Int J Software Informatics, Volume 6, Issue 4 (2012), pp. 535–552
International Journal of Software and Informatics, ISSN 1673-7288

E-mail: ijsi@iscas.ac.cn
http://www.ijsi.org
Tel: +86-10-62661040

Classifying Incomplete Data Using Group Difference Detection with Parimputation Approach

Shichao Zhang[1,2] and Jilian Zhang[1,3]

[1](College of Computer Science and Information Technology, Guangxi Normal University,
Guilin 541004, China)

[2](Faculty of Information Technology, University of Technology Sydney, Australia)

[3](School of Information Systems, Singapore Management University, Singapore)

Email: shichao.zhang@uts.edu.au, yet_zjl@tom.com

Abstract We propose an efficient approach for classifying insufficient dataset with missing data (incomplete data) with group difference detection. Specifically, missing data in an insufficient dataset are first completed with the parimputation strategy. And then, the insufficient dataset is grouped by contrasting with a known dataset (transfer learning). Finally, for assessing the quality of the induced models, empirical likelihood (EL) inference is used to estimate the confidence intervals of structural differences between the insufficient dataset and the known dataset. In such a way of mining, classifying incomplete data can be beneficial to industries as it will provide easier and smarter use of information. This will include evaluating a new medical product by detecting differences between the new product and an old one for pharmaceutical companies and, identifying frauds by detecting abnormal operations. To experimentally illustrate the benefits, we evaluate the proposed approach using UCI datasets, and demonstrate that our method works much better than the bootstrap resampling method on, for example, distinguishing spam from non-spam emails; and the benign breast cancer from the malign one.

Key words: incomplete data; missing data imputation; group difference detection

Zhang SC, Zhang JL. Classifying incomplete data using group difference detection with parimputation approach. *Int J Software Informatics*, Vol.6, No.4 (2012): 535–552. http://www.ijsi.org/1673-7288/6/i140.htm

1 Introduction

Incompleteness of information is ubiquitous in real applications and incomplete information mining is an actual and challenging issue. Incomplete information is mainly classified into two categories: (1) a dataset with missing data, and (2) a dataset with insufficient information, simply called as insufficient dataset. However, existing

---

This work is supported in part by the Australian Research Council (ARC) under large grant DP0985456, the China "1000-Plan" National Distinguished Professorship, the China 863 Program under grant 2012AA011005, the Nature Science Foundation of China (NSFC)under grant 61170131, the China 973 Program under grant 2013CB329404, the Guangxi "Bagui" Teams for Innovation and Research, the Jiangsu Provincial Key Laboratory of E-business at the Nanjing University of Finance and Economics, the the Guangxi Natural Science Foundation for Innovation and Research Teams under grant 2012GXNSFGA060004, and the the Guilin Natural Science Foundation.
Corresponding author: Shichao Zhang, Email: shichao.zhang@uts.edu.au
Received 2012-01-20; Accepted 2012-08-09.

models and algorithms are designed for learning complete yet quality information and do not perform well for processing incomplete information. Some researchers simply discard the data (or instance) with missing values and apply these models and algorithms only to the complete instances in a dataset. However, it often leads the original dataset to be an insufficient dataset, because a datum with missing values often contains many observed attribute values that are valuable in problem solving. If you discard the datum you may lose key features. In particular, it wastes data resources because even data with missing values is very expensive and valuable in some applications. Although we are inundated with vast amounts of information, data is utterly lacking in many real applications. For example, we often lack data for a new disease, a new product, or a dangerous item. Therefore, there is a clear need to learn quality models from incomplete information.

In this paper, we study an efficient approach for classifying insufficient dataset with missing data (incomplete data) with group difference detection. Specifically, missing data in an insufficient dataset are first completed with the parimputation strategy. And then, the insufficient dataset is grouped by contrasting with a known dataset (transfer learning). Finally, for assessing the quality of the induced models, empirical likelihood (EL) inference is used to estimate the confidence intervals of structural differences between the insufficient dataset and the known dataset. In such a way of mining, classifying incomplete data will be beneficial to industries as it will provide easier and smarter use of information. This will include identifying clues from data sources for snaring terrorists; evaluating a new medical product by detecting differences between the new product and an old one for pharmaceutical companies; and identifying frauds by detecting abnormal operations; bridging rule mining for financial companies.

Difference detection is naturally and widely used in scientific research. For example, consider a new medicine B for a specific disease in medical research. Researchers usually compare B with an old medicine A, which has been used effectively to treat the disease for many years. Differences on statistics of the two medicines are studied so that researchers have a clear understanding of the properties of B. The observations of statistics when applying medicine A and B to the disease are called contrast groups (or simply groups, sets, populations), and the main differences considered between the two groups are the mean and distribution, which are referred to as structural differences in this paper. Difference detection for groups is of great importance in data mining and machine learning community, such as exploratory data analysis (EDA), change mining etc. On the other hand, difference detection gains prevalence recently in many real world applications. For instance, in children's health research, the height below/over the standard are 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. As another application in information security, identifying group differences between spam and non-spam emails (they form the two email groups) can provide knowledge to users so that the users can distinguish spams from normal emails. Accordingly, software companies can devise well-performed anti-spam email systems based on these differences, so as to improve the availabilities of email systems for users.

Much research effort has been devoted to detect group differences in different context. For example, there is some research work reported on mining group differences between contrast groups from observational multivariate data[2−4,24] in data mining community. Researchers from the machine learning field also encounter the problems of difference detection. This is witnessed by a few publications on change mining over decision trees etc[8,14,21]. Like difference detection, mining changes assists in intelligent decision support for business managers, such as understanding customer behaviors.

However, contrast groups are only two samples obtained through limited observations or tests over contrasting objects, and sometimes the obtained data may be (1) incomplete, i.e., with missing values, and (2) distorted, i.e., there may be noises or outliers. Therefore, the differences of groups derived from the data are inevitably incurring uncertainties. This generates an urgent need of measuring the uncertainty of structural differences between contrasting groups, when the observations (the data) are incomplete or with noises.

Existing techniques for difference detection and change mining both individually and collectively participate in the goal of association analysis. While Zhang [29] reported difference detection between complete datasets, in this paper we propose an efficient approach for measuring uncertainty of group difference by identifying the confidence intervals of structural differences between contrast groups. Specifically, for a pre-assigned confidence level 1-α, the confidence interval would contain the parameters of interest (refer to the differences of mean and distribution function of the two contrast groups in this article) with probability not smaller than the prescribed confidence level 1-α, which is more reliable than the point estimate of the parameter (as the point estimate does not tell us how far is it away from the true parameter value, or the point estimate does not tell us the lower and upper bound of the parameter value). On the other hand, the derived confidence intervals can be directly applied to test the hypotheses on the parameter of interest. For instance, given a significance level α, if the hypothesis is $H : \theta = \theta_0$, we first construct the confidence interval on$\theta - \theta_0$, then check whether $\theta' - \theta'_0$ lies in the interval or not (here θ and $\theta'$is the parameter of population and observations/samples, respectively). If the answer is yes, under the significance level α, we accept the hypothesis; otherwise, the hypothesis is rejected.

From statistics' point of view, mean and distribution function are important measures of the data, and one can almost have a full understanding of the data if he knows the exact mean and distribution function. We can use statistical methods to obtain the above differences. For instance, for the mean difference Δ, between groups X and Y, one can use the equation $\Delta = E(Y) - E(X)$ to calculate difference of mean, where $E(Y) = \frac{1}{m}\sum_{j=1}^{m} y_j$, $E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$, and x_i, y_i is the sample data of group X and Y, respectively. As for the distribution function difference Δ between X and Y, one can compute it as $\Delta = G_Y(\gamma) - F_X(\gamma)$, where G_Y and F_X are the distribution functions of Y and X respectively, and γ is a reference point for comparing the distribution function of X and Y and is a constant given by the user. Generally, since the exact form of distribution function is difficult to obtain, an empirical form is adopted in practice, i.e., $\hat{G}_Y(\gamma) = \frac{1}{m}\sum_{j=1}^{m} I(y_j \leq \gamma)$, $\hat{F}_X(\gamma) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq \gamma)$, where $I(.)$ is an indicator function and I(x<y)=1 if x<y is true; otherwise I(x<y)=0.

This is called a non-parametric model. If we know the form of either G_Y or F_X in advance, we call it a semi-parametric model.

In real world applications obtained data are sampled from a population, thus the knowledge mined 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 intervals in order to measure their uncertainties, since confidence intervals (CI) can tell people how reliable the derived differences are given two groups X and Y.

We focus on applying the non-parametric model to measure how reliable the differences in mean and distribution function of two groups X and Y are, when there are missing data. We are only taking into account the case in which there are missing values in the data, whereas dealing with the situation that the data contain outliers is out of the scope of this paper, and will not be discussed. Instead, we take it as future work. We experimentally evaluate our approach using UCI datasets, and demonstrate that our method works much better than its competitors on applications, such as distinguishing spam from non-spam emails and the benign breast cancer from the malign one.

The rest of this paper is organized as follows. Section 2 briefly reviews related work and some basic concepts, including the empirical likelihood method, data structure and imputation method for dealing with incomplete dataset. In Section 3, we describe how to build confidence intervals for mean and distribution function by using the empirical likelihood method; the bootstrap method for constructing confidence intervals is also presented in this section. In Section 4, we give extensive experimental results of our method on the simulation dataset as well as UCI datasets. Conclusion is given in Section 5.

2 Preliminary

2.1 Related work

Group difference detection has attracted tremendous interests from researchers around the world. For example, work in Refs. [2-4,24] focused 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, “What is the difference in study hours between History and Computer Science students?” or “What has changed in income level from 1993 through 1998?”

Another direction of related work is change mining, as in Cong & Liu 2002; Liu, et al 2000; Wang, et al 2003. For the change mining problem, there are an old classifier, representing some previous knowledge about 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 various applications such as identifying customer buying behavior[7], association rules[1], items over continuous append-only and dynamic data streams[25], and predicting source code changes[12].

The above work has revealed many interesting results of difference detection. However, they did not consider the situation where missing data may involve, and they also did not measure reliabilities of the derived group differences from statistics’ point of view. Different from the above work, our approach in this paper takes into account (1) the structure of a group (non-parametric), (2) the imputation strategy when contrast groups contain missing data, and (3) employing empirical likelihood (EL) method to build confidence intervals for differences in mean and distribution function of two groups[19,29].

In statistics, the problem of making inference about difference in mean is the well-known Behrens-Fisher problem, if distribution function F and G of group X and Y, respectively, are both normal. This is a very broad topic in applications. In general, both F and G are unknown in advance, thus nonparametric methods are developed to address this problem. In the case of complete observations, related work can be found in Ref. [9], among others.

As a powerful tool to deal with nonparametric settings (i.e. populations or models are not specified into some parametric structures, which describe the cases of complex data systems), the original idea of EL dates back to Ref. [10] in sample survey context. Owen[17] made a systematic study of the empirical likelihood method in the complete data settings. Owen[16] presented a form of data squashing based on empirical likelihood and outlined the differences in data mining, then he showed that empirical likelihood weighting can accelerates the rate at which coefficients are learned.

When making statistical inference it is typically assumed that all the observations in the sample are available. This may not be true in many practical situations since some observations may be missing for various reasons such as unwillingness of some sampled units to supply the desired information, loss of information caused by uncontrollable factors, failure on the part of the investigator to gather correct information, and so on. In fact, missing observations (responses in these examples) are common in opinion polls, market research surveys, mail enquiries, medical studies and other scientific experiments[28,31]. Missing data analysis covers a variety of problems that are often seen in practical applications[13,18,26]. In this situation, the usual inference procedures cannot be applied directly. Existing methods typically employ the parametric likelihood as they assume the data structures are parametric. When there is little knowledge about a population or model and there exist missing data, one often imputes missing data to form a "complete data" set and then uses the EL method to make inference based on the "complete data". Wang and Rao[22,23] first use EL method to construct confidence intervals for the mean of the response variable in a linear model with missing data (they do not specify the form of the error distribution in the linear model, i.e. the error distribution is nonparametric).

Different from the above techniques, in this paper we are interested in using EL method to construct confidence intervals for structural differences, such as the differences in mean and distribution function of two populations X and Y, when there are missing values in the data. The advantage of our model is that we need not to specify the exact distribution forms of X and Y, because in practical applications people usually have no a prior knowledge about the underlying distribution of the data. Thus, we adopt the empirical distributions of X and Y in our model.

2.2 Model and data structure

Let F and G be the distribution function of group X and Y, respectively. We are interested in constructing confidence intervals for structural differences such as differences in mean and distribution function of the two populations. Making inference of difference in mean is the well-known Behrens-Fisher problem if F and G are both normal. In general, both F and G are unknown so that non-parametric methods are developed to address this situation. In the case of complete observations, related work can be found in Refs. [9,11,15].

Let θ_0 and θ_1 be unknown parameters with respect to F and G, respectively. Let the difference between the two parameters be $\Delta = \theta_1 - \theta_0$. The following information is available

$$E\omega_1(x, \theta_0, \Delta) = 0, E\omega_2(y, \theta_0, \Delta) = 0 \tag{1}$$

where ω_i, i=1, 2, are functions of known forms. Some examples that fit Equation (1) are given in the following.

Difference of mean: by defining $\theta_0 = Ex, \quad \theta_1 = Ey$ and$\Delta = \theta_1 - \theta_0$, we get equation $\omega_1(x, \theta_0, \Delta) = x - \theta_0, \quad \omega_2(y, \theta_0, \Delta) = y - \theta_0 - \Delta$.

Difference of distribution function: For a fixed x_0, by defining$\theta_0 = F(x_0)$, $\theta_1 = G(x_0)$ and $\Delta = \theta_1 - \theta_0$, we get equation $\omega_1(x, \theta_0, \Delta) = I(x \le x_0) - \theta_0$, $\omega_2(y, \theta_0, \Delta) = I(y \le x_0) - \theta_0 - \Delta$, where $I(.)$ is an indicator function, $I(x)$=1 if x is true; otherwise $I(x)$=0.

It is interesting to measure the difference$\Delta$for X and Y. To do this, we first construct the confidence interval for difference$\Delta$of two populations. We then compute the difference Δ' using the observation data, i.e., samples. If the difference Δ' falls within the generated interval, we accept the hypothesis that the difference of the respective parameters of two groups is Δ with respect to a pre-specified significance level; otherwise we reject this hypothesis. In this paper, we construct confidence interval based on EL (empirical likelihood) method to solve the two nonparametric population problems.

2.3 Parimputation strategy

Consider the following random samples of incomplete data associated with populations $X = (x_i, \delta_{xi})$, $i = 1,\cdots, m$ and $Y = (y_j, \delta_{yj}), j = 1, \cdots, n$, where

$$\delta_{xi} = \begin{cases} 0, \text{if } x_i \text{ is missing} \\ 1, \text{otherwise} \end{cases}, \delta_{yj} = \begin{cases} 0, \text{if } y_j \text{ is missing} \\ 1, \text{otherwise} \end{cases}$$

There are several missing mechanisms in literature, such as missing completely at random (MCAR), missing at random (MAR), non-negligible, etc. Since it is difficult to identify which missing mechanism a given real dataset is, an MCAR assumption is common and viable. Throughout this paper, we assume that missing data in X and Y are MCAR[13], i.e., $P(\delta_x = 1\,|x) = Prob_1$ and $P(\delta_y = 1\,|y) = Prob_2$, where $Prob_i$ is constant and $0 \le Prob_i \le 1$, i=1, 2 . Note that we also assume that population X and Y are independent.

A common method for handling incomplete data is to impute each missing value and then standard statistical methods are applied on the complete data as if the data consist of true observations. Commonly used imputation methods include deterministic imputation and random imputation. We refer to the reader to Little and Rubin[13] for examples and excellent account of parametric statistical inferences with missing data.

In this paper we adopt a new imputation approach, parimputation (partial imputation, see Refs. [28, 31]), to dealing with missing values. The parimputation strategy is proposed for addressing those missing data in a given dataset in which all the nearest neighbors (nearest neighbor is measured using Euclidean distance) are far from them. From the observed part of an incomplete datum in a dataset, if there are some complete data in a small neighborhood of the incomplete data, we refer it as predictable missing data; otherwise, we refer it as unpredictable missing data. With the observed part of an unpredictable missing data in a dataset, finding the unpredictable missing data is similar to that of detecting outliers (or isolation points) in machine learning and data mining. This means that there are many well-established outlier detection techniques (such as John 1995; Ramaswamy, et al. 2000) that can be applied to determine whether a missing data is unpredictable or not.

Accordingly, the parimputation is defined as: imputing all the predictable missing data in a given dataset and removing all the unpredictable missing data from the dataset. Certainly, the parimputation strategy is simple and easy to be understood and implemented. With the parimputation strategy, we will investigate in the following section how to use EL based method to construct confidence intervals forΔ, given population X and Y.

3 Confidence Interval for Group Difference Δ

In many real world applications, we are always confronted with the problem of deciding whether two objects are coming from a same population or not. This involves using tools from statistics as well as data mining field to identify the difference between the two objects. If the difference is small, we can draw a conclusion that they are coming from a same population with high probability; if the difference between them is large, we believe that it is very unlikely that the two objects are belonging to a same population. To measure how reliable the conclusion is, we resort to confidence intervals (CI) from statistics community.

In this section we present the empirical likelihood (EL) based method for constructing CI for differences of mean and distribution function of population X and Y. In order to compare the performance of EL based method, we choose the bootstrap re-sampling method for comparison, which is widely used in statistics, as well as in data mining community. The reason that we choose bootstrap re-sampling is that it is a simple and effective technique to compute an estimator of the data, when the parametric form of the estimator is not available or difficult to obtain.

3.1 Empirical likelihood (EL) statistic

For two populations $X = \{x_{1,}x_2, ..., x_m\}$ and $Y = \{x_1, x_2, ..., x_n\}$ with size m and n respectively, the empirical likelihood function is defined as

$$\prod_{i=1}^{m} p_i \prod_{j=1}^{n} q_j \tag{2}$$

where $p_i > 0, i = 1, \cdots, m, \sum_i p_i = 1$, and $q_j > 0, j = 1, \cdots, n, \sum_j q_j = 1$. Note that p_i is the probability that observation of the i-th value of X obtains a specific value, i.e., $p_i = P(X_i = x_i)$. The definition of $q_j$is similar. Essentially, based on independent assumption of X and Y, the above empirical likelihood function (Eq. 2) models the overall probability that we get current sample datasets of population X and Y.

Since the empirical likelihood function (Eq. 2) reaches the maximum at the same moment as its logarithm, by taking logarithmic form, we introduce a scale parameter θ and define the log-empirical likelihood ratio statistic as follows

$$R(\Delta) = \sup_{p_i, q_j, i, j} \left\{ \sum_{i=1}^{m} \log(mp_i) + \sum_{j=1}^{n} \log(nq_j) \right\} = \sup_{\theta} R(\Delta, \theta) \tag{3}$$

where

$$R(\Delta, \theta) = \sup_{p_i, q_j} \left\{ \sum_{i=1}^{m} \log(mp_i) + \sum_{j=1}^{n} \log(nq_j) \right\} \tag{4}$$

and p_i, q_j are subject to restrictions:

$$\sum_i p_i \omega_1(x_{I,i}, \theta, \Delta) = 0, \text{ and } \sum_j q_j \omega_2(y_{I,j}, \theta, \Delta) = 0 \tag{5}$$

From Lagrange multipliers, we get the following

$$R(\Delta, \theta) = -\sum_{i=1}^{m} \log\{1 + \lambda_1(\theta)\omega_1(x_{I,i}, \theta, \Delta)\} - \sum_{j=1}^{n} \log\{1 + \lambda_2(\theta)\omega_2(y_{I,j}, \theta, \Delta)\} \tag{6}$$

The empirical likelihood equation could be obtained from the above equations, which is then used to derive the confidence intervals for the group differences Δ.

3.2 Empirical Likelihood (EL) based confidence interval for Δ

The log-empirical likelihood ratio statistic under imputation converges to a weighted Chi-squared distribution[15], which will be used to construct the EL based confidence intervals for Δ.

Let t_α satisfy $P(\chi_1^2 \le t_\alpha) = 1 - \alpha$, where $1 - \alpha$ is the confidence level. An EL based confidence interval on Δ with asymptotically coverage probability $1 - \alpha$ can be constructed as

$$\{\Delta : -2\hat{a}_0^{-1}(\Delta) R(\Delta, \theta_{m,n}) \le t_\alpha\} \tag{7}$$

where $\theta_{m,n}$ is the root of Equation (7).

This result can directly apply to test the hypotheses on Δ. For instance, if the zero hypothesis is $H_0 : \Delta = \Delta_0$ and the alternative is $H_1 : \Delta \neq \Delta_0$, where Δ_0 is a constant. We first construct the confidence interval for $\Delta - \Delta_0$, then check if the sample difference $\Delta' - \Delta_0'$ falls within that confidence interval. If it is true, we accept hypothesis H_0; otherwise, we accept hypothesis H_1.

Note that the result can be applied to the complete data settings (i.e., dataset without missing data). In the complete data situation, i.e., $P(\delta_x = 1\,|x) = \text{Prob}_1 = 1$ and $P(\delta_y = 1\,|y) = \text{Prob}_2 = 1$, we can see that the asymptotic distribution of the EL statistic follows a χ_1^2 distribution. Thus, the EL based confidence interval for Δ is constructed as $\{\Delta : -2R(\Delta, \theta_{m,n}) \le t_\alpha\}$.

3.3 Bootstrap re-sampling based confidence interval for Δ

Bootstrap re-sampling is based on the idea that in the absence of any other prior information about the distribution, and the observed sample contains all the available information about the underlying distribution. On the other hand, if we have only few data in hand due to expensive cost to collect, we can resort to bootstrap re-sampling to generate enough "new" data samples. In order to compare with the EL based method for building CIs, in this paper we only describe how to use the bootstrap methods to construct CIs of structural differences for two groups of data.

Given two groups X and Y (note that the data size of group X may not equal to that of Y), the bootstrap methods is used on each of the group, generating m bootstrap samples for X, say $X_1^*, X_2^*, ..., X_m^*$, and for Y, say $Y_1^*, Y_2^*, ..., Y_m^*$. Then we compute the collection of the mean and distribution function difference estimators for each pair of these m bootstrap samples. We take the mean difference for example. After bootstrap sampling, we get a sequence of differences of mean $\Delta^* = \{\Delta_1^*, \Delta_2^*, ..., \Delta_m^*\}$, where $\Delta_i^* = E(Y_i^*) - E(X_i^*)$. According to the bootstrap re-sampling theory, if Δ^* is approximately normally distributed, we can calculate the $1 - \alpha$ confidence interval for the mean difference Δ, which ranges from $E(\Delta^*) - z_{\alpha/2}Se_{boot}(\Delta^*)$ to $E(\Delta^*) + z_{\alpha/2}Se_{boot}(\Delta^*)$, where z_α is the α critical value of the standard normal distribution, $Se_{boot}(\Delta^*)$ is the standard variance ofΔ^*. However, since making the normal-distributed assumption is contrary to the non-parametric aspect of bootstrap method, instead we can obtain $[\alpha, 1 - \alpha]$ confidence interval (e.g., [0.05, 0.95]) by finding the corresponding quantiles of bootstrapped estimators Δ^* (e.g., the 5^{th} and the 95^{th} values in a sorted list of 100 bootstrap estimators).

To compare the performance of EL based method and Bootstrap method, we use the EL based method to construct CIs on imputed dataset at first, and then utilize the bootstrap method to build CIs on the same dataset.

4 Experimental Study

We have implemented our approaches using MATLAB, and conducted several experiments on real datasets on a DELL Workstation PWS650 with 2G main memory and 2.6GHz CPU. The operating system is WINDOWS 2000.

In order to evaluate the performances of our EL based method in building confidence intervals on real datasets, we designed three kinds of experiments on several datasets extracted from the UCI machine learning repository[5], i.e., one-class experiment, two-class experiment, and multiple-class experiment. The one-class experiment uses data samples that are coming from a same population, i.e., the dataset corresponds to a population. In the two-class experiment, we choose dataset that contains a binary-valued class attribute, which is then divided into two portions based on the class attribute. Each portion is regarded as a population, and data samples are extracted from the two populations. We also consider in the last experiment dataset that contains tuples from multiple classes.

4.1 One-Class experiment

The objective of one-class experiment is to check whether the confidence intervals (CI) constructed are tight enough around the sample differences of groups. The sample difference should approach to zero, since the two groups are coming from a same population. To measure the difference, we construct CI for the difference with a confidence level $1 - \alpha$, where we set $\alpha = 0.05$ in our experiments (Note that $\alpha = 0.05$ is a commonly used parameter in statistical inference, and one may choose other confidence level, other than 0.05, to suit his applications). We use the *abalone* dataset for our one-class experiments, which contains 4177 instances in total and each instance consists of 9 attributes. These 9 attributes give features of abalones, such as *length*, *diameter* and *height*, etc. Some other statistics of *abalone* dataset are listed in Table 1. There are no missing values in *abalone*.

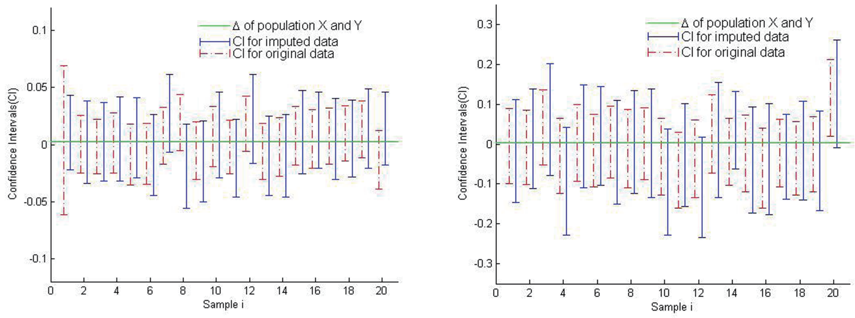
We evenly separate the *abalone* dataset into two parts (groups), denoted as D_1 and D_2, where the two parts have the same size. We then construct the confidence intervals (denoted as CI) for the differences of mean and distribution functions (denoted as DF) of the original complete data in D_1 and D_2. The “incomplete” version of data is generated by using MCAR missing mechanism on the complete data under a missing rate of 20%, and then the parimputation method[30,31] is adopted to impute these “incomplete” data. After imputation, we get the “imputed” data. CIs for the differences of mean and DF are thus built from the imputed data, which are compared with those CIs generated from the original complete data. Theoretically, the difference $\Delta$of the complete data is very close to 0, due to the fact that population D_1 and D_2 have the same mean and distribution (note that D_1 and D_2 are drawn from a same attribute A_i). The CIs built on complete data are compared with those built on imputed data. We repeat this whole process (including random separation, imputation, and CI construction) multiple times (e.g., 50) in order to avoid randomness. Then the result is averaged over these repeated trials.

The experiment results on all the attributes exhibit a similar trend. Thus, for simplicity we only report experiment results on attributes 3, 5 and 6 of the *abalone* dataset, which correspond to the diameter, whole weight, and shucked weight of the abalone, respectively. For each attribute, a fixed percentage, say 10%, of attribute values is randomly sampled from D_1 and D_2, which yields the population X and Y. The processes of constructing CIs on complete and imputed data are the same as described above. For each attribute, we generate 20 random samples from D_1 and D_2 in order to avoid randomness. The constructed CIs for mean and distribution function (DF) on complete and imputed data are presented in Fig. 1. The vertical solid bars denote the CIs on samples from imputed data, whereas the vertical dashed bars are CIs on samples from complete data. The horizontal straight line is the true differenceΔ of the samples from complete data. We can see thatΔ for mean and DF are very close to 0, which conforms the fact that the two populations X and Y are drawn from a same attribute (group), that is, D_1 and D_2 approximately have the same mean and distribution function.

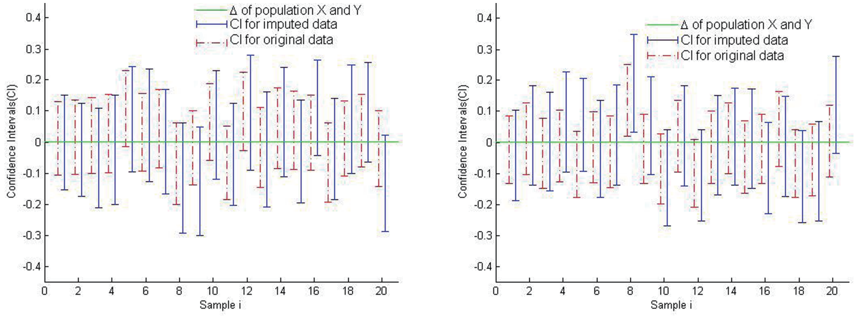
Table 1 Statistics for attributes 3, 5 and 6 of dataset

	Attribute 3	Attribute 5	Attribute 6
Min	0.055	0.002	0.001
Max	0.650	2.826	1.488
Mean	0.408	0.829	0.359

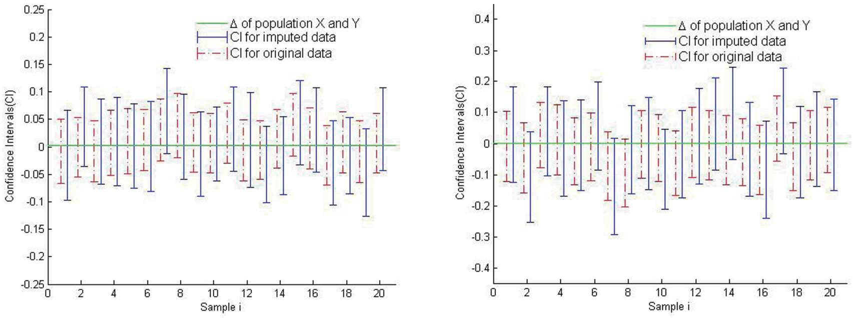
First, we investigate the effect of missing data on the constructed CIs. From Fig. 1(2) and (4) we can see that for the 20^{th} and 8^{th} sample, the CIs (vertical solid bars) for complete data do not include the actual differenceΔ, which means that on the complete data we have a coverage probability of (19/20)*100%=95%. While for CI of sample 20 in Fig. 1(2) on imputed data, its lower endpoint covers the true difference marginally. In contrast, for sample 8 in Fig. 1(4), CI on imputed data does not include the true difference. From all the subfigures above, we can see that the lengths of the EL based CIs on both complete and imputed data are stable around the true differences, and they give us high coverage probability. As expected, CIs on complete data are shorter than CIs on imputed data, because the imputation of missing data may introduce uncertainty or even distort the distribution of the original data. This results in a longer CI. Note that we do not give the results of bootstrap re-sampling method in Fig. 1, for the reason of avoid cluttering the graphs. Instead, we present the detailed experimental results of both EL and bootstrap re-sampling method in Tables 2 and 3.



(1) CIs for mean (attribute 3) (2) CIs for DF (attribute 3, X_0=0.5)



(3) CIs for mean (attribute 5) (4) CIs for DF (attribute 5, X_0=0.5)



(5) CIs for mean (attribute 6) (6) CIs for DF (attribute 6, X_0=0.5)

Figure 1. Comparisons of CIs for mean and DF on the complete and imputed dataset

We compare the EL based and bootstrap re-sampling method in Table 2 and 3, with respect to the average endpoints of CIs and the average length (AL) for mean and DF. We use the abbreviations “Boot” and “EL” for the bootstrap re-sampling method and the EL based method respectively, and “A.3 Ori” and “A.3 Imp” for the complete data and imputed data drawn from attribute 3.

Table 2 Left and Right Endpoints and AL for mean of *abalone*

		LE	RE	AL	CP%
Boot	A.3 Ori A.3 Imp	-0.04463	0.03201	0.07664	100
		-0.03264	0.04705	0.07969	100
EL	A.3 Ori A.3 Imp	-0.02455	0.02996	0.05451	100
		-0.03137	0.03724	0.06861	100
Boot	A.5 Ori A.5 Imp	-0.18264	0.21617	0.39881	100
		-0.2511	0.18025	0.43135	100
EL	A.5 Ori A.5 Imp	-0.10777	0.13994	0.24771	100
		-0.17565	0.16382	0.33947	100
Boot	A.6 Ori A.6 Imp	-0.03623	0.07763	0.11386	95
		-0.08212	0.08857	0.17069	100
EL	A.6 Ori A.6 Imp	-0.04724	0.06413	0.11137	100
		-0.06386	0.07909	0.14295	100

Table 3 Left and Right Endpoints and AL for DF ($\mathbf{X}_0$=0.5) of *abalone*

		LE	RE	AL	CP %
Boot	A.3 Ori A.3Imp	-0.08840	0.10681	0.19521	95
		-0.16142	0.12049	0.28191	100
EL	A.3 Ori A.3 Imp	-0.10371	0.08388	0.18759	95
		-0.13156	0.10676	0.23832	100
Boot	A.5 Ori A.5 Imp	-0.11839	0.12080	0.23919	100
		-0.16970	0.16801	0.33771	100
EL	A.5 Ori A.5 Imp	-0.13295	0.09509	0.22804	95
		-0.15046	0.15084	0.3013	100
Boot	A.6 Ori A.6 Imp	-0.14768	0.08944	0.23712	100
		-0.17232	0.12988	0.3022	100
EL	A.6 Ori A.6 Imp	-0.12886	0.09002	0.21888	100
		-0.15399	0.15015	0.30414	100

From above tables we can see that the average lengths (AL) of derived CIs on imputed data are only slightly longer than that of CIs on complete data. Another observation is that our EL based CIs are generally shorter than those derived by bootstrap re-sampling method, although these two methods both have coverage probabilities that are slightly larger than the pre-specified confidence level 95%.

4.2 Two-Class experiment

Identifying structural differences between samples drawn from two distinct groups is also important in real world applications. It can give us insight into the underlying structural differences of two contrast groups. For example, in medical research on breast cancer, doctors are usually concerned about the differences between the benign and malignant patients. They may ask the questions, such as how large is the mean of tissue radius of a group of benign patients compared with that of the malignant group? How reliable are the differences that we have obtained from these two groups of patients? The solution for these problems is to compute the structural differences of two groups by using statistical methods, and then using EL based method to build confidence intervals for the differences.

We use the *WDBC* (Wisconsin breast cancer) dataset from UCI for our two-group experiment. The *WDBC* contains 569 instances in total and there are 32 features for each instance. Each instance, representing a patient, belongs to either *benign* or *malignant* according to its class label. For simplicity, we only report those expriments on attributes 4, 15 and 27. The reason is that these attributes give the best classification power over the instances[5], and we are interested in how the CIs measure the difference between instances from different classes of a same attribute. We give some statistical information of these features in Table 4, more detailed information about the *WDBC* and its features can be seen in Ref. [5]. First, based on the class attribute we separate *WDBC* into two disjoint portions, one is *benign* group (D_1) which contains 357 instances, and the other *malignant* group (D_2) with 212 instances.

Table 4 Statistics of Attributes 4, 15 and 27 of dataset *WDBC*

	Mean			Distribution function		
	A4	A15	A27	A4 (x_0=15)	A15 (x_0=3)	A27 (x_0=0.1)
Malignant	21.6	4.3239	0.14485	0.0189	0.3392	0.0094
Benign	17.91	2.0003	0.12496	0.2437	0.8907	0.1092
Sample difference Mean	3.69	2.3236	0.01989	-0.2248	-0.5515	-0.0998

We calculate the differences for mean and distribution function for samples drawn from the two groups, then CIs for the differences are constructed. The settings of the two-groups experiment are similar to that of the one-group experiment. At first, for each attribute we generate 20 random samples X's and Y's from D_1 and D_2, respectively, and then the random missing mechanism (MCAR) with 20% missing rate is applied to these samples, which are then imputed by the parimputation method[30,31]. Finally, the EL based method and bootstrap resampling method are utilized to build CIs for structural differences of each sample pair X and Y. The results are given in Tables 5 and 6.

The above tables show that generally the lengths of CIs generated by EL based method are shorter than those derived by bootstrap re-sampling method, which means that with a fixed confidence level, the EL method is preferable to the bootstrap re-sampling method in constructing CIs for group difference. The results also reveal the fact that the *benign* patients are very different from the *malignant* ones with respect to some specific features such as radius, smoothness and perimeter of the breast tumor.

Table 5 Left and Right Endpoints and Average Length (AL) for mean of *WDBC*

		LE	RE	AL	CP %
Boot	A4 Ori	2.6248	4.3506	1.7258	86
	A4 Imp	2.4984	4.6312	2.1328	90
EL	A4 Ori	2.8323	4.2533	1.421	85
	A4 Imp	2.5606	4.3972	1.8366	93
Boot	A15 Ori	1.4788	3.4099	1.9311	92
	A15 Imp	1.2365	3.3782	2.1417	100
EL	A15 Ori	1.5062	3.1558	1.6496	95
	A15 Imp	1.3589	2.9475	1.5886	93
Boot	A27 Ori	-0.2023	0.2167	0.419	100
	A27 Imp	-0.1985	0.2201	0.4186	100
EL	A27 Ori	-0.1972	0.2094	0.4066	100
	A27 Imp	-0.2010	0.2105	0.4115	100

Table 6 Left and Right Endpoints and Average Length (AL) for DF of *WDBC*

		LE	RE	AL	CP %
Boot	A4 Ori	-0.3211	-0.1267	0.1944	94.4
	A4 Imp	-0.3395	-0.1348	0.2047	100
EL	A4 Ori	-0.2996	-0.1156	0.184	95.3
	A4 Imp	-0.3107	-0.1055	0.2052	100
Boot	A15 Ori	-0.6496	-0.4212	0.2284	95.3
	A15 Imp	-0.6388	-0.4035	0.2353	98.5
EL	A15 Ori	-0.6197	-0.4326	0.1871	95
	A15 Imp	-0.6291	-0.4117	0.2174	100
Boot	A27 Ori	-0.2015	0.1909	0.3924	99.6
	A27 Imp	-0.2268	0.2054	0.4322	100
EL	A27 Ori	-0.1890	0.1921	0.3811	100
	A27 Imp	-0.1995	0.2033	0.4028	100

4.3 Multiple-Class experiment

As we have shown in Section 4.1, in real world applications there are some datasets with binary-valued class attribute. However, there are also other problems with multiple-valued class attribute. For instance, in a weather forecasting application, weather can be classified as *sunny*, *cloudy*, *windy* and *rainy*. Thus, we must take into account the problem of building CIs for differences of samples drawn from dataset with multiple-valued class attribute, since samples from different populations (classes) may have distinct statistics such as mean, variance, distribution, etc. We refer this as multiple-class problem. A straightforward way to address the multiple-class problem is to use the methodology in the two-class experiment to construct CIs for data samples from all class pairs. For example, we consider the two populations *sunny* and *cloudy*, or *sunny* and *windy*, where *sunny*, *cloudy* and *windy* are three distinct classes. Although this method can give some information about differences on any permutation of the classes, it is not appropriate to reveal characteristics of the whole dataset. In this paper, we use a simple strategy for our multiple-class experiments.

Specifically, we divide the example weather database DB into two datasets, one is D_1 that contains instances with class value *sunny*, the other is D_2 that contains the rest instances (i.e., with class value *cloudy*, *windy* and *rainy*). Thus we can use our EL based method to construct CIs for differences on D_1 and D_2.

To test our EL based method, we extract several datasets from UCI repository, where each of the datasets contains a class attribute with more than 2 class labels. We conduct experiments on numerical attributes of these datasets, and we found that the results follow a similar trend across different attributes. Since some datasets contain too many attributes, we only select one attribute for each dataset to present in our experiments, which are given in Tables 7 and 8.

Table 7 Experiments on datasets with multiple-valued class attribute for mean

Dataset	Method	LE	RE	AL	CP(%)
Spambase Attr. 24 (2-Classes)	Boot.Ori	0.17980	0.23697	0.05717	95.8
	EL.Ori	0.17879	0.22690	0.04811	96
	Boot.Imp	0.1697	0.2296	0.0599	97.2
	EL.Imp	0.1701	0.2188	0.0487	97.1
Diabetes Attr. 4 (2-Classes)	Boot.Ori	1.4198	3.5641	2.1443	99.4
	EL.Ori	1.5265	3.5713	2.0448	100
	Boot.Imp	1.3588	3.6467	2.2879	100
	EL.Imp	1.4476	3.4745	2.0269	100
Wine Attr. 11 (3-Classes)	Boot.Ori	-3.51207	-2.98109	0.53098	82.5
	EL.Ori	-4.15280	-2.66225	0.89055	100
	Boot.Imp	-3.8012	-2.8916	0.9096	100
	EL.Imp	-4.1029	-2.5688	1.5341	100
Iris Attr. 3 (3-Classes)	Boot.Ori	-3.32600	-2.58400	0.742	80.6
	EL.Ori	-4.05828	-2.31574	1.74254	100
	Boot.Imp	-3.5727	-2.6346	0.9381	85.7
	EL.Imp	-3.8262	-2.4459	1.3803	100
Yeast Attr. 3 (10-Classes)	Boot.Ori	-0.48495	-0.36687	0.11808	88.5
	EL.Ori	-0.46824	-0.30341	0.16483	95.3
	Boot.Imp	-0.5346	-0.3402	0.1944	95
	EL.Imp	-0.5752	-0.3260	0.2492	100

From Table 7, we can see that for dataset *spambase* and *diabetes*, the lengths of EL based CIs for mean are shorter than those derived by bootstrap re-sampling method; whereas for dataset *wine*, *iris*, and *yeast*, EL generates slightly longer CI than that of bootstrap method. However, EL based method always achieves a better coverage probability than bootstrap method. Now we turn to CIs for DF, as presented in Table 8. Although bootstrap re-sampling method derives slighty shorter CIs on *spambase*, *yeast*, and *wine*, the average coverage probability of these CIs is smaller than that generated by EL based method.

Table 8 Experiments on datasets with multiple-valued class attribute for DF

Dataset	Method	LE	RE	AL	CP(%)
Spambase Attr. 24 (2-Classes)	Boot.Ori	0.2414	0.4105	0.1691	89.9
	EL.Ori	0.2388	0.4285	0.1897	93
	Boot.Imp	0.2209	0.4299	0.209	95.4
	EL.Imp	0.2307	0.4356	0.2049	98.7
Diabetes Attr. 4 (2-Classes)	Boot.Ori	0.1040	0.2663	0.1623	100
	EL.Ori	0.1178	0.2602	0.1424	98.5
	Boot.Imp	0.0994	0.2598	0.1604	99.3
	EL.Imp	0.1272	0.2696	0.1424	98.3
Wine Attr. 11 (3-Classes)	Boot.Ori	0.49295	0.56338	0.07043	84.2
	EL.Ori	0.42028	0.60124	0.18096	96.3
	Boot.Imp	0.6012	0.7955	0.1943	100
	EL.Imp	0.4006	0.6255	0.2249	100
Iris Attr. 3 (3-Classes)	Boot.Ori	-0.14999	-0.05000	0.09999	92.4
	EL.Ori	-0.13047	-0.10581	0.02466	89.1
	Boot.Imp	-0.2056	-0.0811	0.1245	100
	EL.Imp	-0.1878	-0.0942	0.0936	94.3
Yeast Attr. 3 (10-Classes)	Boot.Ori	0.16248	0.37997	0.21749	87.6
	EL.Ori	0.17085	0.42803	0.25718	92.0
	Boot.Imp	0.1832	0.3951	0.2119	90.8
	EL.Imp	0.1788	0.4035	0.2247	94.1

4.4 Discussion

As shown in Tables 2, 3, 5 and 6, the empirical likelihood (EL) method is superior to the bootstrap re-sampling method in computing CIs for group difference, both in one-group and two-class experiments. For multiple-class experiments, EL based method generates a slightly longer CIs that of bootstrap re-sampling method. However, EL based method achieves a higher coverage probability than bootstrap. Another observation is that the lengths of CIs on imputed data are slightly longer than the lengths of CIs on complete data, which shows the effectiveness of the parimputation method in dealing with incomplete data, for the purpose of building CI for group differnce.

The implication of the three experiments is that our method, using EL method and parimputation strategy, can robustly construct CIs for group difference either on complete or incomplete dataset, no matter the datasets are coming from a same population or different populations. This means that our method for buildiing CIs for group difference is suitable for a broader range of applications, and as a tool it can help researchers in exploratory data analysis tasks such as medical research, anti-spam email software development, customer behavior analysis, etc.

5 Conclusion

Recognizing the importance of differences between populations (groups), there are many data mining techniques developed for group difference detection in the context of associations[2−4,8,14,21,24], and group interaction detection between complete datasets[29]. In this paper we have incorporated the parimputation strategy and the group interaction approach to mining quality models from incomplete data by identifying the (mean and distribution function) differences between an incomplete dataset and a known dataset, which can be utilized for measuring the quality when one is making inferences on the datasets.

In comparison with the differences of two contrast groups with missing data, we have shown that in most cases the empirical likelihood (EL) based method works better than the bootstrap re-sampling counterpart in building confidence intervals for the mean and distribution function differences. We also showed that this result can directly be used to test the hypotheses on difference Δ.

6 Acknowledgements

This work is supported in part by the Australian Research Council (ARC) under large grant DP0985456, the China "1000-Plan" National Distinguished Professorship, the China 863 Program under grant 2012AA011005, the Nature Science Foundation (NSF) of China under grant 61170131, the Guangxi "Bagui" Teams for Innovation and Research, the Jiangsu Provincial Key Laboratory of E-business at the Nanjing University of Finance and Economics, the the Guangxi Natural Science Foundation for Innovation and Research Teams under grant 2012GXNSFGA060004, and the the Guilin Natural Science Foundation.

References

[1] Au WH, ChanKC. Mining changes in association rules: a fuzzy approach. Fuzzy Sets and Systems, 2005, 149(1): 87–104.

[2] Bay SD, Pazzani MJ. Detecting change in categorical data: mining contrast sets. Procs. of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99). 1999. 302–306.

[3] Bay SD, Pazzani MJ. Characterizing model erros and differences. Procs. of the Seventeenth International Conference on Machine Learning (ICML 2000). 2000. 49–56.

[4] Bay SD, Pazzani MJ. Detecting group differences: mining contrast sets. Data Mining and Knowledge Discovery, 2001, 5(3): 213–246.

[5] Blake C, Merz C. UCI Repository of machine learning database. 1998. `http://www.ics.uci.edu/∼mlearn/MLResoesitory.html`

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

[7] Cho YB, Cho YH, Kim SH. Mining changes in customer buying behavior for collaborative recommendations. Expert Systems with Applications, 2005, 28(2): 359–369.

[8] Cong G, Liu B. Speed-up Iterative Frequent Itemset Mining with Constraint Changes. Procs. of the International Conference on Data Mining (ICDM 2002). 2002.107–114.

[9] Hall P, Martin M. On the bootstrap and two-sample problems. Austral. J. Statist, 1988, 30A: 179–192.

[10] Hartley H, Rao J. A new estimation theory for sample surveys. Biometrika, 1968, 55: 547–557.

[11] Jing BY. Two-sample empirical likelihood method. Statistics and Probability Letters, 1995, 24: 315–319.

[12] Li HF, Lee SY, Shan MK. Online Mining Changes of Items over Continuous Append-only and Dynamic Data Streams. Journal of Universal Computer Science, 2005, 11(8): 1411–1425.

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

[14] Liu B, Hsu W, Han HS, Xia Y. Mining Changes for Real-Life Applications. DaWaK 2000. 2002. 337–346.

[15] Qin YS, Zhang SC. Empirical likelihood confidence intervals for differences between two datasets

with missing data. Pattern Recognition Letters, 2008, 29(6): 803–812.

[16] Owen A. Data squashing by empirical likelihood. Data Mining and Knowledge Discovery, 2003, 7(1): 101–113.

[17] Owen A. Empirical Likelihood. Chapman & Hall, New York. 2001.

[18] Pyle D. Data Preparation for Data Mining. Morgan Kaufmann, 1999.

[19] Qin YS, Zhang SC, Zhu XF, Zhang JL, Zhang CQ. Estimating confidence intervals for structural differences between contrast groups with missing data. Expert Systems With Applications, 2009, 36(3): 6431–6438.

[20] Rao J. On variance estimation with imputed survey data. J. Amer. Statist. Assoc., 1996, 91: 499–520.

[21] Wang K, Zhou SQ, Fu AWC, Yu XJ. Mining changes of classification by correspondence tracing. SDM'03, SIAM International Conference on Data Mining. May 1–3, 2003. San Francisco.

[22] Wang Q, Rao J. Empirical likelihood-based inference in linear models with missing data. Scand. J. Statist., 2002a, 29: 563–576.

[23] Wang Q, Rao J. Empirical likelihood-based inference under imputation for missing response data. Ann. Statist., 2002b, 30: 896–924.

[24] Webb GI, Butler SM, Newlands DA. On detecting differences between groups. Proc. of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'03. 2003. 256–265.

[25] Ying AT, Murphy GC, Raymond TN, Mark CC. Predicting source code changes by mining change history. IEEE Trans. Software Eng., 2004, 30(9): 574–586.

[26] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological Methods, 2002, 7(2): 147–177.

[27] Zhang JL, Zhang SC, Zhu XF, Wu XD, Zhang CQ. Measuring the uncertainty of differences for contrasting groups. Proc. of 22nd National Conference on Artificial Intelligence (AAAI-07). 2007. 1920–1921.

[28] Zhang SC. Shell-Neighbor Method and Its Application in Missing Data Imputation. Applied Intelligence, 2011, 35(1): 123–133.

[29] Zhang SC. Detecting Differences between Contrast Groups. IEEE Trans. on Information Technology in Biomedicine, 2008a, 12(6): 739–745.

[30] Zhang SC. Parimputation: From imputation and null-imputation to partially imputation. IEEE Intelligent Informatics Bulletin, 2008b, 9(1): 32–38.

[31] Zhang SC, Jin Z, Zhu XF. Missing data imputation by utilizing information within incomplete instances. Journal of Systems & Software, 2011, 84(3): 452–459.

[32] Zhu XF, Zhang SC, Jin Z, Zhang ZL, Xu ZM. Missing value estimation for mixed-attribute datasets. IEEE Trans. on Knowledge and Data Engineering, 2011, 23(1): 110–121.