Difference Detection Between Two Contrast Sets*

Hui-jing Huang[1], Yongsong Qin[2], Xiaofeng Zhu[2], Jilian Zhang[2], and Shichao Zhang[3,**]

[1] Bureau of Personnel and Education, Chinese Academy of Sciences, Beijing, China
[2] Deparment of math and Computer Science, Guangxi Normal University, China
[3] Faculty of Information Technology, UTS, PO Box 123, Broadway NSW 2007, Australia
hjhuang@cashq.ac.cn, ysqin@mailbox.gxnu.edu.cn,
xfzhu_dm@163.com, zhangjilian@yeah.net, zhangsc@it.uts.edu.au

Abstract. Mining group differences is useful in many applications, such as medical research, social network analysis and link discovery. The differences between groups can be measured from either statistical or data mining perspective. In this paper, we propose an empirical likelihood (EL) based strategy of building confidence intervals for the mean and distribution differences between two contrasting groups. In our approach we take into account the structure (semi-parametric) of groups, and experimentally evaluate the proposed approach using both simulated and real-world data. The results demonstrate that our approach is effective in building confidence intervals for group differences such as mean and distribution function.

1 Introduction

In intelligent data analysis, identifying the mean and distribution differences between two groups is useful in predicting the properties of a group using one another. In medical research, it is interesting to compare the mean value of prolonging patient's life between a group using a new product (medicine) and a group with another product; In research of children's growth, the height below/over the standard are important, since the median height (near the standard) is associated with normal growth status, it may be meaningful with children's growth to compare two groups on the basis of both below the standard or over the standard of height. In this paper we are interested in constructing confidence intervals for mean and distribution differences between two data groups.

Work in [2, 3, 4, 17] focus on mining contrast sets: conjunctions of attributes and values that differ meaningfully in their distribution across groups. This allows us to answer queries of the form, "How are History and Computer Science students different?" or "What has changed from 1993 through 1998?"

Another kind of related work is change mining in [7, 12, 16]. In the change mining problem, there are an old classifier, representing some previous knowledge about

* This work is partially supported by Australian large ARC grants (DP0449535 and DP0559536), a China NSF major research Program (60496327), a China NSF grant (60463003), a National Basic Research Program of China (2004CB318103), and a National Science Foundation of China (60033020).
** Correspondence author.

A Min Tjoa and J. Trujillo (Eds.): DaWaK 2006, LNCS 4081, pp. 481 – 490, 2006.

classification, and a new data set that has a changed class distribution. The goal of change mining is to find the changes of classification characteristics in the new data set. Change mining has been applied to identifying customer buying behavior [6], association rules [1], items over continuous append-only and dynamic data streams [18], and predicting source code changes [10].

The work of [8] uses the bootstrap approach to measure the uncertainty in link discovery (LD), while most current LD algorithms do not characterize the probabilistic properties of the hypothesis derived from the sample of data. The authors adopt the bootstrap resampling to estimate group membership and their associated confidence intervals, because it makes no assumptions about the underlying sampling distribution and is ideal for estimating statistical parameters.

Different from the above work, our approach takes into account the structure of a group: parametric, semi-parametric, or nonparametric; the imputation method when contrasting groups are with missing data; and confidence intervals for the mean and distribution differences between two groups. Use F and G to denote the distribution functions of groups x and y, respectively. We construct confidence intervals for the mean and distribution differences between contrasting groups x and y using an empirical likelihood (EL) model.

The rest of this paper is organized as follows. Section 2 presents the semi-parametric model, data structure and imputation method. In Section 3, the empirical likelihood ratio statistic and the empirical likelihood (EL) based confidence intervals (CIs) for the mean and distribution function differences are constructed. In Section 4, we give the experimental results both on the simulation data and a real medical dataset. Conclusion and future work are given in Section 5.

2 Semi-parametric Model, Data Structure and Imputation Method

We use $F(x)$ and $G_{\theta_0}(y)$ to denote the distribution functions of groups x and y, respectively, where G is known, F and θ_0 are unknown. This is regarded as Semi-parametric model. We are interested in constructing confidence intervals for some differences of x and y such as the differences of the means and the distribution functions of two groups. In general, either F or G is unknown, or both. So nonparametric methods are developed to address this situation. In the case of complete observations, related work can be found in [9].

For any difference, denoted by Δ , the following information is available:

$$E\ \omega\ (\ x\ ,\theta_0\ ,\Delta\ ) = 0 \tag{1}$$

Where ω is a function in a known form. Some examples that fit (2.1) are given in the following.

Difference of means: Denote $\mu_1=E(x),\mu_2=E(y)=\mu(\theta_0)$ and $\Delta=\mu_2-\mu_1$, Let

$$\omega(x,\theta_0,\Delta)=x-\mu(\theta_0)+\Delta \tag{2}$$

Difference of distribution functions: For fixed x_0, denote $p_1=F(x_0)$, $p_2=G_{\theta_0}(x_0)=p(\theta_0)$ and $\Delta=p_2-p_1$. Let

$$\omega(x,\theta_0,\Delta)=I(x\le x_0)-p(\theta_0)+\Delta \tag{3}$$

Where I(.) is the indicator function. Note that we can assume that F follows exponential or normal distribution in order to construct the model (denote as exponential and normal distribution model respectively).

We use a simple method to represent the data. Consider the following simple random samples of data associated with groups x and y, we denoted them as (x,δ_x) and y respectively,

$$(x_i,\delta_{x_i}),\ i=1,\cdots,m;\qquad y_j,\ j=1,\cdots,n.$$

Where

$$\delta_{x_i}=\begin{cases}0, & if\ x_i\ is\ missing\\ 1, & otherwise,\end{cases}. \tag{4}$$

We assume that x and y are missing completely at random (MCAR) [11], i.e. $P(\delta_x=1|x)=P_1$ (constant) throughout this paper. We also assume that (x,δ_x) and y are independent. Next, an example from real life application is given below in order to illustrate the goal of this paper.

In the medical analysis of a kind of disease, the breast cancer for example, some data are obtained from the patients (see Table 1).

Table 1. Breast Cancer data

Patient ID	Radius	Smoothness	Perimeter	Diagnosis
1	13.5	0.09779	78.04	benign
2	21.16	0.1109	94.74	malignant
3	12.5	0.0806	62.11	benign
4	14.64	0.01078	97.83	benign
…	…	…	…	…

There are two problems that we concerned most. One, what is the difference of the benign and malignant patients with regard to a specified feature? The other is, how reliable the difference is, when we calculated it from the sample data of the benign and malignant patients?

One can compute the difference of a specified feature of two groups by using simple statistical methods or other more sophisticated data mining techniques [2, 3, 4, 17]. While for the second problem, we use the empirical likelihood (EL) method to construct the confidence intervals, under a significance level α, for the difference Δ of two groups with missing data.

A common method for handling incomplete data is to impute a value for each missing value and then apply standard statistical methods to the complete data as if they were true observations. Commonly used imputation methods include deterministic imputation and random imputation [15]. We refer to the reader to [11] for examples and excellent account of parametric statistical inferences with missing data.

Let $r_x = \Sigma_{i=1}^{m} \delta_{x_i}$, $m_x = m - r_x$. Denote the sets of respondents and nonrespondents with respect to x as s_{rx} and s_{ry}, respectively. We use random hot deck imputation method to impute the missing values. We do not use the deterministic imputation as it is improper in making inference for distribution functions [15]. Let x_i^* be the imputed values for the missing data with respect to x. Random hot deck imputation selects a simple random sample of size m_x with replacement from s_{rx}, and then uses the associated x-values as donors, that is, $x_i^* = x_j$ for some $j \in s_{rx}$. Let $x_{I,i} = \delta_{x_i} x_i + (1 - \delta_{x_i}) x_i^*$ represent the 'complete' data after imputation, where $i=1,\cdots,m$, $j=1,\cdots,n$, .

We will investigate the asymptotic properties of the empirical likelihood ratio statistic for Δ based on $x_{I,i}, i=1,\cdots,m$; $y_{I,j}, j=1,\cdots,n$. The results are used to construct asymptotic confidence intervals for Δ.

3 Building CI for Δ Based on Empirical Likelihood

At first, the empirical likelihood ratio statistic is constructed. It is interesting to notice that the empirical likelihood ratio statistic under imputation is asymptotically distributed as a weighted chi-square variable χ_1^2 [13, 14], which is used to construct the EL based confidence interval for Δ. The reason for this deviation from the standard χ_1^2 is that the complete data after imputation are dependent.

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 $\{\Delta : -2\omega \log(R(\Delta, \theta_{m,n})) \le t_\alpha\}$, where ω is the weight [13, 14].

This result can directly apply to test the hypotheses on Δ. For instance, if the hypothesis is $H_0: \Delta = \Delta_0, H_1: \Delta \neq \Delta_0$, we first construct the confidence interval on Δ. Then check if Δ_0 is in the interval. If Δ_0 is in the interval, we accept the hypothesis H_0 and reject H_1; otherwise, H_0 should be rejected and H_1 is accepted.

We also want to notice that the result can apply to the data without missing values. In complete data situation, we can see that the asymptotic distribution of the EL statistic is found to be a standard χ_1^2 distribution. The EL based confidence interval for Δ in complete data case is thus constructed as $\{\Delta : -2\log(R(\Delta, \theta_{m,n})) \le t_\alpha\}$.

4 Experiments

Extensive experiments were conducted on a DELL Workstation PWS650 with 2G main memory and 2.6GHz CPU, the operating system is WINDOWS 2000.

4.1 Simulations Models

We conducted a simulation study on the finite sample performance of EL based confidence intervals on the mean difference $\Delta_1=E(y)-E(x)$, and the distribution function difference $\Delta_2=G_{\theta_0}(x_0)-F(x_0)$ for fixed x_0. For the purpose of simulating the real world data distributions as closely as possible, we generated two groups of $x_i s$ and $y_i s$ from the exponential distributions ($\exp(1)$ and $\exp(2)$) and the normal distributions ($N(2,2)$ and $N(3,2)$) respectively, because these two data distributions are the most popular and common distributions in real world applications. And then the exponential and normal distribution models are running on these different distributed datasets. The following two cases of response probabilities were used under the MCAR assumption (in which the response rates is denoted as P): Case 1: $P_1=0.6$; Case 2: $P_1=0.9$. The response rates in Case 2 were higher than those in case 1, which were chosen to compare the performance of EL confidence intervals under different response rates.

Sample sizes were chosen as (m, n) = (100, 100), and (m, n) = (200, 150) for the purpose to compare the performance of EL confidence intervals under different sample sizes. For each of the cases of different response rates and sample sizes, we generated 1,000 random samples of incomplete data $\left\{(x_i,\delta_{x_i}), i=1,\cdots,m;\ y_j, j=1,\cdots,n\right\}$. For nominal confidence level $1-\alpha$ =0.95, using the simulated samples, we evaluated the coverage probability (CP), the average left endpoint (LE), the average right endpoint (RE) and the average length of the interval (AL) of the empirical likelihood based (EL) intervals.

Tables 2-9 present the performance of proposed method for finding CIs of the mean difference and distribution function with different models on different distributed datasets. More detailed experimental settings can be seen in the table titles.

Table 2. CIs of the mean difference for the exponential distribution model (with exponential distributed data, true difference x_0 =1)

Case	(m,n)	CP(%)	LE	RE	AL
1	(100,100)	100	0.374348	1.664007	1.289659
1	(200,150)	99.69	0.381603	1.552474	1.170871
2	(100,100)	99.78	0.498466	1.548741	1.050275
2	(200,150)	98.77	0.518341	1.454267	0.935926

Table 3. CIs of the mean difference for the normal distribution model (with exponential distributed data, true difference $\Delta_1 = 1$)

Case	(m,n)	CP(%)	LE	RE	AL
1	(100,100)	96.78	0.259185	1.233423	0.974238
1	(200,150)	92.46	0.401004	1.204256	0.803252
2	(100,100)	88.15	0.556322	1.168804	0.612482
2	(200,150)	88.82	0.572257	1.176884	0.604627

Table 4. CIs of the distribution function difference for the exponential distribution model (with exponential distributed data, fixed $x_0 = 2$, true difference $\Delta_2 = -0.2325$)

Case	(m,n)	CP(%)	LE	RE	AL
1	(100,100)	90.98	-0.259507	-0.10194	0.158
1	(200,150)	89.50	-0.253168	-0.126162	0.127
2	(100,100)	85.64	-0.224214	-0.134105	0.091
2	(200,150)	82.86	-0.227793	-0.158197	0.079

Table 5. CIs of the distribution function difference for the normal distribution model (with exponential distributed data, fixed $x_0 = 2$, true difference $\Delta_2 = -0.1915$)

Case	(m,n)	CP(%)	LE	RE	AL
1	(100,100)	92.21	-0.415641	-0.193948	0.222
1	(200,150)	87.62	-0.403683	-0.218695	0.185
2	(100,100)	83.50	-0.401349	-0.266323	0.135
2	(200,150)	84.62	-0.399944	-0.264595	0.135

Table 6. CIs of the mean difference for the exponential distribution model (with normal distributed data, true difference $\Delta_1 = 1$)

Case	(m,n)	CP(%)	LE	RE	AL
1	(100,100)	100	0.304478	2.04389	1.739411
1	(200,150)	99.67	0.28442	1.98484	1.70043
2	(100,100)	100	0.42359	1.8026	1.379
2	(200,150)	98.68	0.38113	1.7308	0.979761

Table 7. CIs of the mean difference for the normal distribution model (with normal distributed data, true difference $\Delta_1 = 1$)

Case	(m,n)	CP(%)	LE	RE	AL
1	(100,100)	98.76	0.362515	1.561752	1.199237
1	(200,150)	99.01	0.453377	1.408632	0.955255
2	(100,100)	98.37	0.475443	1.373007	0.897564
2	(200,150)	94.12	0.599176	1.306111	0.706935

Table 8. CIs of the distribution function difference for the exponential distribution model (with normal distributed data, fixed x_0 =2, true difference Δ_2 =–0.2325)

Case	(m,n)	CP(%)	LE	RE	AL
1	(100,100)	93.64	-0.216152	0.139717	0.355869
1	(200,150)	90.52	-0.175146	0.13745	0.312596
2	(100,100)	88.10	-0.162031	0.111836	0.273867
2	(200,150)	87.58	-0.130788	0.104844	0.261068

Table 9. CIs on the distribution function difference for the normal distribution model (fixed x_0 =2, true difference Δ_2 =–0.1915)

Case	(m,n)	CP(%)	LE	RE	AL
1	(100,100)	91.42	-0.419944	-0.202104	0.2178
1	(200,150)	90.48	-0.39838	-0.228987	0.169
2	(100,100)	88.75	-0.377728	-0.238151	0.13958
2	(200,150)	89.68	-0.379188	-0.270484	0.10870

Tables 2-9 reveal the following results:

For every response rate and sample size, the coverage probabilities (CPs) of all EL-based confidence intervals for mean are close to the theoretical confidence level 95%. In almost all situations, the lengths of CIs also become smaller as the sample size increases. The same trends occur when considering different response rates. While the ALs for distribution function difference fluctuate slightly with respect to different sample size and response rates.

Another interesting phenomenon is that the CIs built by using normal distribution model for mean difference are shorter than those by exponential distribution model, without much loss of coverage accuracy. That is to say, we can use the normal distribution model to construct CIs in real applications when we have no prior knowledge about the distribution of the data.

We can see from above results that the length of CIs will be shorter when the amount of sample data increases, because the information that is useful for building the CIs also increases. So under the same significance level α , the shorter CIs will give the same confidence of the difference. Note that higher response rate means that there are more data available when building CIs than those under lower response rate.

4.2 Experiments on UCI Dataset

We also conducted extensive experiments on real world dataset, due to the fact that the real world data do not fit the ideal statistical distributions exactly. What's more, there may be noises in real world data, which will distort the distribution of the real world data.

We used the medical dataset, Wisconsin Diagnostic Breast Cancer (WDBC), which is downloaded from [5]. It contains 569 instances in total and 32 features for each instance. Each instance, represented a patient, has been classified as benign and malignant according to these features. The WDBC dataset contains 357 benign instances and 212 malignant instances. For interesting of space, we only report the

experimental results of attribute 4 and 27. We give some statistical information of these two features in Table 10, more detailed information about these features can be seen in [19]. In order to verify the effectiveness of our method, we randomly divide WDBC into two parts. One (contains 2/3 instances, denoted as BS) is used to construct the CI, the other (contains 1/3 instances, denoted as VS) is used to verify the coverage probability (CP) of the CI. We then divide the BS into two groups, that is, the Benign and Malignant groups. Let the values of attribute A from Benign group be the group x, and those from Malignant be group y. Then CI is built based on group x and y using the techniques described in Section 3. In the verification process of CP, we divide the VS into two groups (Benign and Malignant) and compute the difference $\hat{\Delta}$ of them with respect to attribute A. Thus we can easily see whether $\hat{\Delta}$ falls into the range of the constructed CI.

Table 10. Statistics for attribute 4 and 27 of Wisconsin Diagnostic breast Cancer

	Mean		Distribution function	
	A4	A27	A4 (x_0=15)	A27 (x_0=0.1)
Malignant	21.6	0.1448	0.0189	0.0094
Benign	17.91	0.1249	0.2437	0.1092
Difference Δ	3.69	0.0198	-0.2248	-0.0998

(A4: Mean texture, A27: Worst smoothness)

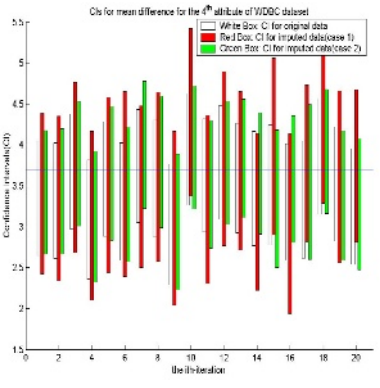

Fig. 2. CIs for attribute 4

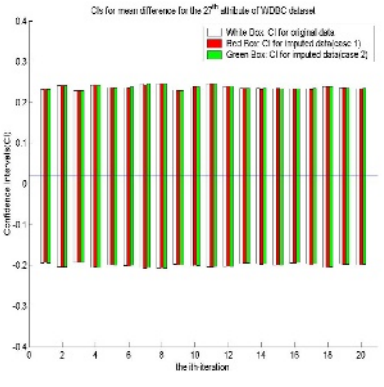

Fig. 3. CIs for attribute 27

Figures 2, 3, and 4 compare the CIs for mean on the complete and imputed dataset WDBC under different missing rates. We give the experimental results of CIs for mean difference of attribute 4 and 27 in Figures 2 and 3. In Figure 2, we can see that the length of CIs built from imputed data (case-1) is much larger than those built from original data (without missing). While the length of CIs built from imputed data (case-2) is very close to the original data's CIs. This means that with a lower missing rate, the length of CIs are shorter. The same phenomenon can be seen in CIs of DF for attribute 4 (see Figure 4). As for attribute 27, the lengths of CIs built from case-1, case-2 and the

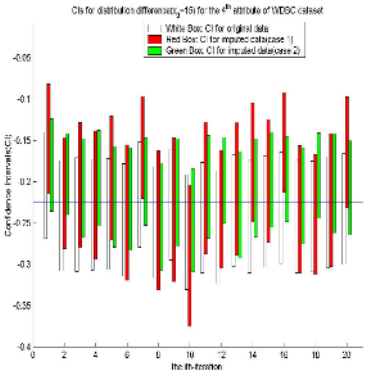

Fig. 4. CIs for distribution function of attribute 4

original data are very close, which almost give the same coverage probabilities. However, we don't present the CIs of DF for attribute 27, due to lack of space.

The average left, right endpoint (LE, RE), length and CP are listed in table 11. An interesting observation is that the value of CP is decreesing from 70% to 60% when the response rate P (note that missing rate=1-P) is increasing from (0.6, 0.7) to (1, 1). Note that the original data has the response rate (P1=1, P2=1). On the other hand, the average length AL is also decreasing when the response rate is increasing, that is, the AL is longer when the groups contain more missing data, which are imputed by random imputation. By combining these two facts, we know that the length of CIs will be shorter when using lower missing rate data, but the CP will be lower. On the contrary, the length of CIs will be longer when using higher missing rate data, resulting in a higher CP.

For group with small range of values, attribute 27 for example, the LE, RE, AL and CP of CIs are comparatively stable under different response rates.

Table 11. Average intervals, ALs and CP for mean

	LE	RE	AL	Average	CP (%)
A.4 (Original)	2.789261	4.198938	1.409677		60
Case 1(0.6, 1)	2.560288	4.617375	2.057087	4.124679	75
Case 2(0.9, 1)	2.781663	4.362656	1.580993		65
A.27 (Original)	-0.19973	0.237031	0.436765		100
Case 1(0.6, 1)	-0.19981	0.237	0.43681	0.022297	100
Case 2(0.9, 1)	-0.19961	0.237172	0.436782		100

5 Conclusions

In this paper we have proposed a new method based on empirical likelihood (EL) for identifying confidence intervals for the mean and distribution differences between two contrasting groups. The mean and distribution differences between two contrasting groups assist in predicting the properties of a group using one another. To extend the applied range, our method takes into account the situation of two contrasting groups, one group is known well, and the other is unknown (for example, having no information about the form of distribution and parameters). In comparing of the differences of two contrasting groups with missing data, we have shown that the EL-based confidence intervals works well in making inference for various differences

between the two groups, especially for the mean and distribution function differences. We have also shown that this result can directly be used to test the hypotheses on the differences, and that the result can apply to the complete data settings.

References

1. Au, W., Chan, K. (2005), Mining changes in association rules: a fuzzy approach. Fuzzy Sets and Systems, 149(1): 87-104.
2. Bay, S., and Pazzani, M. (1999), Detecting Change in Categorical Data: Mining Contrast Sets. KDD'99, pp. 302-306.
3. Bay, S., and Pazzani, M. (2000), Characterizing Model Erros and Differences. ICML'00, pp. 49-56.
4. Bay, S., and Pazzani, M. (2001), Detecting Group Differences: Mining Contrast Sets. Data Mining and Knowledge Discovery, 5(3): 213-246.
5. Blake, C., and Merz,C. (1998). UCI Repository of machine learning database. http:// www.ics.uci.edu/~mlearn/
6. Cho, Y.B., Cho, Y.H., & Kim, S. (2005), Mining changes in customer buying behavior for collaborative recommendations. Expert Systems with Applications, 28(2): 359-369.
7. Cong, G., and Liu, B. (2002). Speed-up Iterative Frequent Itemset Mining with Constraint Changes. ICDM'02, pp 107-114.
8. Adibi, J. Cohen, P., and Morrison, C. (2004), Measuring Confidence Intervals in Link Discovery: A Bootstrap Approach. KDD'04.
9. Hall, P., and Martin, M. (1988), On the bootstrap and two-sample problems. Austral. J. Statist, 30A, pp 179-192.
10. Li, H.F., Lee, S.Y., and Shan, M.K., (2005). Online Mining Changes of Items over Continuous Append-only and Dynamic Data Streams. The Journal of Universal Computer Science, 11(8): 1411-1425 (2005).
11. Little, R. and Rubin, D. (2002), Statistical analysis with missing data. 2nd edition. John Wiley & Sons, New York.
12. Liu, B., Hsu, W., Han, H. and Xia, Y. (2000), Mining Changes for Real-Life Applications. DaWaK'00, pp337-346.
13. Qin, J., (1994). Semi-empirical likelihood ratio confidence intervals for the difference of two sample mean. Ann. Inst. Statist. Math. 46, 117-126.
14. Qin, J. and lawless, J. (1994), Empirical likelihood and general estimating equations. Ann. Statists. 22, 300-325.
15. Rao, J. (1996). On variance estimation with imputed survey data. J. Amer. Statist. Assoc., 91: 499-520.
16. Wang, K., Zhou, S., Fu, A., and Yu, X. (2003). Mining Changes of Classification by Correspondence Tracing. SIAMDM'03, 2003.
17. Webb, G., Butler, S., and Newlands, D. (2003), On detecting differences between groups. KDD'03, pp. 256-265.
18. Ying, A., Murphy, G., Raymond, T., and Mark, C. (2004), Predicting Source Code Changes by Mining Change History. IEEE Trans. Software Eng., 30(9): 574-586.
19. W.N. Street, W.H. Wolberg and O.L. Mangasarian1993. Nuclear feature extraction for breast tumour diagnosis. IS&T/SPIE 1993 volume 1905, pages 861-870, San Jose, CA, 1993.