Second International Conference
Modelling and Development of Intelligent Systems
Sibiu - Romania, September 29 - October 02, 2011

A cluster analysis for recommender systems evaluation metrics

Ionela Maniu, George Maniu

Abstract

Evaluation of recommender systems is a challenging task due to the many possible scenarios in which such systems may be deployed. Comparison between recommender systems it becomes difficult to achieve due to the large diversity of published metrics that have been used to quantitatively evaluate the accuracy of recommender systems. In this paper, we present a cluster analysis whose goal is to offer a classification of binary evaluation function in order to establish standardization within this field.

Keywords: cluster analysis, recommender systems, evaluation, binary metrics

1. Introduction

Recommender systems are define as ones in which "people provide recommendations as inputs", and the system then "aggregates and directs" on appropriate items (Resnick & Varian, 1997). Comparison between recommender systems it becomes difficult to achieve due to the large diversity of published metrics. To illustrate this, we quote some appropriate paragraphs from Herlocker, Konstan, Terveen, and Riedl (2004): "The challenge of selecting an appropriate metric is compounded by the large diversity of published metrics that have been used to quantitatively evaluate the accuracy of recommender systems. This lack of standardization is damaging to the progress of knowledge related to collaborative filtering recommender systems. With no standardized metrics within the field, researchers have continued to introduce new metrics when they evaluate their systems. With a large diversity of evaluation metrics in use, it becomes difficult to compare results from one publication to the results in another publication. As a result, it becomes hard to integrate these diverse publications into a coherent body of knowledge regarding the quality of recommender system algorithms."

In this paper, we present a cluster analysis whose goal is to offer a classification of binary evaluation function in order to establish standardization within this field.

This paper is organized as follows. Section 2 goal consists of achieving a logical framework with a common terminology in recommender systems domain. Section 3 describes the definitions of 27 binary similarity (dissimilarity) measures. Section 4 discusses the grouping of those measures using hierarchical clustering. Section 5 concludes this work.

2. A general framework for recommender systems

Recommender systems main objective is to guide the user to useful/interesting objects. For this, a number I of items are available to be recommended. In order to start the

recommendation process some of those items must be rated, these ratings are obtained explicit or implicit (inferred from other users interactions). Once the recommender system has enough ratings, it can start the process. For each recommendation, a number N < I of objects are chosen by the recommender, and show to the target user. Some recommender systems also rank the marked-out objects in order to show them as an ordered list, and in this case, the user will investigate these items starting at the top of this list.

In order to evaluate the performance of the recommender system for each object shown to a particular user, we must measure how close the utility of the shown object is with respect to the preferences of the user. In the case of an ordered list, additionally, we should take into account, the place that each recommended object has in this list.

In order to measure the closeness of predictions to users' real preferences, a numerical representation is normally used. This representation uses the predictions of a recommender system for every particular user u and item i, and the real preferences of user u for item i.

3. Binary similarity evaluation metrics

Once we have obtained the formal structures, in this section, we will study several metrics that can be applied into this framework. To compute these metrics, a confusion matrix is expected such as the one in table below. This table reflects the four possibilities of any recommendation decisions.

	1 successful recommendation	0 Non - successful recommendation	sum
1 recommended	a	b	a + b
0 not recommended	c	d	c + d
sum	a + c	b + d	n=a +b +c +d

Table 1: Confusion matrix for recommender systems

A binary vector X with N dimensions is defined as $x = (x_1, x_2, \ldots, x_N)$ where x_i has the value 0 or 1 and N represents the number of futures or dimension of the feature vector.

In confusion matrix, a is the number of occurrences of matches 1 in the first pattern and 1 in the second pattern at the corresponding positions (positive matches), b is the number of occurrences of matches 1 in the first pattern and 0 in the second pattern at the corresponding positions, c is the number of occurrences of matches 0 in the first pattern and 1 in the second pattern at the corresponding positions and d is the number of occurrences of matches 0 in the first pattern and 0 in the second pattern at the corresponding positions (negative matches). The diagonal sum a+d represents the total number of matches between patterns and the other diagonal sum b+c represent the total number of mismatches between patterns. The total sum of the table, a+b+c+d is always equal to n.

Numerous binary similarity evaluation metrics have been described in the literature [3]. Many papers discuss their properties and features. Table above lists formulas of 27 binary evaluation metrics used in cluster analysis, in this paper, in the case of recommender systems algorithms.

Metrics	Formula	Metrics	Formula
Precision	$d(x,y) = \frac{d}{b+d}$	Faith	$d(x,y) = \frac{a+0.5d}{n}$
Recall	$d(x,y) = \frac{d}{c+d}$	Gower&Legendre	$d(x,y) = \frac{a+d}{a+0.5(b+c)+d}$

Vari	$d(x,y)=\frac{b+c}{4n}$	Sokal&Michener	$d(x,y)=\frac{a+d}{n}$
Shapedifference	$d(x,y)=\frac{n(b+c)-(b-c)^2}{n^2}$	Sorgenfrei	$d(x,y)=\frac{a^2}{(a+b)(a+c)}$
Patterndifference	$d(x,y)=\frac{4bc}{n^2}$	Otsuka	$d(x,y)=\frac{a}{((a+b)(a+c))^{0.5}}$
Lance&Williams	$d(x,y)=\frac{b+c}{2a+b+c}$	Mountford	$d(x,y)=\frac{a}{0.5(ab+ac)+bc}$
Hellinger	$d(x,y)=\sqrt{(1-\frac{a}{\sqrt{(a+b)(a+c)}})}$	Mcconnaughey	$d(x,y)=\frac{a^2-bc}{(a+b)(a+c)}$
Cosinus	$d(x,y)=\frac{a}{(a+b)(a+c)}$	Tarwid	$d(x,y)=\frac{na-(a+b)(a+c)}{na+(a+b)(a+c)}$
Ochiai-I	$d(x,y)=\frac{a}{\sqrt{(a+b)(a+c)}}$	Driver&Kroeber	$d(x,y)=\frac{a}{2}(\frac{1}{a+b}+\frac{1}{a+c})$
Jaccard	$d(x,y)=\frac{a}{a+b+c}$	Jhonson	$d(x,y)=\frac{a}{a+b}+\frac{a}{a+c}$
Dice	$d(x,y)=\frac{2a}{2a+b+c}$	Simpson	$d(x,y)=\frac{a}{\min(a+b,a+c)}$
Sokal&Sneath-I	$d(x,y)=\frac{a}{a+2b+2c}$	Braun&Banquet	$d(x,y)=\frac{a}{\max(a+b,a+c)}$
Sokal&Sneath-II	$d(x,y)=\frac{2(a+d)}{2a+b+c+2d}$	Sokal&Sneath IV	$d(x,y)=\frac{\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{b+d}}{4}$
MAE	$d(x,y)=\frac{b+c}{a+b+c+d}$		

Table 2: Binary evaluation metrics

4. Hierarchical clustering

Hierarchical clustering is conducted to estimate the similarity among 27 measures collected. The correlation coefficient values between two measures are used to build a *dendrogram*. The average linkage between groups with squared Euclidian distance clustering method is used.

We used data from the well-known MovieLens project(http://movielens.umn.edu). Data sets consists of 100,000 ratings (1-5) from 943 users on 1682 movies. The dataset was divided into training set (90%,80% of the data) and test set (10%, 20% of the data) five times.

The dendrogram in figure 1 provides intuitive semantic grouping of 27 binary evaluation metrics[4], used to measure the prediction results of our algorithms and the real rating.

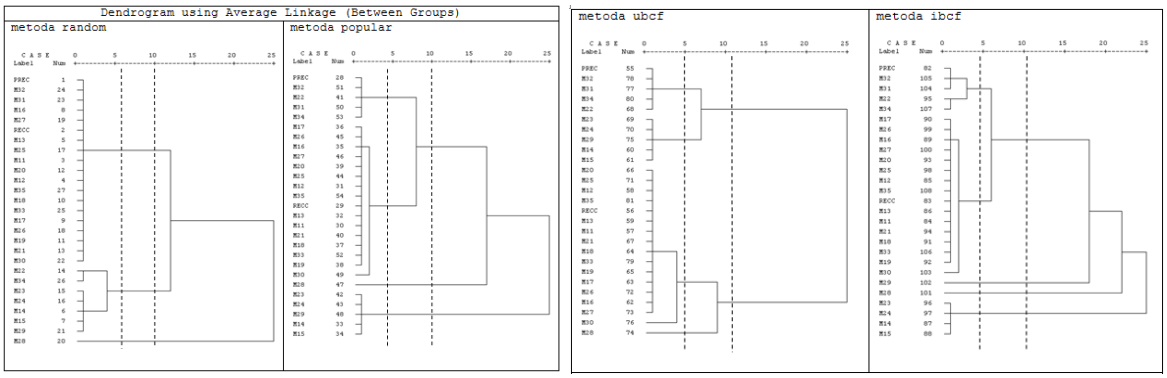

Fig. 1: Classification of 27 binary evaluation metrics, using cluster analysis, the case of random, popular, ubcf, ibcf recommendation algorithm

Binary measures with high correlation: m14-Lance&Williams, m15-Hellinger, m23-Gower&Legendre, m24-Sokal&Michener are categorized in first group, while m11-Vari, m12-Shapedifference, m13-Patterndifference, m16-cosinus, m17-Ochiai I, m18-Jaccard, m19-Dice, m20-Sokal&Sneath I, m21-Sokal&Sneath II, m25-Sorgenfrei, m26-Otsuka, m27-Mountford, m30-Driver&Kroeber, m31-Jhonson, m32-Simpson, m33-Braun, m35-MAE, recc – recall binary measures are categorized in second group, in all 4 different recommender systems algorithms. Apart from this group is m28-Wcconnanghey measure. It is interesting that additive form of negative match measures such as Jaccard, Dice have high correlation with the cosine based measures such as Ochiai I or Sorgenfrei.

5. Conclusions

In this survey, we used 27 binary similarity and distance metrics for evaluating 4 different recommender systems algorithms and classified them (metrics) through hierarchical clustering. In all cases, there are 3 main classes of metrics, at aggregation level smaller than 10, and there are 2 main classes of metrics, at aggregation level smaller than 15. There are differences between clusters elements (metrics) in the case of ubcf algorithm versus case of ibcf, random, popular algorithm.

References

[1] Herlocker, J. L., Konstan, J. A., Terveen, L. G., Riedl, J. T. *Evaluating collaborative filtering recommender systems*. ACM Transactions on Information Systems, 22(1), 5–53,2004
[2] Resnick, P., Varian, H. R. , *Recommender systems*. Communications of the ACM, 1997
[3] Cha, S.H, Tappert, C.C., *Enhancing Binary Feature Vector Similarity Measures,* Journal of Pattern Recognition research I, 2006
[4] Maniu, I., *Mechanisms of economic language modeling with applications in recommender systems*, doctorate dissertation, ASE Bucharest, 2011

Ionela Maniu
Faculty of Science
Department of Informatics
Str. Ion Ratiu, nr. 5-7
Romania
E-mail: mocanionela@yahoo.com

George Maniu
Faculty of Management
Department of Management
Str. Turnului, nr. 7
Romania
E-mail: georgemaniu@yahoo.com