An exact analytical relation among recall, precision, and classification accuracy in information retrieval

Sergio A. Alvarez

Department of Computer Science

Boston College

140 Commonwealth Avenue

Chestnut Hill, MA 02467 USA

e-mail: alvarez@cs.bc.edu   phone: (617) 552-4333

Abstract

From a machine learning perspective, information retrieval may be viewed as a problem of classifying items into one of two classes corresponding to interesting and uninteresting items respectively. A natural performance metric in this context is *classification accuracy*, defined as the fraction of the system's interesting/uninteresting predictions that agree with the user's assessments. On the other hand, the field of information retrieval has two classical performance evaluation metrics: *precision*, the fraction of the items retrieved by the system that are interesting to the user, and *recall*, the fraction of the items of interest to the user that are retrieved by the system. In the present paper we present for the first time a general mathematical relation linking classification accuracy with precision and recall. Our relation provides an analytical tool that promises to be useful in theoretical and applied work in information retrieval. As a significant example in this direction, we show that our relation implies a trade-off between recall and precision under certain conditions on the accuracy.

1 Introduction

Consider an information retrieval system whose task it is to identify among a given collection of data items only those items that would be of interest to a given user. The performance of such a system is often gauged in terms of its precision p and recall r [9]. Precision measures retrieval specificity defined as the proportion of retrieved items that are judged by the user as being relevant; this measure penalizes system retrieval of irrelevant items (false positives) but does not penalize failures by the system to retrieve items that the user considers to be relevant (false negatives). Recall measures retrieval coverage defined as the proportion of the set of relevant items that is retrieved by the system, and therefore penalizes false negatives but not false positives. However, if one views the system as a classifier that attempts to label all data items as being either interesting or uninteresting, then the classification accuracy a, the fraction of the system's labelings that coincide with the user's opinions, is also a natural measure of performance. Unlike precision and recall, the classification accuracy gives equal weight to mislabelings of both types, that is, instances in which a relevant item is judged by the system to be irrelevant, as well as instances in which an irrelevant item is judged by the system to be relevant.

Classification accuracy is widely used as a metric for evaluation of machine learning systems. Machine learning techniques are being used to address many information filtering tasks, for instance recommender systems, e.g. [3], and adaptive web systems [4]. An earlier survey of several machine learning applications in information retrieval appears in [6]. Understanding how classification accuracy relates to more traditional information retrieval metrics is an important issue in the evaluation of such systems. Results in this direction can contribute to further stimulating the interaction between the fields of information retrieval and machine learning. In the present paper we address the connection between the two types of performance measures by establishing a general analytical result that relates recall, precision, and accuracy. We will show that recall r, precision p, and classification accuracy a satisfy the following general equation:

$$\lambda r + (\lambda + a - 1)\, p = 2\lambda p r \tag{1}$$

In Eq. 1, λ is the *generality*, the probability that the user will consider a randomly selected item to be interesting. Eq. 1 may be explicitly solved for any of the three measures r, p, a in terms of the other two and of λ if desired. More importantly, Eq. 1 provides an exact mathematical description of the dependence between these measures. In particular, Eq. 1 quantitatively describes, under certain well-defined conditions, a trade-off between recall and precision.

2 A motivating example

In this section we will present an example that illustrates the connections between accuracy, recall, and precision for a simple stochastic model of data generation. We will lay some groundwork for the discussion to follow, but we will not yet attempt to derive a general relation among these measures. Our objective in this example is largely to provide some familiarity with the performance measures involved in a concrete context. A derivation of the general relation of Eq. 1 will be provided in the next section of the paper.

2.1 Data generation

We assume that data items belong to one of two distinct populations A and B, each with a known prior probability of occurrence and with a known probability distribution for a shared numerical attribute x. This formulation is patterned after a typical Bayesian approach within machine learning; the use of a numerical attribute is intended to simplify the discussion of probability distributions.

The task of an information retrieval or filtering system in this context is to sort arbitrary instances, each defined by a value of x, into one of the two bins A and B. The relevant items may be identified with those that belong to the class A. Fig. 1 depicts one possible pair of (Gaussian) distributions of the attribute for the two classes. In summary, the model for generating data instances in this situation is the following. One of the two populations is first chosen according to the prior probabilities $P(A)$ and $P(B)$ (which sum to 1). A value of the attribute x is then selected according to the distribution of x for the chosen population.

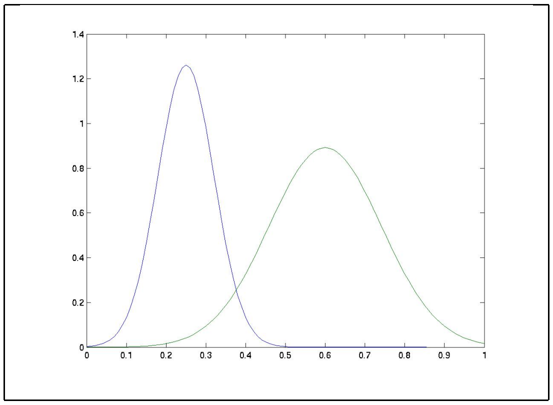

Figure 1: Two–class Gaussian mixture distribution

2.2 Retrieval method

Unless the two class conditional distributions of the attribute x are identical, it is possible to classify an instance simply by comparing its value of x with a threshold value θ. If the distribution of x given A is located below that of x given B as in Fig. 1, such a decision rule would state:

$$\begin{aligned} x < \theta \Rightarrow A \\ x > \theta \Rightarrow B \end{aligned} \tag{2}$$

We will assume in this example that retrieval/filtering is carried out based on the thresholding rule given in Eq. 2. The suitability of this method is of course dependent on the underlying statistics of the data. If one of the distributions were bimodal, for example, then relying on a single threshold would not be reasonable.

2.3 Performance measures

Interpreting class A as containing the relevant instances, the standard information retrieval performance measures may be expressed as functions of the decision threshold θ:

$$\begin{aligned} r(\theta) &= P(x < \theta \mid A) \\ p(\theta) &= P(A \mid x < \theta) \end{aligned}$$

The classification accuracy is given by:

$$\begin{aligned} a(\theta) &= P(A \cap x < \theta) \ + \ P(B \cap x > \theta) \\ &= P(A)P(x < \theta \mid A) \ + \ P(B)P(x > \theta \mid B) \end{aligned}$$

It is convenient to introduce the cumulative distribution functions F_A and F_B of the two classes:

$$\begin{aligned} F_A(\theta) &= P(x > \theta \mid A) \\ F_B(\theta) &= P(x > \theta \mid B) \end{aligned}$$

The performance measures may now be rewritten as follows:

$$\begin{aligned} r(\theta) &= F_A(\theta) \\ p(\theta) &= \frac{P(A)F_A(\theta)}{P(A)F_A(\theta) \ + \ P(B)F_B(\theta)} \qquad (3) \\ a(\theta) &= P(A)F_A(\theta) \ + \ P(B)(1 - F_B(\theta)) \end{aligned}$$

The expression for p in Eq. 3 follows from Bayes' rule:

$$\begin{aligned}P(A \mid x < \theta) &= \frac{P(A \ \cap \ x < \theta)}{P(x < \theta)} \\ &= \frac{P(A)P(x < \theta \mid A)}{P(A)P(x < \theta \mid A) \ + \ P(B)P(x < \theta \mid B)}\end{aligned}$$

2.4 Discussion

The behavior of the performance measures of Eq. 3 as functions of the decision threshold θ is shown in Fig. 2. We assume here that the a priori probabilities of the two classes are equal to 1/2 and that the conditional distributions of the attribute for the two classes are the Gaussian distributions shown in Fig. 1. Fig. 2 was obtained by plotting Eq. 3 using the software package MATLAB. Recall increases monotonically as the decision threshold increases, whereas precision decreases, illustrating the classical recall–precision trade–off. Accuracy, on the other hand, has a well–defined maximum; a machine learning system would normally be tuned to operate near the maximum point. This qualitative behavior of the measures is intuitively reasonable given the unimodal distributions used to generate Fig. 2.

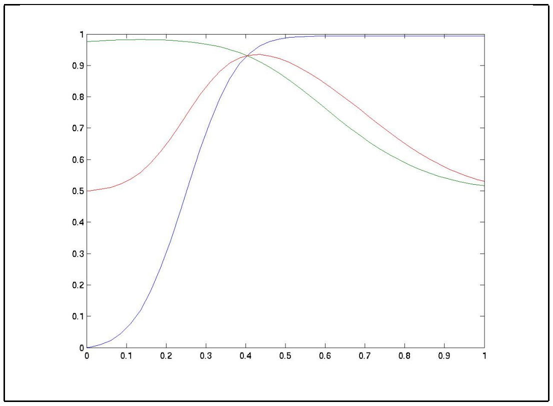

Figure 2: Recall, precision, and accuracy as functions of the decision threshold

Another salient feature of Fig. 2 is the existence of a point at which the curves corresponding to the three measures intersect. Let us try to understand the origin of this feature based on the expressions in Eq. 3. At a point at which recall and precision are equal, we have:

$$F_A(\theta) = \frac{P(A)F_A(\theta)}{P(A)F_A(\theta) \ + \ P(B)F_B(\theta)}$$

Substituting this constraint into the expression for the accuracy, we find:

$$a(\theta) = 2P(A)F_A(\theta) + P(A) - P(B)$$

Since $P(A)$ and $P(B)$ are both equal to $1/2$, we find:

$$a(\theta) = F_A(\theta) = r(\theta) = p(\theta),$$

which confirms the existence of the triple intersection point.

As regards dependencies among the measures considered here, it is clear from Eq. 3 that a relation is plausible, at least in the context of this example. For instance, the denominator in the expression for the precision involves terms similar to those that appear in the expression for the accuracy. However, it is not immediately clear what relation connects the measures. Also, it is not clear to what extent a relation gleaned in this situation would apply in general. In any case, we trust that this discussion, although based on a specific model of data generation and retrieval as explained above, serves to motivate the general treatment to follow.

3 The main formula

We now present a derivation of Eq. 1. We begin with a description of general context and notation.

3.1 Context and notation

Consider a data source (store or stream) that contains n items. The user of this data source wishes to retrieve items that are of interest in a given context. We assume that for each item in the data collection it is possible to determine (perhaps by directly asking the user) if the user would consider the item to be interesting or not. This produces a partition of the data collection into two disjoint subsets:

$$\begin{aligned} U^+ &= \{\text{all items that the user rates as being of interest}\} \\ U^- &= \{\text{all items that the user rates as } \textit{not} \text{ being of interest}\} \end{aligned} \tag{4}$$

We also assume that it can be determined which items would be recommended to the given user by the information filtering system under evaluation. Items not recommended by the system are considered to be implicitly labeled by it as being uninteresting. The system's implicit labelings of the data items produce a partition of the original collection of n items into two disjoint subsets:

$$\begin{aligned} S^+ &= \{\text{all items that the system rates as being of interest}\} \\ S^- &= \{\text{all items that the system does not rate as being of interest}\} \end{aligned} \tag{5}$$

Taken together, the partitions in Eq. 4 and Eq. 5 produce a four-cell partition of the original collection of n items. For brevity we denote the item counts in these cells as follows:

$$n_1 = \#S^+ \cap U^+, \quad n_2 \; = \#S^+ \cap U^-, \quad n_3 = \#S^- \cap U^+, \quad n_4 \; = \#S^- \cap U^- \tag{6}$$

Since the four cells taken together constitute the original data collection, simple accounting leads to the relation:

$$n_1 + n_2 + n_3 + n_4 \;=\; n \tag{7}$$

3.2 Performance measures

We can express our three measures, recall r, precision p, and accuracy a in terms of the numbers n_i defined in Eq. 6:

$$p = \frac{n_1}{n_1 + n_2}, \qquad r \;\; = \frac{n_1}{n_1 + n_3}, \qquad a = \frac{n_1 + n_4}{n} \tag{8}$$

We may also describe the situation in probabilistic notation. This notation is similar to that used in the example of section 2. However, unlike the example, here we do not make any assumptions about what data attributes are available for retrieval purposes, or about what technique (such as thresholding as in the example) is used to carry out the retrieval. The connection between the two notations involves the normalized cell counts:

$$\frac{n_1}{n} = P(S^+ \cap U^+), \quad \frac{n_2}{n} \;\; = P(S^+ \cap U^-), \quad \frac{n_3}{n} = P(S^- \cap U^+), \quad \frac{n_4}{n} \;\; = P(S^- \cap U^-) \tag{9}$$

In probabilistic notation, the performance measures become:

$$r = P(S^+ \mid U^+), \quad p \;\; = P(U^+ \mid S^+), \quad a = P(U^+ \cap S^+) \;+\; P(S^- \cap U^-) \tag{10}$$

Using Bayes' rule, the performance measures of Eq. 10 may be written as follows:

$$\begin{aligned} r &= P(S^+ \mid U^+) \\ p &= \frac{P(U^+)P(S^+ \mid U^+)}{P(U^+)P(S^+ \mid U^+) \;+\; P(U^-)P(S^+ \mid U^-)} \\ a &= P(U^+)P(S^+ \mid U^+) \;+\; P(U^-)P(S^- \mid U^-) \end{aligned} \tag{11}$$

3.3 Derivation of the formula

We now proceed to derive the precision-recall trade-off formula given in Eq. 1. We will use the first notation in terms of cell counts as described above. A basic idea behind the derivation is that the numbers $r,\ p,\ a,\ n$ may be viewed as providing alternate "coordinates" for the cell counts $n_1,\ n_2,\ n_3,\ n_4$. However, there are four numbers in each set of coordinates, and the four n_i values may be chosen independently, so no relation should be expected among the values $r,\ p,\ a,\ n$. Thus, an additional value is needed. We will use the generality λ, the fraction of items in the collection that the user considers interesting:

$$\lambda = \frac{n_1 + n_3}{n} \tag{12}$$

The extended set of five numbers $r,\ p,\ a,\ \lambda,\ n$ provides alternate "coordinates" for the cell counts $n_1,\ n_2,\ n_3,\ n_4$. Since there are five new numbers and only four n_i's, one can expect there to be a relation connecting the new numbers. This relation is the formula we seek. The derivation will proceed by rewriting the four numbers n_i in terms of the new coordinates. Using Eqs. 8 and 12, we obtain an expression for n_1:

$$n_1 = r\lambda n \tag{13}$$

Eq. 12 now yields an expression for n_3:

$$n_3 = (1 - r)\lambda n \tag{14}$$

Expressions for n_2 and n_4 now follow from the appropriate portions of Eq. 8:

$$n_2 = \frac{1-p}{p} r\lambda n, \qquad n_4 \quad = (a - r\lambda)n \tag{15}$$

At this point we invoke the counting constraint that the numbers n_i sum to n:

$$r\lambda n + (1-r)\lambda n + \frac{1-p}{p} r\lambda n + (a - r\lambda)n \;=\; n \tag{16}$$

After canceling the factor n and reorganizing the remaining terms, we arrive at the relation (Eq. 1) that we have been seeking:

$$\lambda r + (\lambda + a - 1)\, p \;=\; 2\lambda p r$$

This relation, equivalently Eq. 1, provides a mutual constraint linking precision p, recall r, and classification accuracy a, assuming that one knows the user's *a priori* probability λ of being interested in a randomly selected item.

3.4 Corollaries

Observe that Eq. 1 allows one to obtain an explicit formula for either one of the two quantities p, r in terms of the other. For example, solving for r we obtain:

$$r \;=\; \frac{(\lambda + a - 1)\, p}{\lambda(2p - 1)} \tag{17}$$

The analogous expression for p is:

$$p \;=\; \frac{\lambda r}{\lambda(2r-1)+1-a} \tag{18}$$

Also, Eq. 1 allows us to express the accuracy in terms of recall and precision:

$$a \;=\; 1 \;-\; \lambda\,\frac{p+r-2pr}{p} \tag{19}$$

The above expressions are clearly useful. For instance, for future reference we note that one may derive from Eq. 17 the following fact relating precision and accuracy:

$$\mathrm{sign}\,(\lambda+a-1) = \mathrm{sign}\,(2p-1) \tag{20}$$

4 Examples

This section contains examples that verify that the relation of Eq. 1 is correct in specific situations.

4.1 Random guessing

Assume that a system randomly labels data items as relevant or not relevant, with each of the two possible labels being chosen with probability 1/2. A key property of such a system is that the labels chosen by it are independent in the probabilistic sense from the user's judgments:

$$\begin{aligned} P(S^+|U^+) = P(S^+), \quad P(S^+|U^-) &= P(S^+) \\ P(S^-|U^+) = P(S^-), \quad P(S^-|U^-) &= P(S^-) \end{aligned} \tag{21}$$

Such a system therefore has the following values for the three performance measures:

$$
\begin{aligned}
r &= P(S^+|U^+) = P(S^+) = 1/2 \\
p &= P(U^+|S^+) = \frac{P(U^+)P(S^+|U^+)}{P(S^+)} = \lambda \\
a &= P(U^+)P(S^+|U^+) \ + \ P(U^-)P(S^-|U^-) = \lambda/2 + (1-\lambda)/2 = 1/2
\end{aligned}
$$

The truth of Eq. 1 in this case may now be verified directly:

$$
\begin{aligned}
\lambda r + (\lambda + a - 1)\, p &= \lambda/2 + (\lambda - 1/2)\lambda \\
&= \lambda^2 \\
&= 2\lambda\lambda 1/2 \\
&= 2\lambda p r
\end{aligned}
$$

4.2 Improved random guessing

A related example is obtained by considering a system that chooses labels randomly but with probabilities that match those of the user. That is, if λ denotes the probability that the user will consider an item to be relevant, then the system behaves according to the following description:

$$
P(S^+|U^+) = P(S^+) = \lambda, \qquad P(S^+|U^-) = P(S^+) = \lambda
$$

The values of the performance measures are found to be the following:

$$
\begin{aligned}
r &= P(S^+|U^+) = P(S^+) = \lambda \\
p &= P(U^+|S^+) = \frac{P(U^+)P(S^+|U^+)}{P(S^+)} = \lambda \\
a &= P(U^+)P(S^+|U^+) \ + \ P(U^-)P(S^-|U^-) = \lambda^2 + (1-\lambda)^2
\end{aligned}
$$

The recall of such a system will be greater than that of the fair coin tossing system of the previous example assuming that λ exceeds 1/2. Precision remains unchanged, and accuracy has also improved since the quadratic expression for a above has a minimum value of 1/2 (the previously attained accuracy) at $\lambda = 1/2$. We now verify Eq. 1 for this system:

$$\begin{aligned}
\lambda r + (\lambda + a - 1)\, p &= \lambda^2 + (\lambda + \lambda^2 + (1-\lambda)^2 - 1)\lambda \\
&= 2\lambda^2 + \lambda^3 + (-2\lambda + \lambda^2)\lambda \\
&= 2\lambda^3 \\
&= 2\lambda p r
\end{aligned}$$

4.3 One–dimensional mixture model

Let us revisit the example discussed above in section 2. We assume that data instances are to be classified into one of two classes A and B according to the value of a single numerical attribute as described there. We will verify that the forms of the measures for this situation as found in Eq. 3 satisfy the general relation of Eq. 1. Instead of going through a long calculation, it suffices to observe that the general expressions for the performance measures in Eq. 11 derived during the proof of the formula of Eq. 1 reduce to the expressions of Eq. 3 for the situation described in the example. Indeed, we have the following identities relating the notations in the two situations:

$$\begin{aligned}
P(S^+|U^+) &= F_A \\
P(S^+|U^-) &= 1 - F_B \\
P(U^+) &= P(A) \\
P(U^-) &= P(B)
\end{aligned}$$

These identities show that the example satisfies the general relation of Eq. 1.

In passing, note that if the a priori class probabilities are 1/2 and if the class conditional distributions of the numerical attribute are uniform, not Gaussian as in the original example, then the accuracy actually remains constant across the interval of values of the decision threshold over which the two distributions overlap. This follows from the fact that the cumulative distribution functions F_A and F_B are piecewise linear in this case and the linear growth terms precisely cancel one another because they have opposite signs in the expression for the accuracy in Eq. 3.

5 Trade-off between recall and precision

It is somewhat of a "folk theorem" in information retrieval that there is a trade-off between precision and recall in a given context, with an increase in one of these two measures leading to a decrease in the other. Design choices may favor high values of either precision or recall at the expense of low values in the other. There is much empirical information supporting this heuristic inverse relation between precision and recall. However, it is clear that in a completely unrestricted setting a trade-off between precision and recall does not necessarily occur, as a system that returns only uninteresting items has both zero recall and zero precision, while the perfect system referred to above has unit recall and unit precision. Thus, any argument leading to a trade-off between precision and recall must necessarily involve some additional constraints. Previous work concerning the origins of the recall-precision tradeoff [5], [7], assumes knowledge of the size of the set of retrieved documents as a fraction of the total number of documents in the database. In [7], the additional concept of *fallout* is also invoked to mediate between recall and precision.

In this section we will show that our general formula of Eq. 1 allows us to precisely describe an inverse dependence or trade-off between recall and precision under certain conditions on the classification accuracy. Under the assumption of constant accuracy, for example, we show that recall–precision curves must be hyperbolic arcs. We establish a concise condition for which a recall–precision trade–off occurs in this context, namely, that the precision of the system remain higher than that of random guessing.

5.1 Analytical form of the recall–precision relation at constant accuracy

Notice that the functional dependence of r on p, or vice-versa, is not completely explicit in formulas such as Eq. 17 and Eq. 18 unless independent information about the accuracy a is available. Happily, even if the accuracy is not completely known in advance, it is still possible to derive a trade-off between recall and precision from Eq. 17 or its reciprocal version, Eq. 18. For example, the accuracy may be known to remain constant, as occurs in situations similar to that discussed in the two–class mixture model example above. Specifically, it may be shown that if the class conditional distributions of the attribute in that example are uniform (instead of being unimodal as discussed), then accuracy remains constant across a range of values of the decision threshold. As we will show, the resulting recall–precision curves may be described simply and elegantly in such situations.

Under the assumption of constant accuracy, Eq. 1 is of the form

$$c_r r + c_p p = cpr, \tag{22}$$

where c_r, c_p, and c are constants independent of p and r. Consider a new system of coordinates (r', p') obtained from the original system (r, p) by a clockwise rotation through an angle of $\pi/4$ radians. The original coordinates may be expressed in terms of the new coordinates as follows:

$$\begin{aligned} r &= \frac{1}{\sqrt{2}}(r' - p') \\ p &= \frac{1}{\sqrt{2}}(r' + p') \end{aligned} \tag{23}$$

Using these expressions, the curve described by Eq. 22 reduces to the form:

$$(r' - c'_r)^2 - (p' - c'_p)^2 = c' \tag{24}$$

The values c'_r, c'_p, c' are constants formed by suitable combinations of the unprimed constants that appear in Eq. 22:

$$\begin{aligned} c'_r &= \frac{2\lambda + a - 1}{2\sqrt{2}\lambda} \\ c'_p &= \frac{1 - a}{2\sqrt{2}\lambda} \\ c' &= \frac{\lambda + a - 1}{2\lambda} \end{aligned} \tag{25}$$

Eq. 24 shows that the recall–precision curve in question is a hyperbolic arc.

5.2 Condition for a recall–precision trade–off

Sample curves for various values of the accuracy are shown in Fig. 3. This figure suggests an inverse dependence between recall and precision. However, for the purpose of rigorously establishing a trade-off between recall and precision, we must determine the precise orientation of these curves. In the original coordinate system (r, p), an inverse relationship between recall and precision requires that, modulo a translation, the curve occupy the "first" and "third" quadrants, that is, the regions in which r and p are of the same sign. In the rotated and translated coordinate system, this corresponds to the condition that the constant c' from Eq. 24 be positive. Using Eq. 25, this condition becomes:

$$\lambda + a - 1 > 0 \tag{26}$$

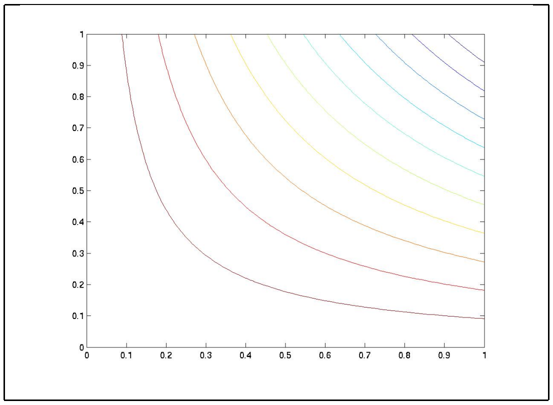

Figure 3: Precision–recall tradeoff at constant accuracy

Invoking Eq. 20, this establishes a definite condition under which a recall-precision trade-off will occur, namely:

$$p > \frac{1}{2} \tag{27}$$

A system that labels items as either relevant or not with equal probability (by tossing a fair coin, so to speak), provides the limiting case $p = 1/2$ of the trade-off condition in Eq. 27. Any system that performs better than random guessing automatically satisfies the condition and therefore must exhibit a tradeoff between recall and precision. A typical system will gradually increase the number of interesting item predictions in time. A transient initial phase may find the system erring in its interesting item predictions, so that the overall precision of the system is less than 1/2; *in this phase, no recall-precision tradeoff need occur*, and both recall and precision may increase together even if the accuracy remains constant. However, once the precision of the system exceeds 1/2, a recall-precision tradeoff is unavoidable if accuracy is constant.

5.3 Case of non–constant accuracy; sensitivity analysis

The above results establish a recall–precision trade–off under the assumption that the classification accuracy remains constant. If the accuracy varies, then it is possible that the inverse dependence between recall and precision will break down. For example, let us revisit the stochastic two–class mixture model for data generation discussed in the Example. Consider now a unimodal distribution for class A (the "target" class) and a bimodal distribution for class B. Assume that the unimodal distribution is nestled roughly in between the two modes of the bimodal distribution as in Fig. 4. Using MATLAB to plot Eq. 3 for the mixture distribution shown in Fig. 4, we obtain the recall, precision, and accuracy curves shown in Fig. 5. As Fig. 5 shows, recall increases from 0 to 1 as the decision threshold (on the horizontal axis) increases, while precision decreases from near 1 to about 0.6; the latter number is just the a priori probability of class A. Accuracy oscillates between 0.4 and 0.65 or so. However, there is a range of values of the decision threshold θ, approximately $0.8 < \theta < 0.9$, over which recall and precision increase together. This range overlaps the upper

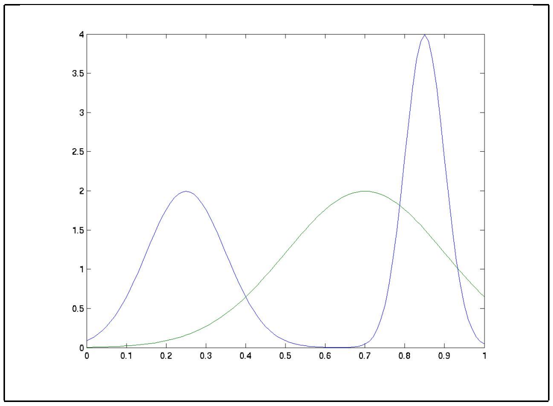

Figure 4: Two–class mixture distribution (bimodal target class)

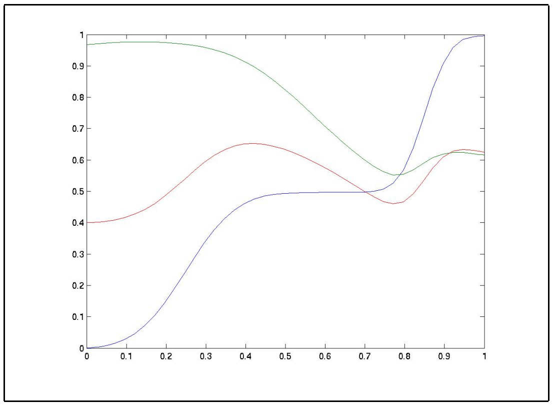

Figure 5: Recall, precision, and accuracy (bimodal distribution for target class)

mode of the bimodal distribution for class A.

However, this does not mean that non–constant accuracy excludes a recall–precision trade–off. In order to address this issue, we perform a sensitivity analysis. We view recall r as a function of the two variables p and a, with λ held constant as before. We then calculate the rate of change of r with respect to each of its two variables considered separately. This allows us to estimate how large a variation in accuracy a relative to the change in precision p will still lead to a recall–precision trade–off.

Taking partial derivatives with respect to p in Eq. 1, we obtain:

$$\lambda \frac{\partial r}{\partial p} + (\lambda + a - 1) = 2\lambda \left(r + p\frac{\partial r}{\partial p} \right) \tag{28}$$

Solving for the partial derivative of r with respect to p, we find:

$$\frac{\partial r}{\partial p} = \frac{\lambda(1-2r) + a - 1}{\lambda(2p-1)} \tag{29}$$

Through a similar process, we also obtain the partial derivative of r with respect to a:

$$\frac{\partial r}{\partial a} = \frac{p}{\lambda(2p-1)} \tag{30}$$

We have the following estimate of the total change Δr in r when both p and a vary by the amounts Δp and Δa respectively:

$$\Delta r = \frac{\partial r}{\partial p}\Delta p \;+\; \frac{\partial r}{\partial a}\Delta a \;+\; o\left(\|(\Delta p, \Delta a)\|\right), \tag{31}$$

where the ratio of the "little o" term divided by the length of the vector of deltas that appears in parentheses approaches 0 as the deltas approach 0 simultaneously. Using Eqs. 29 and 30, we see that the net change in r will still have a sign opposite to the change in p if the following condition holds:

$$p \; \Delta a \; < \; (\lambda(2r-1)+1-a) \; \Delta p \tag{32}$$

Conclusions and future work

We have derived an exact analytical relation (Eq. 1) that must be satisfied by the classification accuracy, recall, and precision of any information retrieval or filtering system. Assuming constant accuracy, the relation implies that recall–precision curves are hyperbolic arcs. The orientation of these curves in the recall–precision plane depend only on whether the system in question performs better than random guessing or not. We have derived from our results the empirically observed phenomenon of recall-precision tradeoff under this mild condition on the performance of the system. No previous work on the relation between recall and precision that we are aware of connects these measures with the concept of accuracy. This may be due to the fact that such work has arisen within information science contexts in which measures other than accuracy, such as fallout, are more commonly used. However, in applications such as recommender systems [8] and others that involve an interaction between machine learning and information retrieval, classification accuracy is of paramount importance. We hope that our results will therefore be of particular interest for ongoing work in such applications. For example, many machine learning systems exhibit during training first an increase and then a decrease in classification accuracy as measured on a set of test data (separate from the training data) as they begin to overfit the training data. Such systems will generally be tuned to operate near the point of maximum test accuracy. If variations in the accuracy near this point are small enough, the results presented in this paper may be used to derive a recall–precision tradeoff in such situations. We have empirically observed such tradeoffs in systems that employ neural networks [1] and plan to address this phenomenon in future work using the results of the present paper.

References

[1] Alvarez, S. A. & Ruiz, C. "Effect of Content Information on Neural Network-Based Collaborative Filtering Systems", *work in progress*

[2] Baeza–Yates, R. & Ribeiro–Neto, B. (1999). *Modern Information Retrieval*, ACM Press

[3] Billsus, D., & Pazzani, M. (1998). "Learning collaborative information filters", *Proc. Int. Conf. on Machine Learning*

[4] Brusilovsky, P., & Maybury, M. T. (2002). "From adaptive hypermedia to the adaptive web", Guest Editors' Introduction, *Communications of the ACM*, 45 (5), 31-33

[5] Buckland, M. K., & Gey, F. (1994). "The relationship between Recall and Precision", *Journal of the American Society for Information Science*, 45 (1), 12-19

[6] Chen, H. (1995) "Machine Learning for information retrieval: neural networks, symbolic learning and genetic algorithms", *Journal of the American Society for Information Science*, 46(3):194-216

[7] Gordon, M. D., & Kochen, M. (1989) "Recall-precision trade-off: a derivation", *Journal of the American Society for Information Science*, 40, 145-151

[8] Resnick P., & Varian, H. R. (1997). "Recommender Systems", Guest Editors' Introduction, *Communications of the ACM*, 40 (3), 56-58

[9] Van Rijsbergen, C. J. (1979) *Information Retrieval*, 2nd edition, Butterworths