STA 794 HW1

Summary for the lecture: Some Recent Developments and Open Challenges

(Presented by Rebecca Steorts on Thursday 7th July 2016)

Dayi Fang

09/27/2016

In this lecture, professor Steorts goes over some basic knowledge of the Record Linkage, provides clear explanations regarding to both major methods and Bayesian methods, and states several inspiring Bayesian findings.

The motivation of the Record Linkage is to remove administrative, medical, and other type of duplicated records. In practice, company or institute can apply Record Linkage algorithms to efficiently filter the valuable information that is shared by different groups. Since Record Linkage is a newly developed topic, the first paper "A Theory for Record Linkage" by Fellegi and Sunter was published in 1969. During 1980 to 2000, the development of Record Linkage focused on applying major methods in health, government, and privacy fields. Recently (2000-2016), modern Bayesian methods have been developed with the progress in related areas, such as machine learning and clustering.

The following chapter of professor Steorts' lecture is focusing on introducing two major methods and explaining their advantages and disadvantages. Hand matching is the most traditional method, which is not only very inefficient but also lack of certain scalability. Another major method is the famous Fellegi and Sunter algorithm, which contains transitive closures but still not scalable. Based on these two methods, professor Steorts demonstrates two concerns to the audiences: scalability and the level of interest. Although scalability is always a critical limitation of the Record Linkage study, uncertainty problem can be addressed though the application of the Bayesian method. However, applying Bayesian algorithms has some limitations, such as scalability, sensitivity to prior and hyperparameter, and difficulty for generalizing based on multiple record linkage.

As the major section of the lecture, professor Steorts exhibits several inspiring Bayesian findings and provides short introductions to each of them. The basic idea of Bayesian graphical record linkage is to link similar pieces of information to the latent "records" instead of linking them with each other. This idea provides more accurate outcomes and help develop better blocking methods. In Steorts' paper "A Bayesian Approach to Graphical Record Linkage and De-duplication (Steorts, Hall, and Fienberg, 2015)", she states an efficient algorithm to split and merge record linkage and de-duplication (SMERED). Through her idea, it is simple to find that Bayesian can provide better structure of doing de-duplication. In another paper " A Comparison of Blocking Methods for Record Linkage (Steorts, Ventura, Sadinle, and Fienberg, 2014)", professor Steorts gives detailed comparison among three popular blocking methods and states the applying requirements and effectiveness of them.

Professor Steorts gives a general picture of the Record Linkage topic and provides very detailed introduction to modern Bayesian methods in this lecture. It is extremely helpful for people that are not very familiar with Record Linkage.