Evaluation of Anomaly Detection Algorithms for the Real-World Applications

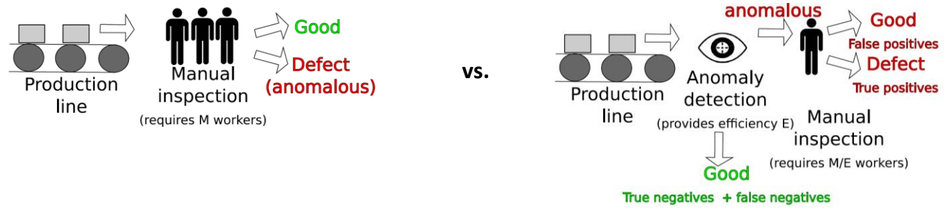
Marija Ivanovska [1], Domen Tabernik [2], Danijel Skočaj [2], Janez Perš [1]

[1]Laboratory For Machine Intelligence, Faculty of Electrical Engineering, University of Ljubljana
[2]Visual Cognitive Systems Laboratory, Faculty of Computer and Information Science, University of Ljubljana

University of Ljubljana

Motivation

- Real-world anomaly data is usually ***highly imbalanced.***
- Recently, the problem of scarcity of anomalies has been succesfully avoided by introduction to various GAN-based models which are trained using ***one class learning*** techniques.
- Evaluation of such models is still carried out in ***the same manner*** as evaluation on well balanced data. Most often used metrics are AUC and AP, which are proved to be ***inappropriate for imbalanced dataset***, according to already published scientific studies.
- In practice anomaly detectors are usually imperfect, but can be used to ***reduce the manual inspection workforce***



Methods

Standard metrics: AUC & AP

Our proposed metric: **%TN@X%TP**

- allows adjustment of X, depending on customer's specifications
- efficiency of the anomaly detection can be directly calculated as:

$$Efficiency = \frac{N}{TP + FP} \approx \frac{N}{FP}$$

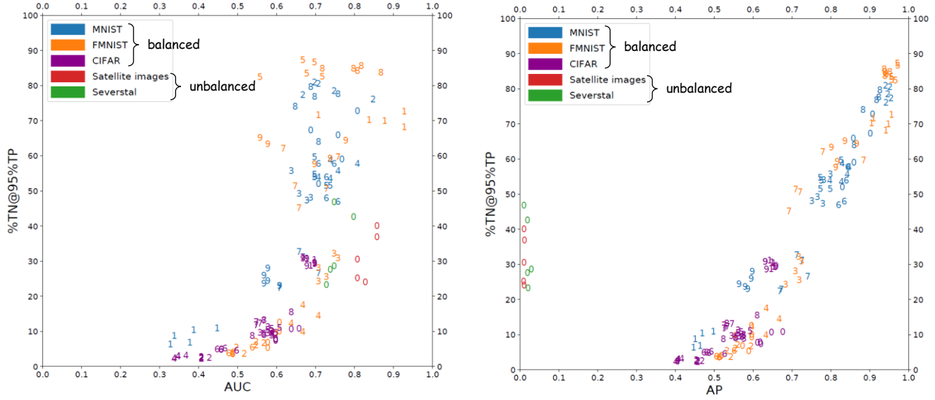
$$\boldsymbol{Efficiency = \frac{100\%}{100\% - \%TN@X\%TP}}$$

Experiments

- 4 SOTA GAN-based methods have been trained and evaluated: GANomaly, skip-GANomaly, f-AnoGAN and OCGAN
- 3 standard datasets were used: MNIST, Fashion-MNIST and CIFAR10, all of them have well balanced testing subsets
- 2 highly imbalanced datasets were additionaly created for testing purposes. In both datasets there is only 1 anomalous testing sample in 3.000 non-anomalous testing samples:

Severstal steel **Satellite images**

Results



Graphs showing correlation between AUC/AP score and %TN@X%TP obtained with GANomaly models.

Analysis of the results using Pearson correlation coefficients show ***high correlation in balanced dataset***, but ***low correlation between AUC/AP and %TN@X%TP in unbalanced dataset***.

	AUC vs %TN@95%TP		AP vs %TN@95%TP	
	balanced	unbalanced	balanced	unbalanced
GANomaly	0.745	0.127	0.961	-0.409
skip-GANomaly	0.738	-0.910	0.933	-0.510
f-AnoGAN	0.816	-0.328	0.942	-0.246
OCGAN	0.790	0.140	0.898	0.171
average correlation	high	low	high	low

Conclusions

- Compared to AUC and AP, %TN@X%TP was both, more robust and more informative quality measure, especially when the data is highly imbalanced.
- Anomaly detectors' performance should be summerized with a ROC curve and then our proposed metric can be used to calculate the actual savings on the workforce
- The ROC curve of anomaly detectors should be averaged and published in conjunction with deviation intervals, calculated for different training runs, under same circumstances.

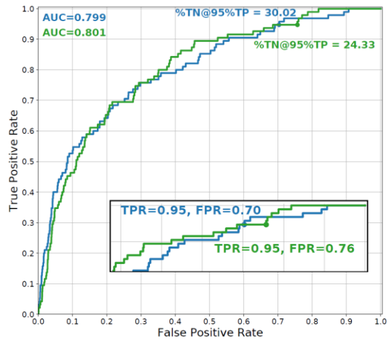


Figure representing two training runs on same data using same parameters' values. Note that they have almost equal AUC but very different %TN@X%TP .

Acknoledgments

This work was in part supported by the ARRS research project J2-9433 (DIVID) and research programmes P2-0214 and P2-0095.