

Mining When Classes are Imbalanced, Rare Events Matter More, and Errors Have Costs Attached



Nitesh V. Chawla
University of Notre Dame
<http://www.nd.edu/~nchawla>
nchawla@nd.edu

Nitesh Chawla, SIAM 2009
Tutorial

Overview

- Introduction
- Sampling Methods
- Moving Decision Threshold
- Classifiers' Objective Functions
- Evaluation Measures

*IEEE ICDM noted "Dealing with Non-static, Unbalanced and Cost-sensitive Data" among the **10 Challenging Problems in Data Mining Research***

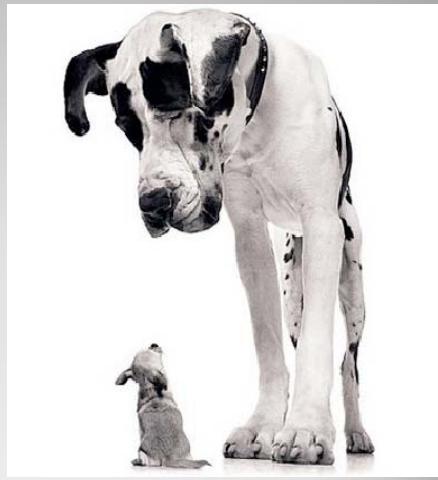
Nitesh Chawla, SIAM 2009
Tutorial

Small Class Matters, and Matters More

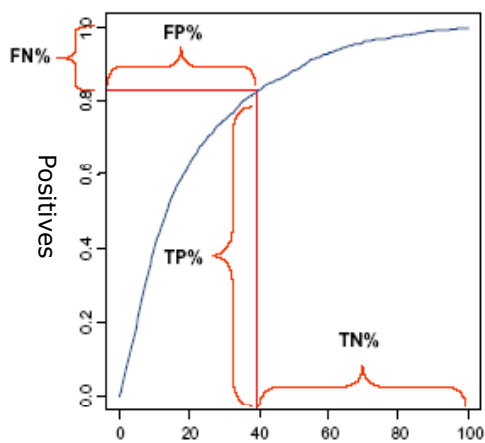
Data set is Imbalanced, if the classes are unequally distributed

Class of interest (minority class) is often much smaller or rarer

But, the cost of error on the minority class can have a bigger bite



Nitesh Chawla, SIAM 2009
Tutorial



	Actual Non-Default	Actual Default
Predict Non-Default	TN	FN
Predict Default	FP	TP

Typical Prediction Model

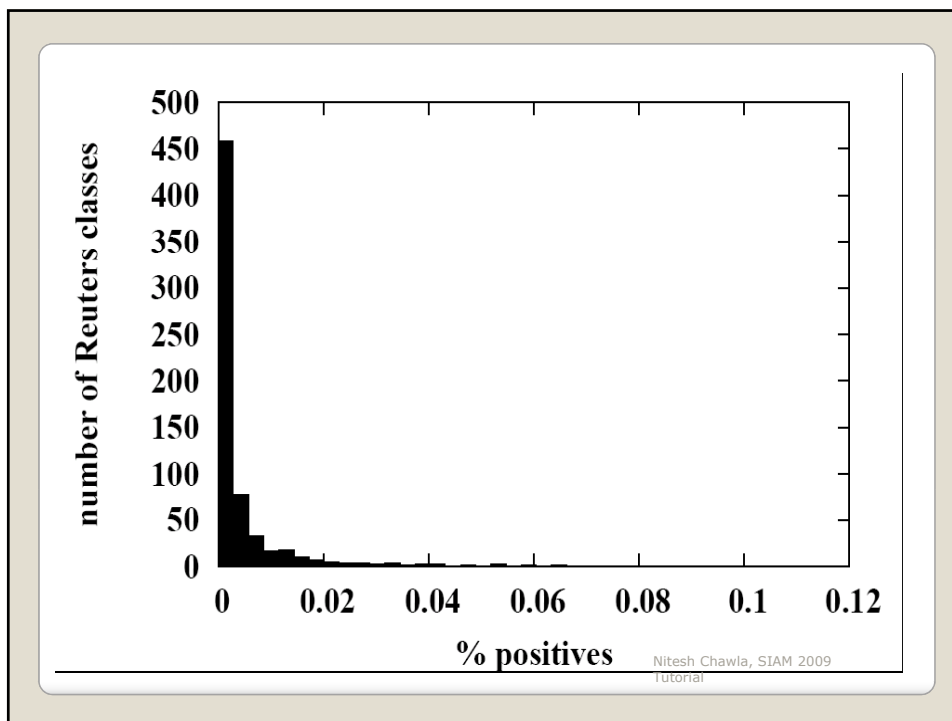
Nitesh Chawla, SIAM 2009
Tutorial

The one in a 100, one in a 1000, one in 100,000, and one in a million event

- Real-world has abundance of scenarios with such imbalance in class distributions
 - Fraud detection
 - Disease prediction
 - Intrusion detection
 - Text categorization
 - Bioinformatics
 - Direct marketing
 - Terrorist attack
 - Physics simulations
 - Climate

Nitesh Chawla, SIAM 2009
Tutorial





Paradox of False Positive

- Imagine a disease that has a prevalence of 1 in a million people. I invent a test that is 99% accurate. I am obviously excited. But, when applied to a million, it returns positive for 10,000 (remember, it is 99% accurate). Priors tell us otherwise. There is one in a million infected --- 99% accurate test is inaccurate 9,999 times out of 10,000.

Yes, measuring performance presents challenges

- A “fruit-bowl” of measures. No more comparing apples and oranges. Take your favorite. But, how do we really compare?
 - Accuracy (CAREFUL)
 - Balanced accuracy (better)
 - AUROC (different ways of computing, potentially)
 - F-measure (requires a threshold)
 - Precision @ Top 20 (where are the positive cases in the ranking)
 - G-mean
 - Probability loss measures such as negative cross entropy and brier score (how well calibrated are the models?)

Fruit for thought. We will return to this.

Nitesh Chawla, SIAM 2009
Tutorial

Countering Class Imbalance: Some Popular Solutions in Data Mining

- Sampling
 - Oversampling
 - Undersampling
 - ...Variations and combinations of the two
- Adapting learning algorithms by modifying objective functions or changing decision thresholds
 - Decision trees
 - Neural Networks
 - SVMs
- Ensemble based methods

(all of the above can also be combined together!)

Nitesh Chawla, SIAM 2009
Tutorial

Overview

- Introduction
- Sampling Methods
- Moving Decision Threshold
- Classifiers' Objective Functions
- Evaluation Measures

Nitesh Chawla, SIAM 2009
Tutorial

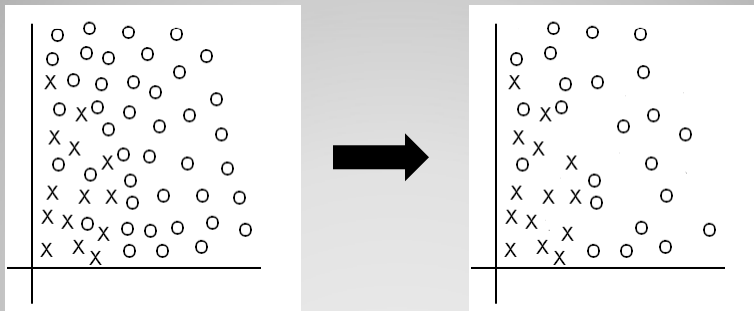
Sampling: Add or remove until satisfactory performance

- Undersampling dictates removal of majority class examples
- Oversampling dictates removal of minority class examples

Nitesh Chawla, SIAM 2009
Tutorial

Undersampling

- Randomly remove majority class examples



Risk of losing potentially important majority class examples, that help establish the discriminating power

Nitesh Chawla, SIAM 2009
Tutorial

What about focusing on the borderline and noisy examples?

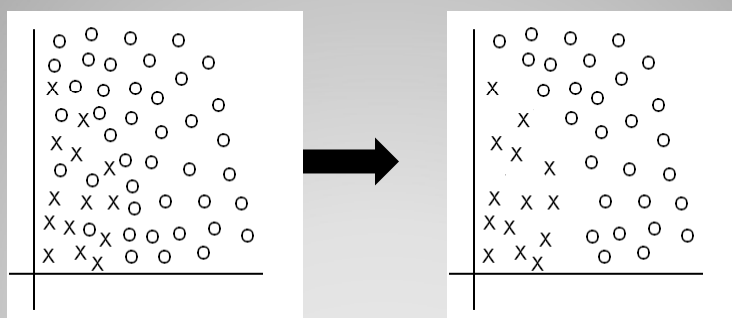
Introducing Tomek Links and Condensed Nearest Neighbor Rule

Nitesh Chawla, SIAM 2009
Tutorial

Tomek links

- To remove both noise and borderline examples
- Tomek link
 - Let E_i, E_j be examples belonging to different classes.
 - Let $d(E_i, E_j)$ is the distance between them.
 - A (E_i, E_j) pair is called a Tomek link if there is no example E_k , such that $d(E_i, E_k) < d(E_i, E_j)$ or $d(E_j, E_k) < d(E_i, E_j)$.

Nitesh Chawla, SIAM 2009
Tutorial



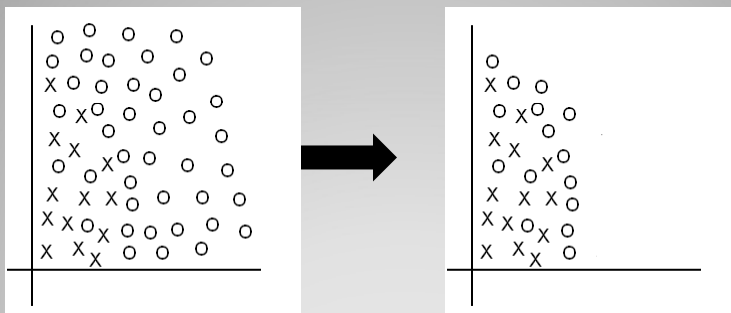
Undersample by Tomek Links

Nitesh Chawla, SIAM 2009
Tutorial

Condensed Nearest Neighbor Rule (CNN rule)

- Find a consistent subset of examples.
 - A subset $E' \subseteq E$ is consistent with E if using a 1-nearest neighbor, E' correctly classifies the examples in E
- The goal is to eliminate examples from the majority class that are much further away from the border

Nitesh Chawla, SIAM 2009
Tutorial

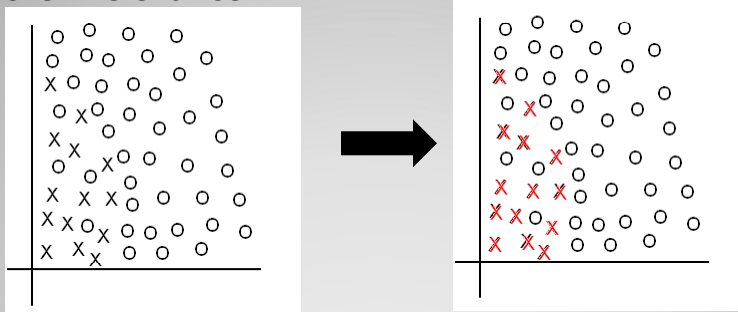


CNN Editing

Nitesh Chawla, SIAM 2009
Tutorial

Oversampling

- Replicate the minority class examples to increase their relevance

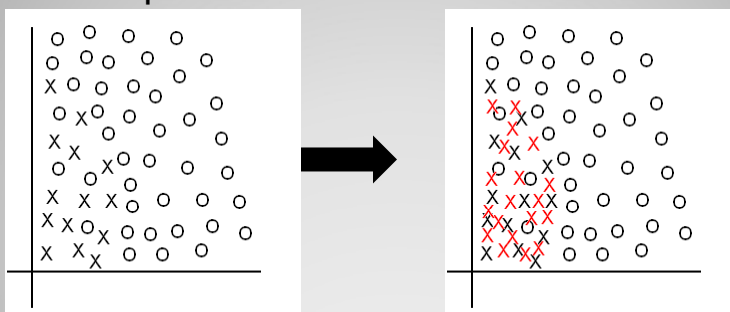


But no new information is being added. Hurts the generalization capacity.

Nitesh Chawla, SIAM 2009
Tutorial

Instead of replicating, let us invent some new instances

- SMOTE: Synthetic Minority Over-sampling Technique



Nitesh Chawla, SIAM 2009
Tutorial

SMOTE

```

 $\forall x \in \text{Minority}$ 
for i = 1 to |x|
  Compute k - NN of  $x_i$ 
  for j = 1 to | $x_i$ |
    if  $x_{ij} \equiv \text{continuous}$ 
       $x_{nj} = (x_i - x_{ki}) * \text{rand}(1)$ 
    else
       $x_{nj} = \arg \max_{x' \in x_j} \sum_{k: x_{jk} = x'} 1$ 
    endif
  endfor
endfor

```

```

k- NN( $x_j, x_j'$ )
if  $j \in \text{continuous}$ 
   $\delta_c = (x_j - x_j')^2$ 
if  $j \in \text{no min al}$ 
   $\delta_n = \sum_{c=1}^c \left| \frac{N_{j, x_j, c}}{N_{j, x}} - \frac{N_{j, x_j', c}}{N_{j, x'}} \right|$  (VDM)
 $\Delta = \delta_c + \delta_n$ 

```

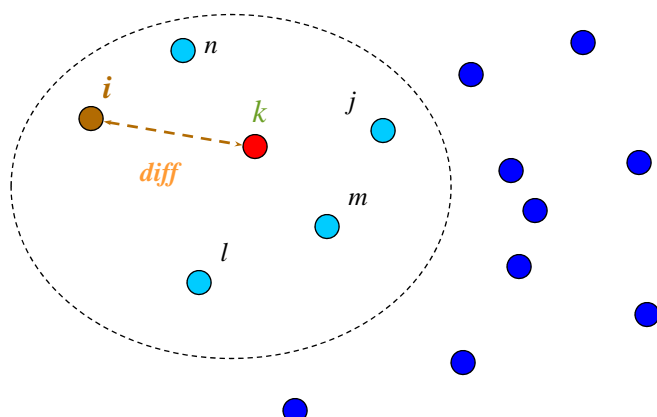
Nitesh Chawla, SIAM 2009
Tutorial

SMOTE

For each minority example k compute nearest minority class examples $\{i, j, l, n, m\}$

Nitesh Chawla, SIAM 2009
Tutorial

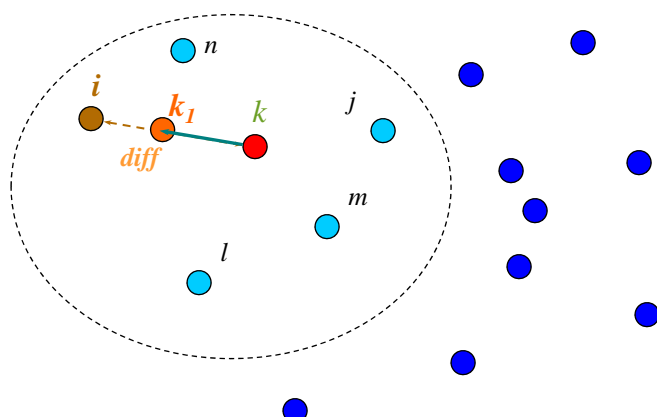
SMOTE



- Randomly choose an example out of 5 closest points

Nitesh Chawla, SIAM 2009
Tutorial

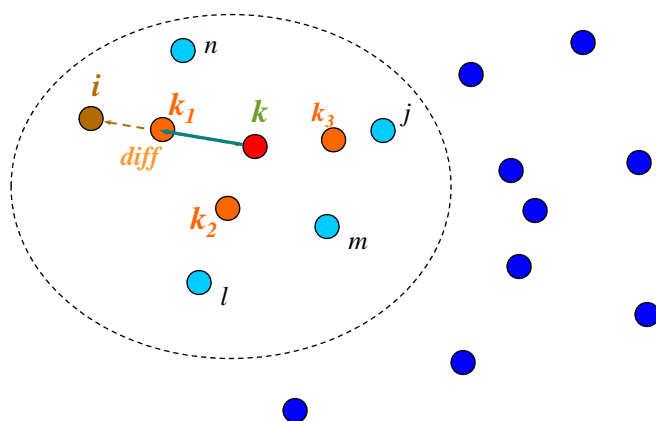
SMOTE



Synthetically generate event k_1 , such that k_1 lies between k and i

Nitesh Chawla, SIAM 2009
Tutorial

SMOTE

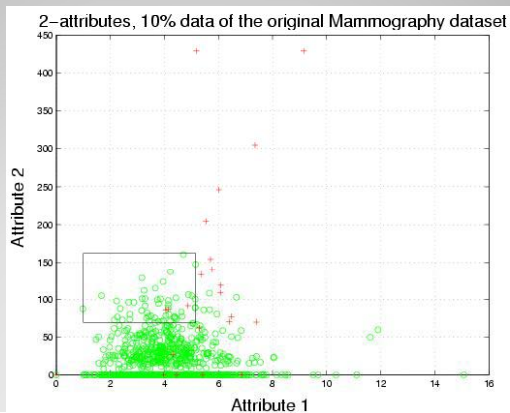


After applying SMOTE 3 times (SMOTE parameter = 300%) data set may look like as the picture above

Nitesh Chawla, SIAM 2009
Tutorial

SMOTE

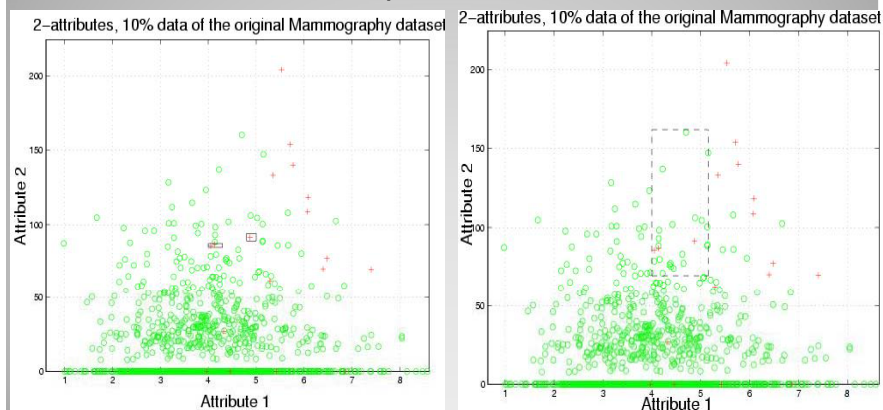
- Generate "new" minority class instances conditioned on the distribution of known instances



Nitesh Chawla, SIAM 2009
Tutorial

SMOTE

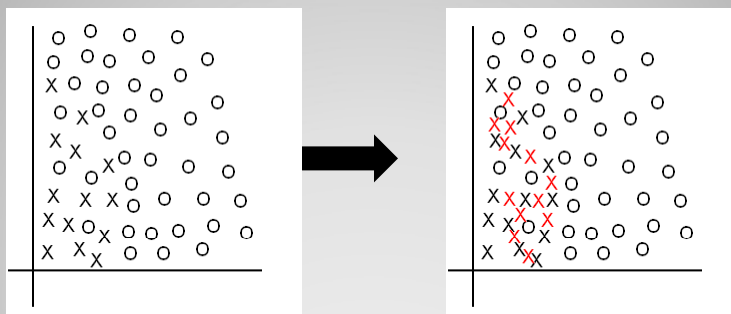
- Generate "new" minority class instances conditioned on the provided instances



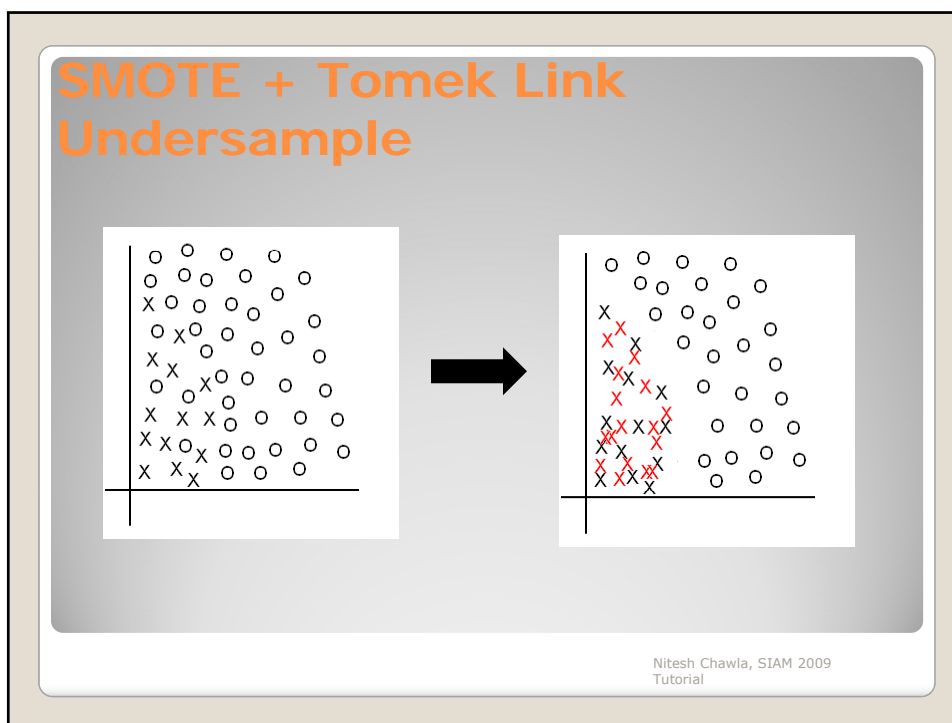
Nitesh Chawla, SIAM 2009
Tutorial

Beware of those lurking majority class examples

- Borderline-SMOTE



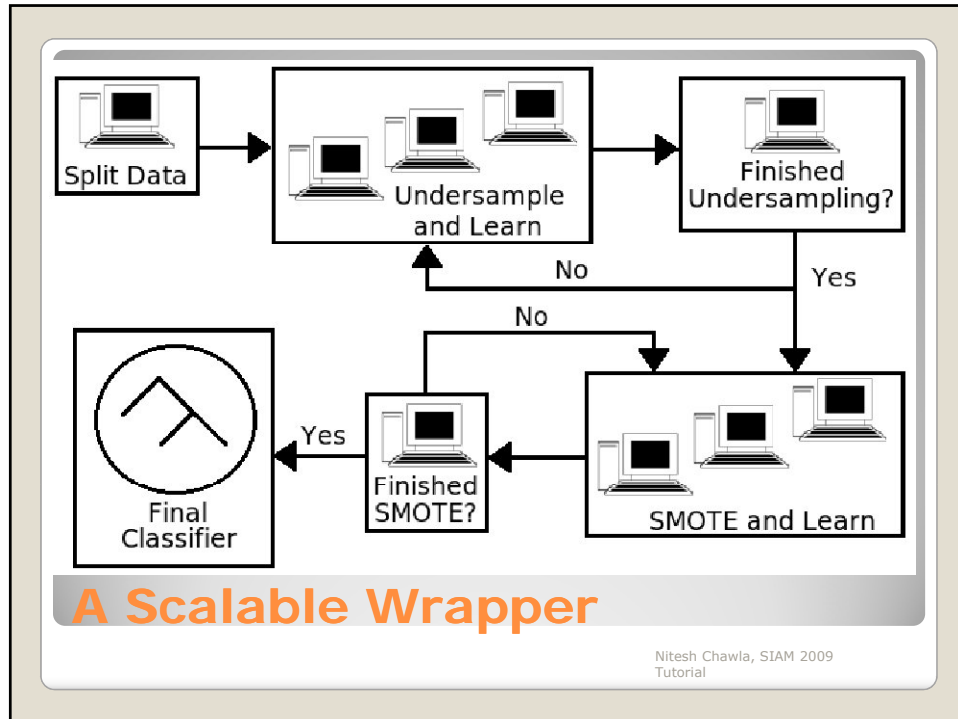
Nitesh Chawla, SIAM 2009
Tutorial



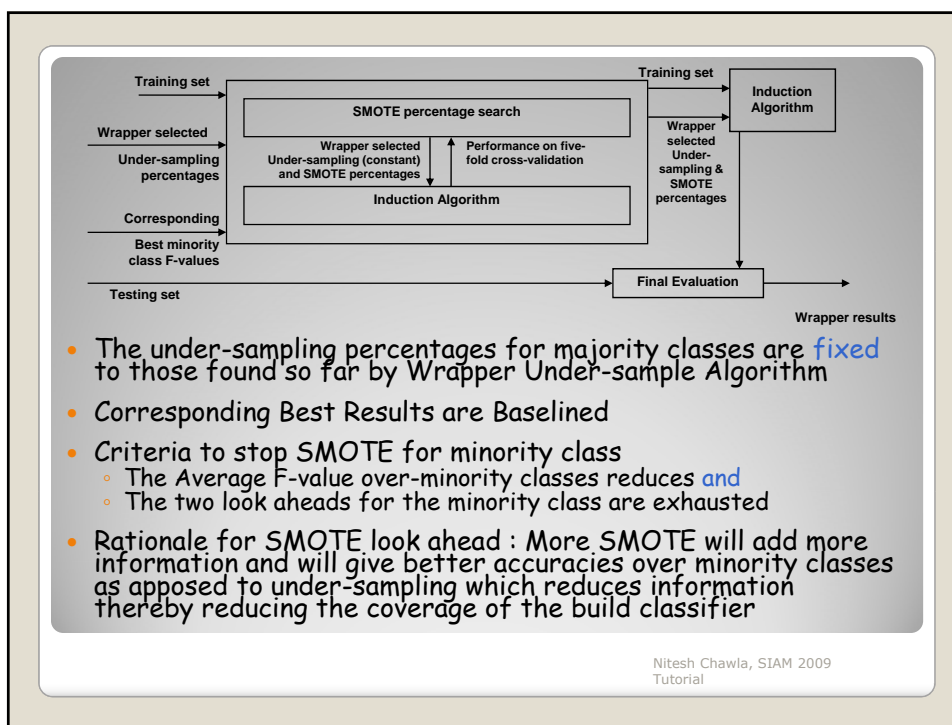
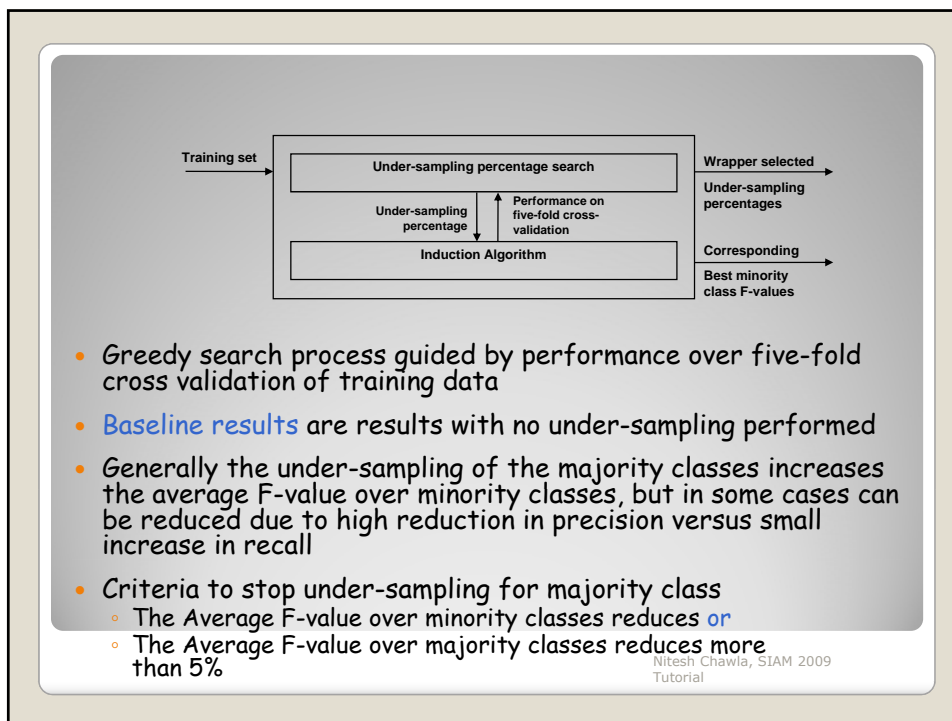
- Two fundamental issues:
 - What is the right sampling method for a given dataset?
 - How to choose the amount to sample?
- Use a wrapper to empirically discover the relevant amounts of sampling

Sampling Methods: Discussions

Nitesh Chawla, SIAM 2009
Tutorial

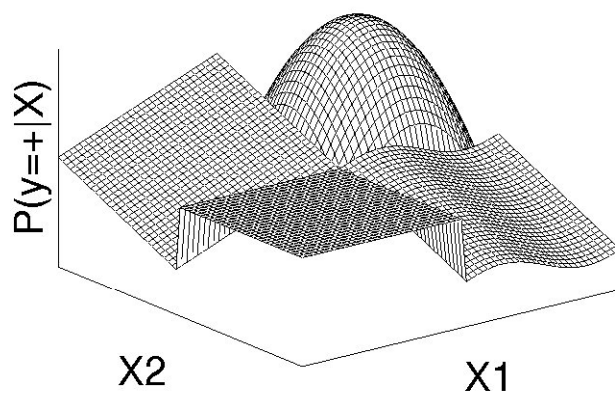


- Testing for each pair of under-sampling and SMOTE percentages is too time consuming
 - So a heuristic is used where searching for under-sampling percentage is done first then followed by search for SMOTE percentage
 - Hypothesis: The under-sampling will first remove the "excess" majority class examples, without much hampering the accuracy on majority classes. Later SMOTE will add synthetic minority class examples which will increase the generalization performance of the classifier over the minority classes
 - Algorithm divided into two parts
 - Wrapper Under-sample Algorithm
 - Wrapper SMOTE Algorithm
 - Our Algorithm can handle multiple minority and majority class problems
 - Uses Five-fold cross-validation over training data as the evaluation function
- Nitesh Chawla, SIAM 2009 Tutorial



Exploiting Locality in Sampling

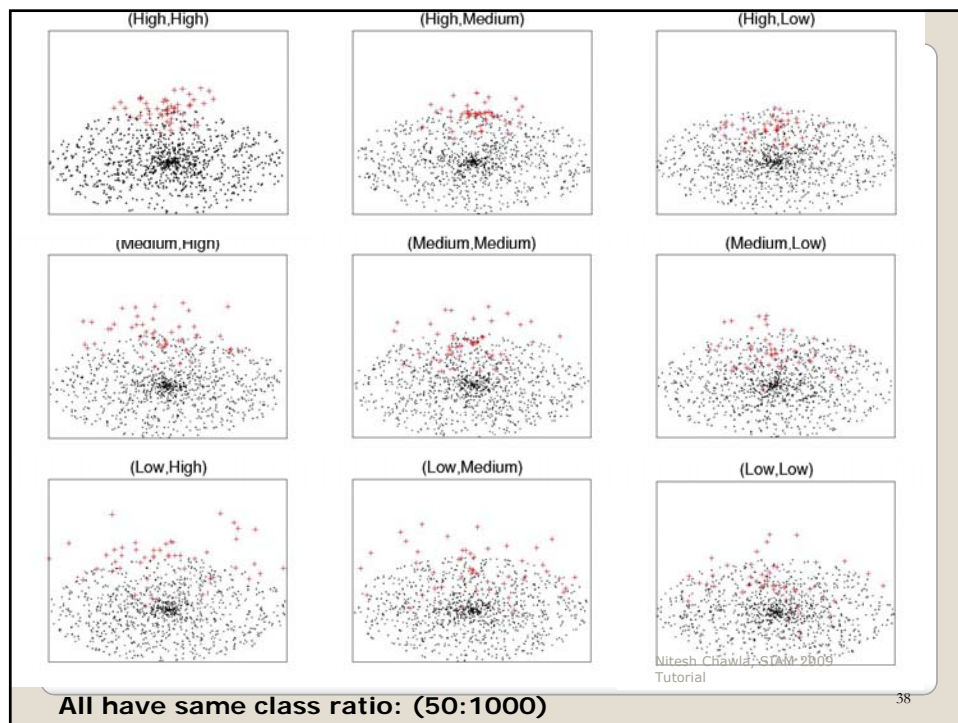
Nitesh Chawla, SIAM 2009
Tutorial



Nitesh Chawla, SIAM 2009
Tutorial

- Class ratio can be important to determining best sampling levels to use
- Other properties may exert greater influence
 - Overlap
 - Density
- Consider the following examples

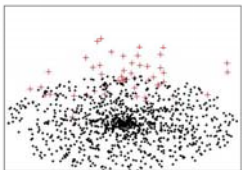
Nitesh Chawla, SIAM 2009
Tutorial



38

(density,separation)	AUROC		Undersample	Smote
	C45	C45+S		
(High, High)	0.926	0.968	40	450
(High, Medium)	0.909	0.942	60	250
(High, Low)	0.898	0.904	60	100
(Medium, High)	0.915	0.961	50	300
(Medium, Medium)	0.878	0.927	40	250
(Medium, Low)	0.814	0.831	70	250
(Low, High)	0.892	0.940	0	150
(Low, Medium)	0.825	0.847	70	350
(Low, Low)	0.705	0.736	20	500

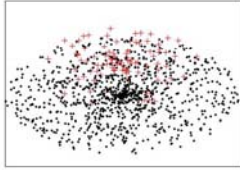
Nitesh Chawla, SIAM 2009 Tutorial



(U,S) =
(50,350)

AUROC =
0.906

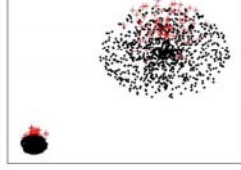
+



(U,S) =
(80,250)

AUROC =
0.812

=

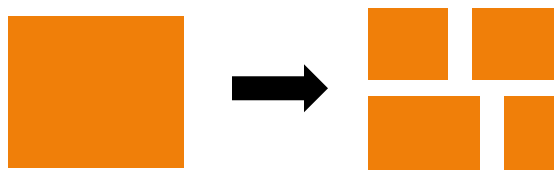


(U,S) =
(??,???)

AUROC =
?????

Nitesh Chawla, SIAM 2009 Tutorial

Step 1: Split the Data



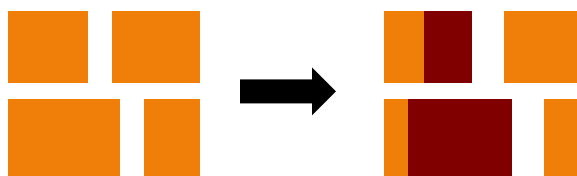
- Could use most supervised and unsupervised methods
- We form 2-level Hellinger distance tree (upcoming)
- Allows localization of diverging class distributions

Nitesh Chawla, SIAM 2009
Tutorial

41

Localized Sampling Framework

Step 2: Sample

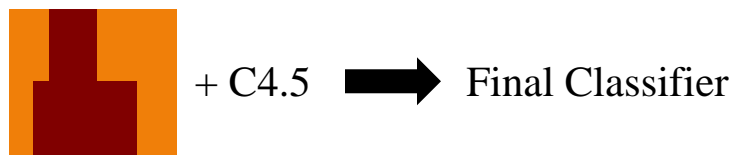


- Sample (SMOTE and undersample) each localization
- Optimize global performance iterating each sample based on minority class size (use wrapper approach)

Nitesh Chawla, SIAM 2009
Tutorial

42

Step 3: Train/predict globally



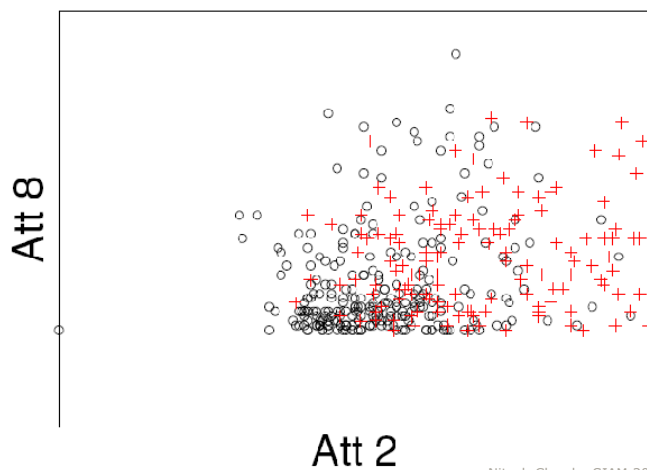
Localized Sampling Framework

Nitesh Chawla, SIAM 2009
Tutorial

[Cieslak, Chawla ICDM 2008]

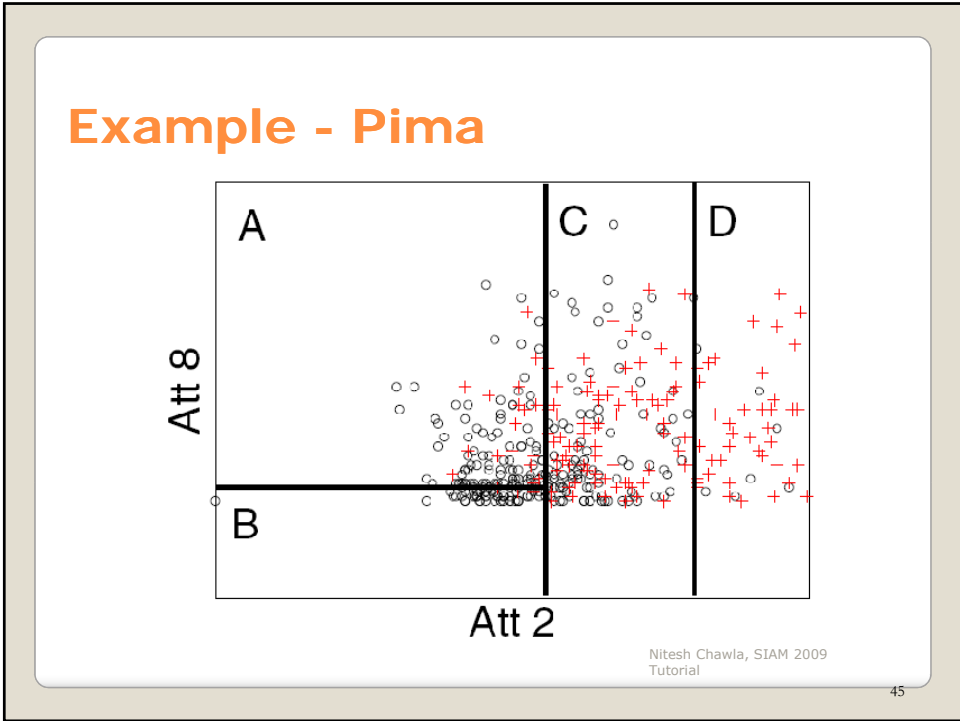
43

Example - Pima

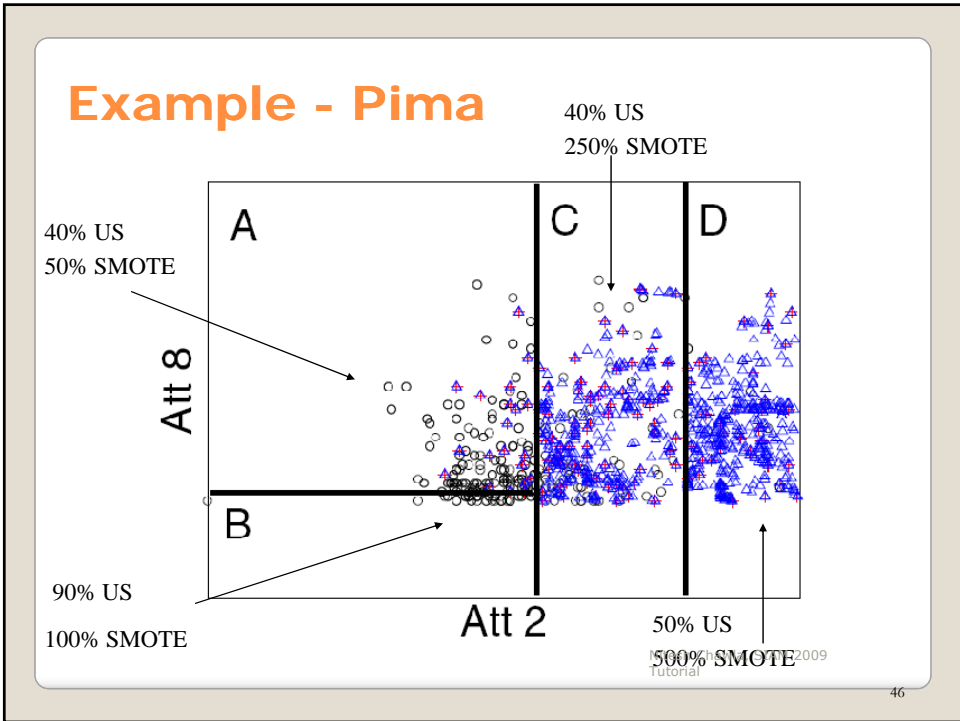


Nitesh Chawla, SIAM 2009
Tutorial

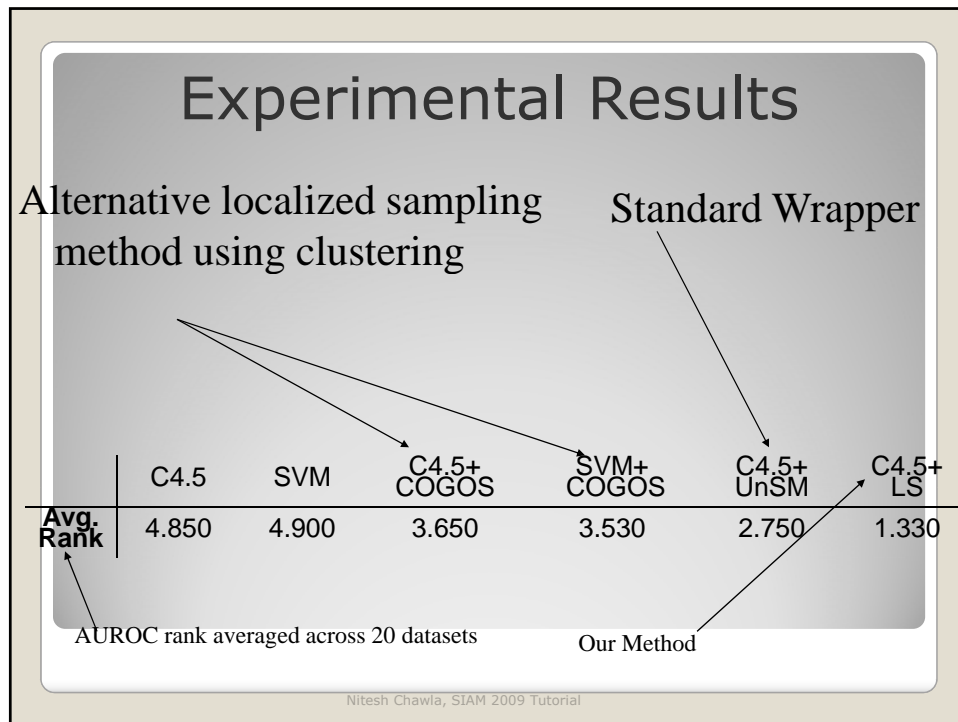
44



45



46



Overview

- Introduction
- Sampling Methods
- Moving Decision Threshold
- Classifiers' Objective Functions
- Evaluation Measures

Nitesh Chawla, SIAM 2009 Tutorial

- Use localized sampling
 - Start Globally, Optimize Locally, and Predict Globally
- Wrapper can be integrated to guide the sampling
- Generally AUC is recommended as the objective function

Recommendations

Nitesh Chawla, SIAM 2009
Tutorial

Changing Decision Thresholds

- Decisions of (scoring) classifiers are typically set at 0.5
 - $P(x > 0.5)$ is class 1 and $P(x \leq 0.5)$ is class 0
- The decision threshold can be moved to compensate for the rate of imbalance
 - Equivalent to different optimization points on the ROC curve
- A wrapper can again be used to optimize threshold

Nitesh Chawla, SIAM 2009
Tutorial

- Quality of probability estimates also becomes important
 - Estimate the quality of estimates using appropriate measures such as negative cross entropy or brier score
- Can also combine sampling methods with threshold moving

Decision Thresholds

Nitesh Chawla, SIAM 2009
Tutorial

Overview

- Introduction
- Sampling Methods
- Moving Decision Threshold
- Classifiers' Objective Functions
- Evaluation Measures

Nitesh Chawla, SIAM 2009
Tutorial

Beyond Sampling: Adapting Classifiers

- Consider
 - Decision Trees
 - SVMs

Nitesh Chawla, SIAM 2009
Tutorial

Decision Trees

- A popular choice when combined with sampling or moving threshold to counter the problem of class imbalance
- The leaf frequencies converted to probability estimates (Laplace or m-estimate smoothing applied, typically)
 - Suggested use is as a PET – Probability Estimation Trees (unpruned, no-collapse, and Laplace)

Nitesh Chawla, SIAM 2009
Tutorial

Converting decision tree leaf predictions into probability estimates

$$P_{freq} = \frac{TP}{TP + FP}$$

$$P_{laplace} = \frac{(TP + 1)}{(TP + FP + C)}$$

$$P_{mest} = \frac{(TP + bm)}{(TP + FP + m)}$$

Nitesh Chawla, SIAM 2009
Tutorial

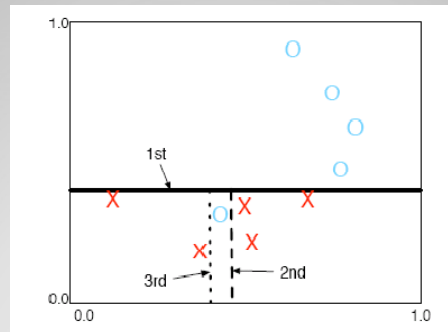
- Dietterich, Kearns and Mansoor (DKM)
- Hellinger distance
- Area under the ROC curve (AUC)
- Minimum squared error of probability estimates (MSEEsplitted)

Some of the skew insensitive metrics proposed

Nitesh Chawla, SIAM 2009
Tutorial

Decision tree (im)purity metrics

Partition feature space to maximize purity at leaves. Recurse



Nitesh Chawla, SIAM 2009
Tutorial

Entropy (Information Gain) as an impurity

(Q, W) classes of interest

N = number of samples

N_i = number of samples in class

N^S = number of samples in S

N_i^S = number of samples in class i in S

$$E = \sum_{i \in (W, Q)} \frac{N_i^L}{N^L} \log_2 \frac{N_i^L}{N^L} + \sum_{i \in (W, Q)} \frac{N_i^R}{N^R} \log_2 \frac{N_i^R}{N^R}$$

Nitesh Chawla, SIAM 2009
Tutorial

Consider a skew insensitive criterion

- Hellinger Distance
 - distance between probability measures independent of the dominating parameters

Nitesh Chawla, SIAM 2009
Tutorial

Properties of Hellinger Distance

$$d_H(P, Q) = \sqrt{\int_{\Omega} (\sqrt{P} - \sqrt{Q})^2 d\lambda}$$

$$d_H(P, Q) = \sqrt{\sum_{\phi \in \Phi} (\sqrt{P(\phi)} - \sqrt{Q(\phi)})^2}$$

- Measures countable space Φ
- Ranges from 0 to $\sqrt{2}$
- Symmetric: $d_H(P, Q) = d_H(Q, P)$
- Lower bounds KL divergence

Nitesh Chawla, SIAM 2009 Tutorial

Formulating for decision tree

Consider a countable space

Consider a two-class problem (W and Q) are the two classes

$$H = \sqrt{\sum_{j=1}^p \left(\sqrt{\frac{N_Q^j}{N_Q}} - \sqrt{\frac{N_W^j}{N_W}} \right)^2}$$

"Distance" in the normalized frequencies space

Nitesh Chawla, SIAM 2009 Tutorial

Inf. Gain vs. Hellinger distance

(Q, W) classes of interest

N

N_i = number of samples in class i

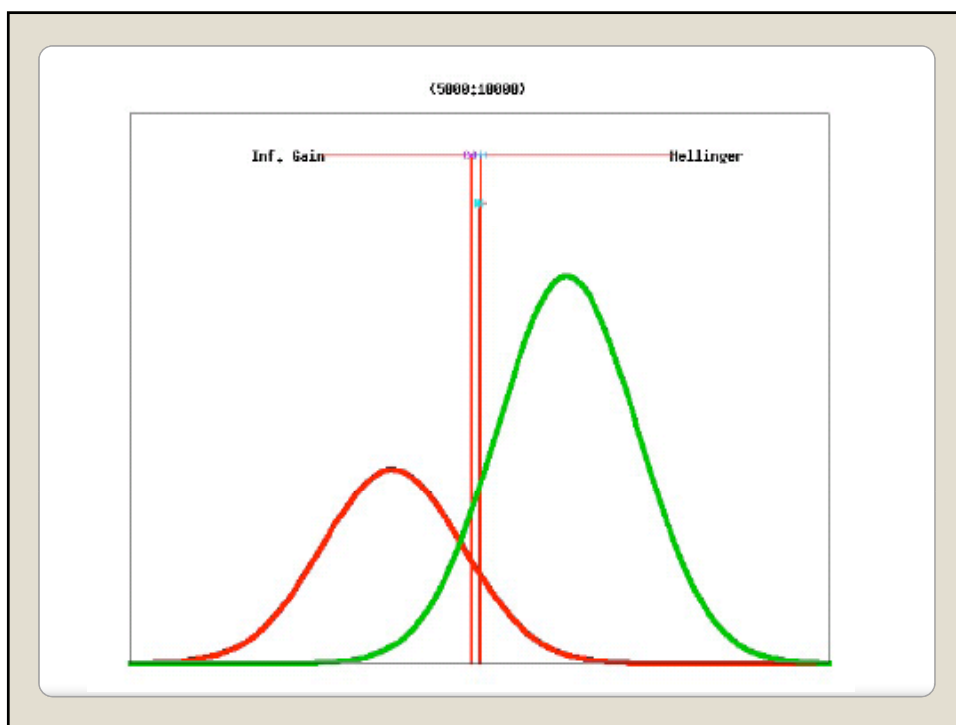
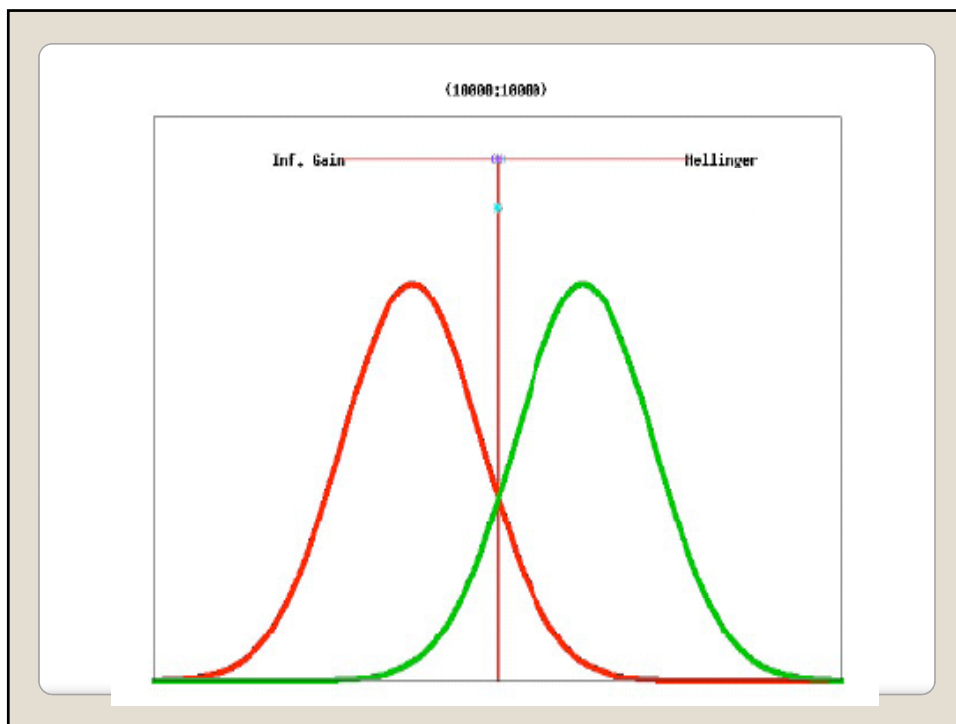
N^S = number of samples in L/R

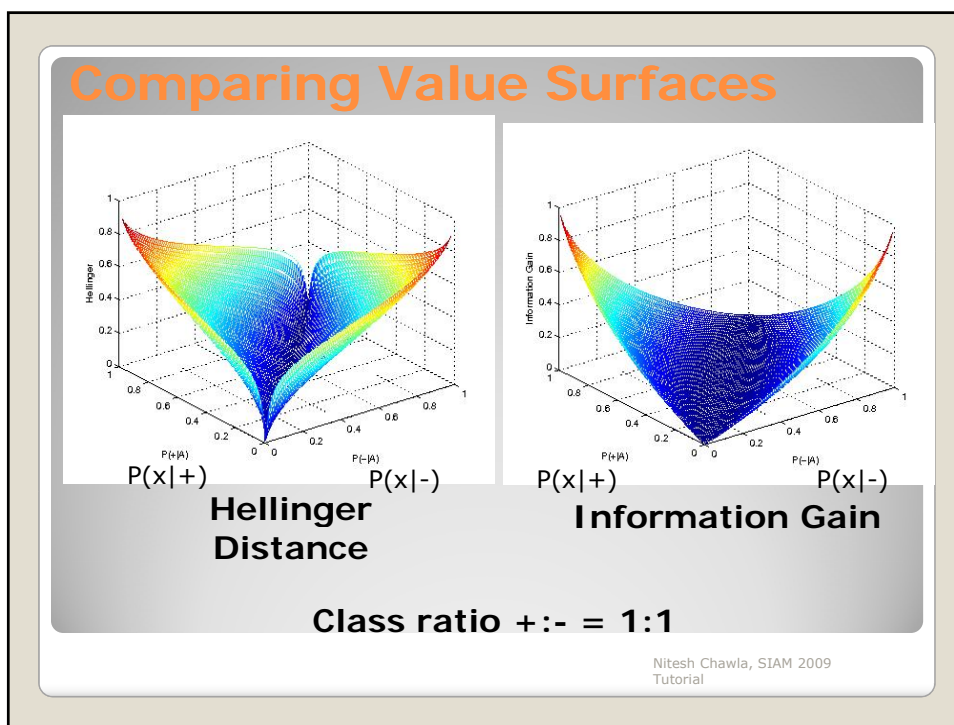
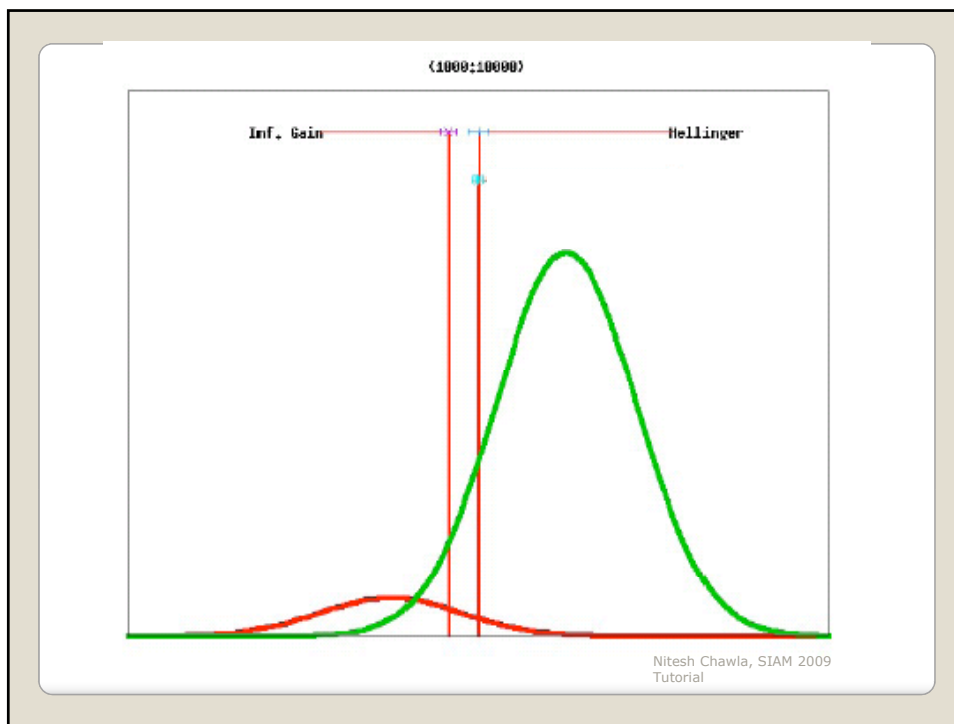
N_i^S = number of samples in class i is L/R split

$$E = \sum_{i \in (W, Q)} -\frac{N_i^L}{N^L} \log_2 \frac{N_i^L}{N^L} + \sum_{i \in (W, Q)} -\frac{N_i^R}{N^R} \log_2 \frac{N_i^R}{N^R}$$

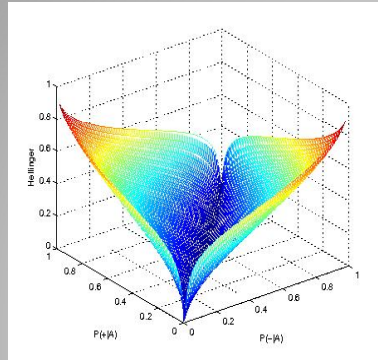
$$H = \sqrt{\left\{ \sqrt{\frac{N_Q^L}{N_Q}} - \sqrt{\frac{N_W^L}{N_W}} \right\}^2 + \left\{ \sqrt{\frac{N_Q^R}{N_Q}} - \sqrt{\frac{N_W^R}{N_W}} \right\}^2}$$

Nitesh Chawla, SIAM 2009 Tutorial

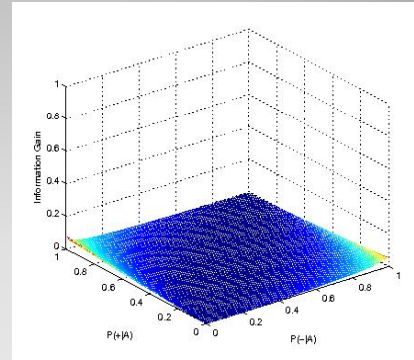




Comparing Value Surfaces



**Hellinger
Distance**



Information Gain

Class ratio +/- = 1:100

Nitesh Chawla, SIAM 2009
Tutorial

Hellinger vs. DKM as decision tree splitting criteria

$$d_{DKM} = 2\sqrt{P(+|A)P(-|A)} - 2P(L|A)\sqrt{P(L|+|A)P(L|-|A)} - 2P(R|A)\sqrt{P(R|+|A)P(R|-|A)}$$

DKM has improved concavity compared to information gain, especially for either very small (relative) class proportions [10].

$$d_H = \sqrt{(\sqrt{P(L|+|A)} - \sqrt{P(L|-|A)})^2 + (\sqrt{P(R|+|A)} - \sqrt{P(R|-|A)})^2}$$

$$d_H = \sqrt{2 - 2\sqrt{P(L|+|A)P(L|-|A)} - 2\sqrt{P(R|+|A)P(R|-|A)}}$$

Nitesh Chawla, SIAM 2009 Tutorial

Algorithm HDDT**Input:** Training Set T , Cutoff size C if $|T| < C$ then

return

end if

for each feature f of T do $H_f = \text{Calc_Hellinger}(T, f)$

end for

 $b = \max(H)$ (best feature)for each branch v of b do $\text{HDDT}(T_{x=b=v}, C)$

end for

Function Calc_Hellinger**Input:** Training set T , Feature f For each value v of f do

$$\text{Hellinger} = \left(\sqrt{\frac{T_{x_f=v, y=+}}{T_{y=+}}} - \sqrt{\frac{T_{x_f=v, y=-}}{T_{y=-}}} \right)^2$$

end for

return $\sqrt{\text{Hellinger}}$ Nitesh Chawla, SIAM 2009
Tutorial

Support Vector Machines

- SVMs are also sensitive to high class imbalance
- Penalty can be specified as a trade-off between the two classes
 - Limitations arise from the Karush Kuhn Tucker conditions

Solutions:

Integrating sampling strategies

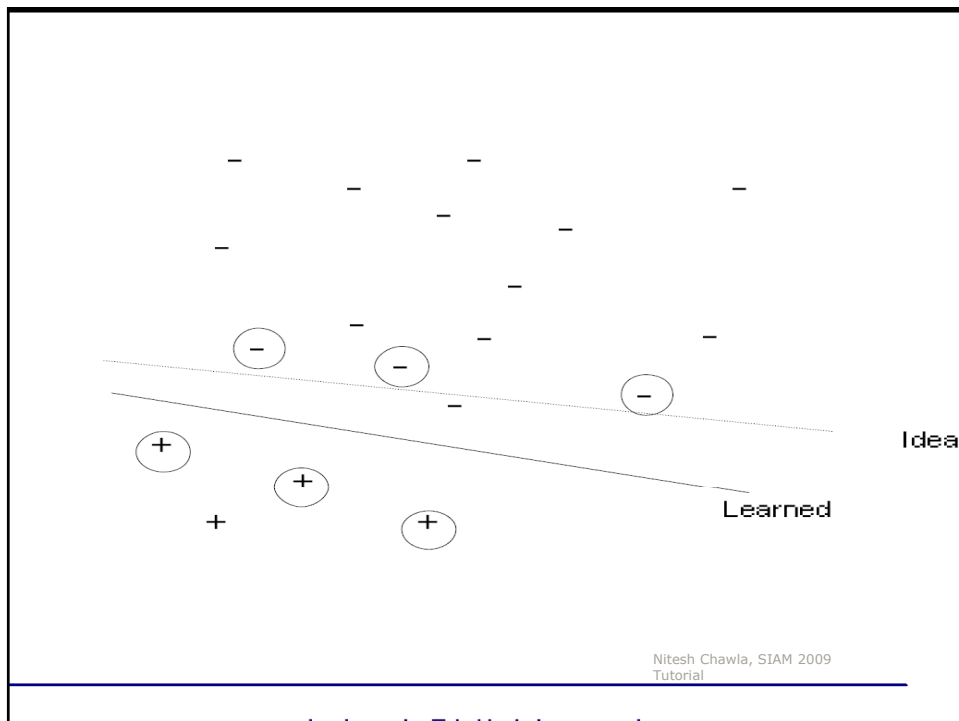
Kernel alignment algorithms

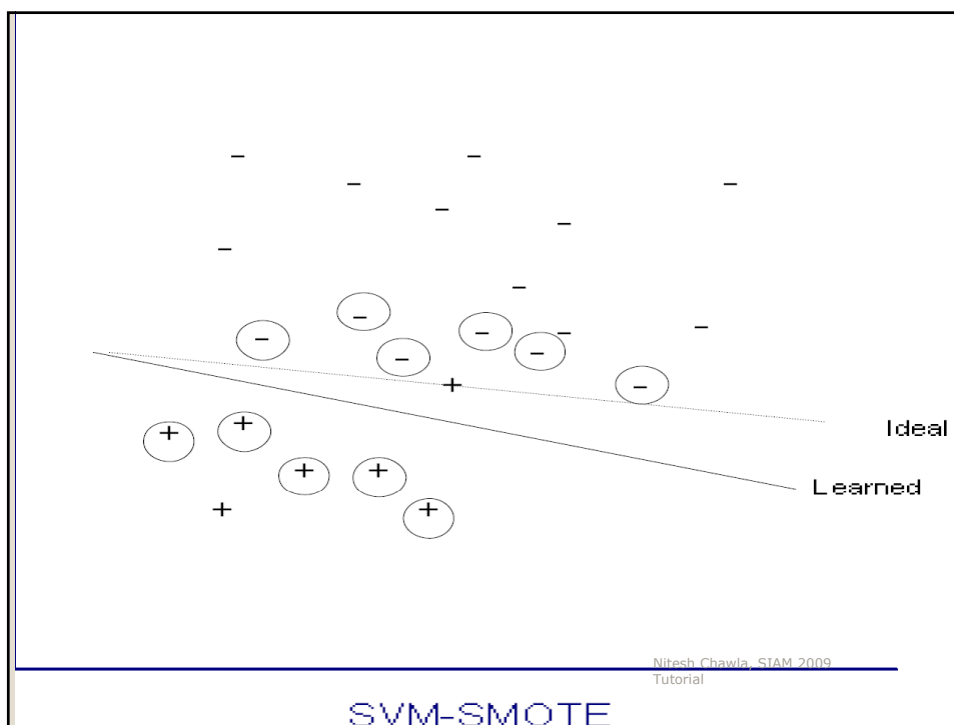
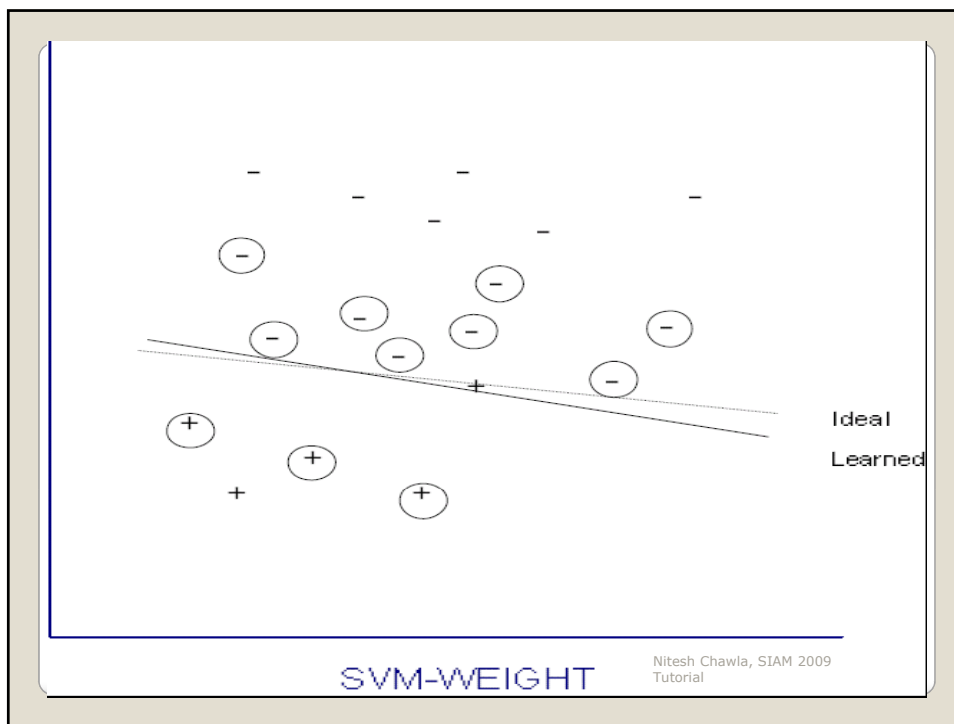
Nitesh Chawla, SIAM 2009
Tutorial

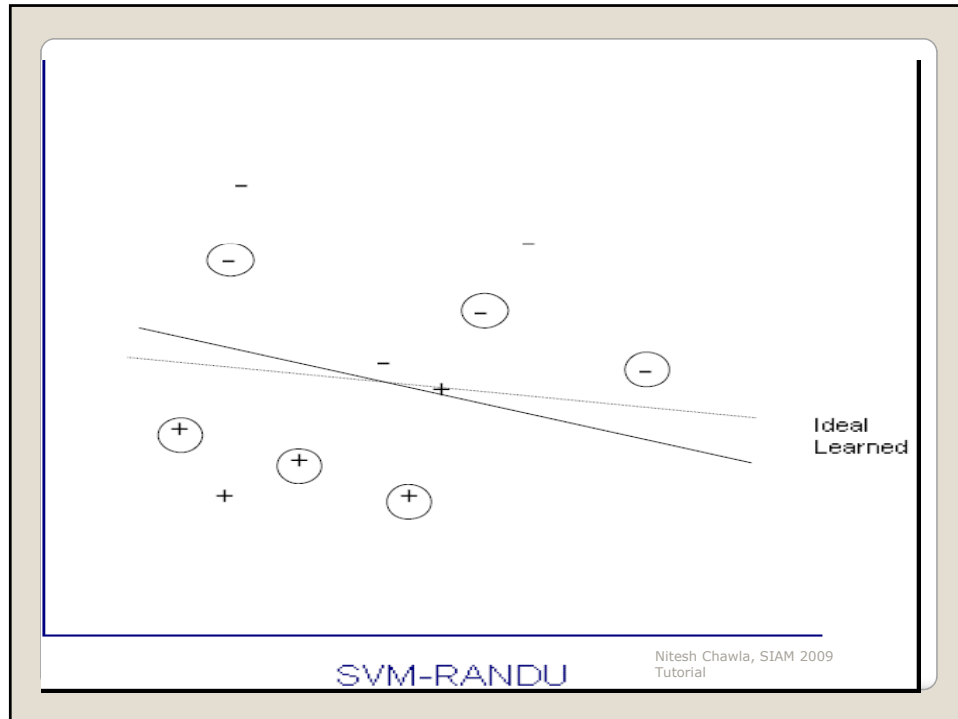
- Hellinger distance is strongly skew insensitive
- More robust for class imbalance as compared to Gini and Information gain
- Recommended decision tree splitting criterion

Recommendations

Nitesh Chawla, SIAM 2009
Tutorial







Kernel Boundary Alignment

- Adaptively modify K (kernel) based on training set distribution
- Addresses
 - Improving class separation
 - Safeguarding overfitting
 - Improving imbalanced ratio

Overview

- Introduction
- Sampling Methods
- Moving Decision Threshold
- Classifiers' Objective Functions
- Evaluation Measures

Nitesh Chawla, SIAM 2009
Tutorial

Back to the Performance Fruit Bowl

- What evaluation measure to use?
- Is there one validation strategy that we can embrace?

Nitesh Chawla, SIAM 2009
Tutorial

		<i>Truth Value</i>	
		<i>Positive</i>	<i>Negative</i>
<i>Prediction Value</i>	<i>Positive</i>	<i>True Positive (TP)</i>	<i>False Positive (FP)</i>
	<i>Negative</i>	<i>False Negative (FN)</i>	<i>True Negative (TN)</i>

Truth Table

Nitesh Chawla, SIAM 2009
Tutorial

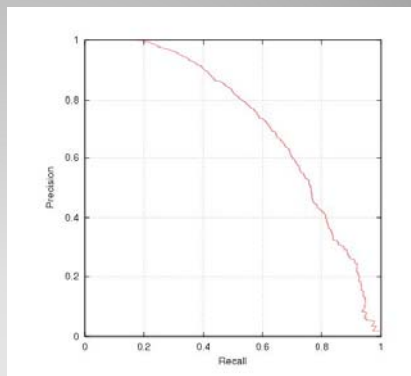
f-measure

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot (\text{precision} \cdot \text{recall}) / (\text{precision} + \text{recall}).$$

- Top 20 Precision
- Top 20 Recall
- Mean averaged precision
- Precision Recall Curves (sweeping across thresholds)



Source of this Figure: Rich Caruana

More on Precision and Recall

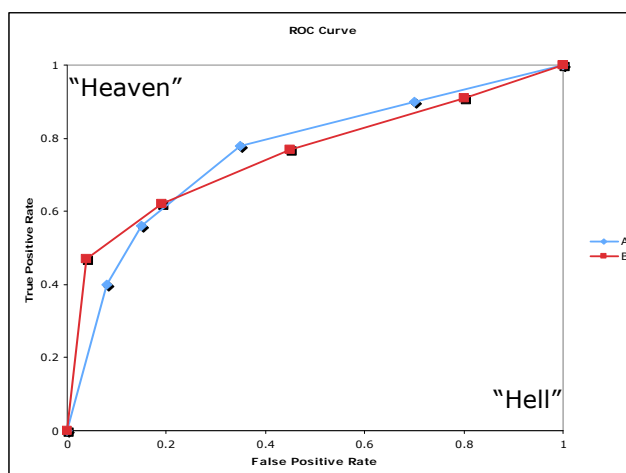
Nitesh Chawla, SIAM 2009
Tutorial

- Balanced Accuracy = $\frac{Accuracy_+ + Accuracy_-}{2}$
- G-mean = $\sqrt{Accuracy_+ \times Accuracy_-}$

Balanced Accuracy and G-mean

Nitesh Chawla, SIAM 2009
Tutorial

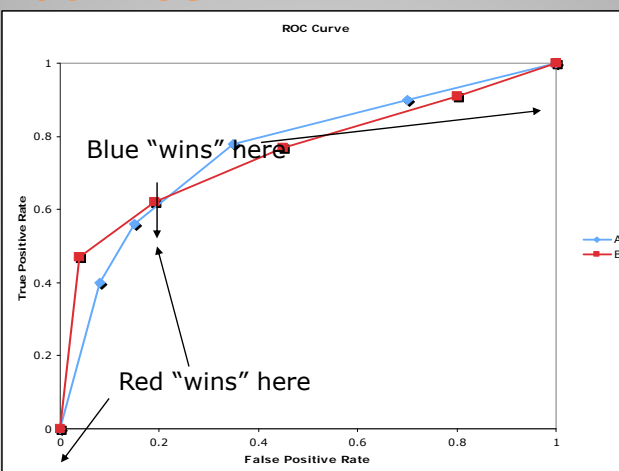
ROC Curves



Nitesh Chawla, SIAM 2009 Tutorial

83

ROC Curves



Nitesh Chawla, SIAM 2009 Tutorial

$$A = \frac{I_1 - n_1(n_1 + 1)/2}{n_0 n_1}$$

I_1 : sum of ranks of all class 1 examples

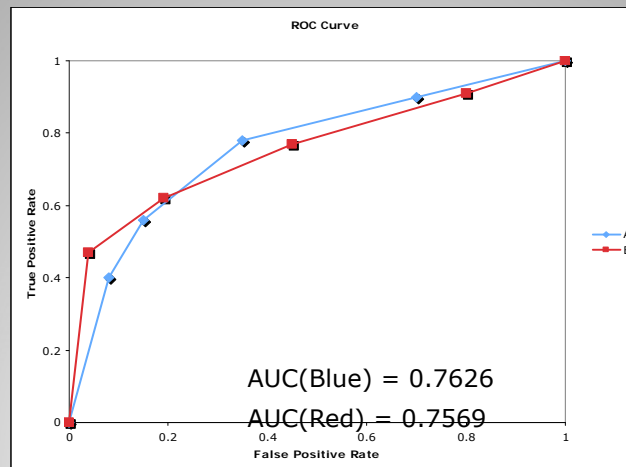
n_0 : number of class 0 examples

n_1 : number of class 1 examples

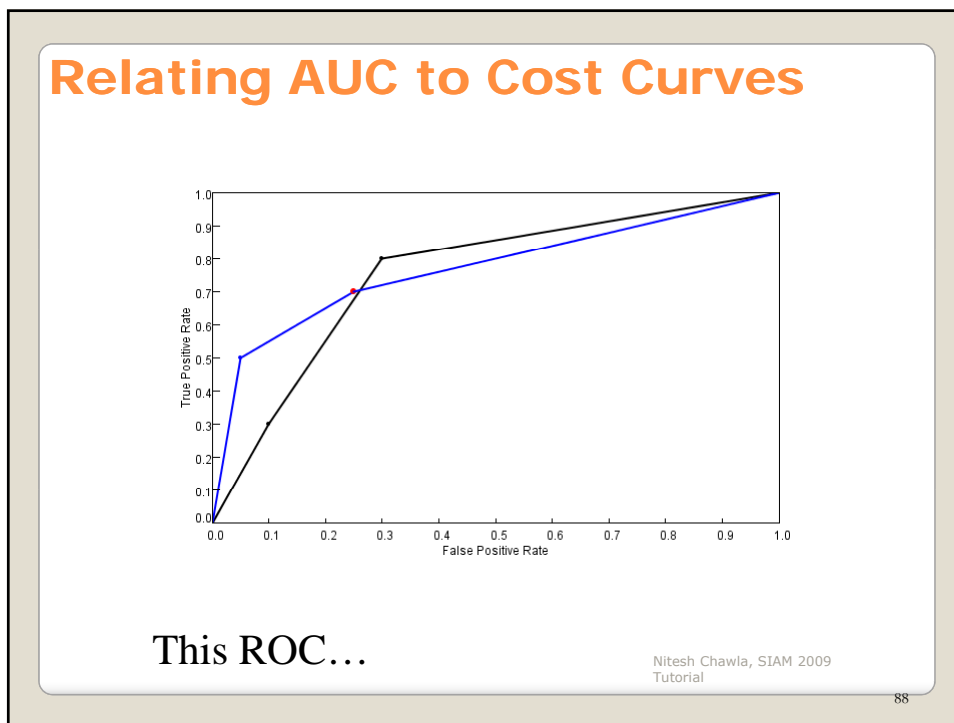
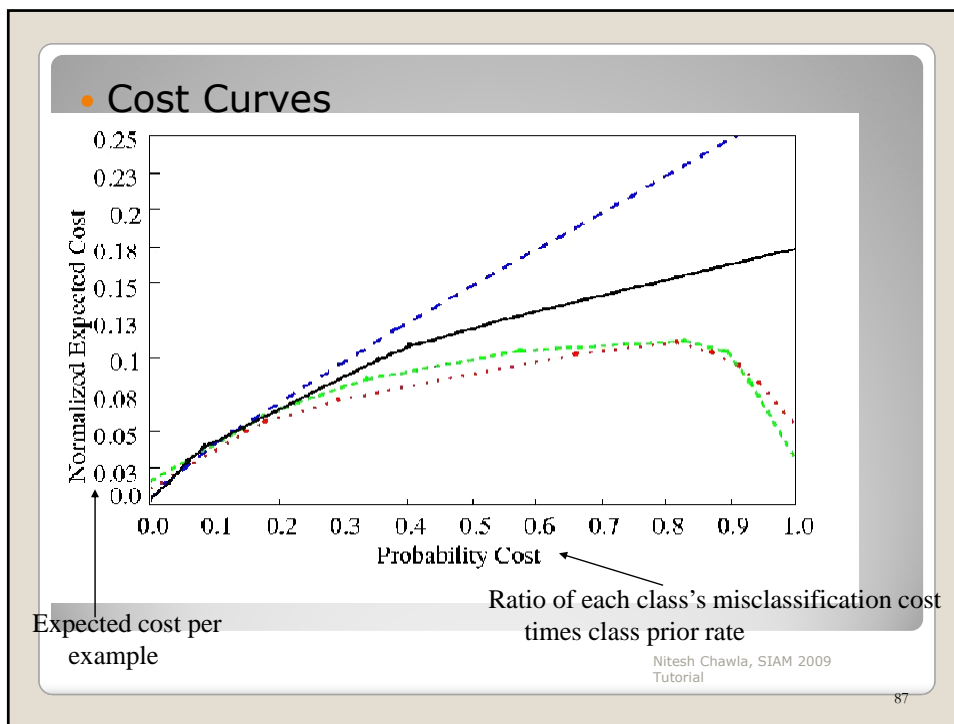
AUC

Nitesh Chawla, SIAM 2009
Tutorial

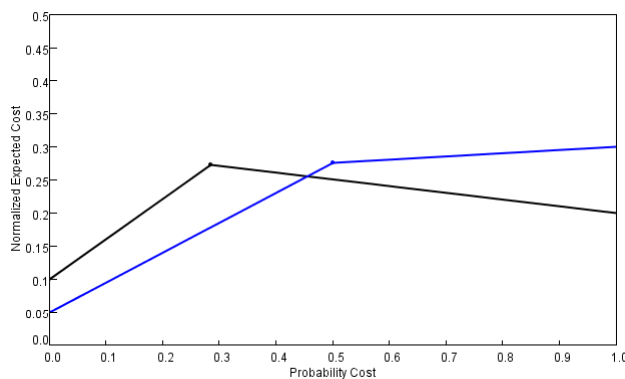
AUC: Area Under the ROC Curve



Nitesh Chawla, SIAM 2009
Tutorial



Relating AUC to Cost Curves



...converts to this Cost Curve Nitesh Chawla, SIAM 2009 Tutorial

89

Cost and Benefits

	Actual Negative	Actual Positive
Predict Negative	b ₀₀	b ₀₁
Predict Positive	b ₁₀	b ₁₁

	Actual Negative	Actual Positive
Predict Negative	TN	FN
Predict Positive	FP	TP

$$B_N = (1 - P_k)b_{00} + P_k b_{01}$$

$$B_P = (1 - P_k)b_{10} + P_k b_{11}$$

Costs

Nitesh Chawla, SIAM 2009 Tutorial

Benefit of Non-Default

$$b_{00}(k, x)(1 - P_k) > (1 - P_k)b_{10} + P_k b_{11} - P_k b_{01}(x)$$

$$b_{00}(k, x) > \frac{(1 - P_k)b_{10} + P_k b_{11} - P_k b_{01}(x)}{(1 - P_k)}$$

$$\therefore NPV = (1 - P_k)b_{00} - (1 - P_k)b_{01} + P_k b_{11} - P_k b_{10}$$

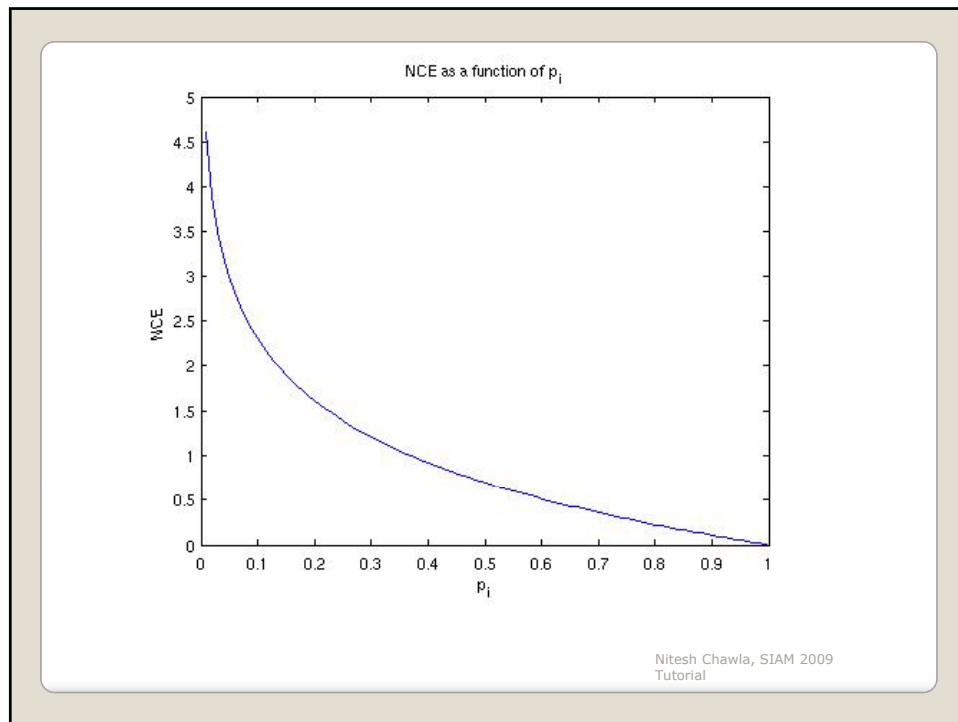
$$\equiv (1 - P_k)b(TN) - (1 - P_k).C(FP) + P_k.b(TP) - P_k.C(FN)$$

Nitesh Chawla, SIAM 2009
Tutorial

Quality of Posterior Probability Estimate

$$NCE = -\frac{1}{n} \left\{ \sum_{i|y=1} \log(p(y=1 | x_i)) + \sum_{i|y=0} \log(1 - p(y=1 | x_i)) \right\}$$

Nitesh Chawla, SIAM 2009
Tutorial

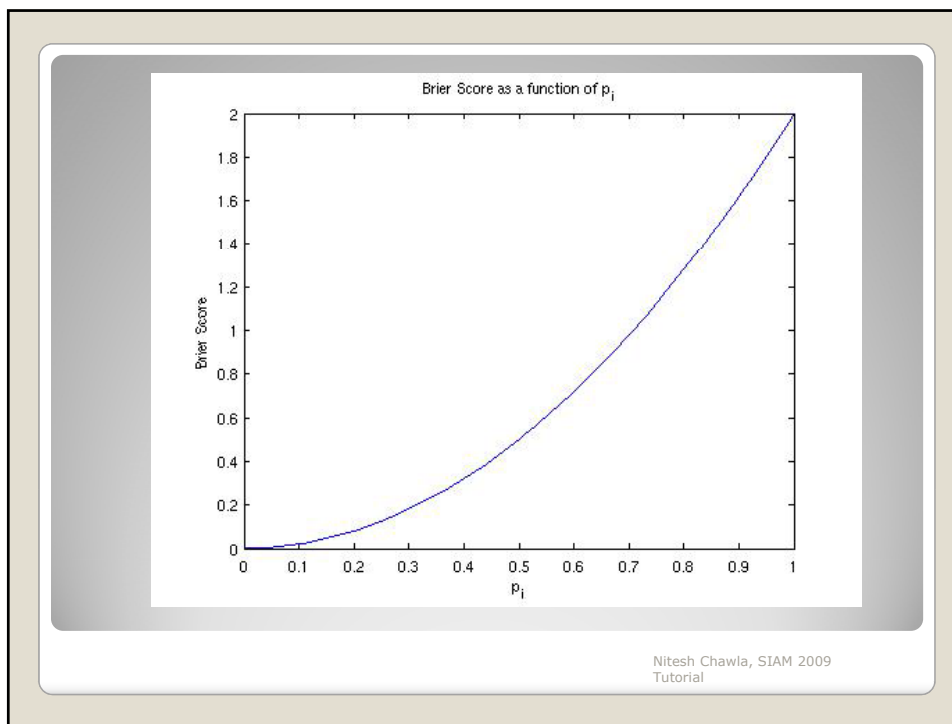


Brier Score

- Average Quadratic Loss on each test instance

$$QL = \frac{1}{n} \sum_{i=1}^n (y_i - p_i)^2$$

- Indicative of best estimates at true probabilities
- accounts for probability assignments to all classes



What are we really evaluating then?

- Rank-order?
- Quality of probability estimates?
- Precision, Recall (and f-measure) at a threshold?
- Balanced accuracy or g-mean (again at a threshold)
- An operating point on ROC curve?
- Costs?

*Different measures have different sensitivities
Call to the community: Let us standardize.*

Nitesh Chawla, SIAM 2009
Tutorial

Step one, choosing the validation strategy

Nitesh Chawla, SIAM 2009
Tutorial

Step two, comparing and contrasting evaluation measures

Nitesh Chawla, SIAM 2009
Tutorial

Step three, computing statistical significance

Nitesh Chawla, SIAM 2009
Tutorial

Step four, some recommendations and call to the community

Nitesh Chawla, SIAM 2009
Tutorial

Discussion

- Need for larger datasets
 - A benchmark repository
- Need for many positives and march towards parts-per-million
- Need for standardization in evaluation
- Need for full parameter disclosure in papers

Nitesh Chawla, SIAM 2009
Tutorial

Datasets and Software

- Available via
 - <http://www.nd.edu/~dial>
 - Email me: nchawla@nd.edu

Nitesh Chawla, SIAM 2009
Tutorial

Acknowledgements

- Kevin Bowyer, David Cieslak, George Forman, Larry Hall, Nathalie Japkowicz, Philip Kegelmeyer, Alek Kolcz

Nitesh Chawla, SIAM 2009
Tutorial

Let neither measurement without theory
Nor theory without measurement dominate
Your mind but rather contemplate
A two-way interaction between the two
Which will your thought processes stimulate
To attain syntheses beyond a rational
expectation!

Contributed by A. Zellner.

Nitesh Chawla, SIAM 2009
Tutorial