Recommender System

Nutan Kulkarni, *Binghamton University, New York*

***Abstract*— Incomplete data is an omnipresent issue in today's world. Especially in those fields where research studies, productivity or profit revolves around data. Such partial data can hamper application of data analysis and data mining techniques on crucial data. Matrix completion is a process of recovering missing data, based on a set of sample data. In this paper, a novel approach to fill such missing data is presented. A recommender system that uses collaborative filtering for matrix completion is presented in this paper.**

***Index Terms*— recommender system, matrix completion, pearson coefficient, collaborative filtering**

I. Introduction

THIS paper discusses a prominent approach used to create a recommender system for matrix completion problem. A brief scenario of real-world application where matrix completion is required and used is recovering corrupted data in Sensor Networks such as weather data [1]. Unforeseen events such as forest fire and earthquakes would be difficult to detect if available sensory data is incomplete or erroneous. Matrix completion is used for such applications. The application used in this paper is one where we have 'm' different users and 'n' different items. The recommender system, after checking the user preferences of ratings will predict a rating for an item 'j' if a user 'i' has not yet given a rating for an item 'j'.

II. Recommender System

Recommender system discussed uses a filtering technology and predicts a particular item rating. It uses a partial matrix to calculate correlation between different users and rating is predicted for a missing entry in the matrix. Collaborative filtering is a well-known technique to perform this similarity.

A. Collaborative filtering

Collaborative filtering algorithms are used for recommending items based on preferences of users. Collaborative filtering is based on concept of similarity. The similarity measure used in this recommender system is user-based. A person who liked a similar item in the past will mostly likely like similar item in the future as well. Among the various collaborative filtering algorithms, Pearson's correlation coefficient was used to find similarity between users. Pearson`s coefficient is the covariance of two variables divided by product of their standard deviation. This coefficient is denoted by r and ranges between -1 to +1. Value +1 implies positive correlation and indicates that the variables are strongly related. Value -1 implies negative correlation and indicates that the variables are loosely correlated. Zero value implies no correlation.

B. Pearson`s correlation coefficient formula [3]

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is the sample size
- x_i, y_i are the single samples indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

C. Algorithm

This algorithm takes a partially filled [m x n] matrix as input where m is total number of users and n is number of items. Each record that exists in the input matrix also consists of a rating ranging from 1 to 5. This recommender systems solves the matrix completion problem and fills the matrix. The recommender system, after checking the user preferences of ratings will predict a rating 5 if strong correlation is found, 1 if negative correlation or no correlation is observed.

Input: Partially filled Data matrix M of m users and n items
Output: Complete Data matrix M
1: Compute average rating per user.
2: Create a new matrix to save ratings minus average ratings.
3: Compute current rating of user for a item minus average rating of that user.
4: Create Coefficient Matrix R using Pearson's Correlation Coefficient.
5: Predict the missing ratings by computing numerator, denominator and (numerator/denominator) using the input matrix and Coefficient matrix.

D. Evaluation of algorithm

This algorithm was evaluated based on a training set in which 80% training data and 20% test data was used.

III. References

[1] Kun Xie, Xueping Ning, Xin Wang, Dongliang Xie, Jiannong Cao, Gaogang Xie, Jigang Wen, "Recover Corrupted Data in Sensor Networks: A Matrix Completion Solution", Journal IEEE Transactions on Mobile Computing, Volume 16 Issue 5, May 2017.
[2] Shiming Yang, Konstantinos Kalpakis, Colin F. Mackenzie, Lynn G. Stansbury, Deborah M. Stein, Thomas M. Scalea and Peter F. Hu ,"Online recovery of missing values in vital signs data streams using low-rank matrix completion", Journal IEEE Transactions on Mobile Computing, Volume 16 Issue 5, May 2017
[3] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient