Summary 21: Learning to Discover Social Circles in Ego Networks

Susan Cherry

This paper presents a machine learning task for identifying users' social circles on social media. The authors attempt to automatically discover users' social circles instead of relying on naive or manual methods. The problem is formulated as a clustering problem. User u is referred to as the ego and the nodes v_i are referred to alters. The task is to identify the circles that each v_i belongs to. The circles are modeled as latent variables and the similarity between alters is modeled as a function of common profile information. This paper proposes unsupervised methods to learn which dimensions of profile similarity lead to circles that are linked densely. They predict hard assignment of a node to multiple circles and propose a parameterized definition of profile similarity.

The paper begins by describing a generative model for friendships in social circles. Nodes within circles should have common aspects, different circles should be formed by different aspects, circles should be allowed to overlap, and profile information and network structure should be leveraged in order to identify the circles. They authors describe a model of social circles that treats circle membership as a latent variable. Nodes with a common circle are allowed to form an edge, which leads to hierarchical overlapping circles. They then devise an unsupervised algorithm that jointly optimizes the latent variables and the profile similarity parameters to best explain the observed data.

Section 3 formally describes how to optimize node circle memberships. They maximize the regularized log likelihood by coordinate descent. Details are given on page 4. They repeat two optimization steps until convergence. To choose the optimal number of circles, they choose K to minimize an approximation of the Bayesian Information Criterion. Then the describe the dataset that they will use. They then construct features from user profiles. Profile information is represented as a tree. They describe a difference vector to capture the relationship between two profiles. In summary, they identify four ways to represent the similarity between two profiles and consider two ways of constructing a difference vector and two ways of capturing compatibility.

Next, they move to experiments. They evaluate on ground-truth data by examine the maximum likelihood assignments of latent circles after convergence. Ideally, latent variables would align closely with human labeled ground truth. To evaluate, they compute the Balanced Error Rate, which assigns equal importance to false positives and false negatives. They compute optimal match by linear assignments by maximizing Equation 15. Their method outperforms baselines on all datasets by a significant margin. They note that baseline model performance depends on predicting hard memberships to multiple circles. They also find that all algorithms perform better on facebook than Google+ or Twitter, probably because the Facebook data is complete and the profile categories are more informative.