Learning from Imbalanced Datasets: Evaluating the Predictive Accuracy of Minority Classes

Adithi Deborah Chakravarthy
CS Graduate Workshop 2019
University of Nebraska at Omaha

Agenda

- Problem Statement
- Dataset
- Design of Study
- Experiment
- Results
- Discussion

What is the Class Imbalance Problem?

- Total number of a positive class (minority) of data is far less than total number of negative class (majority) of data
  - Fraud detection, anomaly detection, medical diagnosis, etc.
- Learning from highly imbalanced data
  - High predictive power over majority class (obviously!)
- Synthetic Minority Over-sampling (SMOTE), boosting, ensembles
- Evaluate predictions on minority class using Random Over-Sampling Examples (ROSE)

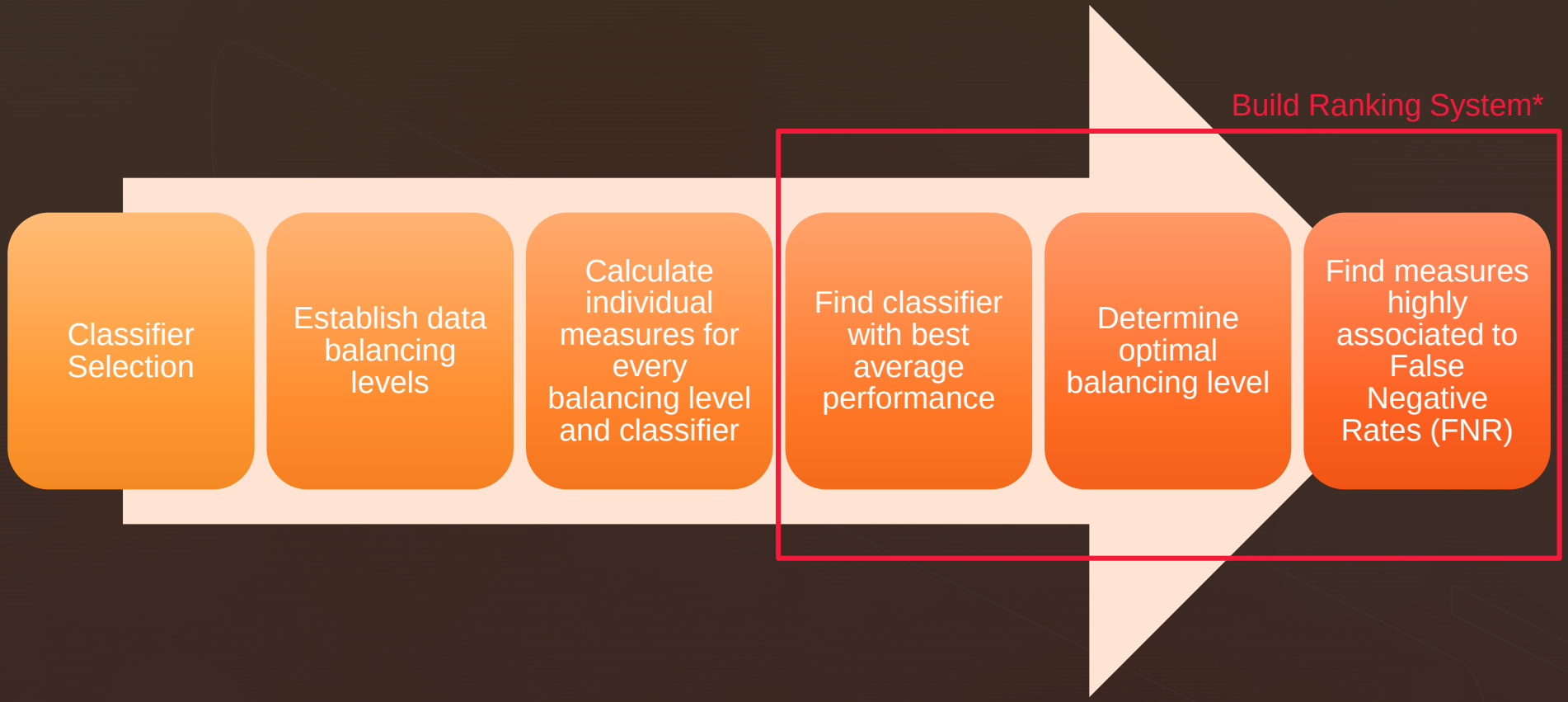
Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *ACM Sigkdd Explorations Newsletter*, *6*(1), 30-39.

Open Corrosion Dataset

- Impact of environmental changes on concrete structures
- Non-destructive testing (NDT) gives detailed insight into affected areas
- 7 measurements from intact and defect areas using NDT methods
- Binary classification
- Highly imbalanced data – 96% intact and 4% defective

Völker, Christoph & Kruschwitz, Sabine & Ebell, Gino & Shen, Jiawei. (2018). Towards Data Based Corrosion Analysis of Concrete with Supervised Machine Learning.

Design of Study



*Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863-905.

Experiment – 10 fold Cross validation

Classifier Selection

- C5.0
- K Nearest Neighbor (KNN)
- Neural Net (NN)
- Random Forest
- Support Vector Machine (SVM)

Balancing Level Selection

- Original
- Oversampling 1:2
- Oversampling 1:3
- Undersampling 2:1
- Undersampling 3:1
- Hybrid (both over and under sampling)

Measures To Evaluate Classifier Performance

- Accuracy – assumes uniform class distribution
- What else can we learn from the Confusion Matrix?

	Predicted Intact	Predicted Defect
Actual Intact	TN (true negative)	FP (false positive)
Actual Defect	FN (false negative)	TP (true positive)

- Sensitivity
- Specificity
- Geometric Mean
- Balanced Accuracy
- Precision
- F-Measure
- Informedness
- Matthews Correlation Coefficient

Measures For Balancing Level and Classifier

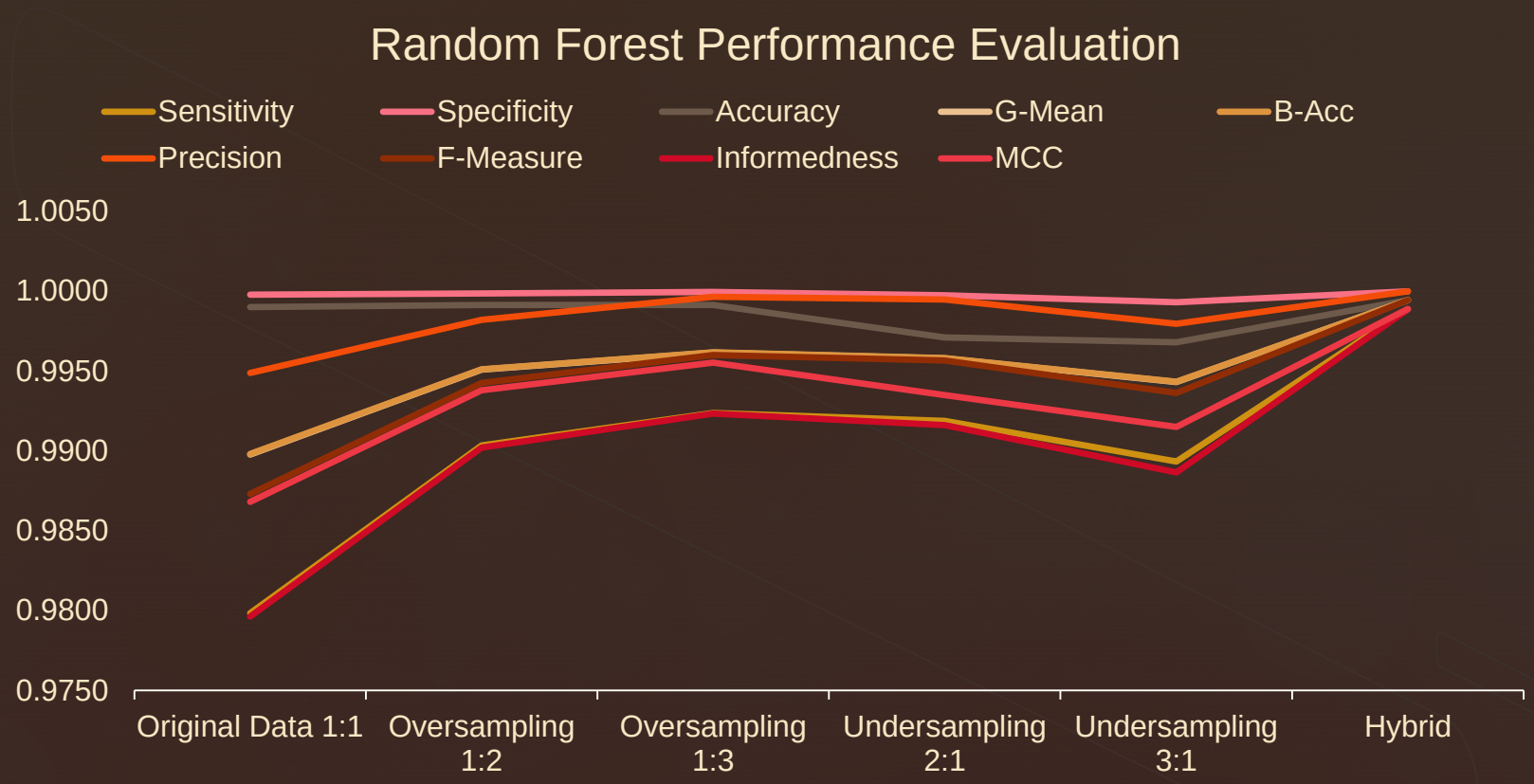
Classifier	Experiment	G-Mean	B-Acc	Precision	F-Measure	Informedness	MCC
C5.0	Original Data 1:1	0.99055469	0.99059472	0.9877175	0.98469388	0.98118945	0.98407488
	Oversampling 1:2	0.99523015	0.99524094	0.99846469	0.9945202	0.99048189	0.99408199
	Oversampling 1:3	0.99695182	0.99695603	0.99880587	0.99642614	0.99391206	0.99599283
	Undersampling 2:1	0.99693114	0.99693232	0.99692938	0.99616466	0.99386463	0.99424545
	Undersampling 3:1	0.99481369	0.99482166	0.9964176	0.99362082	0.98964332	0.99149342
	Hybrid	0.99935593	0.99935614	1	0.99935572	0.99871227	0.99872169
KNN	Original Data 1:1	0.97522807	0.97552129	0.98618219	0.96858507	0.95104258	0.96745175
	Oversampling 1:2	0.98391515	0.98404219	0.9982088	0.98299105	0.96808438	0.98171954
	Oversampling 1:3	0.98717499	0.9872567	0.99965882	0.98694737	0.9745134	0.98543392
	Undersampling 2:1	0.98953782	0.98958989	0.99948823	0.9893617	0.97917978	0.98407611
	Undersampling 3:1	0.98683454	0.98691243	0.99795292	0.98609355	0.97382487	0.98150853
	Hybrid	0.99726117	0.99726492	1	0.99725742	0.99452985	0.99453899
Neural Net	Original Data 1:1	0.97432245	0.9746177	0.96571136	0.9581112	0.9492354	0.95642707
	Oversampling 1:2	0.97962311	0.97980077	0.98106448	0.97100165	0.95960154	0.96867829
	Oversampling 1:3	0.98458387	0.9846794	0.98686455	0.97884941	0.96935881	0.97628794
	Undersampling 2:1	0.99086907	0.99089788	0.99692938	0.99008895	0.98179575	0.98513478
	Undersampling 3:1	0.9888197	0.98885687	0.99232344	0.98626653	0.97771374	0.98168402
	Hybrid	0.99675925	0.99676438	0.9999598	0.99675403	0.99352875	0.99354513
Random Forest	Original Data 1:1	0.98976502	0.9898153	0.99488229	0.9873032	0.97963061	0.98681416
	Oversampling 1:2	0.99509226	0.9951036	0.9982088	0.99426532	0.99020721	0.99380631
	Oversampling 1:3	0.99616165	0.99616885	0.99965882	0.99600578	0.99233771	0.9955252
	Undersampling 2:1	0.99580082	0.99580859	0.99948823	0.99566658	0.99161718	0.99350383
	Undersampling 3:1	0.99431809	0.99433058	0.99795292	0.99363057	0.98866116	0.99151173
	Hybrid	0.99944051	0.99944067	1	0.99944036	0.99888134	0.99888145
SVM	Original Data 1:1	0.93147021	0.93322461	0.76305015	0.81564551	0.86644922	0.81076632
	Oversampling 1:2	0.93389094	0.93515309	0.79810645	0.84002155	0.87030618	0.82908084
	Oversampling 1:3	0.93849568	0.93931164	0.82139202	0.85897779	0.87862327	0.8437632
	Undersampling 2:1	0.95160907	0.95187543	0.9493347	0.93924051	0.90375085	0.90849043
	Undersampling 3:1	0.93050748	0.93214101	0.96315251	0.91804878	0.86428202	0.89079276
	Hybrid	0.96256493	0.96269703	0.97929962	0.96274948	0.92539405	0.9249078

Classifier With Best Average Performance

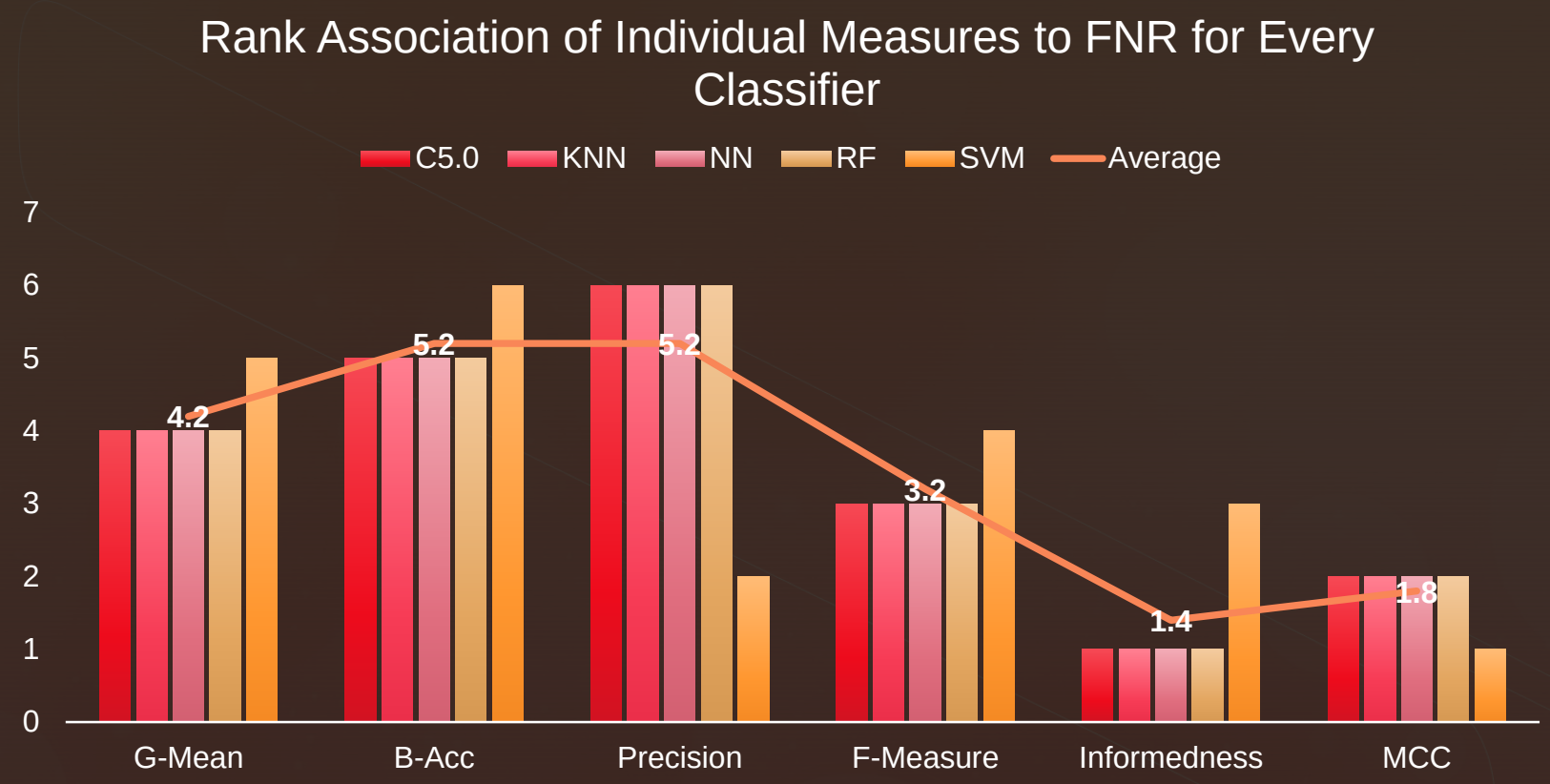
C5.0	G-Mean	B-Accuracy	Precision	F-Measure	Informedness	MCC
Original Data 1:1	5	5	4	4	5	4
Oversampling 1:2	5	5	5	5	5	5
Oversampling 1:3	5	5	3	5	5	5
Undersampling 2:1	5	5	2	5	5	5
Undersampling 3:1	5	5	3	4	5	4
Hybrid	4	4	3	4	4	4
Average Rank	**4.83**	**4.83**	**3.33**	**4.50**	**4.83**	**4.50**

Random Forest	G-Mean	B-Accuracy	Precision	F-Measure	Informedness	MCC
Original Data 1:1	4	4	5	5	4	5
Oversampling 1:2	4	4	3	4	4	4
Oversampling 1:3	4	4	4	4	4	4
Undersampling 2:1	4	4	4	4	4	4
Undersampling 3:1	4	4	4	5	4	5
Hybrid	5	5	3	5	5	5
Average Rank	**4.17**	**4.17**	**3.83**	**4.50**	**4.17**	**4.50**

Optimal Balancing Level



Measure Association to FNR



Conclusion and Future Directions

- Preliminary results for
  - Classifier, Balancing level, Optimal measure evaluation
- Future directions
  - Validate ranking system on more imbalanced datasets
  - Build an ensemble classifier for skewed data
  - Original composite measure to calculate the predictive power of minority class more accurately

Thank You!