**Combination of neighborhood-based algorithm and soft clustering for collaborative filtering**

Loc Nguyen

Loc Nguyen’s Academic Network, Vietnam

Ali A. Amer

TAIZ University, Yemen

**Abstract**

**1. Introduction**

[Amer: writing, referring (Ungar & Foster, 1998) and previous papers]

**2. Methodology**

The main problem of neighborhood-based algorithm is how to determine an appropriate set of *k*-top neighbors, which causes two drawbacks:

1. Time cost is high because it is necessary to calculate similarities of rating vectors over entire rating matrix in order to determine high enough similarities corresponding to *k*-top neighbors.
2. The set of *k*-top neighbors is affected by biases in calculate similarities with existence of missing values.

This research proposes a method of the combination of neighborhood-based algorithm and soft clustering in order to alleviate two drawbacks of traditional neighborhood-based algorithm. The proposed method has two main steps:

1. Soft clustering by mixture model and expectation maximization (EM) algorithm, in which probability of rating vectors is modeled by multivariate multinomial distribution. Hence, missing values are denoted by special nominal value (NaN).
2. Applying neighborhood-based algorithm into each cluster instead of entire rating matrix as before.

The proposed method is called *clustered neighborhood* (*CN*) algorithm. By partitioning rating matrix into many clusters, CN algorithm reduces searching space for traditional neighborhood-based algorithm. As a result, the time cost is reduced too, which alleviates the first drawback of neighborhood-based algorithm. Moreover CN algorithm determines clusters based on probabilistic distribution and statistical estimation, which avoids biases in calculate similarities with existence of missing values. As a result, the second drawback of CN algorithm is alleviated. Note that CN algorithm does not use hard clustering with distance measures to determine clusters because rating vectors contain missing values in incomplete dataset. When distance measures can be considered as inverses of similarity measures, it is easy to recognize that CN algorithm is also combination of numerical calculation and probabilistic calculation. Hence, in CN algorithm, probabilistic calculation is for soft clustering and numerical calculation is for making prediction on missing value.

[Nguyen: producing and proving mixture model with **multivariate multinomial distribution**, referring (Nguyen & Shafiq, Mixture Regression Model for Incomplete Data, 2018) and (Nguyen, Tutorial on EM Algorithm, 2018)]

[Nguyen: describing CN algorithm in detail with equations]

Suppose there are *c* clusters, CN algorithm is improved by selecting a *representative vector* for each cluster and then a missing rating is estimated as weighted sum of *c* representative vectors over *c* clusters along with similarity values. With improved CN algorithm, accuracy is decreased but speed is increased because c is much smaller than the number of rating vectors. If *c* is small enough, improved CN algorithm can make recommendation in real time. The problem of improved CN algorithm is how to select representative vectors when soft clustering with mixture model does not have cluster centers. The easiest solution is to select the vector whose number of missing values are least among vectors within a cluster as the representative vector for such cluster. In other words, the representative vector is longest with maximum number of rating values.

[Nguyen: describing improved CN algorithm in detail]

**3. Experimental design**

[Amer: writing, referring previous papers]

**4. Result and discussion**

[Nguyen: testing and reporting]

[Amer: writing]

**5. Conclusion**

**References**

Nguyen, L. (2018). *Tutorial on EM Algorithm* (1st ed.). (H. Bachoo, Ed.) Long Xuyen, An Giang, Vietnam: Lambert Academic Publishing. Retrieved from https://www.morebooks.shop/store/gb/book/tutorial-on-em-algorithm/isbn/978-613-9-81971-3

Nguyen, L., & Shafiq, A. (2018, December 31). Mixture Regression Model for Incomplete Data. (L. d. Istael, Ed.) *Revista Sociedade Científica, 1*(3), 1-25. doi:10.5281/zenodo.2528978

Ungar, L. H., & Foster, D. P. (1998). Clustering Methods for Collaborative Filtering. In H. Kautz (Ed.), *AAAI Workshop on Recommender Systems.* *1*, pp. 114-129. Madison: AAAI Press. Retrieved from http://www.aaai.org/Papers/Workshops/1998/WS-98-08/WS98-08-029.pdf