

Statistical Inference for Recall, Precision and Average Precision

Wanhua Su (PDF)

Dept. of Mathematical and Statistical Science
University of Alberta, Canada

Joint work with Dr. Peng Zhang

Outline

- 1 Rare Target Detection
 - Characteristics
 - Applications
 - Ranking Methods
 - Performance Measures
- 2 Statistical Inference on Recall, Precision and Average Precision
 - Some Existing Methods
 - Our Approach
 - Sampling Distribution of Recall, Precision and Average Precision Under Random Selection
 - Exact Distribution of Average Precision Under Random Selection
 - A More General Approach
- 3 Simulation Study
- 4 Conclusion

Rare Target Detection

- Set-up: $\{\mathbf{x}_i, y_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of predictors, $y_i \in \{0, 1\}$ is class label. Class-1 observations are called **targets**, class-0 the background.
- Class frequencies: **extremely unbalanced**. Most of the observations belong to background class, only a few targets, say 2%.
- Objective: rank the items so that targets are retrieved at the beginning of the list.

Applications

- Drug Discovery: \mathbf{x}_i is a vector of chemical descriptors for a compound;

$$y_i = \begin{cases} 1 & \text{the compound is an active drug agent} \\ 0 & \text{otherwise} \end{cases}$$

- Credit Card Fraud Detection: \mathbf{x}_i is a vector of descriptors for a credit card transaction;

$$y_i = \begin{cases} 1 & \text{the transaction is fraudulent} \\ 0 & \text{otherwise} \end{cases}$$

- Direct Marketing: \mathbf{x}_i is a vector of descriptors for a potential customer;

$$y_i = \begin{cases} 1 & \text{the potential customer will respond} \\ 0 & \text{otherwise} \end{cases}$$

Ranking Algorithms for Rare Target Detection

All methods used for classification, such as

- Logistic regression: $P(y = 1) = \frac{\exp\{\mathbf{x}^T \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}^T \boldsymbol{\beta}\}}$.
- K -nearest Neighbor (KNN):
$$P(y = 1) = \frac{\# \text{ of nearest neighbors s.t. } y=1}{K}.$$
- Support vector machines (SVM): solve for a hyperplane with the maximum margin to separate the two classes and then rank the items by their signed distances to the hyperplane.
- LAGO (Zhu *et al.* 2006, Technometrics): a computationally efficient kernel method. Score is in the form of radial basis function network.

Performance Measures

- Misclassification rate used in classification problem is not suitable.
- It is all about ranking. Borrow the performance measures widely used in **information retrieval**. A selected item is called a **hit** if it is a target. Given the ranking, define hit function $h(t) = \sum_{i=1}^t y(i)$, where

$$y(i) = \begin{cases} 1 & \text{the } i\text{th object is a hit;} \\ 0 & \text{the } i\text{th object is a miss.} \end{cases}$$

- **Recall** $r(t) = \frac{h(t)}{m}$: proportion of hits that are retrieved.
- **Precision** $p(t) = \frac{h(t)}{t}$: proportion of retrieved that are hits.
- **Average precision** $AP = \frac{1}{m} \sum_{t=1}^n y(t) \frac{h(t)}{t} = \frac{1}{m} \sum_{t=1}^n y(t) p(t)$: average of the precisions for those hits. An approximation to the area under the **recall-precision curve**.

Performance Measures: Example

Suppose we need to rank eight items, among which only three are targets, i.e., $n = 8, m = 3$.

$$r(t) = \frac{h(t)}{m}, \quad p(t) = \frac{h(t)}{t}, \quad \text{AP} = \frac{1}{m} \sum_{t=1}^n y_{(t)} p(t)$$

Rank t	Algorithm 1				Algorithm 2			
	$y_{(t)}$	$h(t)$	$r(t)$	$p(t)$	$y_{(t)}$	$h(t)$	$r(t)$	$p(t)$
1	1	1	1/3	1/1	0	0	0/3	0/1
2	1	2	2/3	2/2	1	1	1/3	1/2
3	0	2	2/3	2/3	0	1	1/3	1/3
4	1	3	3/3	3/4	0	1	1/3	1/4
5	0	3	3/3	3/5	1	2	2/3	2/5
6	0	3	3/3	3/6	1	3	3/3	3/6
7	0	3	3/3	3/7	0	3	3/3	3/7
8	0	3	3/3	3/8	0	3	3/3	3/8

$$\text{AP}_1 = \frac{1}{3} \left(\frac{1}{1} + \frac{2}{2} + \frac{3}{4} \right) = 0.9167, \quad \text{AP}_2 = \frac{1}{3} \left(\frac{1}{2} + \frac{2}{5} + \frac{3}{6} \right) = 0.4667$$

Questions of Interest

Given the performance measure of two ranking systems, we are interested in answering these two questions:

- Whether both two ranking systems significantly outperform random selection.
- Whether one ranking system is more effective than the other one.

Possible solution: Find the sampling distributions of the performance measures which enable us to make statistical inference.

Brief Review: Little Has Been Done

- Pair- t , signed and Wilcoxon test for comparing recall and computationally intensive randomization test for precision. (Yeh, 2000)
- Goutte and Gaussier (2005) argued the joint of (TP, FN, FP, TN) \sim multinomial, $TP|TP+FP \sim \text{Binomial}(TP+FP, p)$, $TP|TP+FN \sim \text{Binomial}(TP+FN, r)$. Adopt conjugate Beta prior to p and r and make inference on the posterior. For comparing two ranking systems, each object falls into one of the following categories:
 - System 1 gives the correct assignment, system 2 fails;
 - System 2 gives the correct assignment, system 1 fails;
 - Both two systems yield the same assignment.

Let π_1, π_2 and π_3 be the probabilities that the above events happen. Assign Dirichlet prior and calculate the posterior distribution. Calculate $Pr(\pi_1 > \pi_2)$ by Monte Carlo method.

Sampling Distribution of Recall, Precision and Average Precision Under Random Selection

$y_{(i)} \sim \text{Bernoulli}(\pi)$, $i = 1, 2, \dots, n$ with

$E(y_{(i)}) = P(y_{(i)} = 1) = \pi = \frac{m}{n}$, $\text{Var}(y_{(i)}) = \pi(1 - \pi)$. As a result,

- $h(t) = \sum_{i=1}^t y_{(i)} \sim \text{Binomial}(t, \pi)$.
- $\text{Var}(h(t)) = t\pi(1 - \pi) + t(t - 1)\pi\left(\frac{m-1}{n-1} - \pi\right)$.
- $\text{Cov}(h(t_i), h(t_j)) = \text{Var}(h(t_i)) + t_i(t_j - t_i)\pi\left(\frac{m-1}{n-1} - \pi\right)$.

By the Central Limit Theorem,

- $r(t) \sim N\left(\pi \frac{t}{m}, \sqrt{\frac{t\pi(1-\pi)}{m^2} + \frac{t(t-1)}{m^2}\pi\left(\frac{m-1}{n-1} - \pi\right)}\right)$.
- $p(t) \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{t} + \frac{(t-1)\pi\left(\frac{m-1}{n-1} - \pi\right)}{t}}\right)$.
- $\text{AP} \sim N(\pi, \sigma_{ap})$, with

$$\sigma_{ap} = \sqrt{\frac{1}{m^2} \left[\sum_{i=1}^m \frac{\pi(1-\pi) + (t_i - 1)\pi\left(\frac{m-1}{n-1} - \pi\right)}{t_i} + 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m \left(\frac{\pi(1-\pi) + (t_j - 1)\pi\left(\frac{m-1}{n-1} - \pi\right)}{t_j} \right) \right]}$$

Distribution of Average Precision Under Random Selection

Let T_1, T_2, \dots, T_m be the ranks (locations) of the i th hits, average precision $AP = \frac{1}{m} \sum_{i=1}^m \frac{h(t_i)}{t_i}$ can be rewritten as

$$AP = \frac{1}{m} \left(\frac{1}{t_1} + \frac{2}{t_2} + \dots + \frac{m}{t_m} \right). \quad (1)$$

- The probability that the j th hits appears at the k th position,

$$P(T_j = k) = \binom{k-1}{j-1} \binom{n-k}{m-j} / \binom{n}{m}.$$

- The joint distribution of T_i and T_j for $i < j$,

$$P(T_i = s, T_j = t) = \binom{s-1}{i-1} \binom{t-s-1}{j-i-1} \binom{n-t}{m-j} / \binom{n}{m}.$$

- $E(AP) = \frac{1}{m} \sum_{i=1}^m E\left(\frac{i}{T_i}\right)$, $\text{Var}(AP) = E(AP^2) - E^2(AP)$ with

$$E(AP^2) = \frac{1}{m^2} \left[\sum_{i=1}^m E\left(\frac{i^2}{T_i^2}\right) + 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m E\left(\frac{ij}{T_i T_j}\right) \right]$$

A More General Approach

Suppose we treat a hit as an event and the ranks of hits as event times, t_1, \dots, t_m . The survival function $S(t)$ can be estimated by

$$S(t) = P(T > t) \approx \frac{\sum_{i=1}^m \mathbb{1}(T_i > t)}{m} = 1 - \frac{h(t)}{m} = 1 - r(t).$$

Therefore,

- $r(t) = 1 - S(t)$, $\text{Var}(r(t)) = \text{Var}(S(t))$
- $p(t) = \frac{m}{t}(1 - S(t))$, $\text{Var}(p(t)) = \frac{m^2}{t^2} \text{Var}(S(t))$
- $\text{AP} = \frac{1}{m} \sum_{i=1}^m p(t_i) = \sum_{i=1}^m \frac{r(t_i)}{t_i}$,

$$\begin{aligned} \text{Var}(\text{AP}) &= \text{Var} \left\{ \sum_{i=1}^m \frac{r(t_i)}{t_i} \right\} = \sum_{i=1}^m \text{Var} \left\{ \frac{r(t_i)}{t_i} \right\} + 2 \sum_{i < j} \text{Cov} \left\{ \frac{r(t_i)}{t_i}, \frac{r(t_j)}{t_j} \right\} \\ &= \sum_{i=1}^m \text{Var} \left\{ \frac{S(t_i)}{t_i} \right\} + 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m \text{Cov} \left\{ \frac{S(t_i)}{t_i}, \frac{S(t_j)}{t_j} \right\}. \end{aligned}$$

Calculation of $S(t)$ and $Var(S(t))$

We can interpret the survival function as $S(t) \approx \frac{1}{m} \sum_{i=1}^m \pi_i$, where $\pi_i = Pr(\text{the } i\text{th hit appears after rank } t)$. In general,

$$\pi_i = \sum_{j:t_j > t} H_{ij}, i = 1, \dots, m,$$

where H_{ij} is the transition probability that the i th hit appears at rank j . As a result, $Var(S(t)) = \frac{1}{m^2} \sum_{i=1}^m \pi_i(1 - \pi_i)$.

- Under random selection: $H_{ij} = \frac{1}{m}$, $\pi_i = \frac{x}{m}$, where $x = \#$ of $T_i > t$. Then

$$S(t) = \frac{x}{m}, \quad Var(S(t)) = \frac{x(m-x)}{m^3}.$$

- Under general case: H_{ij} are not the same and might be difficult to estimate.

Simulation: Under Random Selection

10000 random permutations of $m = 245$ 1's and $n - m = 2755$ 0's are generated. For recall $r(t)$, treat the average value of the performance measures across 10000 replications as the true; for average precision, we know the true as well from the exact distribution.

Rank t	True/Empirical		Our Method		Coverage	
	$r(t)$	$\text{Std}(r(t))$	$E(r(t))$	$E[\text{Std}(r(t))]$	80%	95%
50	0.01680	0.00791	0.01680	0.00793	0.7407	0.9241
100	0.03341	0.01108	0.03341	0.01130	0.7763	0.9160
245	0.08168	0.01682	0.08168	0.01739	0.8014	0.9406
500	0.16675	0.02289	0.16675	0.02373	0.8206	0.9563
1500	0.49948	0.03052	0.49948	0.03188	0.8208	0.9538
2500	0.83296	0.02290	0.83296	0.02375	0.8197	0.9541
AP	0.08399	0.00561	0.08395	0.00514	0.7642	0.9320
	(0.08164)	(0.00497)			0.7586	0.9359

Conclusion

- Give a brief introduction to three performance measures for rare target detection problem: Recall $r(t) = \frac{h(t)}{m}$, precision $p(t) = \frac{h(t)}{t}$ and average precision $AP = \frac{1}{m} \sum_{i=1}^m p(t_i)$.
- Derive the sampling distribution for recall, precision and average precision under random selection.
- Provide a way to calculate the sampling distribution for recall, precision and average precision for any ranking algorithm.

References

- Cyril Goutte and Eric Gaussier (2005), A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation, *Advances in Information Retrieval*, 345–359, Springer.
- Alexander Yeh (2000), More Accurate Tests for the Statistical Significance of Result Differences, *Proceedings of the 18th conference on Computational Linguistics*, Volume 2, 947–953.
- Zhu, M., Su, W. and Chipman, H. A. (2006), LAGO: A Computationally Efficient Approach for Statistical Detection, *Technometrics*, **48**, 193–205.

Acknowledgements

- My PhD supervisors Dr. Mu Zhu at University of Waterloo and Dr. Hugh Chipman at Acadia University for introducing the concept of average precision.
- Dr. Peter Hooper at University of Alberta for funding.