CS 399 Paper

Mar 21, 2014

Bill Chickering (bchick)
Jamie Irvine (jirvine)

Abstract

Will write this when the paper is done

Introduction

The similarity of two items is a valuable measurement for a number of data-mining applications, such as recommender systems. With large amounts of rating data, collaborative filtering is a good method for calculating similarity. However, with sparse rating data, this technique gives noisy and unreliable approximations for similarity. Even worse, low confidence measurements of similarity are indistinguishable from high confidence ones. In this paper, we develop a novel method which incorporates confidence into the similarity score to produce more accurate estimates of similarity.

Item-based collaborative filtering is a common technique for measuring the similarity of two items. There are different forms of collaborative filtering. For this paper, we use a traditional approach. Each item is represented as a vector of ratings. The similarity score of two items is computed by measuring the similarity of the two rating vectors.

There are a few ways to compute the similarity of two vectors. One approach is to use Cosine-similarity, defined as the cosine of the angle between the two vectors:

$$CosSim(A,B) = \frac{\sum\limits_{u \in U_{AB}} r_{u,A} r_{u,B}}{\sqrt{\sum\limits_{u \in U_A} r_{u,A}^2}\sqrt{\sum\limits_{u \in U_B} r_{u,B}^2}} \tag{1}$$

where U_I is the set of all users who rated item I, U_{AB} is the set of all users who rated both item A and item B and $r_{u,I}$ is the rating user u gave to item I. Another popular approach is to use Pearson correlation:

$$PearsSim(A,B) = \frac{\sum\limits_{u \in U_{AB}} r_{u,A} r_{u,B}}{\sqrt{\sum\limits_{u \in U_{AB}} r_{u,A}^2}\sqrt{\sum\limits_{u \in U_{AB}} r_{u,B}^2}} \tag{2}$$

Other similarity functions exist, such as one-sided similarities, but Pearson correlation and Cosine-similarity are the most popular and will be the two measurements used in this paper. [1]

Both similarity measurements primarily leverage information from common users, that is users who have rated both items. Because of this, they perform well with a large number common users, but give unreliable results when the items have few common users. In an extreme case, if only one user has rated item A and item B, $PearsSim(A, B) = 1$ or -1. Not only is this unlikely to be an accurate assessment of the true similarity of A and B, it also gives the most extreme results possible, without giving any indication that this is a low-confidence calculation.

In this paper, we construct a more accurate similarity score by incorporating the number of common users, n, into the similarity function.

[1] Note that the difference between the two similarities is how they handle unpaired ratings; that is ratings from a user who has not rated the other item. Cosine-similarity considers the unknown rating from an unpaired rating to be 0 and then calculates the cosine of the two vectors. Pearson correlation simply throws away all unpaired ratings and calculates the cosine of the two modified vectors.

Problem

The goal is to calculate a modified similiary function that leverages the number of commonn users to better estimate the true similarity. Formally, we design $ModSim$ such that for two items A and B with n users in common

$$ModSim(A, B, n) \approx TrueSim(A, B) \tag{3}$$

where $TrueSim$ is the true similarity of the two items. Of course, there is no way to actually know the true similarity of two items, but for a good enough similarity function, such as $PearsSim$ or $CosSim$ and a large enough n, we can get a good estimate. Thus, we model $TrueSim$ as

$$TrueSim(A, B) = lim_{n \to \infty} Sim(A, B) \tag{4}$$

where Sim is the most appropriate generic similarity function for the dataset.

Since $TrueSim$ dependends on the choice of Sim, we abstract away the details of Sim and use its output directly. In all, the goal is to construct $ModSim$ such that

$$ModSim(Sim(A, B), n) \approx lim_{n \to \infty} Sim(A, B) \tag{5}$$

Model

We model the problem probabilistically. Let Y be a random variable representing the $TrueSim$ of a randomly chosen pair of items. Let X_n be a random variable representing the calculated Sim of a pair of items with n common users.

We model Y as a Normal distribution

$$Y \sim N(\mu, \sigma_1^2) \tag{6}$$

where μ and σ_1^2 are the average similarity score and the variance of similarity scores of all pairs of items, respectively. Since X_n represents a noisy reading of the true similarity Y, we model $X_n|Y$ as a Normal distribution around Y:

$$(X_n|Y = y) \sim N(y, \sigma_{2,n}^2) \tag{7}$$

Note that $\sigma_{2,n}^2$ represents how noisy the estimation of Sim is when there are n common users. This should decrease as n increases.

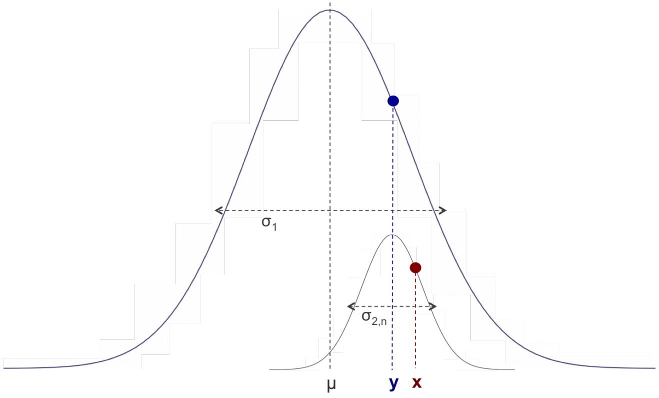

Figure 1: The blue curve represents Y, the distribution of true similarities between a random pair of items. The red curve represents $X|Y$, the estimated similarity when a small number of common users exist.

In probabilistic terms, for a calculated value X_n, we want to find the value of Y that most likely produced

X_n. Thus we desire the Maximum Likelihood Estimate of $Y|X_n$

$$\begin{aligned}
\operatorname*{argmax}_y P(Y=y|X_n=x) &= \operatorname*{argmax}_y \left[\frac{P(X_n=x|Y=y)P(Y=y)}{P(X_n=x)}\right] && (8)\\
&= \operatorname*{argmax}_y \left[P(X_n=x|Y=y)P(Y=y)\right] && (9)\\
&= \operatorname*{argmax}_y \left[\frac{1}{\sigma_{2,n}\sqrt{2\pi}}\exp\left(\frac{-(x-y)^2}{2\sigma_{2,n}^2}\right)\frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(\frac{-(y-\mu)^2}{2\sigma_1^2}\right)\right] && (10)\\
&= \operatorname*{argmin}_y \left[\frac{(x-y)^2}{2\sigma_{2,n}^2}+\frac{(y-\mu)^2}{2\sigma_1^2}\right] && (11)\\
&= \operatorname*{argmin}_y \left[\left(\sigma_1^2+\sigma_{2,n}^2\right)y^2-2\left(\sigma_1^2x+\sigma_{2,n}^2\mu\right)y\right] && (12)
\end{aligned}$$

We find the exact minimum by taking the derivative with respect to y and setting it to zero:

$$\begin{aligned}
\frac{d}{dy}\left[\left(\sigma_1^2+\sigma_{2,n}^2\right)y^2-2\left(\sigma_1^2x+\sigma_{2,n}^2\mu\right)y\right] &= 2\left(\sigma_1^2+\sigma_{2,n}^2\right)y-2\left(\sigma_1^2x+\sigma_{2,n}^2\mu\right) && (13)\\
&= 0 && (14)
\end{aligned}$$

Solving for y:

$$y=\frac{\sigma_1^2x+\sigma_{2,n}^2\mu}{\sigma_1^2+\sigma_{2,n}^2} \tag{15}$$

which can be rewritten as:

$$(y-\mu)=\frac{\sigma_1^2}{\sigma_1^2+\sigma_{2,n}^2}(x-\mu) \tag{16}$$

The model suggests that there is a linear correlation between X_n, the similarity calculated with n common users, and Y, the true similarity. Moreover, the best predicted y is a linear combination of the mean and the observed similiatiry. Since $\sigma_{2,n}^2 \geq 0$, the slope is always ≤ 1, meaning that the calculated distance from μ (the average similarity) overestimates the true distance. This makes sense, since there is a prior distribution of true similarity weighted around μ.

The parameters μ, σ_1^2, and $\sigma_{2,n}^2$ all can be approximated using the method of moments over training data where there are enough common users that $TrueSim$ can be approximated. This method requires approximating $\sigma_{2,n}^2$ for every value of n separately. Alternatively, we can model $\sigma_{2,n}^2$ as a function of n. Recall that $\sigma_{2,n}^2$ is the noisiness of the measurement Sim about $TrueSim$ and $TrueSim = lim_{n\to\infty}Sim$. Thus Sim can be thought of as a sampling n users from the infinite set used to calculate $TrueSim$. Although Sim may not be linear, we the Central Limit Theorem motivates the intuition that the variance of Sim would decrease as the $sqrt(n)$. Thus we model $\sigma_{2,n}^2$ as:

$$\sigma_{2,n}^2=\frac{\alpha}{\sqrt{n}} \tag{17}$$

Yielding the single parameter model:

$$(Y-\mu)=\frac{\sigma_1^2}{\sigma_1^2+\frac{\alpha}{\sqrt{n}}}(X_n-\mu) \tag{18}$$

In terms of a modified similarity function, as desired earlier, we now have:

$$ModSim(Sim,n)=\frac{\sigma_1^2}{\sigma_1^2+\frac{\alpha}{\sqrt{n}}}(Sim-\mu)+\mu \tag{19}$$

Technique

Based on this maodel, we try three different techniques to predict the true similarity from a measured similarity. The techniques vary in their fidelity to the model. For each, we limit our data to all pairs of items that have enough common users that we can approximate $TrueSim \approx Sim$. Then we pick subsets of the common users of various sizes n and calculate Sim_n. With examples of Sim, n and the resulting $TrueSim$, we can implement supervised learning.

The first technique is the least faithful to the model. It simply runs a number of linear regressions between Sim and $TrueSim$ for each n. This ignores the data-specific parameters μ, σ_1 and $\sigma_{2,n}$ to find the best linear fit. It also treats each n completely independently. This technique necessarilty has the lowest training error of all linear techniques, potentially at the risk of overfitting.

The second technique is more general. It approximates μ and σ_1 using the method of moments on the training data of all $TrueSim$ scores. Then, $\sigma_{2,n}$ is also approximated using a method of moments on the training data of all Sim_n for each n separately. Here $\sigma_{2,n}$'s are also considered independently of each other.

The third and most general technique models $\sigma_{2,n}$ as a function of n using equation 19. As in the previous technique, μ, σ_1 and each $\sigma_{2,n}$ are approximated using the method of moments over the training data. Then α is calculated that minimizes the sum of squared errors between $\sigma_{2,n}^2$ and $\frac{\alpha}{\sqrt{n}}$. This α is used to construct the three-parameter prediction function of equation 19.

Experiment

Results

Conclusion