CIBuilder: A System Prototype for Building Confidence Intervals for Group Differences (demonstration proposal)

Shichao Zhang
Department of Automatic Control
Beijing University of Aeronautics and Astronautics
Beijing 100083, China
zsc@buaa.edu.cn

Xindong Wu
Department of Computer Science
University of Vermont
Burlington, Vermont 05405, USA
xwu@cs.uvm.edu

Yongsong Qin, Jilian Zhang, Xiaofeng Zhu
Department of Computer Science
Guangxi Normal University
Guilin, China
ysqin@mailbox.gxnu.edu.cn
zhangjilian@yeah.net
xfzhu_dm@163.com

ABSTRACT

Mining the differences between contrasting groups is an important and challenging task in real world applications such as medical research, social network analysis and link discovery. Yet another important issue that has received less attention is to measure the differences between groups. For many applications, the data obtained are sampled from a population, thus the knowledge mined out and hypotheses derived from these data are probabilistic in nature and such uncertainty has to be measured. For this reason, we have developed a system prototype, called CIBuilder, for constructing confidence intervals (CI) for structural differences, such as the mean and distribution function, between two contrasting groups. The CIBuilder is based on the empirical likelihood (EL) method, and is efficient in identifying confidence intervals for differences of groups. Generally, data may have missing values in real world applications, which can degrade the performance of most mining algorithms. Our CIBuilder can also efficiently deal with datasets with missing values, under different missing rates.

Keywords

Empirical likelihood method, Confidence intervals, Group differences, Missing values.

1. INTRODUCTION

A common question in exploratory research is: "How do several contrasting groups differ?" Group difference detecting (or comparing) is a central problem in many domains [4]. For example, in medical research, it is interesting to compare the mean value of prolonging patients' life between a group using a new product (e.g., medicine) and a group with another product; in children's growth research, the height below or above the standard is important, since the median height (near the standard) is associated with a normal growth status, it may be meaningful with children's growth to compare two groups on the basis of both below the standard or above the standard of height. Therefore, detecting group differences has been an important data mining technique designed specifically to compare differences between contrasting groups from observational multivariate data [2,3,4]. From the statistical perspective, the mean and distribution function are very important for characterizing the data, and one will have a full understanding of the data if he knows the mean and distribution function exactly.

People usually are interested in finding what are the differences for mean or distribution function of two data groups, say X and Y, because this information is useful for decision-makers to make decisions or predictions. One can use statistical methods to obtain the above differences. For the mean difference Δ between groups X and Y, one can use the equation $\Delta = E(Y) - E(X)$ to calculate it, where $E(Y) = \frac{1}{m}\sum_{j=1}^{m} y_j$ and $E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the mean of Y and X respectively. As for the distribution function difference Δ between X and Y, one can use the equation $\Delta = G_Y(\alpha) - F_X(\alpha)$, where G_Y and F_X are the distribution functions of Y and X respectively; α is a reference point for comparing the distribution function of X and Y and it is a constant given by the user. Generally, the exact form of the distribution function is difficult to obtain, so the empirical form is adopted in practice, i.e., $\hat{G}_Y(\alpha) = \frac{1}{m}\sum_{j=1}^{m} I(y_j \le \alpha)$, $\hat{F}_X(\alpha) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le \alpha)$, where $I(.)$ is an indicator function, and $I(X)=1$ if X is true, otherwise $I(X)=0$. This is called the non-parametric model. If we know the exact form of distribution function of G_Y (or F_X) in advance, we then call this the semi-parametric model. Currently our CIBuilder system only deals with the semi-parametric model.

Yet in real world applications, the data obtained are sampled from a population, thus the knowledge mined out and hypotheses derived from these data are probabilistic in nature, and such uncertainty has to be measured. Just like the differences calculated above, we must resort to statistical tools to build confidence

The copyright of these papers belongs to the paper's authors. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

KDD'06.

intervals in order to better measure their uncertainties. The confidence interval (CI) can tell how reliable the derived differences of two groups X and Y are. Related work can be seen in [1].

The empirical likelihood (EL) is one of the most popular methods in statistics, which can be used to build confidence intervals [5,7,8,9]. Researchers have used the EL method to construct confidence intervals for complete datasets (where there are no missing values). But in most cases, the dataset may have missing values, which will degrade the performance of the most mining methods. In [11], we have proposed an EL based model for building CI for mean and distribution function differences of two semi-parametric groups of data when there are missing values, and also developed a system prototype that is based on our previous work [11], called CIBuilder, for building confidence intervals for the mean and distribution function differences between two groups that contain missing values. However, we would like point out that CIBuilder can also deal with complete datasets.

The rest of this paper is organized as follows. We present the architecture of the CIBuilder system in Section 2. Section 3 discusses the functional components of CIBuilder briefly. In Section 4, we give conclusions and a demonstration plan.

2. SYSTEM ARCHITECTURE

CIBuilder consists of several components, including *Data Preprocessor*, *Confidence Interval Builder*, and *Confidence Interval Visualizer*, which are shown in Figure 1. At first, the data of two groups X and Y are fed to the *Data Preprocessor*, which can fill up the missing values and generate a 'complete' dataset. In the *Confidence Interval Builder* module, the differences of means and distribution functions of the two datasets are built based on our empirical likelihood method. Finally, the obtained CIs are presented in the form of graphs to the users.

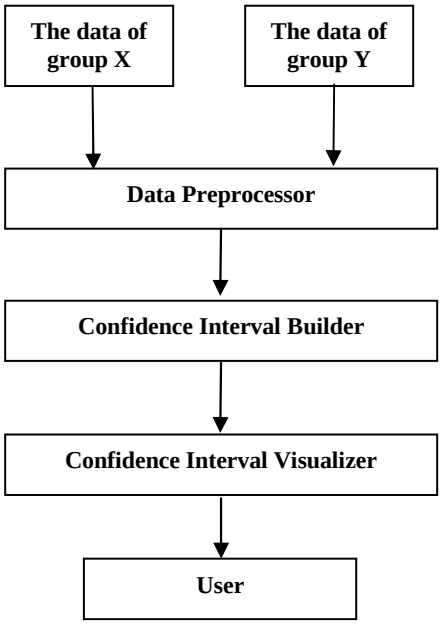

Figure 1. The Architecture of CIBuilder

3. THE FUNCTIONAL COMPONENTS

3.1 Data Preprocessor

In real world applications, data may have missing values before preprocessing, which can hinder the user from getting robust results. As a consequence, a data preprocessing procedure must be in place before the data analysis. Currently, there are many powerful methods in dealing with missing data for data preprocessing; these include mean substitution, complete case analysis and other missing value imputation methods from statistics domain and machine learning filed as well. Focusing mainly on the construction of CI, we do not take many imputation methods into account. Instead, we only adopt the most common method, the mean substitution method, for tackling the missing values in our system. However, one can combine other powerful methods into the *Data Preprocessor* module for missing value imputation [6,10].

3.2 Confidence Interval Builder

In this module, we use the EL based method to construct the confidence intervals of differences of means and distribution functions for two data groups. The empirical likelihood ratio statistic for the differences of means and distribution functions of two groups is proven to be a weighted chi-squared distribution [11]. So the main task of this module is to solve the empirical likelihood equations [11], and search for the left and right endpoints of the confidence intervals in an iterative way. A screenshot of the CIBuilder is presented in the appendix.

3.3 Confidence Interval Visualizer

When the confidence intervals of the differences for two groups are obtained, we present them to the user in the form of graphs, which is the main task of the *Confidence Interval Visualizer* module. We give an example of output image from this module in Figure 2.

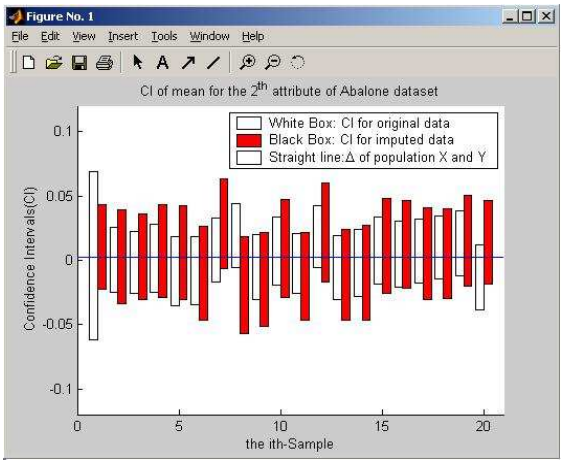

Figure 2. Visualization of CIs

4. CONCLUSION AND DEMONSTRATION PLAN

We have designed and implemented a system prototype called CIBuider for constructing confidence intervals for differences of

means and distribution functions of two data groups. It can efficiently identify the confidence intervals in order to measure the uncertainty of the differences, providing the user with more information about the differences of two data groups.

We implemented the CIBuilder system in MATLAB 6.1 on a Dell Workstation, and the operating system is WINDOWS 2000. It has been tested on synthetic datasets as well as the UCI datasets in order to verify its effectiveness and efficiency. In the demo session of KDD'06, we will demonstrate the functions of our CIBuilder system to build confidence intervals, and to visualize the building results. We encourage other researchers to use our system during the session and welcome comments and constructive advice.

5. REFERENCES

[1] Adibi, J., Cohen, P. R., and Morrison, C., Measuring Confidence Intervals in Link Discovery: A Bootstrap Approach. KDD'04, 2004.

[2] Bay, S. D. & Pazzani, M. J. Detecting Change in Categorical Data: Mining Contrast Sets. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'99, pp.302-306. 1999.

[3] Bay, S. D. & Pazzani, M. J. Characterizing Model Errors and Differences. In Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), pp.49-56. 2000.

[4] Bay, S. D, Pazzani, M. J. Detecting Group Differences: Mining Contrast Sets. Data Mining and Knowledge Discovery, 5(3): 213-246. 2001.

[5] Jing, B. Y. Two-sample empirical likelihood method. Statistics and Probability Letters, 24: 315-319. 1995.

[6] Little, R. & Rubin, D. *Statistical analysis with missing data*. 2nd edition. John Wiley & Sons, New York, 2002.

[7] Qin, Y. S. & Zhao, L. C. Empirical likelihood ratio intervals for various differences of two populations (in Chinese) Systems Science and Mathematical Sciences , 13 :23-30,2000.

[8] Owen, A. Data Squashing by Empirical Likelihood. Data Mining and Knowledge Discovery, 7(1): 101–113, 2003.

[9] Owen, A. *Empirical likelihood*. Chapman & Hall, New York, 2001.

[10] Rao, J., Sitter, R., & Chen, J., Efficient random imputations for missing data in complex surveys. Statistica Sinica, 10(4): 1153-1169. 2000

[11] Huang, H., Qin, Y. S. et al, Difference Detection Between Two Contrast Sets. In Proceedings of the 8th International Conference on Data Warehousing and Knowledge Discovery, DaWak'06.

Appendix