Information Sciences 283 (2014) 241–257

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Unsupervised consensus analysis for on-line review and questionnaire data

Stephen L. France [a,*], William H. Batchelder [b,1]

[a] Lubar School of Business, University of Wisconsin – Milwaukee, Milwaukee, WI 53201, United States
[b] School of Social Sciences, University of California, Irvine, Irvine, CA 92697, United States

ARTICLE INFO

Article history:
Received 15 January 2013
Received in revised form 4 January 2014
Accepted 9 June 2014
Available online 30 June 2014

Keywords:
Clusterwise
Consensus
k-Means
Maximum likelihood

ABSTRACT

We describe a set of Cultural Consensus Theory (CCT) models for analyzing review and questionnaire data. The basic single culture/cluster model can be used to estimate user competencies, user biases, and aggregate review scores. The model is unsupervised and only utilizes the input review scores. A maximum likelihood approach is used to estimate the model. We expand existing work by developing a clusterwise multi-culture continuous CCT model, for which we use the acronym CONSCLUS (CONSensus CLUStering). The original single culture CCT model is a special one-cluster case of CONSCLUS. We show that when all user competencies are equal, CONSCLUS is equivalent to *k*-means clustering. CONSCLUS is estimated using an alternating least squares variant of the algorithm for *k*-means clustering, which we denote as CCT-Means. CONSCLUS is a partitioning clustering technique. We describe extensions to CONSCLUS to incorporate fuzzy clustering and overlapping clustering.

We run a series of simulation experiments using generated data with random error. We test both the single cluster and multiple cluster models. These experiments show that CONSCLUS is able to recover aggregate rating values and latent cluster assignments better than a range of other aggregation methods. The performance increase over the other aggregation methods is particularly strong when the users have varying competencies. We give an illustrative example using the Movielens dataset. We give a set of recommendations for the practical implementation of CONSCLUS on real world data and show how the user competencies can be used to gain insight into these data that cannot be gained from simple partitioning clustering.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Recent years have seen an explosion in the amount of user generated content available on the internet. In particular, there has been an increase in the amount of on-line review data generated by consumers. Given the very large volumes of data, there is great scope for the use of data mining models and algorithms. Consider a data set containing multiple product reviews. The reviews can be stored in a sparse n user $\times$ m product matrix. There has been a large amount of work under the banner of recommender systems for data of this type [1,9,35]. Given a set of reviews, a recommender system would

* Corresponding author. Tel.: +1 4142294596.
E-mail address: france@uwm.edu (S.L. France).

[1] The second author acknowledges the support of a grant from the Army Research Office (ARO) and a Fellowship from the Oak Ridge Institute for Science and Education (ORISE).

http://dx.doi.org/10.1016/j.ins.2014.06.015
0020-0255/© 2014 Elsevier Inc. All rights reserved.

typically predict scores for {user, product} pairs for which reviews are not available. Recommender system techniques can be placed under the broad categories of content filtering [44] and collaborative filtering [19,32,37,59]. Content filtering techniques utilize product information, while collaborative filtering techniques utilize review scores and correlational measures between users [47] or between items [38,51]. These categories are not mutually exclusive and recommender systems can incorporate both content filtering and collaborative filtering [40].

Our proposed set of techniques model an aggregate answer or knowledge component for each question/review item and model user competency with respect to the knowledge components for each user. Here, competency is defined as a measure of inverse error variance. Consumer preferences and social effects are accounted for by clustering consumers and modeling cluster level preferences. Given the cluster level preferences, each cluster can be considered to have its own knowledge or culture. We describe both a simple single cluster model and a multiple cluster model. The multiple cluster model utilizes cultural consensus theory and uses a clusterwise scheme to model both cluster membership and cluster level effects.

The models described in this paper utilize only review scores, so using the etymology of recommender systems can be thought of as collaborative techniques rather than as content techniques and are particularly useful when review content information is not available or is of poor quality. A simple collaborative method to analyze review scores is to average the review scores. In fact, the average review score is usually reported on review websites. However, an average review score does not account for differing reviewer competencies. For a set of reviews, where there are more reviewers with idiosyncratic preferences than reviewers who are unbiased and competent, the arithmetic mean may give a poor overall score.

2. Cultural consensus theory

Cultural Consensus Theory (CCT) [11,12,14,15,48] is an approach to information pooling (aggregation, data fusion). Since its inception, CCT has been utilized across the social and behavioral sciences, especially cultural anthropology. The initial application of CCT [12,48] was to estimate folk medical beliefs in a sample of Guatemalan women using two questions about each of a set of diseases: 1. is the disease contagious and 2. does the disease require a hot or cold remedy? While exact answers for disease contagiousness are known to modern science and the characterization of diseases as requiring hot or cold remedies comes from ancient folk medical beliefs, isolated cultures may share different beliefs and thus have different consensus answers for such questions. The disease data were analyzed using the classical CCT implementation for dichotomous data and the model is estimated using a maximum likelihood estimation scheme, where the bias is fixed [13,17,48]. The output parameters were the disease classifications ({hot,cold} and {contagious,non-contagious}) and estimates of how competent the women were with respect to the belief system of their culture.

The primary goal of CCT is to estimate consensus answers from a set of raters (respondents), each of whom provides responses to questions about some aspect of their shared knowledge or beliefs. CCT consists of a set of parametric, cognitive models, each corresponding to a different questionnaire format, e.g., true/false, multiple choice, ordered categories, continuous responses. CCT specifies the consensus answers to the questions as latent variables rather than as known a priori to the researcher. In addition, the models specify latent parameters for the competences (degrees of cultural knowledge) and response biases of the informants.

The continuous CCT model [14] is designed to analyze numerical ratings. For a set of users giving numerical ratings for a set of items, the continuous CCT model gives a set of user competencies, a competency weighted aggregate value for each item, and optionally, a set of user biases. Continuous CCT can be applied to any set of numerical ratings data where multiple users rate multiple items. An exam grading application for continuous CCT is described in [14]. Here, CCT was used to analyze ratings for 50 essays, each rated by 14 raters. Each rater was given the essay prompt, a grading rubric, and some example graded essays. Utilizing CCT for essay grading allows raters to be evaluated and the competency weighted aggregate values give greater weight to more competent raters than to less competent raters.

We describe the continuous single culture CCT model and extend it to account for multiple cultures. We implement a fixed point estimation procedure for the basic model and combine this procedure with a *k*-means like clustering procedure for the multiple culture model. We run a series of Monte Carlo experiments to test how well both the basic model and the extended model recover model parameters from error perturbed data. We give an illustrative example, showing how CCT can be applied to on-line review data. Our work provides two major advances in the development of CCT. First, the extended CCT model allows for multiple clusters of users, each with its own latent consensus answer pattern. Second, our estimation approach is able to handle large amounts of missing data, as is often the case with on-line reviews. Most on-line review scores are collected as ordinal scale data using a Likert scale. However, we treat the data as continuous. We do this for three reasons. The first reason is precedence. Most collaborative filtering techniques take on-line review scores as continuous and utilize continuous correlational measures, as researchers have found that these measures give better predictive performance than ordinal, rank order measures [32,59]. The second reason is that there is evidence that when respondents are asked for a specific score (for example, a direct Likert rating scale of 1–10), the underlying latent continuous distances between categories are almost even [36] and thus the data can be considered to be approximately interval scale. And third, a previous CCT paper dealing with a different CCT model for continuous response data showed that it was a good approximation to Likert scale data [15].

Download English Version:

https://daneshyari.com/en/article/6857714

Download Persian Version:

https://daneshyari.com/article/6857714

Daneshyari.com