
- Some have constructed a 3-dimensional surface from the diagonals of the conditional probability matrix, ignoring off-diagonals to calculate Volume Under the ROC Surface (VUS)
- Some treat *either* cost or class prevalence as threshold parameters, but not *both*
- Note: no assumptions on independence are required for any of these measures, as they will be for ideas presented in this thesis

Outline

Previous Work
Thorsen and Oxley, SPIE 2006

- Thorsen and Oxley suggested ROC analysis minimizing Bayes Risk:

Previous Work

Thorsen and Oxley, SPIE 2006

- Thorsen and Oxley suggested ROC analysis minimizing Bayes Risk:
- $$\min_{\mathbf{q} \in Q} \int_{\Gamma} \langle \mathbf{q}, \hat{\gamma} \rangle \, \mathrm{W}(\gamma) \, \mathbf{d}\gamma$$

  where Q is a collection of ROC vectors and $\mathrm{W}(\gamma)$ is a *joint* weighting function on a cost-prior Hadamard Product *vector* $\hat{\gamma} = \mathbf{c} \odot \hat{\mathbf{p}}$, either a probability density function or belief function

Previous Work

Thorsen and Oxley, SPIE 2006

- Thorsen and Oxley suggested ROC analysis minimizing Bayes Risk:
- $$\min_{\mathbf{q} \in Q} \int_{\Gamma} \langle \mathbf{q}, \hat{\gamma} \rangle \, \mathrm{W}(\gamma) \, \mathbf{d}\gamma$$

  where Q is a collection of ROC vectors and $\mathrm{W}(\gamma)$ is a *joint* weighting function on a cost-prior Hadamard Product *vector* $\hat{\gamma} = \mathbf{c} \odot \hat{\mathbf{p}}$, either a probability density function or belief function
- Requires knowledge of joint distribution of costs and prior probabilities

Previous Work

Thorsen and Oxley, SPIE 2006

Risk-Based Comparison of Classification Systems

2d Lt Seth Wagenman

Background
Binary ROC Analysis
Definitions & Assumptions
ROC Analysis Review

ROC Risk Functional (RRF)
Development and Definition
Unknown Priors
Partially Known Priors

Example Application
Uniform Priors
Beta Distributed
Results

Summary

- Thorsen and Oxley suggested ROC analysis minimizing Bayes Risk:
- $$\min_{\mathbf{q}\in Q} \int_{\Gamma} \langle \mathbf{q}, \hat{\gamma} \rangle \, \mathrm{W}(\gamma) \, \mathbf{d}\gamma$$
  where Q is a collection of ROC vectors and $\mathrm{W}(\gamma)$ is a *joint* weighting function on a cost-prior Hadamard Product *vector* $\hat{\gamma} = \mathbf{c} \odot \hat{\mathbf{p}}$, either a probability density function or belief function
- Requires knowledge of joint distribution of costs and prior probabilities
- *Expected value* of risk $\langle \mathbf{q}, \hat{\gamma} \rangle$ is Bayes Risk

Framework of ROC Risk Functional

Using matrices and dropping notation for A_θ,

$$E(R) \equiv E\left[\left\langle \mathbf{Q}, (\mathbf{C} \odot \mathbf{P}) \right\rangle_F\right] \text{ becomes:}$$

$$\int\int\int \langle \mathbf{Q}, (\mathbf{C} \odot \mathbf{P}) \rangle_F \, W_{\mathbf{Q},\mathbf{C},\mathbf{P}} \, \mathbf{dQ} \, \mathbf{dC} \, \mathbf{dP}$$

$$= \int\int\int \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \left[q_{i|j} (c_{i|j} \, p_j) \right] \right) W_{\mathbf{Q},\mathbf{C},\mathbf{P}} \, \mathbf{dQ} \, \mathbf{dC} \, \mathbf{dP}$$

$$= \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \left(E\left[q_{i|j}\right] \left[E\left[c_{i|j}\right] E\left[p_j\right] \right] \right) \right]$$

$$= \left\langle \mathbf{E}\left[\mathbf{Q}_A\right], \left(\mathbf{E}\left[\mathbf{C}_A\right] \odot \mathbf{E}\left[\mathbf{P}_A\right] \right) \right\rangle_F$$

where $\mathbf{E}[\cdot]$ is a *matrix* $\left[E(\cdot)\right]_{ij}$ of expected values

ROC Risk Functional

- Given an acceptable estimate $\widehat{\mathbf{Q}_{A,\theta}}$ of the conditional probability matrix $\mathbf{Q}_{A,\theta}$, we may substitute $\mathbf{E}\left[\widehat{\mathbf{Q}_{A,\theta}}\right] \approx \mathbf{Q}_{A,\theta}$, and we may also consider costs to be subjectively fixed

ROC Risk Functional

- Given an acceptable estimate $\widehat{\mathbf{Q}_{A,\theta}}$ of the conditional probability matrix $\mathbf{Q}_{A,\theta}$, we may substitute $\mathbf{E}\left[\widehat{\mathbf{Q}_{A,\theta}}\right] \approx \mathbf{Q}_{A,\theta}$, and we may also consider costs to be subjectively fixed
- Define the ROC Risk functional as that threshold parameter θ such that the classification system $A_\theta \in \mathbb{A}$ minimizes Bayes risk over the family $\mathbb{A}$:

$$\arg\min_{\theta\in\Theta}\left\{E\left[R_{A,\theta}\right]\right\} = \arg\min_{\theta\in\Theta}\left\{\left\langle \widehat{\mathbf{Q}_{A,\theta}},\ \mathbf{C}_A \odot \mathbf{E}[\mathbf{P}]\right\rangle_F\right\}$$

ROC Risk Functional

- Given an acceptable estimate $\widehat{\mathbf{Q}_{A,\theta}}$ of the conditional probability matrix $\mathbf{Q}_{A,\theta}$, we may substitute $\mathbf{E}\left[\widehat{\mathbf{Q}_{A,\theta}}\right] \approx \mathbf{Q}_{A,\theta}$, and we may also consider costs to be subjectively fixed
- Define the ROC Risk functional as that threshold parameter θ such that the classification system $A_\theta \in \mathbb{A}$ minimizes Bayes risk over the family $\mathbb{A}$:

$$\arg\min_{\theta\in\Theta}\left\{E\left[R_{A,\theta}\right]\right\} = \arg\min_{\theta\in\Theta}\left\{\left\langle \widehat{\mathbf{Q}_{A,\theta}},\ \mathbf{C}_A \odot \mathbf{E}[\mathbf{P}]\right\rangle_F\right\}$$

- Now we may estimate priors under fixed costs and acceptable conditional probability estimates

Outline

Uniform Distribution

- We will illustrate applications with a three-class scenario

Uniform Distribution

- We will illustrate applications with a three-class scenario
- Since 3 prevalences sum to one, the last one drawn is a function of the first two (order of assignment also randomized for uniform case)

Uniform Distribution

- We will illustrate applications with a three-class scenario
- Since 3 prevalences sum to one, the last one drawn is a function of the first two (order of assignment also randomized for uniform case)
- The joint distribution of the first two has constant pdf $(3-1)!$, general case , $(n-1)!$

Uniform Distribution

- We will illustrate applications with a three-class scenario
- Since 3 prevalences sum to one, the last one drawn is a function of the first two (order of assignment also randomized for uniform case)
- The joint distribution of the first two has constant pdf $(3-1)!$, general case , $(n-1)!$
- The marginal distribution of the last random variable may be found using the method of distributions or transformations, etc.

Uniform Distribution

- We will illustrate applications with a three-class scenario
- Since 3 prevalences sum to one, the last one drawn is a function of the first two (order of assignment also randomized for uniform case)
- The joint distribution of the first two has constant pdf $(3-1)!$, general case , $(n-1)!$
- The marginal distribution of the last random variable may be found using the method of distributions or transformations, etc.
- Expected value for each class prevalence is $\frac{1}{3}$, general case $\frac{1}{n}$

Uniform Distribution

- We will illustrate applications with a three-class scenario
- Since 3 prevalences sum to one, the last one drawn is a function of the first two (order of assignment also randomized for uniform case)
- The joint distribution of the first two has constant pdf $(3-1)!$, general case , $(n-1)!$
- The marginal distribution of the last random variable may be found using the method of distributions or transformations, etc.
- Expected value for each class prevalence is $\frac{1}{3}$, general case $\frac{1}{n}$
- Assumption of equal priors implies this uniform distribution

Method of Choosing First Two Prevalences

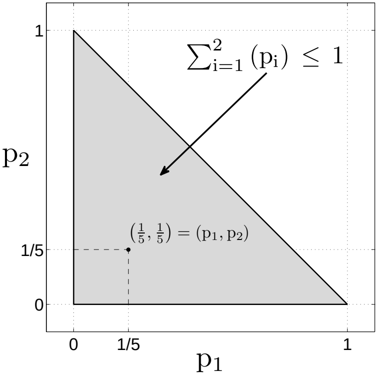

Figure: $\Delta_2 \equiv \left\{ \mathbf{p} \in [0,1]^2 : \sum_{i=1}^{2} p_i \leq 1 \right\}$

Method of Choosing Third Class Prevalence

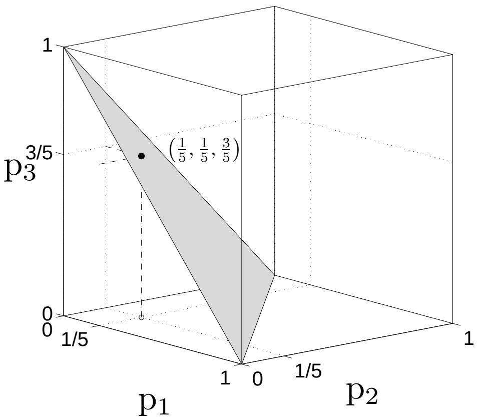

Figure: $\Delta_3 \equiv \left\{ \mathbf{p} \in [0,1]^3 : \sum_{i=1}^{3} p_i \leq 1 \right\}$

Outline

Beta Distribution as Normal Approximation

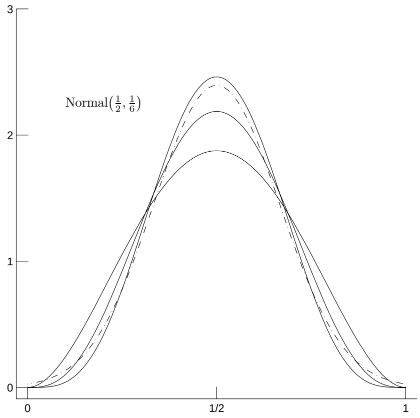

Figure: Beta Approximation to Normal$(\frac{1}{2}, \frac{1}{6})$

Beta Distribution as Normal Approximation

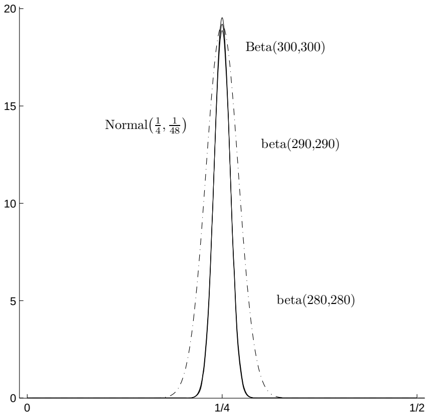

Figure: Beta Approximation to Normal($\frac{1}{4}$, $\frac{1}{48}$)

Bivariate Beta Joint Prevalence Distribution

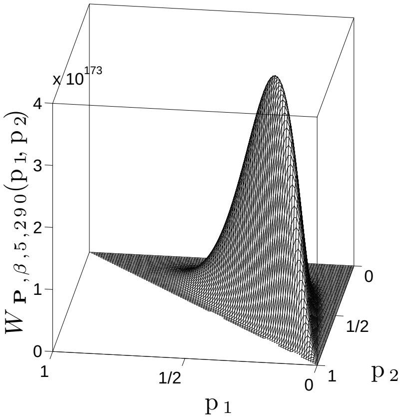

Figure:

$$W_{\mathbf{P},\beta,5,290}(p_1, p_2) \approx \frac{1.197\times 10^{178}\left(p_1 - p_1^2\right)^4 \left(p_2 - p_1 p_2 - p_2^2\right)^{289}}{\left(1-p_1\right)^{579}}$$

Notes on Beta Distribution Implementation

- General Beta distributions are very flexible fixed-support distributions

Notes on Beta Distribution Implementation

- General Beta distributions are very flexible fixed-support distributions
- Class distribution of third prevalence too difficult to derive

Notes on Beta Distribution Implementation

- General Beta distributions are very flexible fixed-support distributions
- Class distribution of third prevalence too difficult to derive
- Order of choosing classes not randomly assigned because distribution has shape

Notes on Beta Distribution Implementation

- General Beta distributions are very flexible fixed-support distributions
- Class distribution of third prevalence too difficult to derive
- Order of choosing classes not randomly assigned because distribution has shape
- In general, tried to maximize overall size of training data

Notes on Beta Distribution Implementation

- General Beta distributions are very flexible fixed-support distributions
- Class distribution of third prevalence too difficult to derive
- Order of choosing classes not randomly assigned because distribution has shape
- In general, tried to maximize overall size of training data
- Could use a non-standard weighting, such as belief functions

Notes on Testing of Assumptions

- Using Probabilistic Neural Network and Lachenbruch Holdout Method to estimate class-conditional probabilities as average over repeated random draws from prevalence distribution

Notes on Testing of Assumptions

- Using Probabilistic Neural Network and Lachenbruch Holdout Method to estimate class-conditional probabilities as average over repeated random draws from prevalence distribution
- To test assumption of independence, assumed only necessary to test between a conditional probability estimate $q_{i|j}$ and the prevalence p_j of the class upon which it is conditioned

Notes on Testing of Assumptions

- Using Probabilistic Neural Network and Lachenbruch Holdout Method to estimate class-conditional probabilities as average over repeated random draws from prevalence distribution
- To test assumption of independence, assumed only necessary to test between a conditional probability estimate $q_{i|j}$ and the prevalence p_j of the class upon which it is conditioned
- Power curve analysis through Monte Carlo simulation:

Notes on Testing of Assumptions

- Using Probabilistic Neural Network and Lachenbruch Holdout Method to estimate class-conditional probabilities as average over repeated random draws from prevalence distribution
- To test assumption of independence, assumed only necessary to test between a conditional probability estimate $q_{i|j}$ and the prevalence p_j of the class upon which it is conditioned
- Power curve analysis through Monte Carlo simulation:
  - Alternative two-tailed hypothesis: Kendall's Tau $\geq$ 0.4 in absolute value

Notes on Testing of Assumptions

- Using Probabilistic Neural Network and Lachenbruch Holdout Method to estimate class-conditional probabilities as average over repeated random draws from prevalence distribution
- To test assumption of independence, assumed only necessary to test between a conditional probability estimate $q_{i|j}$ and the prevalence p_j of the class upon which it is conditioned
- Power curve analysis through Monte Carlo simulation:
  - Alternative two-tailed hypothesis: Kendall's Tau $\geq$ 0.4 in absolute value
  - P-values of ≤ 0.15 assumed significant

Notes on Testing of Assumptions

- Using Probabilistic Neural Network and Lachenbruch Holdout Method to estimate class-conditional probabilities as average over repeated random draws from prevalence distribution
- To test assumption of independence, assumed only necessary to test between a conditional probability estimate $q_{i|j}$ and the prevalence p_j of the class upon which it is conditioned
- Power curve analysis through Monte Carlo simulation:
  - Alternative two-tailed hypothesis: Kendall's Tau $\geq$ 0.4 in absolute value
  - P-values of ≤ 0.15 assumed significant
  - 99% confidence interval for 37-sample power: $(0.8083, 0.8282)$

Notes on Training of Classifier

- Probabilistic Neural Net (PNN) uses spread parameter for thresholds

Notes on Training of Classifier

- Probabilistic Neural Net (PNN) uses spread parameter for thresholds
- PNN yields stable results in replication (unlike Feed-Forward)

Notes on Training of Classifier

- Probabilistic Neural Net (PNN) uses spread parameter for thresholds
- PNN yields stable results in replication (unlike Feed-Forward)
- With enough samples, PNN approaches Bayes Optimal Classifier

Notes on Training of Classifier

- Probabilistic Neural Net (PNN) uses spread parameter for thresholds
- PNN yields stable results in replication (unlike Feed-Forward)
- With enough samples, PNN approaches Bayes Optimal Classifier
- Used Fisher Iris Data, equal class populations of 50 each

Notes on Training of Classifier

- Probabilistic Neural Net (PNN) uses spread parameter for thresholds
- PNN yields stable results in replication (unlike Feed-Forward)
- With enough samples, PNN approaches Bayes Optimal Classifier
- Used Fisher Iris Data, equal class populations of 50 each
- Stored resulting class prevalences and conditional probabilities for correlation testing

Notes on Training of Classifier

- Probabilistic Neural Net (PNN) uses spread parameter for thresholds
- PNN yields stable results in replication (unlike Feed-Forward)
- With enough samples, PNN approaches Bayes Optimal Classifier
- Used Fisher Iris Data, equal class populations of 50 each
- Stored resulting class prevalences and conditional probabilities for correlation testing
- Used standardized form (pre-processed) data, so spread varied from ≈ 0 to 1 in 10 steps

Notes on Cost Regimes Used

Cost Set 1	Actual Class: 1	Actual Class: 2	Actual Class: 3
Labeled Class: 1	0	1	1
Labeled Class: 2	2	0	1
Labeled Class: 3	2	1	0

Cost Set 2	Actual Class: 1	Actual Class: 2	Actual Class: 3
Labeled Class: 1	1	10	2
Labeled Class: 2	2	1	2
Labeled Class: 3	2	10	1

Outline

Results of Risk Minimization, Uniform Cost 1

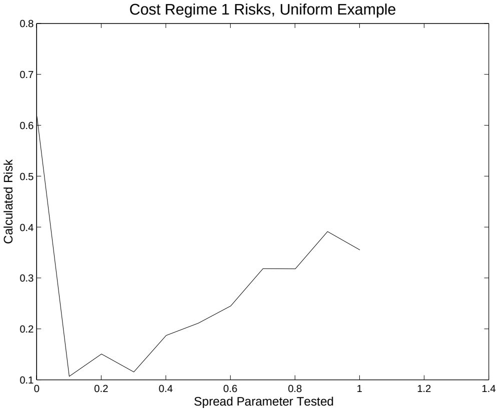

Figure: Risk for Cost Regime 1, Uniform Application

Results of Risk Minimization, Uniform Cost 2

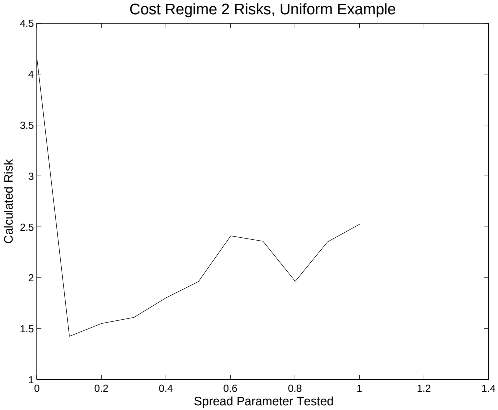

Figure: Risk for Cost Regime 2, Uniform Application

Results of Uniform Distribution Experiment

Spread parameter $\theta = 0.101$ minimized cost over two separate fixed cost scenarios

Mean Absolute Value of Kendall's Tau Coefficients:

Kendall's Tau	Actual Class: 1	Actual Class: 2	Actual Class: 3
Labeled Class: 1	0.34	0.14	0.06
Labeled Class: 2	0.34	0.66	0.58
Labeled Class: 3	0.00	0.59	0.63

Mean p-Value matrix for Kendall's Tau Coefficients:

p-values	Actual Class: 1	Actual Class: 2	Actual Class: 3
Labeled Class: 1	0.16	0.59	0.91
Labeled Class: 2	0.16	0.00	0.09
Labeled Class: 3	1.00	0.09	0.00

Outline

Results of Risk Minimization, beta Cost 1

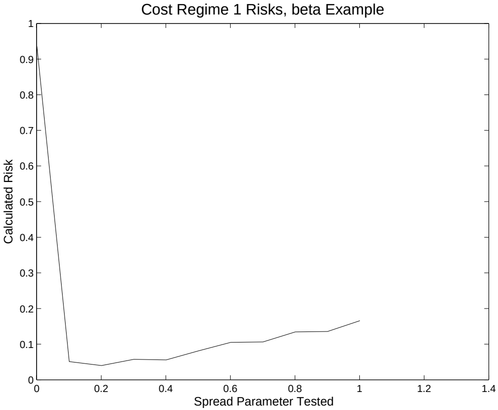

Figure: Risk for Cost Regime 1, beta Application

Results of Risk Minimization, beta Cost 2

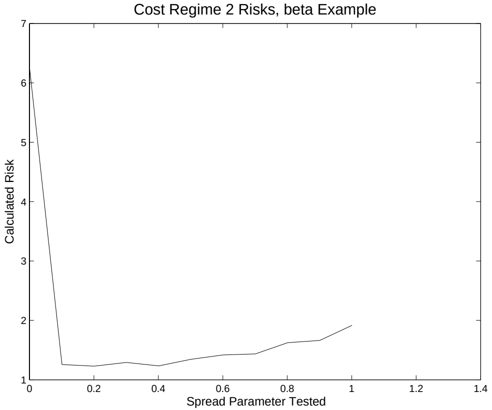

Figure: Risk for Cost Regime 2, beta Application

Results of beta Distribution Experiment

Spread parameter $\theta = 0.201$ minimized cost over two separate fixed cost scenarios

Mean Absolute Value of Kendall's Tau Coefficients:

Kendall's Tau	Actual Class: 1	Actual Class: 2	Actual Class: 3
Labeled Class: 1	0.13	0.18	0.05
Labeled Class: 2	0.13	0.26	0.15
Labeled Class: 3	0.00	0.20	0.20

Mean p-Value matrix for Kendall's Tau Coefficients:

p-values	Actual Class: 1	Actual Class: 2	Actual Class: 3
Labeled Class: 1	0.58	0.48	0.91
Labeled Class: 2	0.58	0.20	0.39
Labeled Class: 3	1.00	0.31	0.30

Outline

Discussion

- Assumptions met better in beta application

Discussion

- Assumptions met better in beta application
- Assumptions allow risk calculation without knowing joint distribution of all variables

Discussion

- Assumptions met better in beta application
- Assumptions allow risk calculation without knowing joint distribution of all variables
- How important is independence?

Summary

- The mathematical framework of ROC Analysis is necessary to enable precise quantification of risk when competing classification systems

Summary

- The mathematical framework of ROC Analysis is necessary to enable precise quantification of risk when competing classification systems
- The ROC Risk Functional (RRF) is a precise, general, and flexible measure of performance

Summary

- The mathematical framework of ROC Analysis is necessary to enable precise quantification of risk when competing classification systems
- The ROC Risk Functional (RRF) is a precise, general, and flexible measure of performance
- The application of the RRF is simple and straightforward

Summary

- The mathematical framework of ROC Analysis is necessary to enable precise quantification of risk when competing classification systems
- The ROC Risk Functional (RRF) is a precise, general, and flexible measure of performance
- The application of the RRF is simple and straightforward
- Outlook

Summary

- The mathematical framework of ROC Analysis is necessary to enable precise quantification of risk when competing classification systems
- The ROC Risk Functional (RRF) is a precise, general, and flexible measure of performance
- The application of the RRF is simple and straightforward
- Outlook
  - Belief functions may be more accessible to normal users

Summary

- The mathematical framework of ROC Analysis is necessary to enable precise quantification of risk when competing classification systems
- The ROC Risk Functional (RRF) is a precise, general, and flexible measure of performance
- The application of the RRF is simple and straightforward
- Outlook
  - Belief functions may be more accessible to normal users
  - Other probability distributions beyond the Beta distribution may be better suited to certain situations

Summary

- The mathematical framework of ROC Analysis is necessary to enable precise quantification of risk when competing classification systems
- The ROC Risk Functional (RRF) is a precise, general, and flexible measure of performance
- The application of the RRF is simple and straightforward
- Outlook
  - Belief functions may be more accessible to normal users
  - Other probability distributions beyond the Beta distribution may be better suited to certain situations
  - Distributions of costs and ROC information may be analyzed separately

Summary

- The mathematical framework of ROC Analysis is necessary to enable precise quantification of risk when competing classification systems
- The ROC Risk Functional (RRF) is a precise, general, and flexible measure of performance
- The application of the RRF is simple and straightforward
- Outlook
  - Belief functions may be more accessible to normal users
  - Other probability distributions beyond the Beta distribution may be better suited to certain situations
  - Distributions of costs and ROC information may be analyzed separately
  - Method of training classifiers may merit wider application