Joseph Blass & Vijay Murganoor
Machine Learning
Spring 2014, Northwestern University
Prof. Doug Downey
joeblass@u.northwestern.edu; vijaym123@u.northwestern.edu

I Agree with You About Adventure and Sex, but You Don't Know Anything About Humor or Animals: Predicting User Ratings Based on User Agreement with Third-Party Raters on Individual Attributes

An unsolved problem that Machine Learning seems ideally suited to is predicting whether a person will like a particular work of art, or, put differently, which particular works of art will appeal to a particular person. Existing recommendation systems are often based on far too general categories like genre, or far too specific attributes (e.g. "foreign romantic dystopian science fiction comedies about pastry chefs with a strong female lead"), or the nebulous concept of "you liked A; this other person liked A; this other person liked B; you will like B". They have trouble recognizing that you might like the movie Love Actually, which has some very dark moments filled with pathos, and not like the movie Notting Hill, which features no darkness or pathos, despite the fact that they are both romantic comedies, are British, share a lead actor, a director, etc. This is all despite the fact that someone who knows you extremely well and knows what you like will be able to make highly accurate movie recommendations to you.

Critics are supposed to fill this role: they provide a comprehensive dataset; they tend to be less biased by preference and mood; and they tend to be more internally consistent. But the fact that critics disagree with each other (and with their readers) shows that they cannot serve as an objective guide. We noticed, however, that there are certain critics with whom we do not always agree, but from whose reviews we can tell whether we will enjoy a particular movie, even if that does not square with the critic's own enjoyment of it.

So we have all this knowledge about movies, knowledge about what individuals like, and knowledge about what critics like, and it is all useless! Our goal is to combine all of these things to create an accurate predictor of whether a particular person will enjoy a particular movie. A fair amount of work has already been done on this problem, but as far as we know our approach has not been tried. Our program uses the extent to which a user's taste aligns with critic's tastes on particular attributes to recommend movies, including those currently in theaters or that have had advance reviews. Eventually, such a system could be expanded to other media, such as music or visual arts, which can be classified by attribute and have a healthy amount of mainstream criticism attached to them.

We compared three techniques, two of which have been used before, and a new one of our own design. Our new technique involves using k-nearest neighbor algorithms to predict user ratings based on agreement with critics about individual attributes; we compared this approach to matching users simply with critics simply based on overall movie ratings, and by averaging user ratings of attributes to predict a new film (not using critic ratings at all). In the first set of conditions we used those critics the user agrees with on the particular attributes of target films to generate a prediction; in the second set of conditions, we took only overall agreement with critics into account; in the third, we took only user ratings of attributes, not the extent to which they agreed with critics about those attributes. In the first two cases we tried 1-nearest, 3-nearest, and 5-nearest neighbor, for a total of seven conditions.

Movie attributes include things such as when the film takes place; whether there is sex; gore; if a family is at the center of the film; if the film prominently features sports; general genre; etc. For each critic we determined the critic's rating of a particular attribute by averaging that critic's ratings of all films with that have that particular attribute. We do the same thing for a user's rating of that attribute. In our experimental condition we use nearest-neighbor to match the user to the k critics whose rating *for that attribute* most closely matches the user's rating. A rating for a new movie is predicted by averaging the ratings of the critics that user matches with on the attributes of the new movie. In our two control conditions, ratings were predicted either by using nearest-neighbor to find which critics a user tended to agree with on overall movie ratings, or by simply taking the user ratings of each attribute and averaging across the attributes present in the target film.

Though it is clear that certain companies already have a lot of information about film attributes and critic ratings, that information is not available to us as students. Therefore we built our own test database of approximately 160 films and hand-tagged them on around 40 attributes. We collected critic ratings for these films from Metacritic; we also collected the metascore, on the assumption that some people generally align with public opinion. Not every critic rated every movie; we removed critics who rated too few movies and movies for which we had too few ratings, leaving us with a dozen critics (including the metascore) and approximately 135 films. We then had several users rate all the movies they had seen (since our technique does not rely on a large number of users, we only tested this on 5 users who provided ratings). We ran our various algorithms on these users, attempting to predict which movies they would like from among those they had not seen (the test data), as well as predicting ratings for approximately 10% of the films they had seen that were left out of training (the validation data).

We found that none of the permutations of our three approaches significantly outperformed the others, nor did any of them do particularly well. All 7 approaches we tested provided ratings that were, on average, around 40 points off. Note that this is not 40 *percent* off, but 40 *points* off: our system would predict a user would rate a movie at a 40 and the movie would actually have been rated at an 80. There was a general trend of k-nearest neighbor approaches performing better with higher values for k; we got the lowest error from our experimental condition using 5-nearest neighbors. Figure 1 shows our aggregate results, significantly zoomed in to highlight the trend.

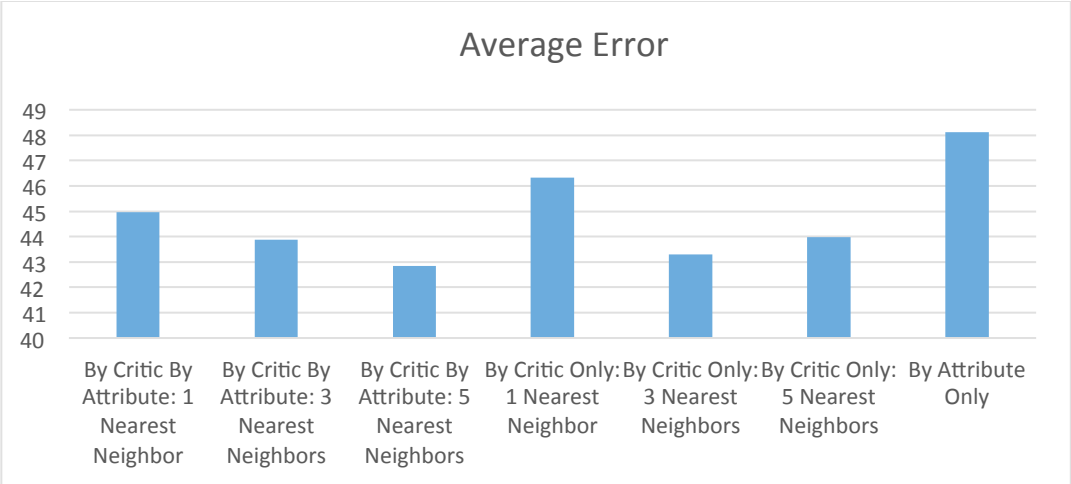

Figure 1: Average Error (lower is better; note significantly enhanced scale)

However, the average error alone does not tell the whole story. There was a wide range of performance for each algorithm, that varied relatively consistently by user. Figure 2 shows how each approach performed for each individual user that we tested. Across the board, these algorithms seem to work better for some people than for others.

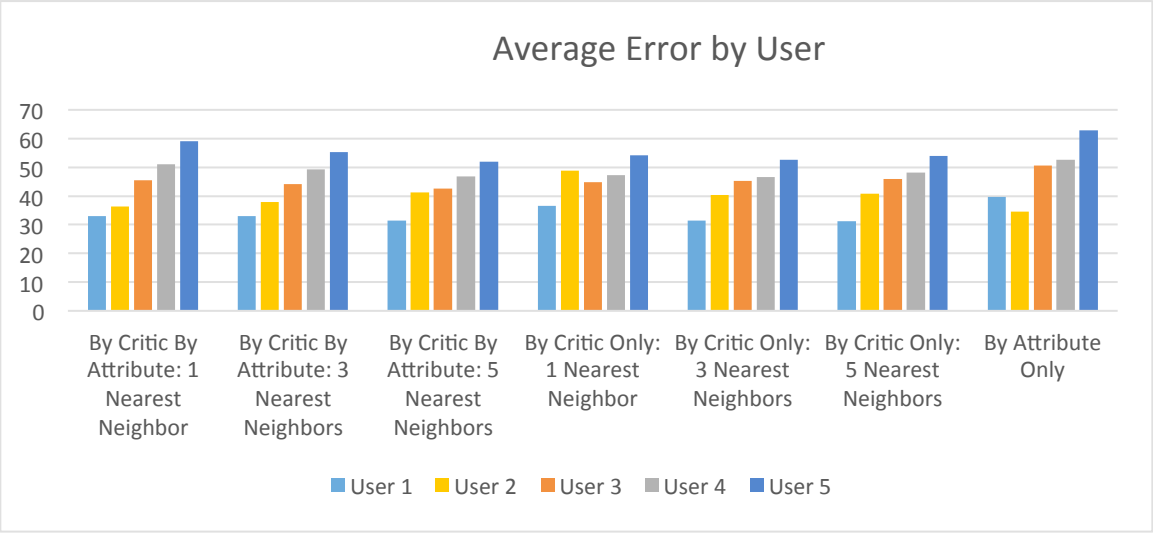

Figure 2: Average Error by Approach by User

The trend generally aligns with the number of movies a user rated, so it is likely that this only serves to prove that more data is better.

There are many reasons our approaches would have performed poorly and equivalently to each other. First and foremost, our dataset is extremely limited: with only 130 films and 40 attributes, the attributes are not well distributed and represented among films. We also may be missing crucial attributes that matter to peoples' enjoyment. The sparsity of data is exacerbated by the fact that not every user had seen every film: if our dataset only contains five documentaries, and the user has only seen two of them, and one of those documentaries is withheld for the validation set, then the user's enjoyment of documentaries as a whole will be determined on the basis of only one film! This will hugely skew our findings. Furthermore, with only a dozen critics, the nearest-neighbor to a given user might not be so near. In one instance, a user rated a film as a 30; most critics rated it at 80 or above, and only one critic rated it at a 60. This critic was assigned as the nearest neighbor of the user, but there is still a large difference between a 60 and a 30! The nearest neighbor can sometimes be rather far away. In addition, a 1-100 scale is extremely broad, and exaggerates errors; in fact, aggregation website ratings are often extrapolated from 1-5 or 1-10 in the first place. Scaling our predictions down to a 1-5 star recommendation score would solve that problem and mitigate the distant-nearest-neighbor problem by collapsing distances. Finally, even with a complete dataset and lots of critics, different factors might matter more or less to a person's enjoyment of a movie (i.e., if the movie is violent may strongly negatively correlate with enjoyment, but if the movie has animals may have a very weak correlation or even be orthogonal to enjoyment). Weighting each feature for each user would be a further task.

Despite these setbacks and a failure to accurately and significantly predict user ratings of films based on critic evaluations of those films, we believe we have demonstrated a powerful concept in predicting user ratings. It remains to be seen, with a more powerful dataset, whether this technique can truly work if scaled up. But we have demonstrated that, on our limited dataset, it is at least as effective as previously attempted techniques in the field. Breaking down user agreement on the basis of features had not been previously attempted, and doing so shows promise for the future of recommendation systems.