Measuring the Uncertainty of Differences for Contrasting Groups

Jilian Zhang[1], Shichao Zhang[1,*], Xiaofeng Zhu[1], Xindong Wu[2], Chengqi Zhang[3]

[1] Department of Computer Science, Guangxi Normal University, Guilin, China
[2] Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA
[3] Faculty of Information Technology, University of Technology Sydney, P.O. Box 123, Broadway, NSW 2007, Australia
zhangjilian@yeah.net; {zhangsc, chengqi}@it.uts.edu.au; xwu@cs.uvm.edu; xfzhu_dm@163.com

Abstract: In this paper, we propose an empirical likelihood (EL) based strategy for building confidence intervals for differences between two contrasting groups. The proposed method can deal with the situations when we know little prior knowledge about the two groups, which are referred to as non-parametric situations. We experimentally evaluate our method on UCI datasets and observe that proposed EL based method outperforms other methods.

Introduction

Mining the differences between contrasting groups is an important and challenging task in many real world applications such as medical research, social network analysis and link discovery (Bay 99,01 and Webb 03). For example, finding out differences that can distinguish spam from non-spam emails or the benign breast cancer from the malign one will benefit to people, because researchers or companies can utilize this information to design powerful anti-spam software or new medicine for curing breast cancer. 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 (Adibi 2004). That is, when the difference is obtained, the important thing would be that one might want to know how reliable the answer is.

From the statistical perspective, the mean and distribution function (DF) are very important for characterizing a group of data, and one will almost have a full understanding of the data if he knows the mean and distribution function exactly (we refer to the differences of mean and DF as structural differences). Thus, people usually are interested in finding what are the differences for mean or DF of two data groups, say X and Y, because this information is useful for decision-makers to make decisions or predictions. Mathematically, for the mean difference Δ between groups X and Y, one can use the equation $\Delta = E(Y) - E(X)$ to calculate it, where

$$E(Y) = \frac{1}{m}\sum_{j=1}^{m} y_j \text{ and } E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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 DF of X and Y and it is a constant given by the user. Generally, the exact form of the DF 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 $I(X<a)=1$ if $X<a$, otherwise $I(X<a)=0$. This is called the non-parametric model. If we know the exact form of the DF of G_Y (or F_X) in advance, we then call this semi-parametric model.

Building Confidence Intervals

Researchers have been used the bootstrap method to construct confidence intervals (CI) in link discovery (Adibi, Cohen, & Morrison 2004). Superior to the bootstrap method, the empirical likelihood (EL) method has many valuable features in practice and is popular in statistics and other fields (Owen 01).

In our previous work (Huang 2006) we have proposed a model that adopted the EL method to deal with the problem of measuring the differences of two contrasting groups under semi-parametric assumption. The model assumes that the distribution function of one of the two contrasting groups is known in advance, thus this information can be utilized in constructing confidence intervals. A more accurate result will be obtained if the assumption is approximately in accordance with the data. But in many real world applications, people have little priori knowledge about the data, and they can't specify the exact form of the DF of the data in the model. Generally, a misspecified modal may produce inaccurate or misleading results. Aiming to solve this problem, in this paper we improve the model based on our previous work, which can deal with the situation that the distribution function of the data can't be obtained in advance, i.e., the two data groups are non-parametric.

Similar to our previous work, we first formulate the mean and DF differences of two contrasting groups as mentioned in introduction, that is, $\Delta = E(Y) - E(X)$ and $\Delta = G_Y(\alpha) - F_X(\alpha)$ respectively. We define the empirical likelihood function for the two contrasting groups as

$$\prod_{i=1}^{m} p_i \prod_{j=1}^{n} q_j$$

Copyright © 2007, Association for the Advancement of Artificial Intelligence(www.aaai.org). All rights reserved.
*Corresponding author: Shichao Zhang

where

$$p_i > 0, i = 1, \cdots, m, \quad \sum_i p_i = 1$$
$$\mathrm{q_j} > 0, \mathrm{j} = 1, \cdots, n, \quad \sum_{\mathrm{j}} \mathrm{q_j} = 1$$

Then the log-empirical likelihood ratio statistic R(delta) is defined according to the empirical likelihood theory (). The log-empirical likelihood ratio statistic converges to a weighted Chi-squared distribution (), which will be used to construct the EL based confidence intervals for Δ. Let t_α satisfy

$$P(\chi_1^2 \le t_\alpha) = 1 - \alpha$$

We can construct an EL based confidence interval on Δ with coverage probability $1-\alpha$, that is $\{\Delta : -2\log(R(\Delta, \theta_{m,n})) \le t_\alpha\}$, where t_α is the critical value of chi-squared distribution under confidence level α. For the non-parametric model, as we assume that there is little prior knowledge available about the DF of the contrasting groups, the empirical forms for the DF are adopted, i.e,

$$\hat{G}_Y(\alpha) = \tfrac{1}{m}\sum_{j=1}^{m} I(y_j \le \alpha), \ \hat{F}_X(\alpha) = \tfrac{1}{n}\sum_{i=1}^{n} I(x_i \le \alpha)$$

These empirical forms of DF are used in the log-empirical likelihood ratio equations, in order to build confidence intervals for non-parametric contrasting groups.

Experimental Analysis

We conducted extensive experiments on various datasets from the UCI machine learning repositories to evaluate the performance of our method. The bootstrap method was adopted in the experiment for comparison, since it has been successfully used to build CI (Adibi 2004). We design two kinds of experiments, one is to construct CI for differences of data samples that come from a same group (denoted as one-group experiment), while the other is building CI for differences of two data samples that come from two different groups (denoted as two-groups experiment) respectively.

In real world applications, we are always confronted with the problem of make a decision of whether two objects belong to a same group. This involves using tools from statistical field as well as data mining field to identify the difference of these two objects. If the difference is minor, then we can say that they are the same kind of things. While if the difference between them is significant, we may believe that these objects are two kinds of things. The differences can be measured using CI at some confidence level $1-\alpha$ specified by user. This is the main goal of the one-group experiment. For the two-group experiment, we aim to identify and measure the differences between two distinct contrasting groups. The *abalone* dataset is selected in the one-group experiment, which is randomly divided into two parts. Then experiments are conducted on two data groups sampled from the two parts respectively. Results show that our method can build CI with slightly shorter length and higher probability under 95% confidence level than the bootstrap method, indicating that the two data groups come from a same data population. In the two-group experiment, the *spambase* and *wisconsin breast cancer* datasets are used, which are separated into two disjoint parts according to their class labels respectively. These two parts of data are expected to differ significantly over some features. For example, the spam emails may contain more special words, such as 'buy', than the non-spam ones. The experimental results reveal that our method can build CIs with shorter length and higher probability than the bootstrap method in most cases. But when a same numerical attribute of the two contrasting data groups differs significantly and has a large variance, the bootstrap method performs more inferior than our EL based method. This can be seen in the experimental results both in the *spambase* and *Wisconsin breast cancer* datasets.

Conclusions and Future Work

In this paper we have proposed a model for building confidence intervals for the mean and DF differences between two non-parametric contrasting groups. Extensive experiments show that our method outperforms the bootstrap method. One of the main directions of our future work will be to utilize the derived confidence intervals, along with the differences, to make predictions about the properties of the contrasting groups.

Acknowledgement

This work is partially supported by Australian Research Council Discovery Projects (DP0449535, DP0559536 and DP0667060), a China NSF major research Program (60496327), China NSF grants (60463003, 10661003), an Overseas Outstanding Talent Research Program of Chinese Academy of Sciences 06S3011S01), an Overseas-Returning High-level Talent Research Program of China Human-Resource Ministry, and Innovation Plan of Guangxi Graduate Education (2006106020812M35).

References

Adibi, J., Cohen, P. and Morrison, C. (2004). Measuring Confidence Intervals in Link Discovery: A Bootstrap Approach. *KDD04*.

Bay, S. & Pazzani, M. (1999). Detecting Change in Categorical Data: Mining Contrast Sets. *KDD99*, 302-306.

Bay, S. &, Pazzani, M. (2001). Detecting Group Differences: Mining Contrast Sets. *Data Mining and Knowledge Discovery*, 5(3): 213-246.

Huang, H., Qin, Y. S. et al. (2006). Difference Detection Between Two Contrast Sets. *DaWak06*, pp 481-490.

Owen, A. (2001). *Empirical likelihood*. Chapman & Hall, New York.

Webb, G., Butler, S., and Newlands, D. (2003). On detecting differences between groups. *KDD03*, 256-265.