

Situational Awareness at Internet Scale: Detection of Extremely Rare Crisis Periods

2008 Sandia Workshop on Data Mining and Data Analysis



David Cieslak, dcieslak@cse.nd.edu, <http://www.nd.edu/~dcieslak/>, 8962

Philip Kegelmeyer, wpk@sandia.gov, csmr.ca.sandia.gov/~wpk, 8962

(Presented by Philip Kegelmeyer, July 22, 2008)

The Conclusion

When building **ensembles** of **decision trees** to classify extremely **skew** data, such as **Border Gateway Protocol (BGP)** data, use the **Hellinger Distance** metric, rather than the traditional **Infogain** metric, for increased accuracy.

(And be sure to use a **sensible metric** for “accuracy”.)

Skew vs. Supervised Machine Learning

When data is skew,
the safe choice is
“accurate” —
and useless.



Minority class overwhelmed.

Accuracy (A) vs Class-Averaged Accuracy (A_c)

Classification confusion matrices.

		Truth	
		Q	W
	Q	80	10
	W	20	90

.....

$$\begin{aligned} A &= \frac{80 + 90}{100 + 100} \\ &= \frac{170}{200} = \mathbf{0.85} \end{aligned}$$

$$\begin{aligned} A_c &= \frac{1}{2} \left(\frac{80}{100} + \frac{90}{100} \right) \\ &= \mathbf{0.85} \end{aligned}$$

Balanced Data

Accuracy (A) vs Class-Averaged Accuracy (A_c)

Classification confusion matrices.

		Truth	
		Q	W
	Q	80	10
	W	20	90

$$A = \frac{80 + 90}{100 + 100}$$

$$= \frac{170}{200} = 0.85$$

$$A_c = \frac{1}{2} \left(\frac{80}{100} + \frac{90}{100} \right)$$

$$= 0.85$$

Balanced Data

		Truth	
		Q	W
	Q	800	10
	W	200	90

$$A = \frac{800 + 90}{1000 + 100}$$

$$= \frac{890}{1100} = \mathbf{0.81}$$

$$A_c = \frac{1}{2} \left(\frac{800}{1000} + \frac{90}{100} \right)$$

$$= \mathbf{0.85}$$

Skew Data

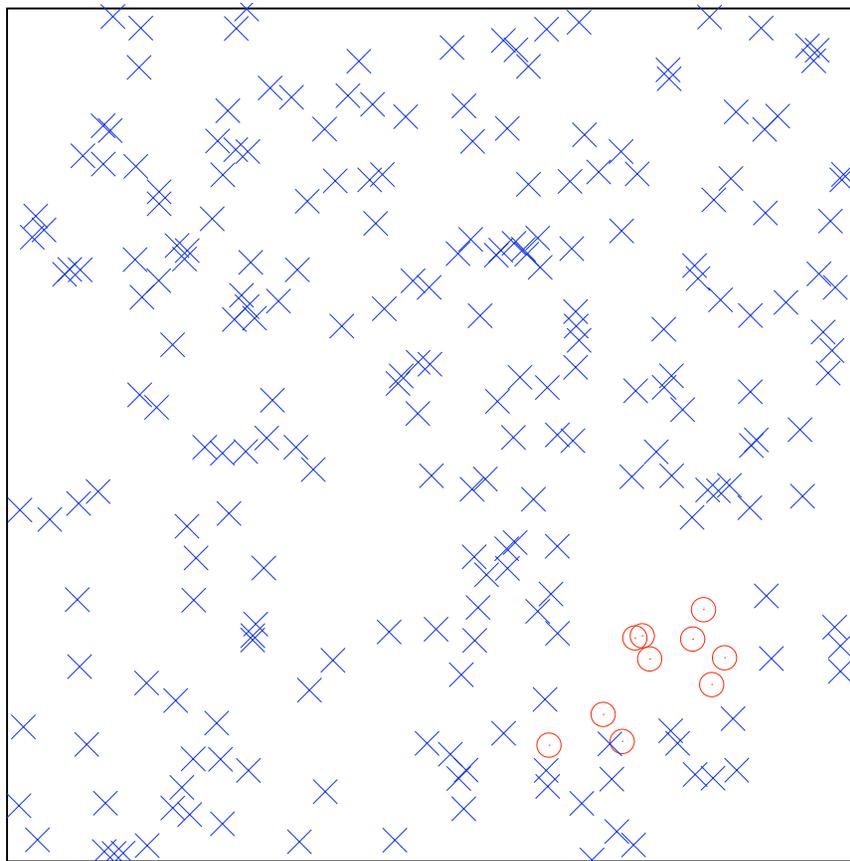
Accuracy (A) vs Class-Averaged Accuracy (A_c)

Classification confusion matrices.

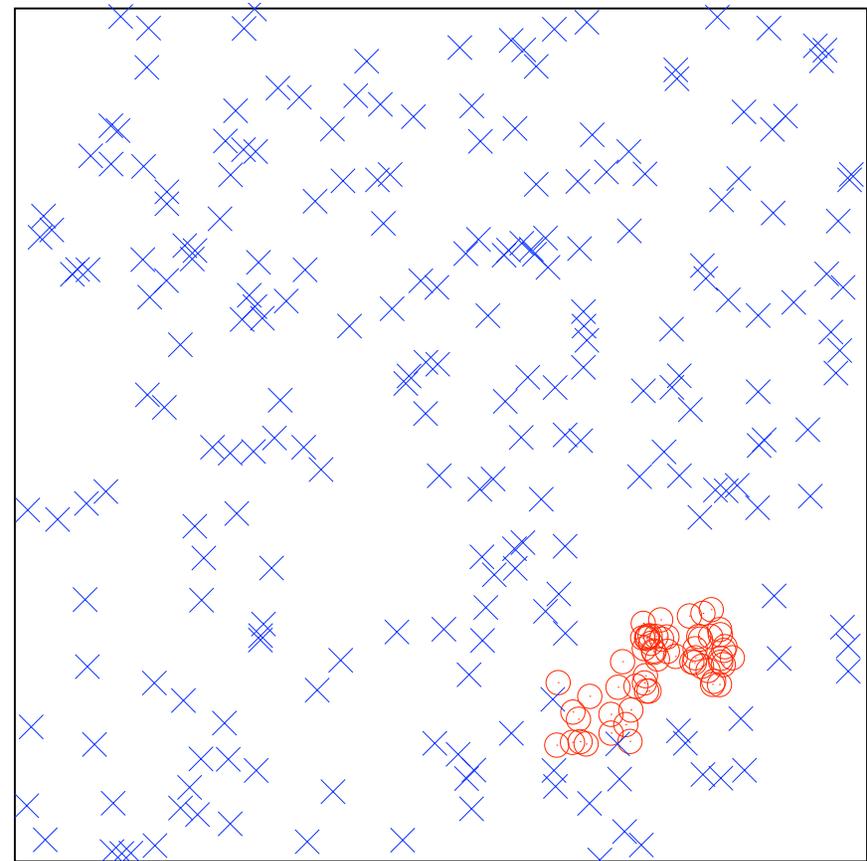
	Truth			Truth			Truth	
	Q	W		Q	W		Q	W
Q	80	10	Q	800	10	Q	800	100
W	20	90	W	200	90	W	200	0
.....								
A	$= \frac{80 + 90}{100 + 100}$		A	$= \frac{800 + 90}{1000 + 100}$		A	$= \frac{800 + 0}{1000 + 100}$	
	$= \frac{170}{200} = 0.85$			$= \frac{890}{1100} = 0.81$			$= \frac{800}{1100} = \mathbf{0.72}$	
A_c	$= \frac{1}{2} \left(\frac{80}{100} + \frac{90}{100} \right)$		A_c	$= \frac{1}{2} \left(\frac{800}{1000} + \frac{90}{100} \right)$		A_c	$= \frac{1}{2} \left(\frac{800}{1000} + \frac{0}{100} \right)$	
	$= 0.85$			$= 0.85$			$= \mathbf{0.40}$	
	Balanced Data			Skew Data			Minority Class Overwhelmed	

Previous Best Solution: SMOTE for Skew Data

Not enough minority data? Invent some!



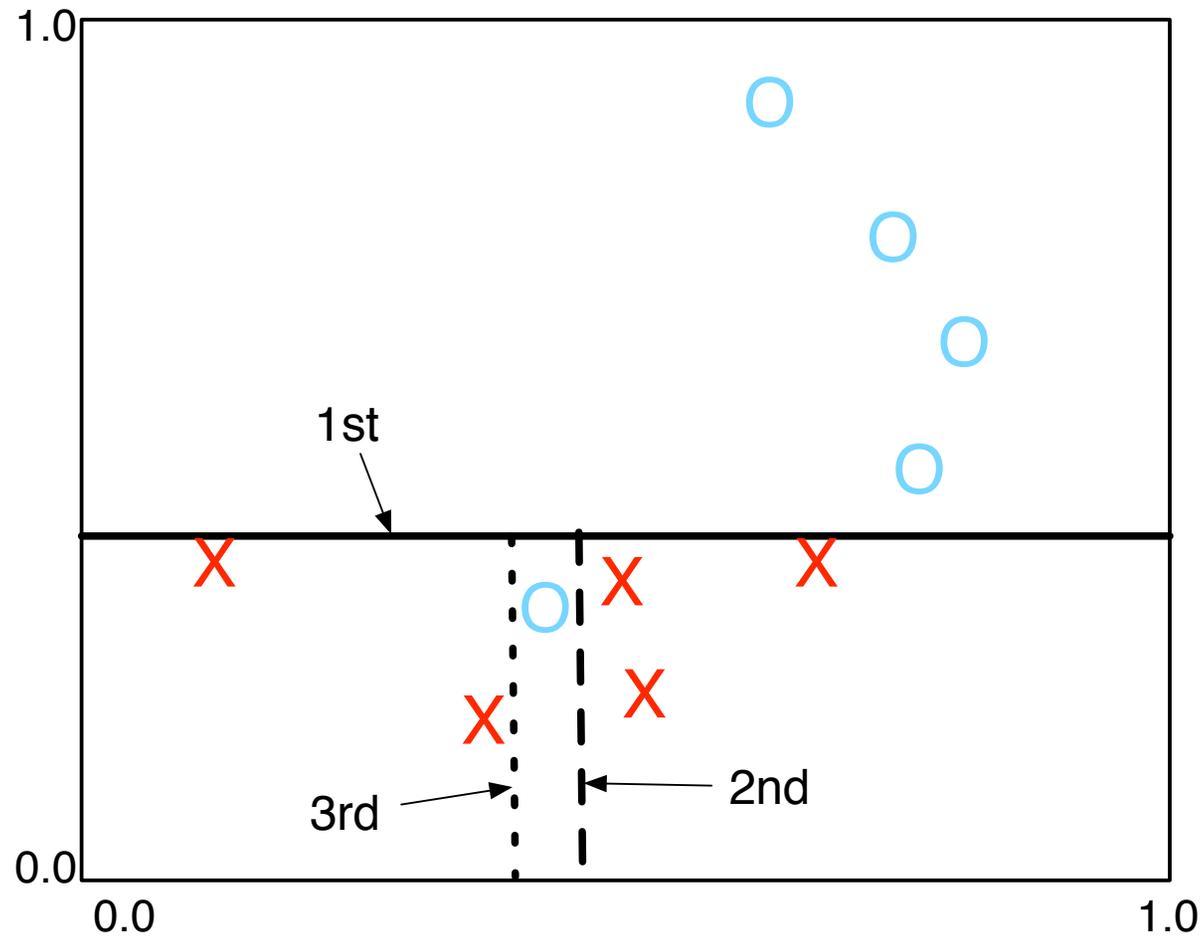
Minority class overwhelmed



Minority class filled out by SMOTE[2]

Decision Trees and Purity Metrics

Partition attribute space to maximize purity of children. Recurse.



Infogain is the Traditional Purity Metric

W, Q are the classes of interest

N = the total number of samples

N_i = number of samples in class i

N^s = total number of samples in the L/R split

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

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

Hellinger Pays Attention to Class Statistics

W, Q are the classes of interest

N = the total number of samples

N_i = number of samples in class i

N^s = total number of samples in the L/R split

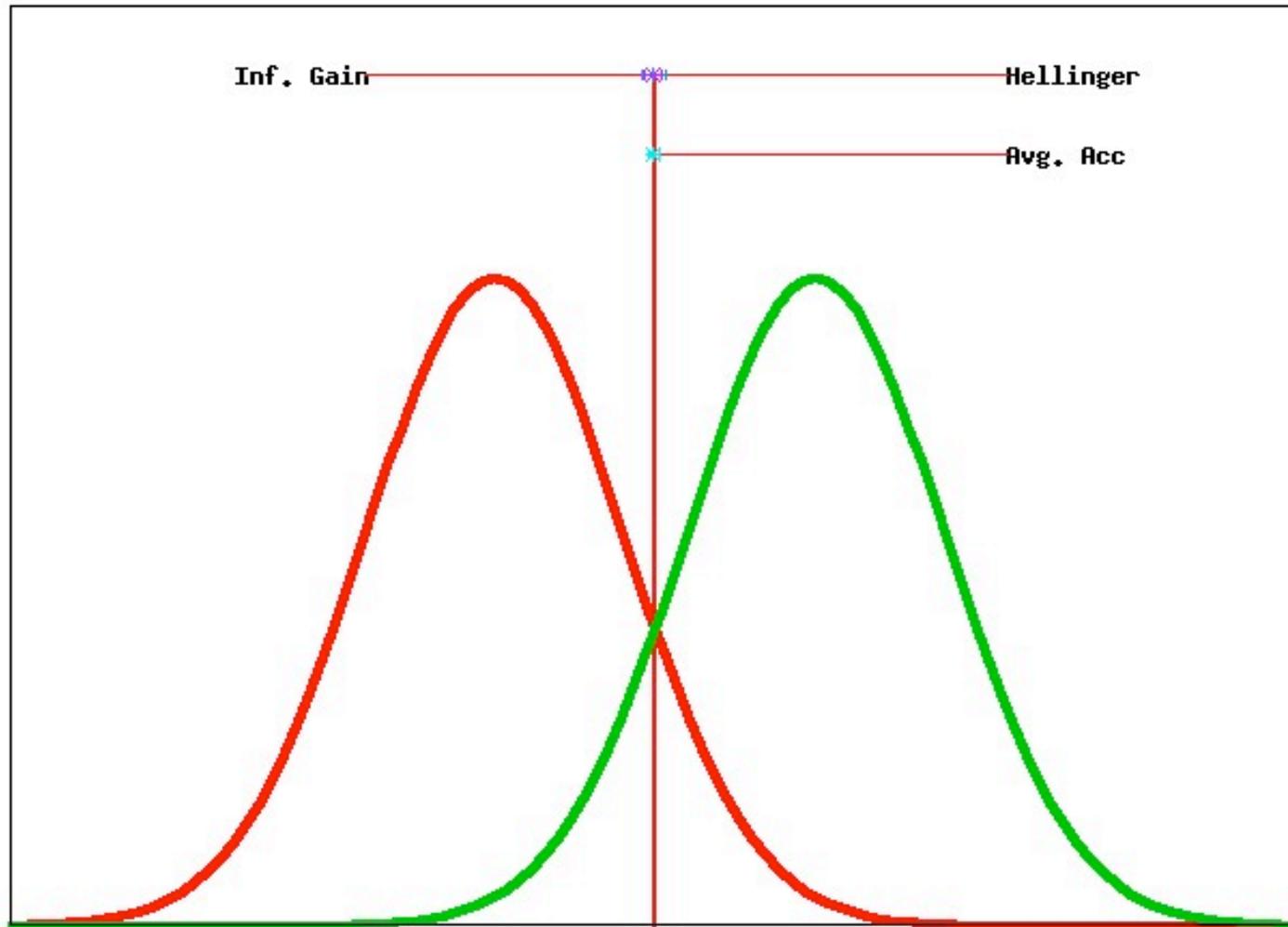
N_i^s = number of samples in class i in L/R split

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

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

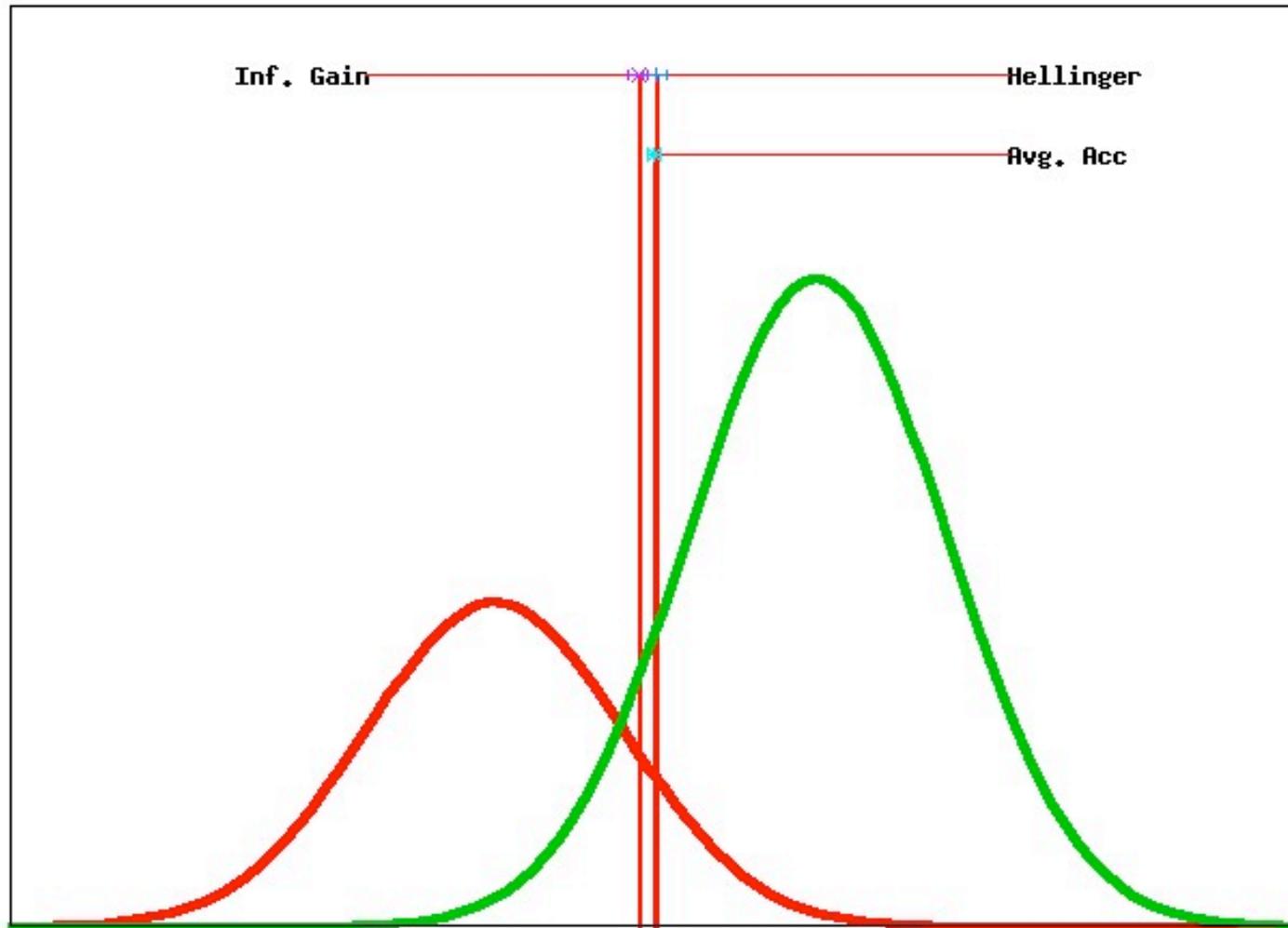
Infogain vs Hellinger on Gaussian Data

(10000:10000)



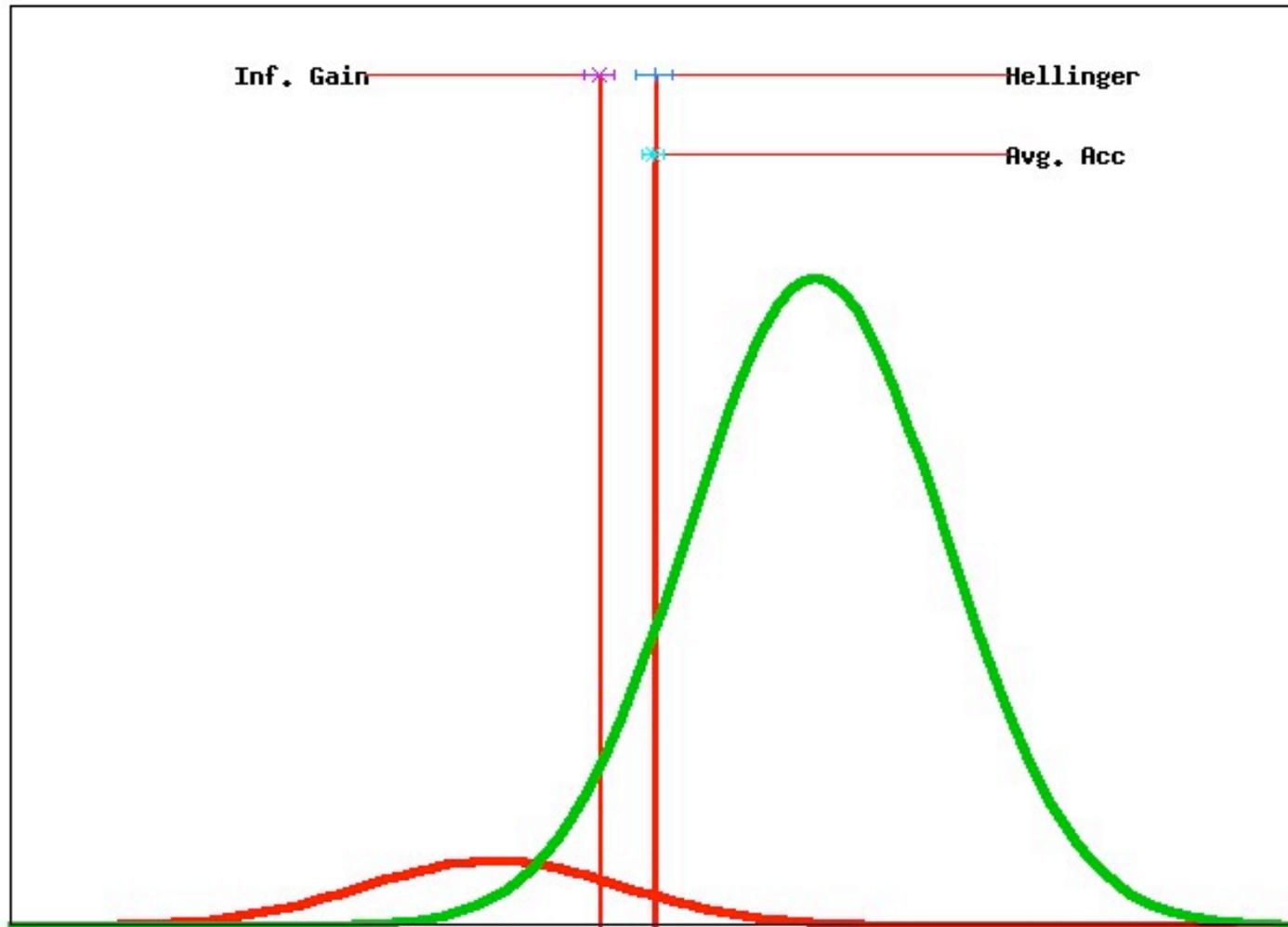
Infogain vs Hellinger on Gaussian Data

(5000:10000)



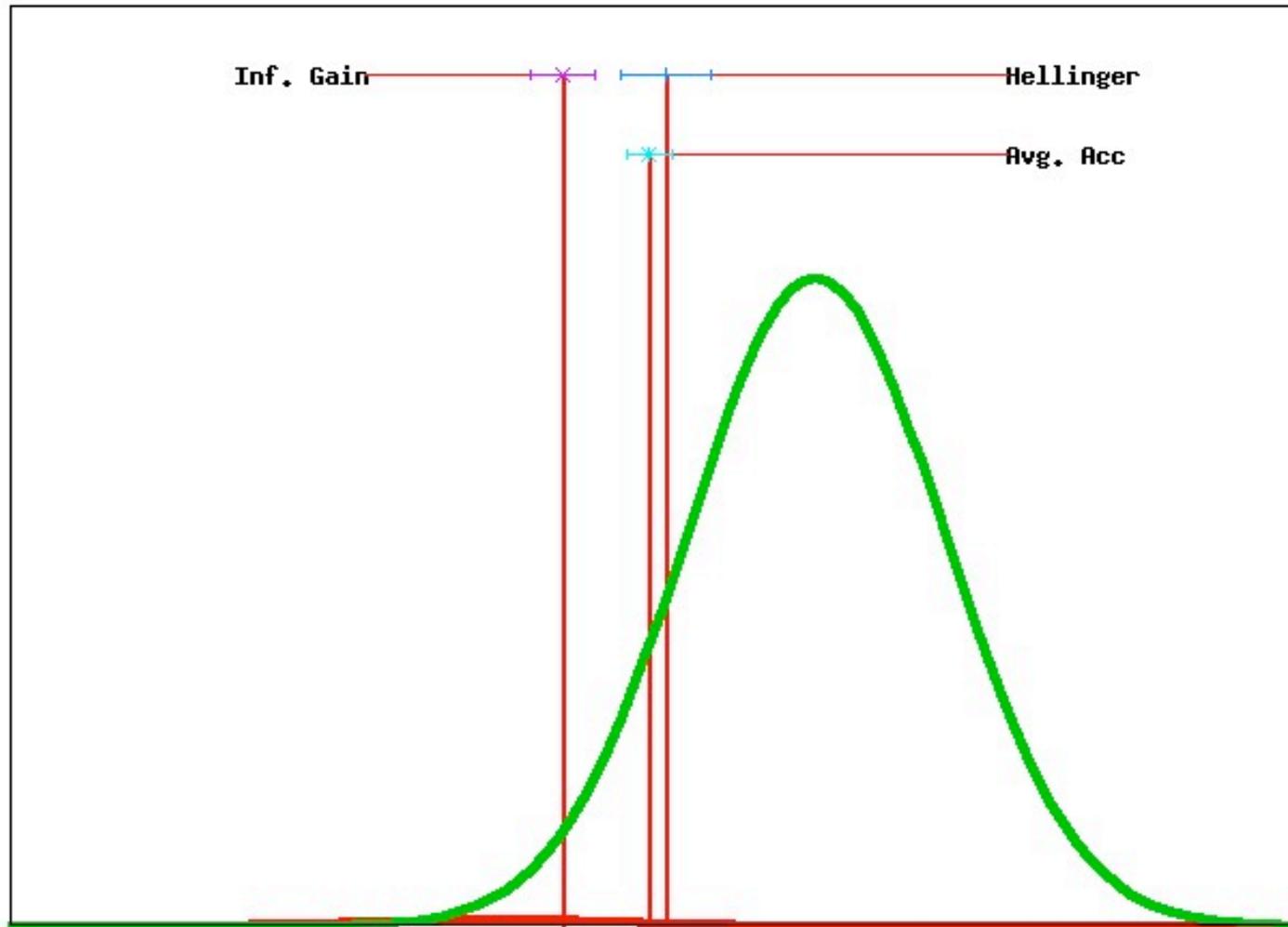
Infogain vs Hellinger on Gaussian Data

(1000:10000)



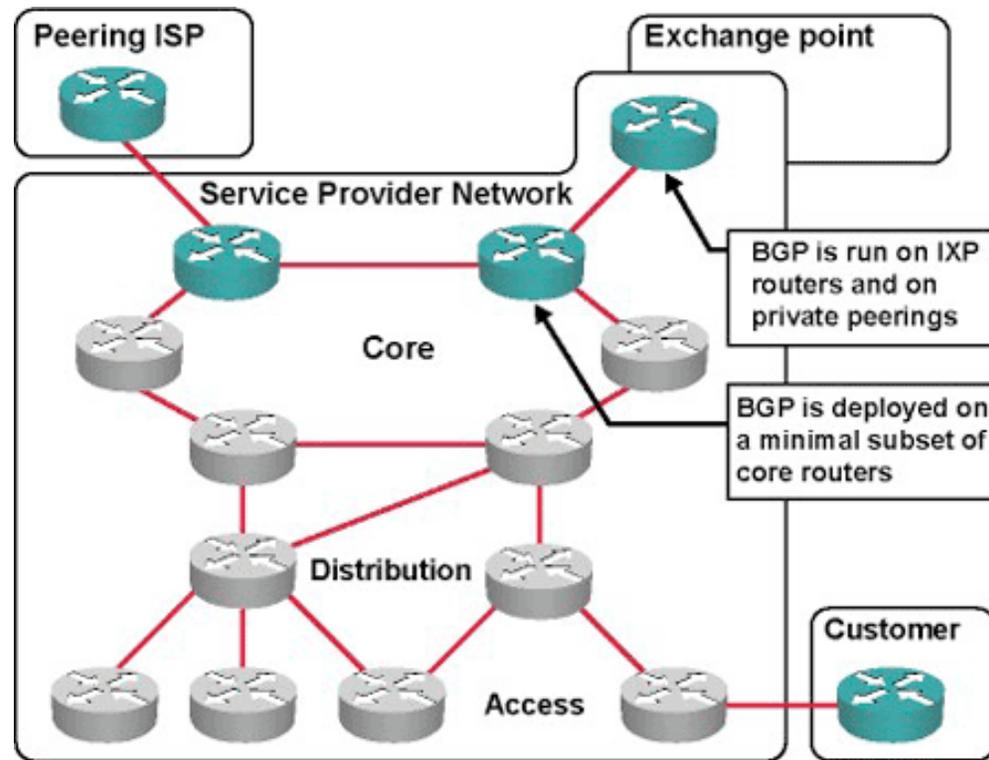
Infogain vs Hellinger on Gaussian Data

(100:10000)



Border Gateway Protocol (BGP) Data

Count “withdrawals” and “announcements” over 30 second intervals.



(Our thanks to Max Planck, New Mexico Institute of Mining and Technology, for data acquisition and feature extraction.)

Example: Event 34, Internet Worm

- Start: 5:28 AM, 01/25/2003.
- End: Noon, 01/26/2003.
- \Rightarrow 2224 worm event samples.
- What happens as skew increases?



Proportion of worm samples	0.114	0.011	0.002	0.001
Infogain Bagging, A_c	0.897	0.833	0.773	0.757
Hellinger Bagging, A_c	0.899	0.838	0.783	0.769
Advantage of Hellinger	0.002	0.005	0.010	0.014

New, Improved Event 34, With Extra SMOTE

- Start: 5:28 AM, 01/25/2003.
- End: Noon, 01/26/2003.
- \Rightarrow 2224 worm event samples.
- What happens as skew increases?



Proportion of worm samples	0.114	0.011	0.002	0.001
Infogain Bagging, A_c	0.897	0.833	0.773	0.757
SMOTE + Infogain Bagging, A_c	0.933	0.929	—	—
Hellinger Bagging, A_c	0.899	0.838	0.783	0.769
SMOTE + Hellinger Bagging, A_c	0.931	0.934	—	—

Access to AvatarTools and Hellinger Trees

- Use the code on the ICC:
 - For \$CLUS equal to tbird, shasta, or spirit:
 - * Add `/projects/ascd/avatar/$CLUS/current/bin` to PATH
 - * Add `/projects/ascd/avatar/$CLUS/current/man` to MANPATH
- Or build it yourself:
 - `www.ca.sandia.gov/avatar`
 - Standard Unix process; unpack tarball, configure, make.
 - Builds and passes tests on Mac, Linux, and Solaris.
 - Includes tutorial and example data.

The Conclusion

When building **ensembles** of **decision trees** to classify extremely **skew** data, such as **Border Gateway Protocol (BGP)** data, use the **Hellinger Distance** metric, rather than the traditional **Infogain** metric, for increased accuracy.

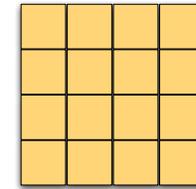
(And be sure to use a **sensible metric** for “accuracy”.)

References

- [1] BANFIELD, R. E., HALL, L. O., BOWYER, K. W., BHADORIA, D., KEGELMEYER, W. P., AND ESCHRICH, S. A comparison of ensemble creation techniques. In *Proceedings of the Fifth International Conference on Multiple Classifier Systems, MCS2004* (2004), J. K. F. Roli and T. Windeatt, Eds., vol. 3077 of *Lecture Notes in Computer Science*, Springer-Verlag.
- [2] CHAWLA, N. V., BOWYER, K. W., HALL, L. O., AND KEGELMEYER, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.
- [3] CHAWLA, N. V., HALL, L. O., BOWYER, K. W., AND KEGELMEYER, W. P. Learning ensembles from bites: A scalable and accurate approach. *Journal of Machine Learning Research* 5 (2004), 421–451.
- [4] CIESLAK, D. A., AND CHAWLA, N. V. Learning decision trees for unbalanced data. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (Antwerp, Belgium, September 2008).
- [5] DOU, D., LI, J., QIN, H., KIM, S., AND ZHONG, S. Understanding and utilizing the hierarchy of abnormal BGP events. In *Proceedings of the 2007 SIAM International Conference on Data Mining* (2007), pp. 467–472.
- [6] RIPE NCC. RIPE routing information service raw data. www.ripe.net.

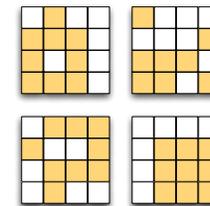
Ensembles: Efficient, Robust, Optimal Accuracy

Traditional: Use 100% of training data to build a sage.



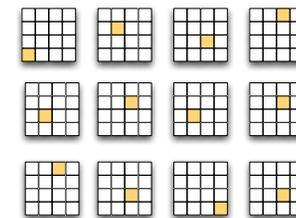
Sage sees all the data.

Ensembles: Use randomized 100% of training data to build an expert. Repeat to build many experts. Vote them.



Each expert sees 2/3rds of the data.

Sandia: Use a semi-random 1% of the training data to build a “bozo”. Repeat to build very many bozos. Vote them.



Each bozo sees a tiny fraction.

The experts beat the sage[1].
The bozos beat the experts[3].