

Statistical Inference on Recall, Precision and Average Precision under Random Selection

Peng Zhang

Department of Mathematics
Zhejiang University
Hangzhou, Zhejiang, 310027 China
pengzzju@gmail.com

Wanhua Su

Department of Mathematics and Statistics
Grant MacEwan University
Edmonton, Alberta, Canada T5J 4S2

Abstract—The objective of a rare target detection problem is to identify the rare targets as early as possible. Recall, precision and average precision are three popular performance measures for evaluating different detection methods. However, there is little literature on the statistical properties of these three measures. We develop a framework for conducting statistical inference on recall, precision and average precision through establishing their asymptotic properties. Simulations are used to illustrate the idea. The proposed methods can also be applied in other areas where ranking systems need to be evaluated, such as information retrieval.

I. INTRODUCTION

A. Information Retrieval

Information retrieval is the science of searching for information relevant to a given topic from a huge database. The information can be in the form of text documents, photos, videos, audio clips or the combinations of these media. An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines (e.g., Google). The system then computes a numeric score on how well each object in the database matches the query. This score is called relevant score. The objects are ranked and shown to the user according to the relevant score. Objects with larger relevant score appear earlier in the list.

B. Rare Target Detection

In a typical rare target detection problem, we have data $\{y_i, \mathbf{x}_i\}$, where $\mathbf{x}_i$ is the predictor vector of the i th observation and $y_i = \{0, 1\}$ is its class label. The objective is to identify the rare class-1 observations (targets) as early as possible. In general, those selected targets will be passed to the next stage for further investigation. Three typical applications of rare target detection are:

- 1) Drug Discovery: Here $\mathbf{x}_i$ is a vector of descriptors for a chemical compound and y_i indicates whether the compound is considered an active drug agent for a certain disease. Most compounds are inactive and we are interested in detecting the active ones.
- 2) Credit Card Fraud Detection [1]: Here $\mathbf{x}_i$ is a vector of descriptors for a credit card transaction and y_i indicates whether the transaction is fraudulent. Most transactions

are not fraudulent and we are interested in catching the frauds.

- 3) Direct Marketing: Here $\mathbf{x}_i$ is a vector of descriptors for a potential customer and y_i indicates whether the customer will respond to the advertisement or not. Most customers will not respond and we are interested in identifying those most likely to respond.

Rare target detection is analogous to information retrieval by treating the rare targets as relevant objects.

C. Performance Measures

Many measures for evaluating the performance of information retrieval systems have been proposed, among which recall and precision are the most popular ones. Suppose there are n objects in the database, among which m are relevant. Let $s_1, \dots, s_n$ be the relevant scores of the objects, $s_{(1)}, \dots, s_{(n)}$ be the corresponding sorted scores in a descending order, and $y_{(1)}, \dots, y_{(n)}$ be binary variables indicating whether the i th ranked document is relevant to the query or not. That is,

$$y_{(i)} = \begin{cases} 1 & \text{the } i\text{th object is relevant;} \\ 0 & \text{the } i\text{th object is irrelevant.} \end{cases}$$

A retrieved item is called a hit if it is a relevant object. Define the hit function $h(t)$ as the number of hits among the top t objects, i.e., $h(t) = \sum_{i=1}^t y_{(i)}$. Recall at rank t , $r(t)$, is defined as the fraction of relevant objects that are successfully retrieved among the top t objects. Precision at rank t , $p(t)$, is the fraction of the retrieved objects that are relevant to the query. That is,

$$r(t) = \frac{h(t)}{m}, \quad p(t) = \frac{h(t)}{t}. \quad (1)$$

Recall and precision can be also used in rare target detection problems. Similar to information retrieval, the first step to solve a rare target detection problem is to rank the items by a score function which can be modelled by K -nearest neighbours (KNN, e.g., [2], [3]), support vector machines (SVM, e.g., [4]), LAGO [5] or other methods. Then according to one's budget, the top Q items with the largest scores are classified as class 1 and the remaining items are classified as class 0. The classification result is usually presented in a confusion matrix presented in Table I.

True label	classification result	
	1	0
1	TP (true positive)	FN (false negative)
0	FP (false positive)	TN (true negative)

TABLE I
CONFUSION MATRIX OF CLASSIFICATION

In the context of rare target detection,

$$\text{recall} = \frac{TP}{TP + FN}, \quad \text{precision} = \frac{TP}{TP + FP}.$$

Given Q , the ranking system gives larger $r(Q)$ or $p(Q)$ is the better. However, there is a trade-off between recall and precision in such a way that recall $r(t)$ generally increases with t while precision $p(t)$ decreases with t . More details about this relationship between recall and precision can be found in [6] and [7]. Therefore, recall and precision should be taken into account simultaneously when we evaluate or compare different ranking systems. One common way to combine recall and precision is the average precision (AP), which is defined as the average of the precisions of those hits. That is,

$$\text{AP} = \frac{1}{m} \sum_{t=1}^n y(t) \frac{h(t)}{t} = \frac{1}{m} \sum_{t=1}^n y(t) \frac{\sum_{j=1}^t y(j)}{t}. \quad (2)$$

Suppose we need to rank eight items, among which only three belong to class 1. Table II shows the calculation of hit function $h(t)$, recall $r(t)$ and precision $p(t)$ at each rank $t = 1, \dots, 8$.

Rank (t)	Result at Each Rank							
	1	2	3	4	5	6	7	8
Hit list	1	1	0	1	0	0	0	0
Hit function $h(t)$	1	2	2	3	3	3	3	3
Recall $r(t)$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{3}{3}$	$\frac{3}{3}$
Precision $p(t)$	$\frac{1}{1}$	$\frac{2}{2}$	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{3}{5}$	$\frac{3}{6}$	$\frac{3}{7}$	$\frac{3}{8}$

TABLE II
CALCULATION OF HIT FUNCTION, RECALL AND PRECISION

Recall that average precision is calculated as the average of precisions over the hits. Since hits appear at the first, second and fourth places, the average precision is

$$\text{AP} = \frac{1}{3} \left(\frac{1}{1} + \frac{2}{2} + \frac{3}{4} \right).$$

Plotting $p(t)$ against $r(t)$ gives the recall-precision curve, average precision is an approximation to the area under the curve (e.g., [8]).

D. Existing methods

To our best knowledge, little has been done on making statistical inference on the performance measure. Yeh [9] suggested using pair-t, signed rank and Wilcoxon test for comparing recalls and using computationally intensive randomization test for precision, nothing mentioned about average precision.

Goutte & Gaussier [10] handled this problem from a Bayesian aspect, they argued that the joint distribution of (TP, FN, FP, TN) follows a multinomial distribution; and the conditional distribution of true positive given the total number of class-1 observation, $TP|TP+FP \sim \text{Binomial}(TP+FP, p)$; the conditional distribution of true positive given the number of retrieved items, $TP|TP+FN \sim \text{Binomial}(TP+FN, r)$. Adopting conjugate Beta priors to precision p and recall r , inference on p and r can be made on the posterior distribution. 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. A Dirichlet prior is assigned to (π_1, π_2, π_3) , and the posterior distribution of π_1, π_2, π_3 is another Dirichlet distribution. By drawing samples from the posterior distribution, one can calculate the probability that system 1 is superior to system 2, i.e., $Pr(\pi_1 > \pi_2)$ empirically by Monte Carlo method.

The rest of the paper is organized as follows. Section II gives the distributions of recall, precision and average precision under random selection. Our approach to making inference on recall, precision and average precision is also presented. Section III presents a simulation study. We end with a discussion in section IV.

II. DISTRIBUTIONS OF RECALL, PRECISION AND AVERAGE PRECISION UNDER RANDOM SELECTION

A. Distributions of Recall and Precision

As given in Equation (1), recall and precision at rank t are defined as

$$r(t) = \frac{h(t)}{m}, \quad p(t) = \frac{h(t)}{t},$$

where the hit function $h(t)$ is the number of hits among the top t retrieved objects. Suppose there are totally n objects, among which m are targets (class 1). Under random selection, every item has the same probability of being retrieved among the top- t list; therefore, the number of hits among the top t items, $h(t)$, follows a hypergeometric distribution with parameters (n, m, t) . As a result,

$$E\{h(t)\} = \frac{tm}{n}, \quad \text{Var}\{h(t)\} = \frac{tm}{n} \frac{(n-m)(n-t)}{n(n-1)}.$$

Hence, the expected value of recall $r(t)$ and precision $p(t)$ are

$$E\{r(t)\} = \frac{t}{n}, \quad E\{p(t)\} = \frac{m}{n}; \quad (3)$$

and correspondingly the variance of $r(t)$ and $p(t)$ are

$$\text{Var}\{r(t)\} = \frac{t(n-m)(n-t)}{mn^2(n-1)}, \quad \text{Var}\{p(t)\} = \frac{m(n-m)(n-t)}{tn^2(n-1)}. \quad (4)$$

B. The Distribution of AP Under Random Selection

Under random selection, the first two moments of average precision can be exactly calculated by establishing the marginal and joint distributions of the location of hits. Recall that average precision is calculated as the average of precisions over the hits. Since hits appear at the first, second and fourth places, the average precision is

$$AP = \frac{1}{3} \left(\frac{1}{1} + \frac{2}{2} + \frac{3}{4} \right).$$

Let $L_1, L_2, \dots, L_m$ be the ranks (locations) of the i th hits, average precision given in (2) can be rewritten as

$$AP = \frac{1}{m} \left(\frac{1}{L_1} + \frac{2}{L_2} + \dots + \frac{m}{L_m} \right). \quad (5)$$

In our example, $L_1 = 1, L_2 = 2, L_3 = 4$. Viewing average precision in this way, we can calculate the moments of average precision by finding the marginal distributions of L_i 's. Suppose there are totally n observations among which m objects belong to class 1. Fixing L_i at the location k , the total distinct ways to arrange these m class-1 objects is $\binom{n-1}{m-1}$. The first hit L_1 can take values of $1, 2, \dots, n-m+1$, the second hit L_2 takes $2, 3, \dots, n-m+2$, the m th hit L_m takes $m, \dots, n$.

In general, the probability that the j th hits appears at the k th position, P_{jk} , is given by

$$P_{jk} = P(L_j = k) = \frac{\binom{n-k}{m-j} \binom{k-1}{j-1}}{\binom{n-1}{m-1}} \quad (6)$$

This is due to the fact that the j th hits at the k th position implies $j-1$ hits appearing in the previous $k-1$ locations and the remaining $j-1$ hits appearing in the rest $k-1$ positions. Similarly, the joint distribution of L_i and L_j for $i < j$ can be calculated as

$$P(L_i = s, L_j = t) = \frac{\binom{s-1}{i-1} \binom{t-s-1}{j-i-1} \binom{n-t}{m-j}}{\binom{n-2}{m-2}}, \quad (7)$$

which means $i-1$ hits appearing among the first $s-1$ locations, $j-i-1$ hits between the s th and t th positions, and $m-j$ among the last $n-t$ positions. The denominator is the total number of arranging $m-2$ hits among $n-2$ locations, fixing L_i and L_j . Given the marginal distribution (6) and the joint distribution (7), it is straight forward to calculate the expectation and variance of AP given in (5). That is

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

where

$$E(AP^2) = \frac{1}{m^2} \left\{ \sum_{i=1}^m E\left(\frac{i^2}{L_i^2}\right) + 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m E\left(\frac{ij}{L_i L_j}\right) \right\}$$

with

$$E\left(\frac{1}{L_i}\right) = \sum_{t=1}^n \frac{1}{t} P(L_i = t), \quad E\left(\frac{1}{L_i^2}\right) = \sum_{t=1}^n \frac{1}{t^2} P(L_i = t),$$

$$E\left(\frac{1}{L_i L_j}\right) = \sum_{s=1}^n \sum_{t=1}^n \frac{1}{st} P(L_i = s, L_j = t).$$

This method, however, is computationally expensive when m and n are large. The following approach based on an approximate normal distribution is appropriate then. We present the result here and leave the detailed derivation in an appendix.

With straight forward algebra we have,

$$E(L_i) = \frac{i n}{m}, \quad (8)$$

$$E(L_i^2) = \frac{in(in+i+n-m)}{m(m+1)},$$

$$Var(L_i) = \frac{in(in+i+n-m)}{m(m+1)} - \frac{i^2 n^2}{m^2},$$

$$E(L_i L_j) = \frac{i(n-1)(nj+n-m)}{m(m-1)}, \quad (9)$$

$$Cov(L_i, L_j) = E(L_i L_j) - E(L_i)E(L_j). \quad (10)$$

where $E(L_i)$, $E(L_j)$ and $E(L_i L_j)$ are given by equations (8) and (9).

The moments of $\frac{1}{L_i}$ are obtained with a second order Taylor expansion of functions $f(x) = \frac{1}{x}$ and $g(x) = \frac{1}{x^2}$ at $x_0 = E(L_i)$. The $Cov(L_i, L_j)$ is obtained using delta method at $E(L_i), E(L_j)$.

$$E\left(\frac{1}{L_i}\right) \approx \frac{1}{E(L_i)} \left\{ 1 + \frac{Var(L_i)}{E^2(L_i)} \right\}, \quad (11)$$

$$E\left(\frac{1}{L_i^2}\right) \approx \frac{1}{E^2(L_i)} \left\{ 1 + \frac{3Var(L_i)}{E^2(L_i)} \right\},$$

$$Var\left(\frac{1}{L_i}\right) \approx \frac{Var(L_i)}{E^4(L_i)} \left\{ 1 - \frac{Var(L_i)}{E^2(L_i)} \right\},$$

$$Cov\left(\frac{1}{L_i}, \frac{1}{L_j}\right) \approx \frac{Cov(L_i, L_j)}{E^2(L_i)E^2(L_j)}.$$

where $E(L_i)$, $E(L_j)$ and $Cov(L_i, L_j)$ are given by equations (8) and (10) respectively.

Because $AP = \frac{1}{n} \sum_{i=1}^m \frac{E(L_i)}{L_i}$ by equation (8), and $E\left\{\frac{E(L_i)}{L_i}\right\} \approx 1 + \frac{Var(L_i)}{E^2(L_i)}$ by equation (11), which goes to 1 when m goes to ∞ , it can be shown, using central limit theorem on weakly dependent variables, that AP is approximately normally distributed, $AP \stackrel{L}{\sim} N(\mu, \sigma^2)$ where

$$\mu = \frac{1}{m} \sum_{i=1}^m E\left(\frac{i}{L_i}\right) \quad (12)$$

$$\sigma^2 = \frac{1}{m^2} \sum_{i=1}^m Var\left(\frac{i}{L_i}\right) + \frac{2}{m^2} \sum_{i < j} Cov\left(\frac{i}{L_i}, \frac{j}{L_j}\right) \quad (13)$$

III. SIMULATION STUDY

We carry out a simulation study to verify the validity of our approach on recall, precision and average precision presented in Section II. We choose two settings, with $m = 100, n =$

1000 and $m = 500, n = 2000$ and go through the following steps:

- 1) Generate a random permutation,
- 2) Calculate the value of recall, precision when $t = m$ (they are the same) and average precision,
- 3) Repeat steps 1) and 2) 10000 times, obtain a sampling distribution of recall (precision) and average precision,
- 4) Find the empirical mean and variance of those 10000 recalls and average precisions,

The theoretical means and variances of recall and precision are calculated with equations (3) and (4), which give $E\{r(100)\} = E\{p(100)\} = 0.1$ and $Var\{r(100)\} = Var\{p(100)\} = 0.00081$ for $m = 100, n = 1000$ and $E\{r(500)\} = 0.25$ and $Var\{r(500)\} = 0.00028$ for $m = 500, n = 2000$. The corresponding empirical means and variances, obtained from the simulation, are 0.099985, 0.000824 for $m = 100, n = 1000$ and 0.250023, 0.000284 for $m = 500, n = 2000$, respectively. Table III lists the theoretical and empirical means and variances, as well as selected quantiles of the distribution of the simulated average precisions, with the first two rows showing the first setting and the next two rows presenting results for the second setting. The histogram of the simulated average precisions of the second setting is presented in Figure 1 with normal density curve of the limiting distribution overlaid. It is shown clearly that the result is satisfactory.

	mean	variance	2.5%	50%	97.5%
Empirical	.1037	.0001286	.0876	.1044	.1321
Theoretical	.1037	.0001392	.0806	.1037	.1268
Empirical	.2527	.000096	.2347	.2521	.2731
Theoretical	.2522	.000142	.2288	.2522	.2755

TABLE III
COMPARISON BETWEEN EMPIRICAL AND THEORETICAL DISTRIBUTIONS
OF SIMULATED AVERAGE PRECISIONS, ON MEANS, VARIANCES AND
QUANTILES.

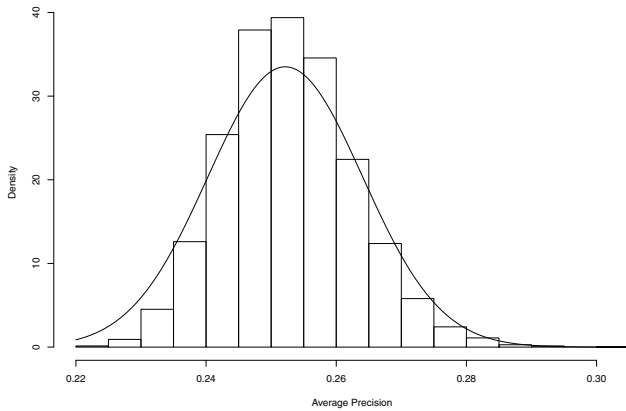


Fig. 1. Histogram of simulated average precisions with theoretical density.

IV. CONCLUSION

Recall, precision and average precision are three performance measures in evaluating different ranking algorithms for rare target detection problems. Little has been done on the statistical properties of these three measures. In this article, we derived the sampling distribution for recall, precision and average precision under random selection. More important, we provided a theoretical way to calculate the limiting distribution for average precisions when the number of hits is large. Based on the limiting distribution, hypothesis testing can be conducted to test whether one ranking algorithm significantly outperforms the other in detecting the rare targets.

APPENDIX

1) Derivation of moments of L_i :

$$\begin{aligned}
 E(L_i) &= \sum_{k=i}^{n-m+i} \frac{k \binom{n-k}{m-j} \binom{k-1}{j-1}}{\binom{n-1}{m-1}} \\
 &= \sum_{k=i}^{n-m+i} \frac{i \binom{k}{j} \binom{n-k}{m-j}}{\frac{m}{n} \binom{n}{m}} \\
 &= \frac{in}{m}
 \end{aligned}$$

$$\begin{aligned}
 E(L_i^2) &= \sum_{k=i}^{n-m+i} \frac{k^2 \binom{n-k}{m-j} \binom{k-1}{j-1}}{\binom{n-1}{m-1}} \\
 &= \sum_{k=i}^{n-m+i} \frac{k(k+1) \binom{n-k}{m-j} \binom{k-1}{j-1}}{\binom{n-1}{m-1}} \\
 &\quad - \sum_{k=i}^{n-m+i} \frac{k \binom{n-k}{m-j} \binom{k-1}{j-1}}{\binom{n-1}{m-1}} \\
 &= \sum_{k=i}^{n-m+i} \frac{i(i+1) \binom{k+1}{j+1} \binom{n-k}{m-j}}{\frac{m(m+1)}{n(n+1)} \binom{n+1}{m+1}} - E(L_i) \\
 &= \frac{in(i+1)(n+1)}{m(m+1)} - \frac{in}{m} \\
 &= \frac{in(in+i+n-m)}{m(m+1)} \\
 Var(L_i) &= E(L_i^2) - E^2(L_i) \\
 &= \frac{in(in+i+n-m)}{m(m+1)} - \frac{i^2n^2}{m^2}
 \end{aligned}$$

$$E(L_i L_j)$$

$$\begin{aligned}
&= \sum_s \sum_t \frac{st \binom{s-1}{i-1} \binom{t-s-1}{j-i-1} \binom{n-t}{m-j}}{\binom{n-2}{m-2}} \\
&= A + B - C \\
A &= \sum_s \sum_t \frac{s(t-s) \binom{s-1}{i-1} \binom{t-s-1}{j-i-1} \binom{n-t}{m-j}}{\binom{n-2}{m-2}} \\
&= \sum_s \sum_t \frac{i(j-i) \binom{s}{i} \binom{t-s}{j-i} \binom{n-t}{m-j}}{\frac{m(m-1)}{n(n-1)} \binom{n}{m}} \\
&= \frac{in(j-i)(n-1)}{m(m-1)} \\
B &= \sum_s \sum_t \frac{s(s+1) \binom{s-1}{i-1} \binom{t-s-1}{j-i-1} \binom{n-t}{m-j}}{\binom{n-2}{m-2}} \\
&= \sum_s \sum_t \frac{i(i+1) \binom{s}{i} \binom{t-s-1}{j-i-1} \binom{n-t}{m-j}}{\frac{m(m-1)}{n(n-1)} \binom{n-1}{m-1}} \\
&= \frac{in(i+1)(n-1)}{m(m-1)} \\
C &= \sum_s \sum_t \frac{s \binom{s-1}{i-1} \binom{t-s-1}{j-i-1} \binom{n-t}{m-j}}{\binom{n-2}{m-2}} \\
&= \sum_s \sum_t \frac{i \binom{s}{i} \binom{t-s-1}{j-i-1} \binom{n-t}{m-j}}{\frac{m-1}{n-1} \binom{n-1}{m-1}} \\
&= \frac{i(n-1)}{m-1} \\
E(L_i L_j) &= \frac{i(n-1)(nj+n-m)}{m(m-1)}
\end{aligned}$$

2) Moments of $\frac{1}{L_i}$:

The second order Taylor expansion of $f(x) = \frac{1}{x}$ at $x_0 = E(L_i)$ is given by

$$f(x) \approx \frac{1}{x_0} - \frac{x - x_0}{x_0^2} + \frac{(x - x_0)^2}{x_0^3}.$$

Hence

$$\frac{1}{L_i} \approx \frac{1}{E(L_i)} - \frac{L_i - E(L_i)}{E^2(L_i)} + \frac{\{L_i - E(L_i)\}^2}{E^3(L_i)}$$

and

$$\begin{aligned}
E\left(\frac{1}{L_i}\right) &\approx \frac{1}{E(L_i)} + \frac{\text{Var}(L_i)}{E^3(L_i)} \\
&= \frac{1}{E(L_i)} \left\{ 1 + \frac{\text{Var}(L_i)}{E^2(L_i)} \right\}.
\end{aligned}$$

The second order Taylor expansion of $g(x) = \frac{1}{x^2}$ at $x_0 = E(L_i)$ is given by

$$g(x) \approx \frac{1}{x_0^2} - \frac{2(x - x_0)}{x_0^3} + \frac{3(x - x_0)^2}{x_0^4}.$$

Hence

$$\frac{1}{L_i^2} \approx \frac{1}{E^2(L_i)} - \frac{2\{L_i - E(L_i)\}}{E^3(L_i)} + \frac{3\{L_i - E(L_i)\}^2}{E^4(L_i)}$$

and

$$\begin{aligned}
E\left(\frac{1}{L_i^2}\right) &\approx \frac{1}{E^2(L_i)} + \frac{3\text{Var}(L_i)}{E^4(L_i)} \\
&= \frac{1}{E^2(L_i)} \left\{ 1 + \frac{3\text{Var}(L_i)}{E^2(L_i)} \right\}.
\end{aligned}$$

Then

$$\begin{aligned}
\text{Var}\left(\frac{1}{L_i}\right) &= E\left(\frac{1}{L_i^2}\right) - E^2\left(\frac{1}{L_i}\right) \\
&= \frac{\text{Var}(L_i)}{E^4(L_i)} \left\{ 1 - \frac{\text{Var}(L_i)}{E^2(L_i)} \right\}
\end{aligned}$$

REFERENCES

- [1] Bolton, R. J. and Hand, D. J. (2002), Statistical Fraud Detection: A Review, *Statistical Science*, **17**, 235–255.
- [2] Fix, E. and Hodges, J. L. (1951), Discriminatory Analysis Nonparametric Discrimination: Consistency Properties, Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas.
- [3] Cover, T. and Hart, P. (1967), Nearest Neighbor Pattern Classification, *IEEE Transactions on Information Theory*, **IT-13**, 21–27.
- [4] Cortes, C. and Vapnik, V. (1995), Support Vector Networks, *Machine Learning*, **20**, 273–294.
- [5] Zhu, M., Su, W. and Chipman, H. A. (2006), LAGO: A Computationally Efficient Approach for Statistical Detection, *Technometrics*, **48**, 193–205.
- [6] Michael Buckland and Fredric Gey (1994), The Relationship Between Recall and Precision, *Journal of the American Society for Information Science*, **45**(1), 12–19.
- [7] Zhu, Mu. (2004), Recall, Precision and Average Precision, unpublished technical report, Department of Statistics and Actuarial Science, University of Waterloo.
- [8] Javed A. Aslam, Emine Yilmaz, and Virgiliu Pavlu (2005), A Geometric Interpretation of R-precision and Its Correlation with Average Precision, *Proceedings of the 28th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 573 - 574, Salvador, Brazil.
- [9] 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.
- [10] 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.