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

Data, Inference, Analysis and Learning Lab @ ND

Overview

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

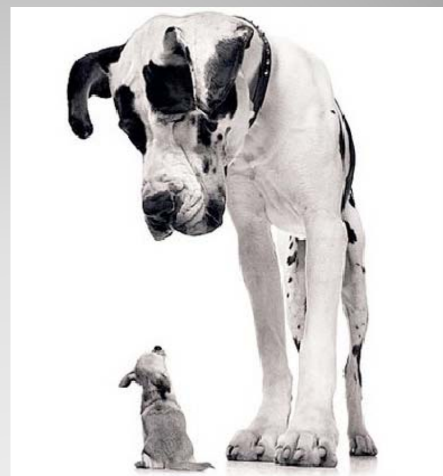
IEEE ICDM noted "Dealing with Non-static, Unbalanced and Cost-sensitive Data" among the **10 Challenging Problems in Data Mining Research**

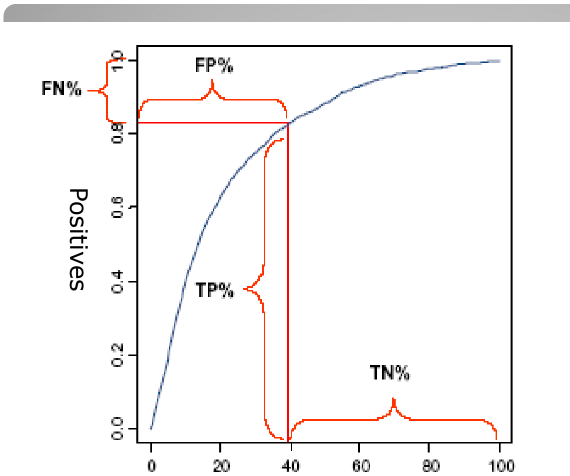
Small Class Matters, and Matters More

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





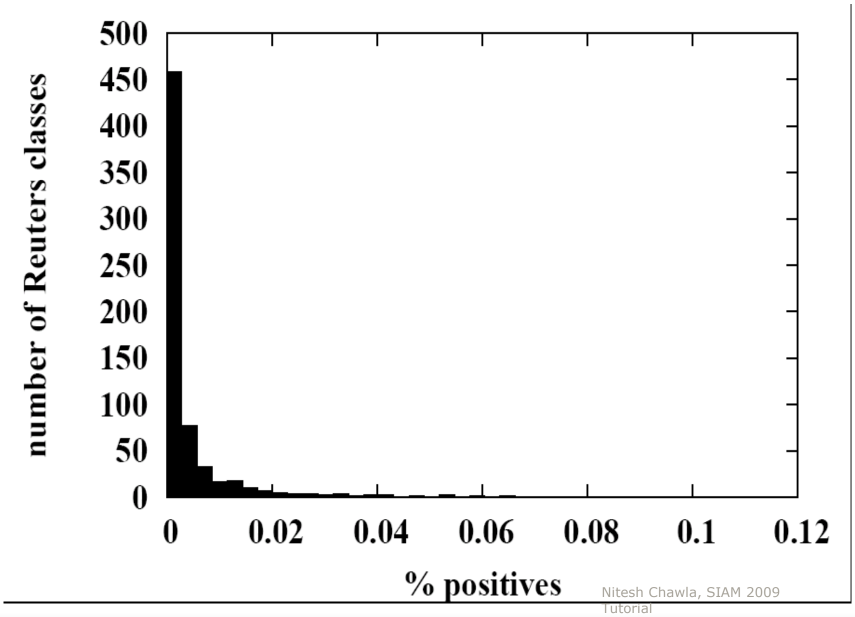
	Actual Non-Default	Actual Default
Predict Non-Default	TN	FN
Predict Default	FP	TP

Typical Prediction Model

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



Nitesh Chawla, SIAM 2009 Tutorial

Paradox of False Positive

- Imagine a disease that has a prevalence of 1 in a mllion 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 10,0000.

Nitesh Chawla, SIAM 2009 Tutorial

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.

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!)

Overview

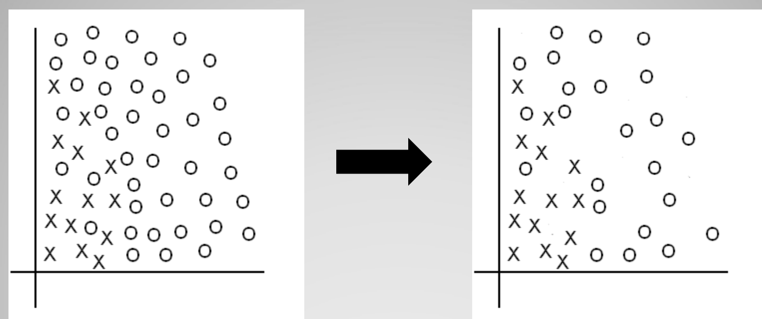
- Introduction
- Sampling Methods
- Moving Decision Threshold
- Classifiers' Objective Functions
- Evaluation Measures

Sampling: Add or remove until satisfactory performance

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

Undersampling

- **Randomly remove majority class examples**



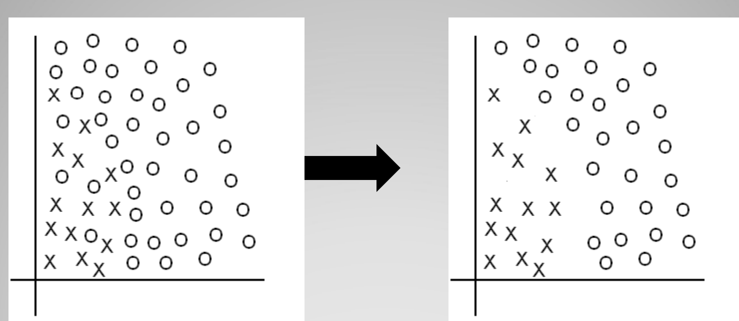
Risk of losing potentially important majority class examples, that help establish the discriminating power

What about focusing on the borderline and noisy examples?

Introducing Tomek Links and Condensed Nearest Neighbor Rule

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 $d(E_j, E_k) < d(E_i, E_j)$.

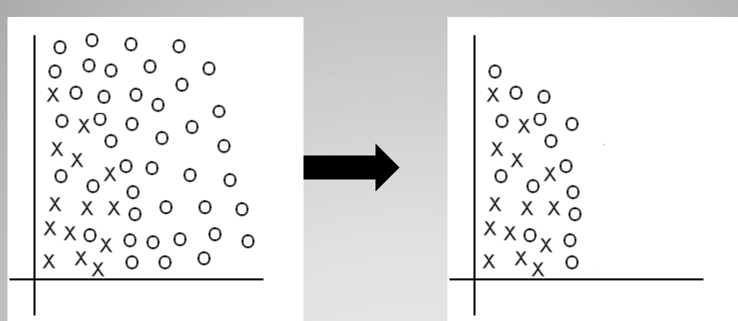


Undersample by Tomek Links

Condensed Nearest Neighbor Rule (CNN rule)

- Find a consistent subset of examples.
  - A subset E'⊆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

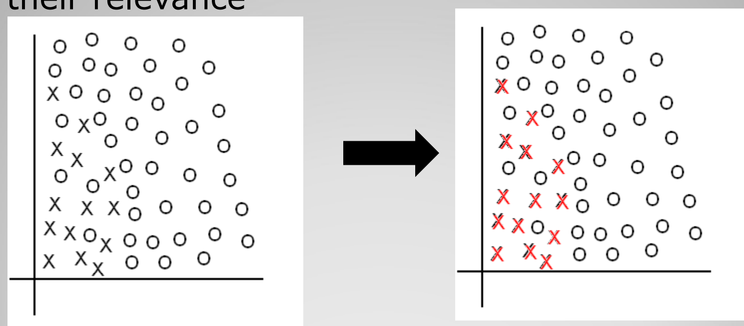


CNN Editing

Nitesh Chawla, SIAM 2009 Tutorial

Oversampling

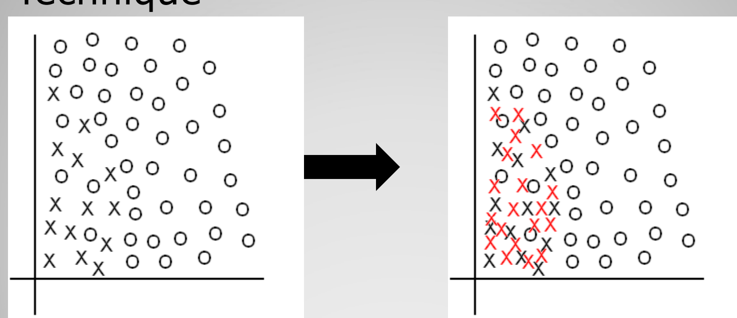
- Replicate the minority class examples to increase their relevance



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

Instead of replicating, let us invent some new instances

- SMOTE: Synthetic Minority Over-sampling Technique



SMOTE

$\forall x \in Minority$

for $i = 1$ to $|x|$

Compute k - NN of x_i

for $j = 1$ to $|x_i|$

if $x_{ij} \equiv$ continuous

$x_{nj} = (x_i - x_{ki}) * \mathrm{rand}(1)$

else

$$x_{nj} = \underset{x' \in x_j}{\arg\max} \sum_{k: x_{jk} = x'} 1$$

endif

endfor

endfor

k- NN(x_j, x_j')

if $j \in continuous$

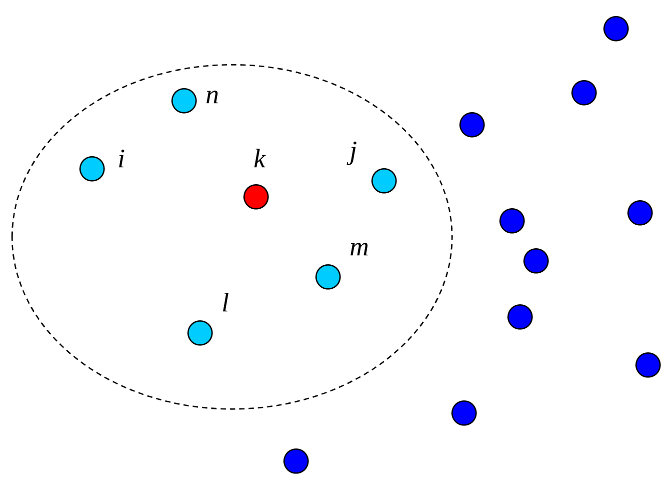
$$\delta_c = (x_j - x_j')^2$$

if $j \in nominal$

$$\delta_n = \sum_{c=1}^{C} \left| \frac{N_{j, x_j, c}}{N_{j, x}} - \frac{N_{j, x_j', c}}{N_{j, x'}} \right| \text{(VDM)}$$

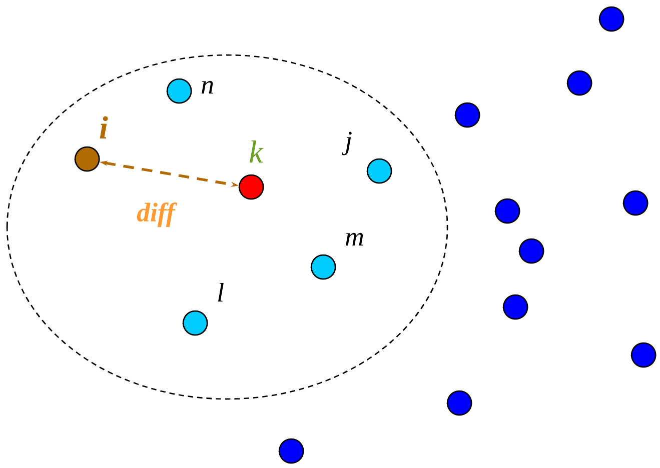
$$\Delta = \delta_c + \delta_n$$

SMOTE



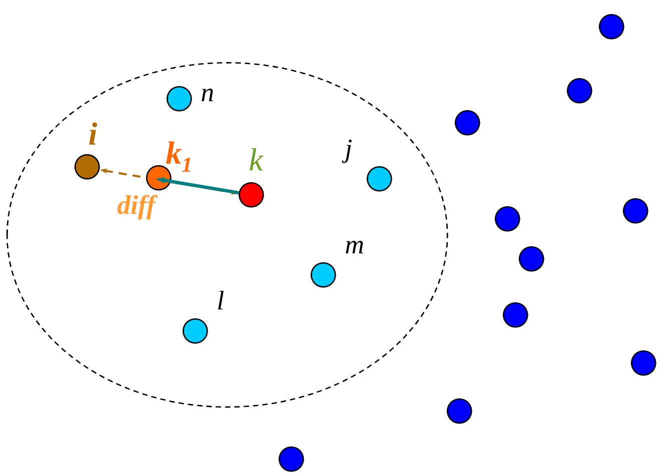
For each minority example *k* compute nearest minority class examples *{i, j, l, n, m}*

SMOTE



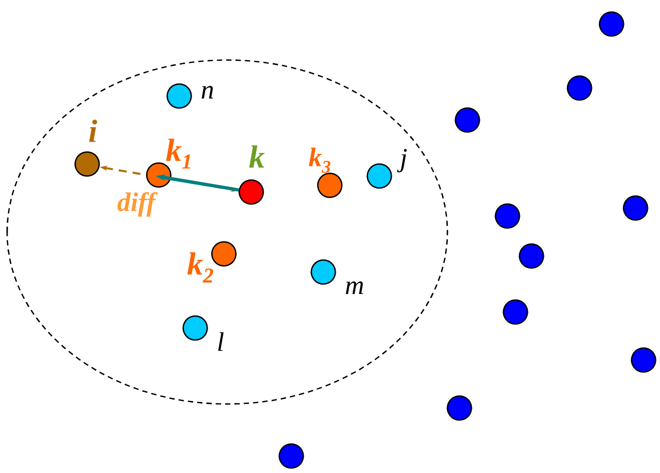
- **Randomly choose an example out of 5 closest points**

SMOTE



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

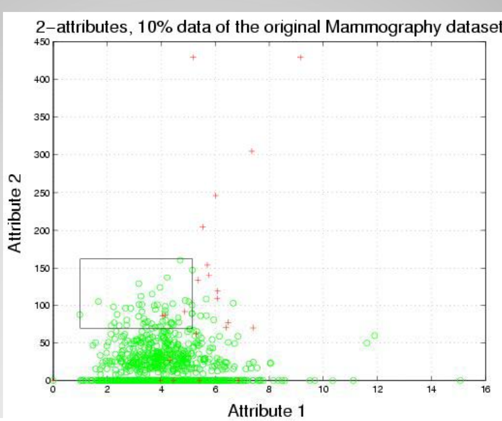
SMOTE



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

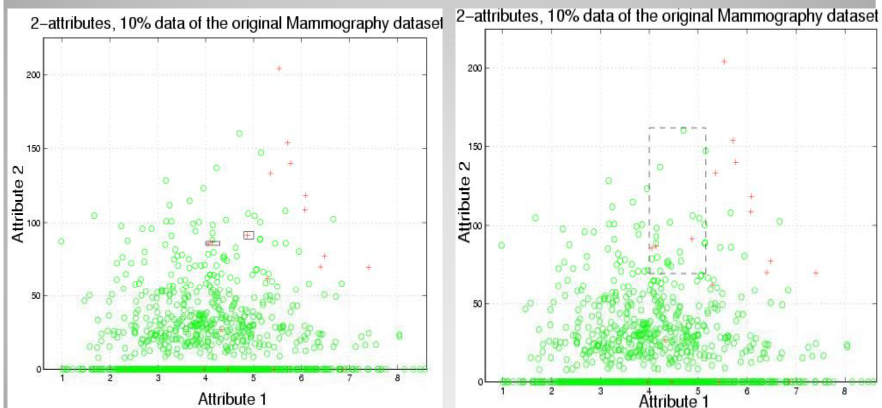
SMOTE

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



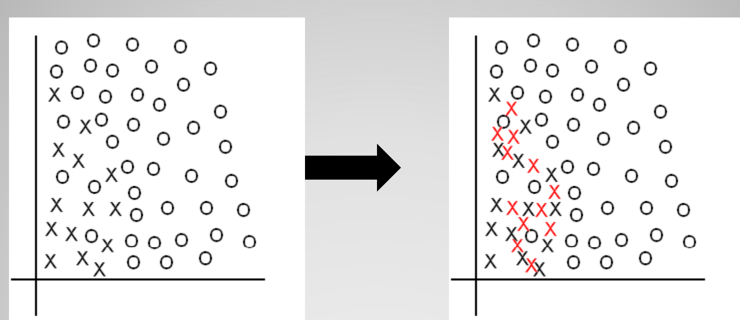
SMOTE

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



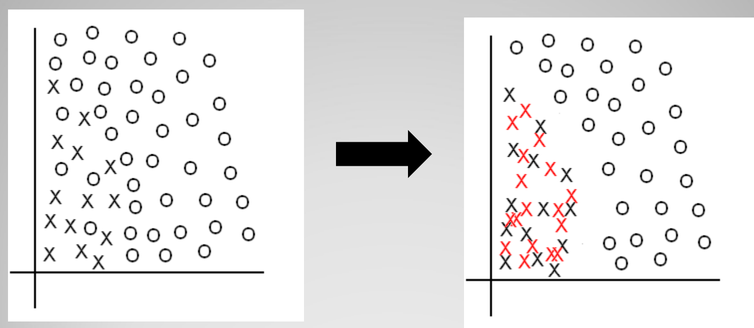
Beware of those lurking majority class examples

- Borderline-SMOTE



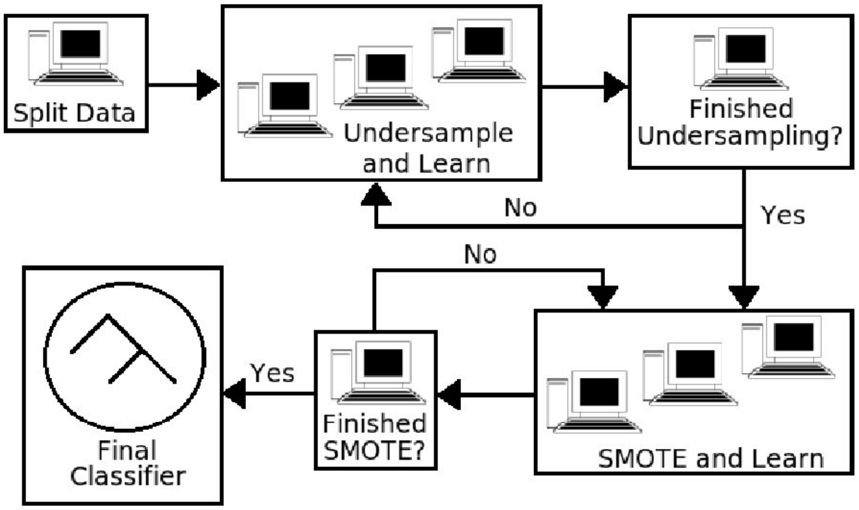
Nitesh Chawla, SIAM 2009 Tutorial

SMOTE + Tomek Link Undersample



- 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

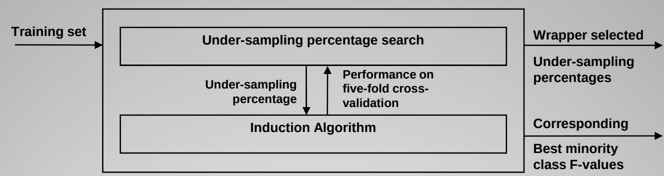


A Scalable Wrapper

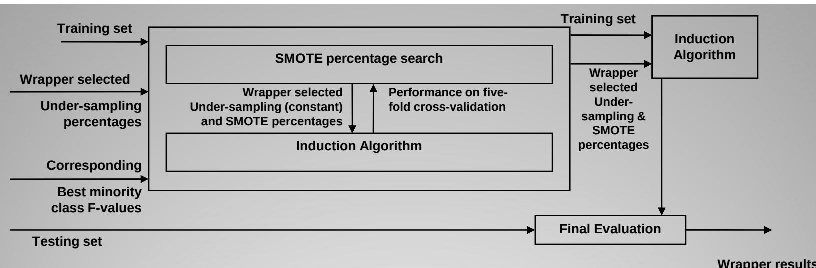
Nitesh Chawla, SIAM 2009 Tutorial

- 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

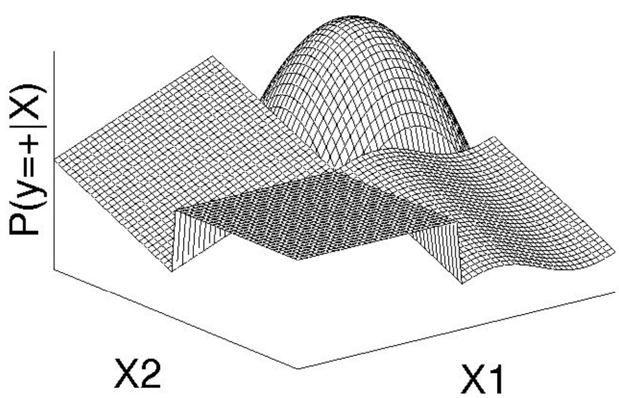


- Greedy search process guided by performance over five-fold cross validation of training data
- Baseline results are results with no under-sampling performed
- Generally the under-sampling of the majority classes increases the average F-value over minority classes, but in some cases can be reduced due to high reduction in precision versus small increase in recall
- Criteria to stop under-sampling for majority class
  - The Average F-value over minority classes reduces or
  - The Average F-value over majority classes reduces more than 5%

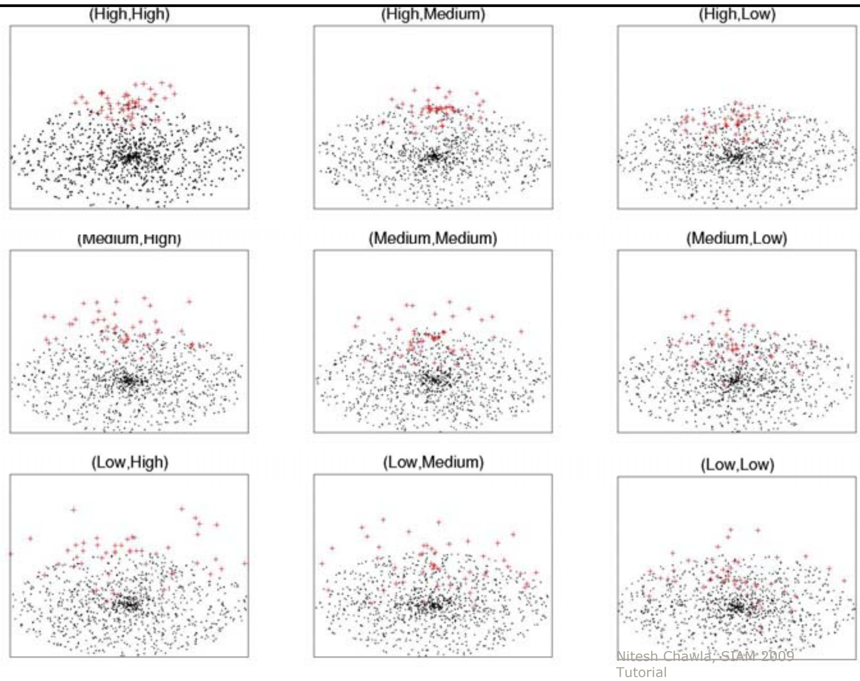


- The under-sampling percentages for majority classes are fixed to those found so far by Wrapper Under-sample Algorithm
- Corresponding Best Results are Baselined
- Criteria to stop SMOTE for minority class
  - The Average F-value over-minority classes reduces and
  - The two look aheads for the minority class are exhausted
- Rationale for SMOTE look ahead : More SMOTE will add more information and will give better accuracies over minority classes as apposed to under-sampling which reduces information thereby reducing the coverage of the build classifier

Exploiting Locality in Sampling

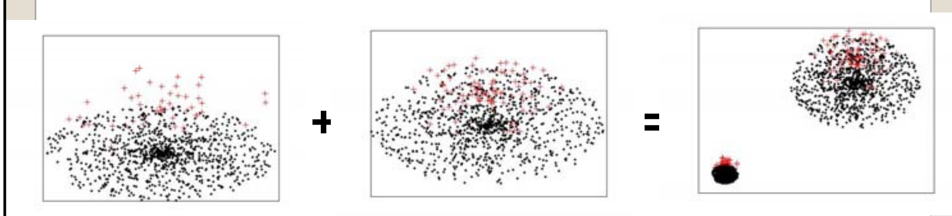


- Class ratio can be important to determining best sampling levels to use
- Other properties may exert greater influence
  - Overlap
  - Density
- Consider the following examples



All have same class ratio: (50:1000)

(density,separation)	AUROC C45	AUROC C45+S	Undersample	Smote
(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



(U,S) = (50,350)

AUROC = 0.906

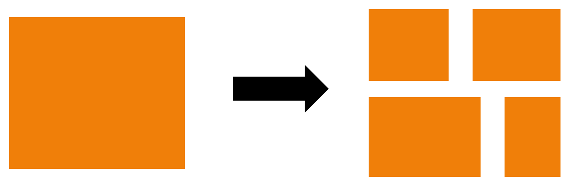
(U,S) = (80,250)

AUROC = 0.812

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

AUROC = ?????

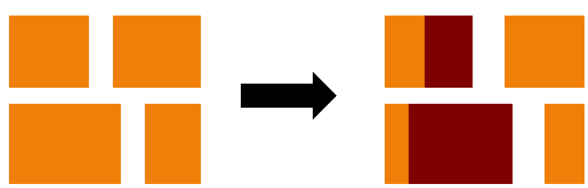
Step 1: Split the Data



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

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)

Step 3: Train/predict globally

+ C4.5 → Final Classifier

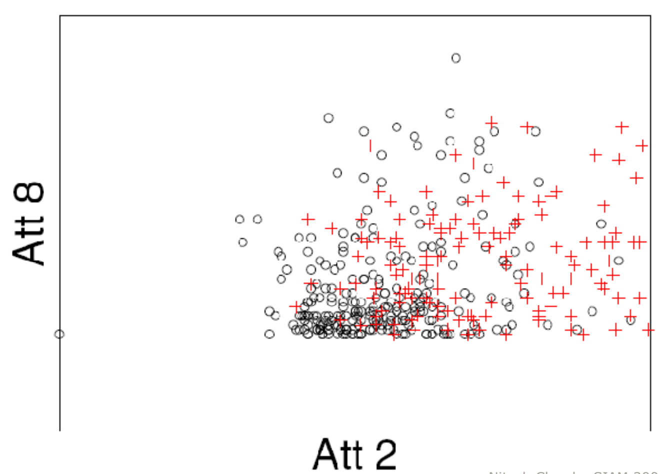
Localized Sampling Framework

Nitesh Chawla, SIAM 2009 Tutorial

[Cieslak, Chawla ICDM 2008]

43

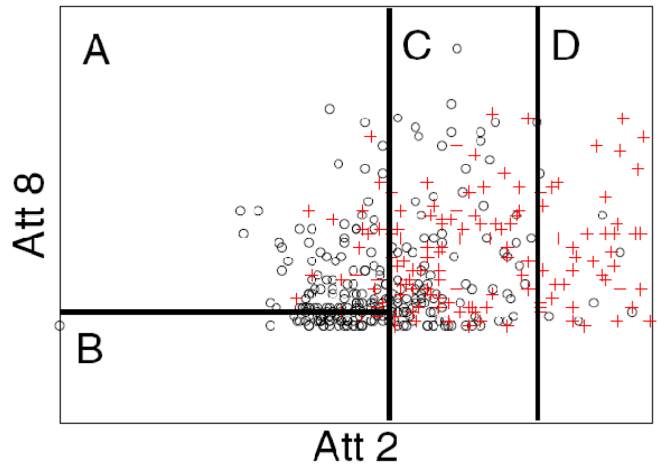
Example - Pima



Nitesh Chawla, SIAM 2009 Tutorial

44

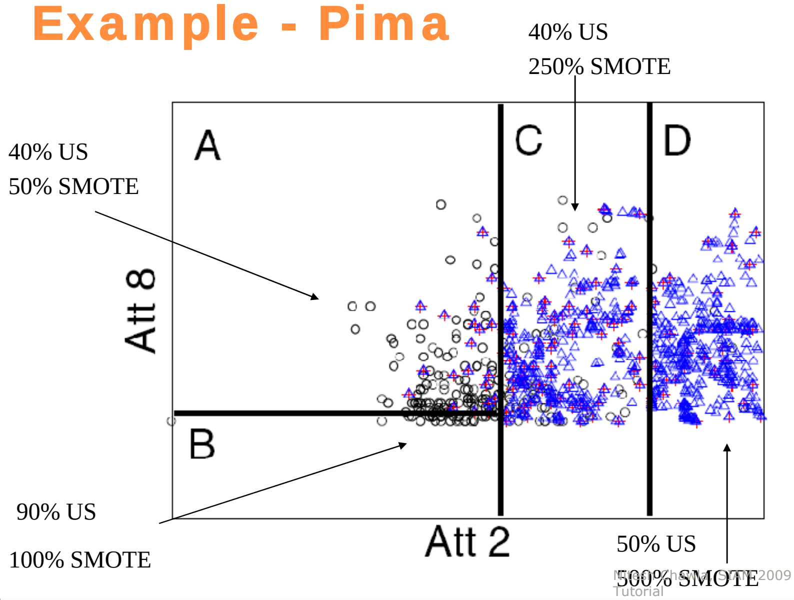
Example - Pima



Nitesh Chawla, SIAM 2009 Tutorial

45

Example - Pima



Nitesh Chawla, SIAM 2009 Tutorial

46

Experimental Results

Alternative localized sampling method using clustering

Standard Wrapper

	C4.5	SVM	C4.5+ COGOS	SVM+ COGOS	C4.5+ UnSM	C4.5+ LS
Avg. Rank	4.850	4.900	3.650	3.530	2.750	1.330

AUROC rank averaged across 20 datasets

Our Method

Overview

- Introduction
- Sampling Methods
- <u>Moving Decision Threshold</u>
- Classifiers' Objective Functions
- Evaluation Measures

- 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

Changing Decision Thresholds

- Decisions of (scoring) classifiers are typically set at 0.5
  - $P(x > 0.5)$ is class 1 and $P(x <= 0.5)$ 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

- 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

Overview

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

Beyond Sampling: Adapting Classifiers

- Consider
  - Decision Trees
  - SVMs

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)

Converting decision tree leaf predictions into probability estimates

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

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

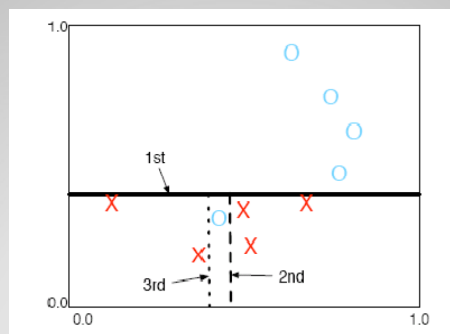
$$P_{mest} = \frac{(TP + bm)}{(TP + FP + m)}$$

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

Some of the skew insensitive metrics proposed

Decision tree (im)purity metrics

Partition feature space to maximize purity at leaves. Recurse



Entropy (Information Gain) as an impurity

(Q,W) classes of interest

N = number of samples

N_i = number of samples in class i

N^S = number of samples in L/R split

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

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

Consider a skew insensitive criterion

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

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

Formulating for decision tree

Consider a countable space

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

"Distance" in the normalized frequencies space

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

Inf. Gain vs. Hellinger distance

(Q,W) classes of interest

N

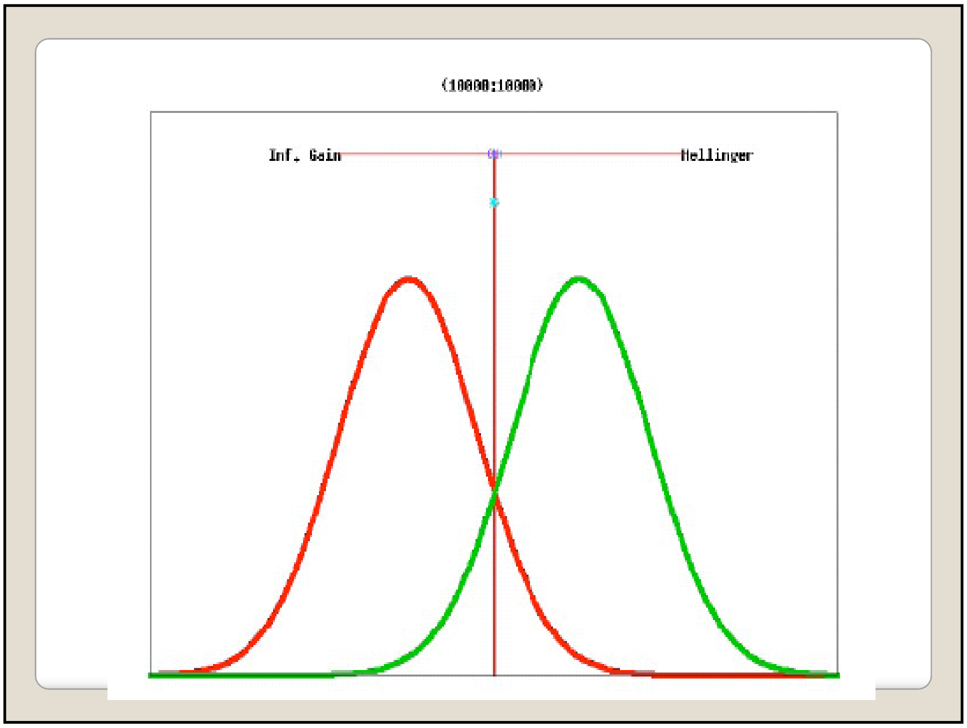
N_i = number of samples in class i

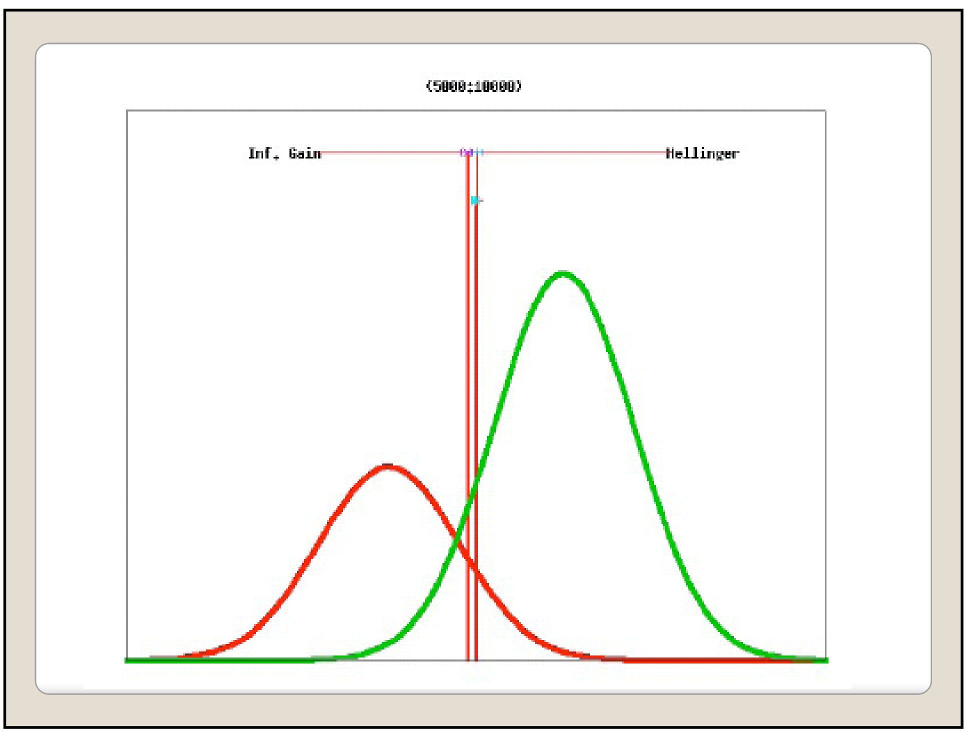
N^S = number of samples in L/R

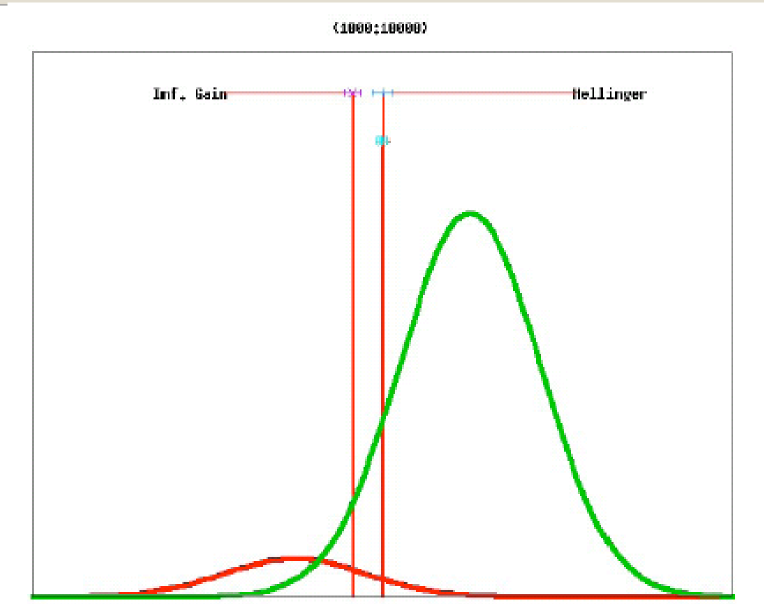
N_i^S = number of samples in class i is L/R split

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

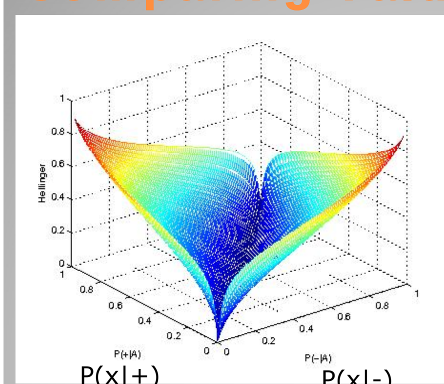
$$H = \sqrt{\left\{\sqrt{\frac{N_Q^L}{N_Q}} - \sqrt{\frac{N_W^L}{N_W}}\right\}^2 + \left\{\sqrt{\frac{N_Q^R}{N_Q}} - \sqrt{\frac{N_W^R}{N_W}}\right\}^2}$$


(10000:10000)
Inf. Gain
Hellinger

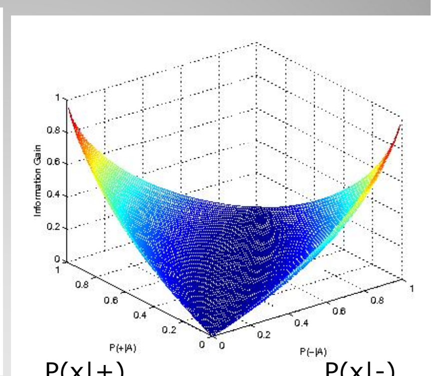

(5000:10000)
Inf. Gain
Hellinger



Comparing Value Surfaces



Hellinger Distance

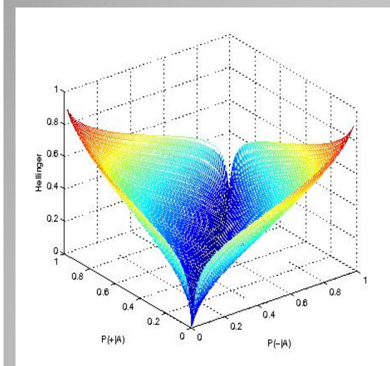


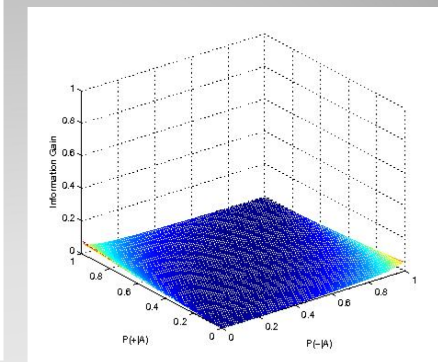
Information Gain

Class ratio +:- = 1:1

Nitesh Chawla, SIAM 2009 Tutorial

Comparing Value Surfaces





Hellinger Distance

Information Gain

Class ratio +:- = 1:100

Hellinger vs. DKM as decision tree splitting criteria

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

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

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

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

Algorithm HDDT

Input: Training Set *T*, Cutoff size *C*

if |*T*| < *C* **then**

return

end if

for each feature *f* of *T* **do**

$H_f = Calc\_Hellinger(T,f)$

end for

b = max(H) (best feature)

for each branch *v* of *b* **do**

$HDDT(T_{xb=v}, C)$

end for

Function Calc_Hellinger

Input: Training set *T*, Feature *f*

For each value *v* of *f* **do**

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

end for

return √Hellinger

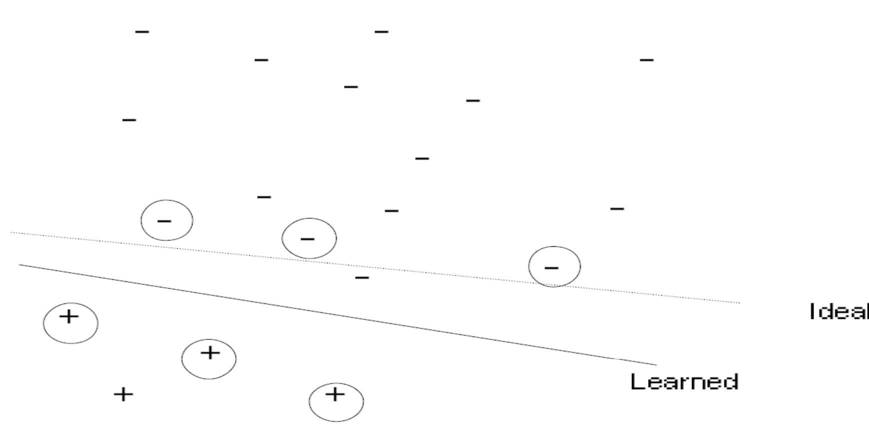
Support Vector Machines

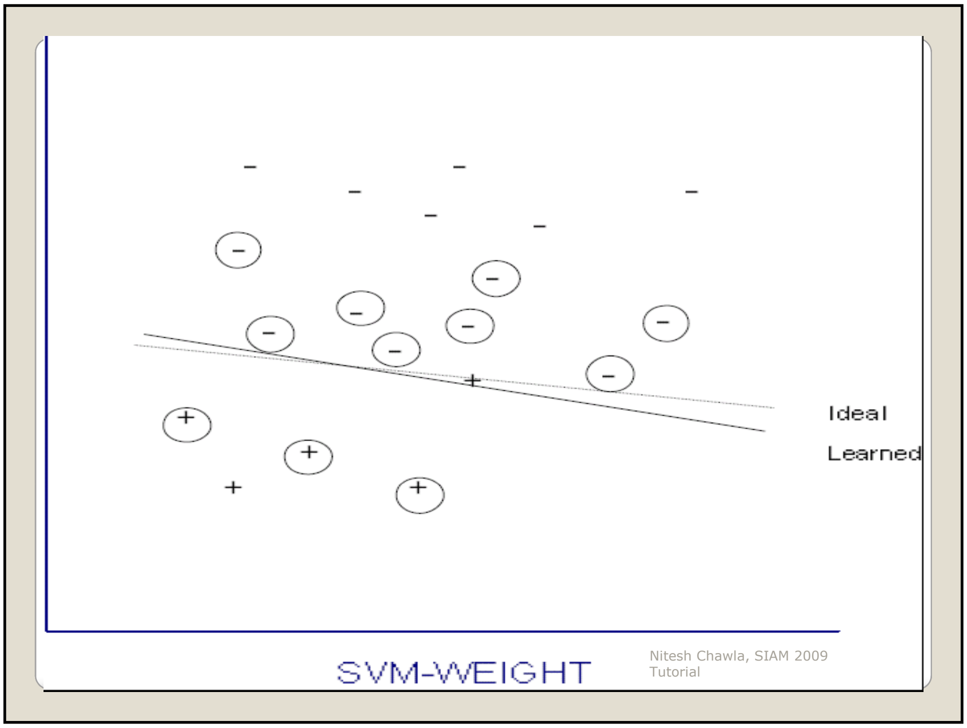
- SVMs are also sensitive to high class imbalance
- Penalty can be specified as a trade-off between the two classes
  - Limitations arise from the Karush Kuhn Tucker conditions

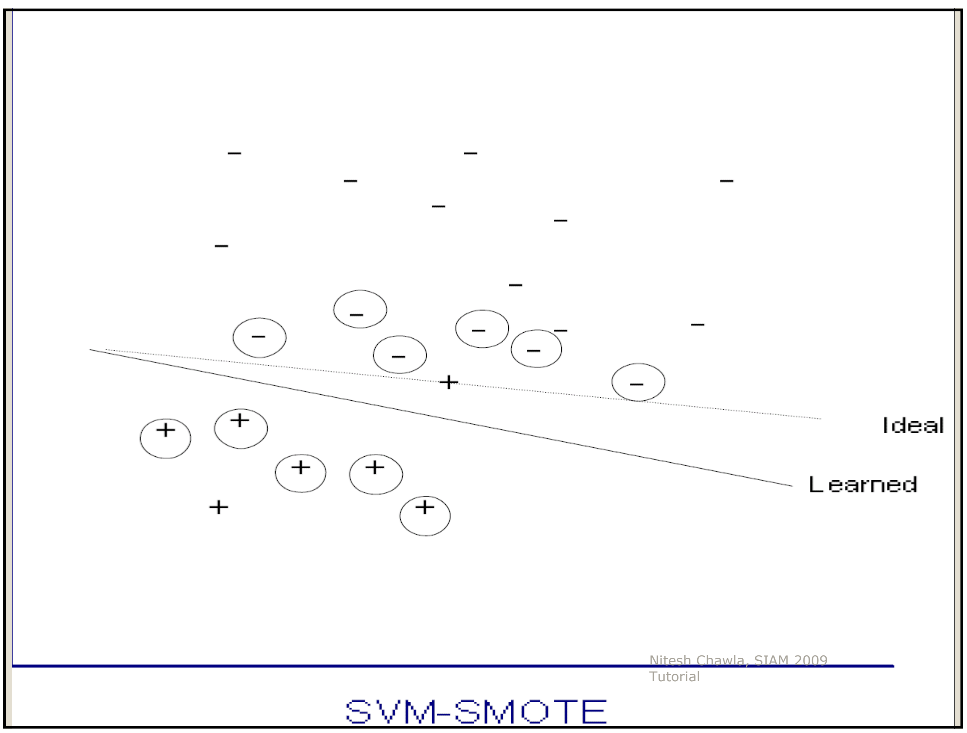
<u>Solutions</u>:
Integrating sampling strategies
Kernel alignment algorithms

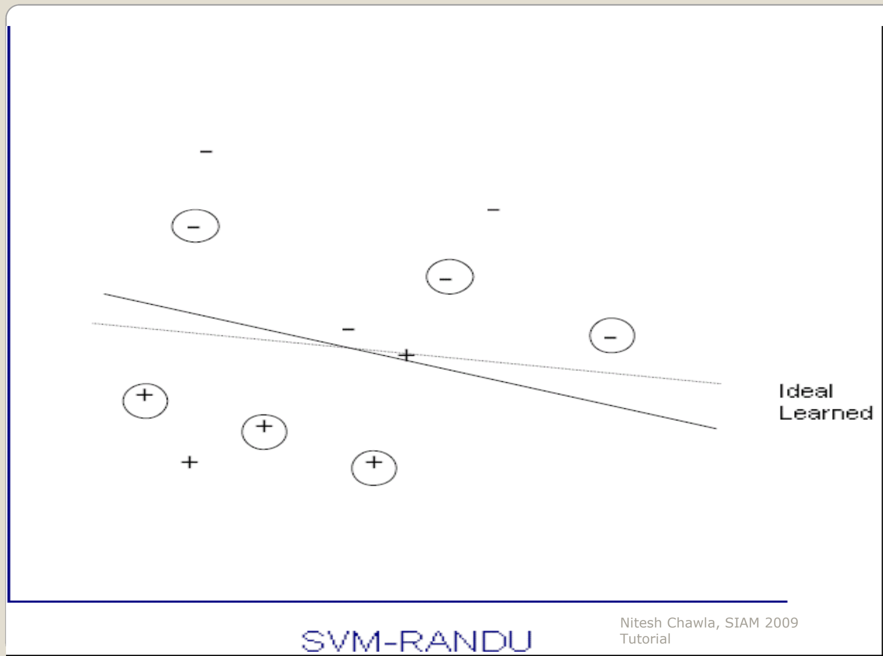
- Hellinger distance is strongly skew insensitive
- More robust for class imbalance as compared to Gini and Information gain
- Recommended decision tree splitting criterion

Recommendations




Ideal
Learned
SVM-WEIGHT
Nitesh Chawla, SIAM 2009 Tutorial


Ideal
Learned
Nitesh Chawla, SIAM 2009 Tutorial
SVM-SMOTE



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

Nitesh Chawla, SIAM 2009 Tutorial

Overview

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

Back to the Performance Fruit Bowl

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

Truth Table

		Truth Value	
		Positive	Negative
Prediction Value	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

f-measure

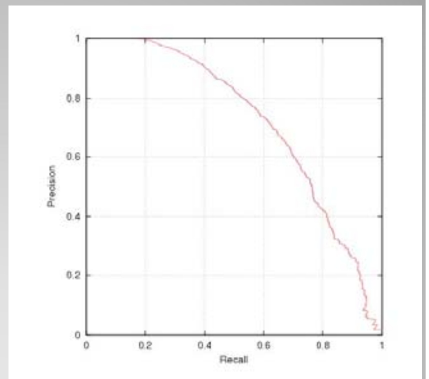
$$\text{Precision} = \frac{tp}{tp + fp}$$

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

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

More on Precision and Recall

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

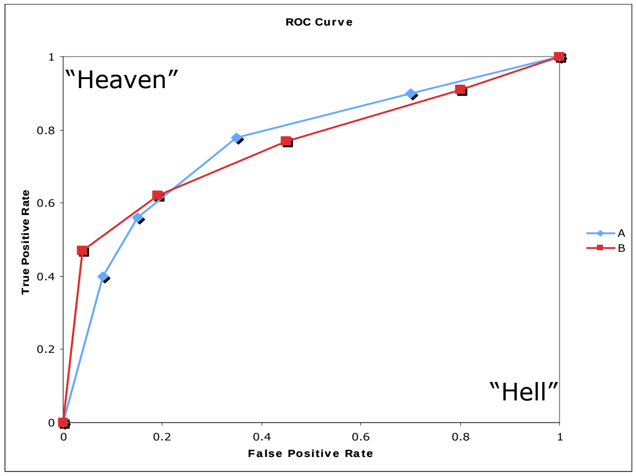


Source of this Figure: Rich Caruana

Balanced Accuracy and G-mean

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

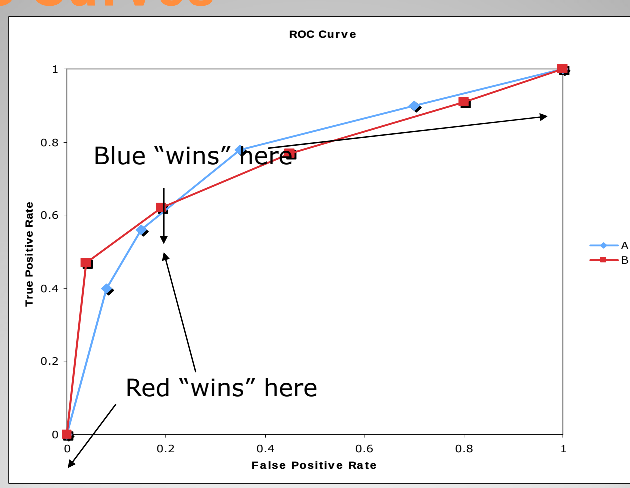
ROC Curves



Nitesh Chawla, SIAM 2009 Tutorial

83

ROC Curves



Nitesh Chawla, SIAM 2009 Tutorial

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

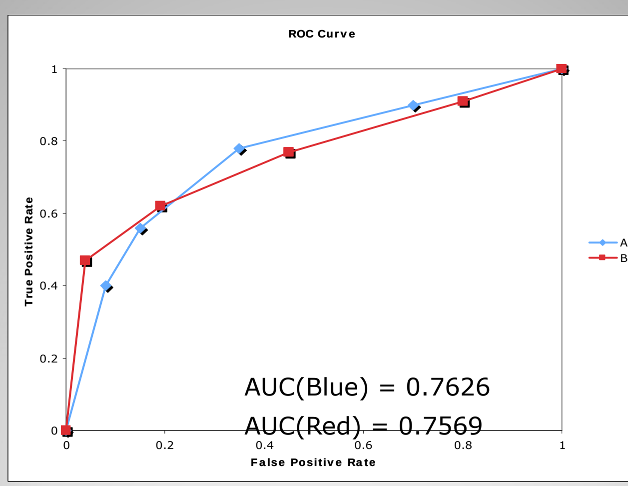
I_1 : sum of ranks of all class 1 examples

n_0 : number of class 0 examples

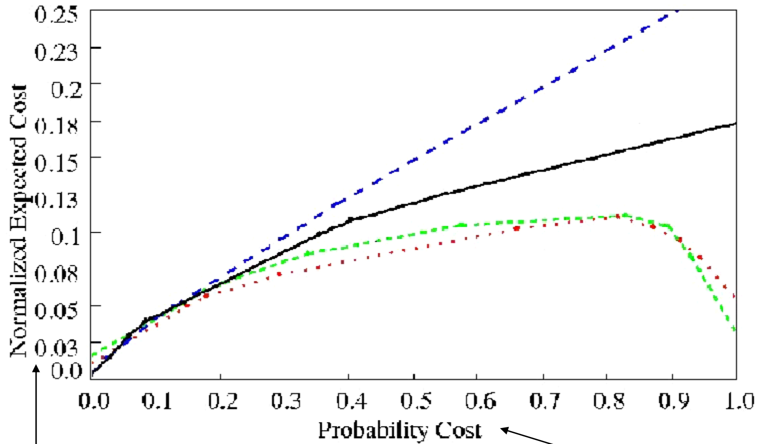
$n1$:number of class 1 examples

AUC

AUC: Area Under the ROC Curve



- Cost Curves



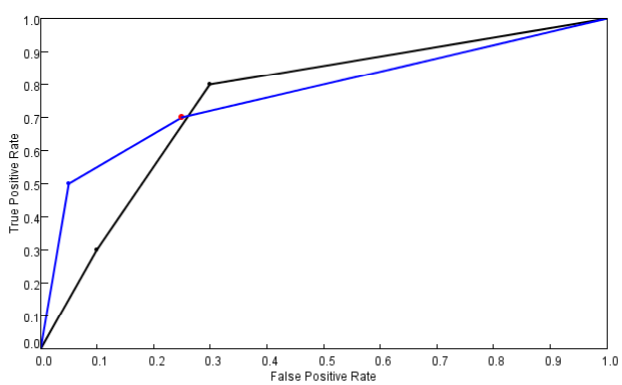
Expected cost per example

Ratio of each class's misclassification cost times class prior rate

Nitesh Chawla, SIAM 2009 Tutorial

87

Relating AUC to Cost Curves

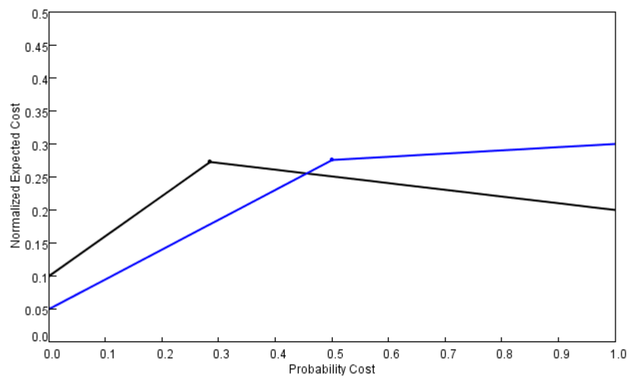


This ROC…

Nitesh Chawla, SIAM 2009 Tutorial

88

Relating AUC to Cost Curves



…converts to this Cost Curve

Cost and Benefits

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

	Actual Negative	Actual Positive
Predict Negative	b00	b01
Predict Positive	b10	b11

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

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

Costs

Benefit of Non-Default

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

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

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

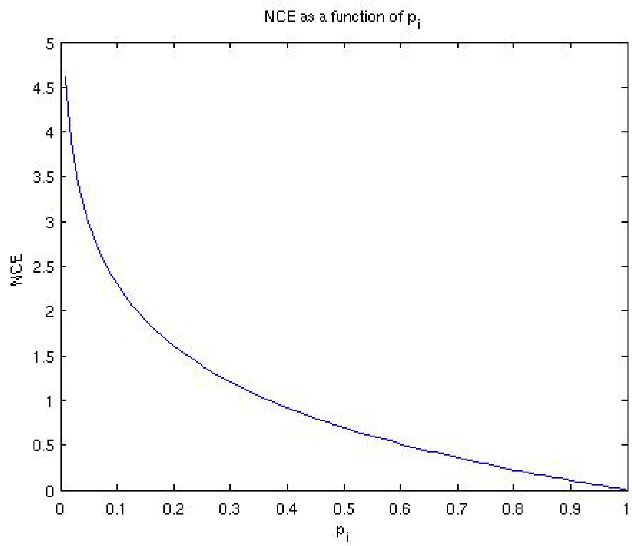
$$\equiv (1-P_k)b(TN) - (1-P_k).C(FP) + P_k.b(TP) - P_k.C(FN)$$

Nitesh Chawla, SIAM 2009 Tutorial

Quality of Posterior Probability Estimate

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

Nitesh Chawla, SIAM 2009 Tutorial



Nitesh Chawla, SIAM 2009 Tutorial

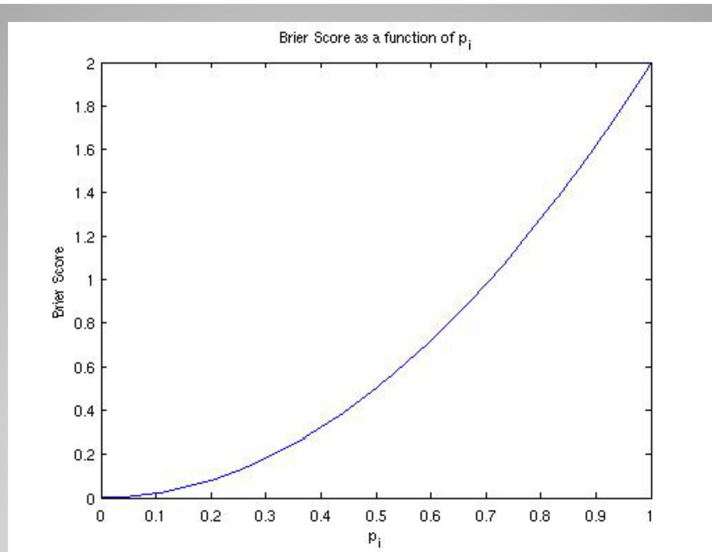
Brier Score

- 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

Nitesh Chawla, SIAM 2009 Tutorial



Nitesh Chawla, SIAM 2009 Tutorial

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

Step two, comparing and contrasting evaluation measures

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

Datasets and Software

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

Acknowledgements

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

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.