Multiple Data Source Discovery with Group Interaction Approach

Wu Hao
Liuzhou Railway VocationalTechnology College
Liuzhou, China, 545007
E-mail: lywh88@126.com

***Abstract*—Medical researchers seek to identify and predict profit (or effectiveness) potential in a new medicine B against a specified disease by comparing it to an existing medicine A, which has been used to treat the disease for many years, called medicine assessment. Applying traditional data mining techniques to the medicine assessment, one can discover patterns, such as A.X=a → B.Y=b, which are identified at the attribute-value level. These patterns are useful in predicting associated behaviors at the attribute-value level. However, to evaluate B against A, we have to obtain globally useful relations between B and A at an attribute level. Therefore, this paper proposes a group interaction approach for multiple data source discovery. Group interactions include, such as rules, differences, and links between datasets. These group interactions are discovered at the attribute level. For example, R(A.X, B.Y), where R is a relationship, or a predication. Some examples are presented for illustrating the use of the group interaction approach.**

Keywords-Data mining; multiple data source mining; interaction; difference detection

I. INTRODUCTION

How to efficiently discover useful patterns from multiple data sources (MDS) it is an important and challenging research topic in data mining and machine learning. The discovery of useful patterns from MDS 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. MDSs are of heterogeneity, incomplete, and large in size which could make the discovery of useful patterns difficult, expensive, and perhaps even impossible to implement. It is well known that data mining tools greatly enhance the ability of an analyst to make data-driven discoveries. The usage of MDSs leads to new data mining system architectures, especially in the field of distributed systems.

There is a significant unmet need to identify patterns in incomplete data from different sources. For example, medical researchers seek to identify the effectiveness of a new medicine B, recognizing its use for a specific disease. The data for applying B to the disease can be incomplete because (1) tests are expensive or impossible, and (2) applications are at risk. Knowledge discovery from incomplete data is relevant in such area as new product evaluation, disease treatment, and fraud detection. A natural solution is to compare B with an existing medicine A, which has been used to treat the disease for many years. This allows the application of a pre-existing conceptual structure to new problems and domains, and hence supports the rapid learning of new systems. In this kind of incomplete data discovery, similarity plays a central role in extant machine-learning approaches. While in human acquisition of knowledge by analogy, difference detection for two contrasted instances is a further underlying measure. Therefore, this research will design computational techniques to address the key problem of efficiently mining incomplete data.

The principal research aim of this paper is designing a group interaction approach for MDS discovery.

Influenced by traditional data mining applications, existing methods to discover group patterns between datasets are focused on association-rule like patterns with two measures: support and confidence [1, 14]. These patterns are identified at the attribute-value level. For example, A(X=a) → B(Y=b) is a pattern at the attribute-value level, where A and B are datasets, X and Y are two attributes. This pattern is useful in predicting some change in behaviors. Group interactions proposed in this paper will include the differences, links, and bridging rules between datasets. These group interactions are discovered at the attribute level. For example, R(A.X, B.Y), where R is a relationship, or a predication.

For simplicity, the group interaction discussed in this paper is the difference at attribute level. In the following descriptions, group interactions and differences are used as two interchangeable concepts.

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. 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 a confidence interval estimation for Group Differences. Conclusion is given in Section 5.

II. RELATED WORK

A. Research into multiple mata source discovery

With massive amounts of data being collected by many businesses, government agencies and research projects, research and development of new techniques are of increasing importance and are the focus of many data mining

Published by Atlantis Press, Paris, France.

projects [2]. These techniques will enable efficient and automatic sharing of large databases between organizations. Therefore, there are three main research directions on multiple data source discovery.

The first direction is to utilize mono-database mining techniques after pooling all the data from multiple data sources to create a huge dataset. The disadvantages are the complexity of the resulting data and loss of some useful patterns, (e.g. the pattern that 70% of the brokers agree that people with lower education background like trading stocks through brokers [15]).

The second direction is similar to the first. It selects the relevant data sources based on the specific application, and then puts them together to mine the knowledge by using mono-database mining techniques [9]. The disadvantages of this approach are its application dependence and the need for multiple scans for each application.

The third direction is called local pattern analysis [17]. The idea of local pattern analysis is to firstly cluster data sources, then mine the knowledge from individual relevant data sources, and finally integrate the knowledge from the data sources. Compared to the other two approaches, local pattern analysis can overcome their disadvantages. It is low complexity because it only mines relevant individual data sources. It is application-independent which means the data source clustering can be used without any specific application. It is able to find special patterns (such as the pattern in above example) that cannot be found by the above two methods.

All of the above approaches are based on the assumption that the data to be mined is of high quality. And they focus only on identifying useful patterns at the attribute-value level. This paper studies the problem of knowledge discovery from data sources at the attribute level.

B. Research into Identifying Group Interactions

In intelligent data analysis, detecting (or comparing) group differences is a central issue in many domains. For example, in a medical research proposal, it is useful to compare the mean value of prolonging patient's life between a group applying a new product (e.g., medicine) and one applying an existing or alternate product. Identifying group differences between spam and non-spam emails can be distinguished in anti-spam email applications. Furthermore, software companies can devise well-performed anti-spam email systems based on these differences. Therefore, there are some research reports on mining group differences between contrast groups from observational multivariate data [1, 14].

Detecting group differences is also very important in social science research. For example, the Integrated Public Use Microdata Series (IPUMS) project [12] has expended great effort standardizing federal census data to allow researchers to compare demographic groups over different time periods. Some of the research conducted with this data involves comparing different racial groups [4] or examining trends in divorce rates [12]. As another example, the Department of Urban and Regional Planning at UCI conducts an annual survey of people in Orange County. The goal is to compare "the quality of life and local government ratings in Orange County with Los Angeles County" and to "analyse the impact of changing demographics by contrasting survey responses of Latinos, Asians, and non-Hispanic whites".

Defined by Bay and Pazzani [1] contrast set discovery seeks to find all contrast sets whose support differs meaningfully across groups. This is defined as seeking all contrast sets cset that satisfy both

$$\exists ij P(cset|Gi) \neq P(cset|Gj) \quad (1)$$

and

$$\mathrm{Max}_{ij}|\mathrm{support}(cset,Gi)-\mathrm{support}(cset,Gj)| \geq a \quad (2)$$

Where Gi and Gj are two distinct groups; a is a user-defined threshold called the minimum support-difference. Contrast sets, for which Eq. 1 is statistically supported, are called significant and those for which Eq. 2 is satisfied are called large.

Another kind of related work is change mining [3]. In the change mining problem, there is 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 identifying customer buying behaviour, association rules and items over continuous append-only and dynamic data streams, and predicting source code changes [8].

In contrast to the above work, my approaches in this paper take into account the structure of a group (nonparametric); imputation of missing data when contrast groups have them; and confidence intervals for the mean and distribution differences between the two groups. As a result we will discover useful patterns from data sources at the attribute level. Specifically, we will use F and G to denote the distribution functions of groups X and Y, respectively; we will construct confidence intervals for the mean and distribution differences between contrast groups X and Y using an empirical likelihood (EL) method. Compared to extant methods [7, 10], we do not specify the exact distribution forms of X and Y because, in practical applications, ones usually have no prior knowledge about the underlying distribution of the data that are being processed, instead we adopt empirical distributions of X and Y.

III. GROUP INTERACTION APPROACH

Use F and G to denote the distribution functions of groups X and Y, respectively. 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 populations. Making inference for the mean difference is the well-known Behrens-Fisher problem if F and G are both normally distributed. In general, both F and G are unknown so that nonparametric methods are developed to address this situation. In the case of complete observations, related work can be found in (Hall and Martin 1988; Jing 1995; Qin and Zhao 2000).

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

$$E\omega_1(x,\theta_0,\Delta)=0, E\omega_2(y,\theta_0,\Delta)=0 \qquad (2)$$

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: Denote $\theta_0 = Ex$, $\theta_1 = Ey$ and $\Delta = \theta_1 - \theta_0$, so we can define $\omega_1(x,\theta_0,\Delta) = x - \theta_0$, $\omega_2(y,\theta_0,\Delta) = y - \theta_0 - \Delta$.

Difference of distribution function: For fixed x_0 , denote $\theta_0 = F(x_0),\ \theta_1 = G(x_0)$ and $\Delta = \theta_1 - \theta_0$, we can also define $\omega_1(x,\theta_0,\Delta) = I(x \le x_0) - \theta_0,\ \omega_2(y,\theta_0,\Delta) = I(y \le x_0) - \theta_0 - \Delta$. Where *I(.)* is indicator function, *I(x)*=1 if *x* is true, otherwise *I(x)*=0.

It may be interesting to test the mean difference Δ of X and Y. To do so, we can first construct the confidence interval for Δ . If Δ is in the generated interval, we accept the hypothesis Δ ; otherwise we reject this hypothesis. In this paper, we construct confidence interval based on EL method to solve the two nonparametric population problems.

IV. Confidence Interval for Group Differences

Existing techniques for difference detection and change mining both individually and collectively participate in the goal of association analysis. Distinguishing from them, in this paper we propose an efficient approach for measuring this uncertainty by identifying the confidence intervals of structural differences between contrast groups. Specifically, for a pre-assigned confidence level (in this article, the confidence level is 1- α , and we take α =0.05), the confidence interval would contain the parameter of interest (the differences of the means and distribution functions of two contrast groups in this article) with probability not smaller than the prescribed confidence level 1-α , which is more precise than the point estimate (a single value) of the parameter (as the point estimate does not tell us how far it is away from the true parameter value (as we do not know the true value)). On the other hand, the result of confidence interval can directly apply to test the hypotheses on the parameter of interest. For instance, if the hypothesis is $H : \theta = \theta_0$, we first construct the confidence interval on θ . Then check whether θ_0 is in the interval or not. If θ_0 is in the interval, we accept the hypothesis; otherwise, the hypothesis should be rejected.

From the statistical perspective, the mean and distribution function are very important for characterizing the data, and one can almost have a full understanding of the data if he knows the mean and distribution function exactly. We can use statistical methods to obtain the above differences. For instance, one can use statistical methods to obtain the differences between contrast groups. 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 $\Delta = G_Y(\alpha) - F_X(\alpha)$ to calculate it, where GY and FX 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 an empirical form is adopted in practice, i.e., $\widehat{G_Y(a)} = \frac{1}{m}\sum_{j=1}^{m} I(y_j \le \alpha)$, $\widehat{F_X(a)} = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le \alpha)$, where I(.) is an indicator function, and I(x<y)=1 if x<y, otherwise I(x<y)=0. This is called a non-parametric model. If we know the form of GY or FX in advance, we call it a 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 intervals in order to better measure their uncertainties. The confidence interval (CI) can tell people how reliable the derived differences for two groups X and Y are.

This paper focuses on applying the non-parametric model to identify how reliable are the differences of mean or distribution of two data groups, X and Y, because this information is useful for decision-makers to make decisions or predictions. Our approach is designed significantly against most of those applications that we do not know their exact data distribution, or, in particular, the data with missing values. We experimentally evaluate our 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.

V. Conclusion

We have proposed a strategy for identifying confidence intervals for the mean and distribution function differences between two contrast groups, which can be utilized for measuring the uncertainties when one is making inferences on the groups. In comparison with the differences of two contrast groups with missing data, we have shown that the EL-based method works well in building confidence intervals for the mean and distribution function differences. We have also shown that this result can directly be used to test the hypotheses on Δ , and that the result can apply to the complete data settings.

VI. ACKNOWLEDGEMENT

This work was supported in part by China NSF (61170131) and Guangxi NSF (2012GXNSFGA060004).

REFERENCES

[1] Au, W.H. and Chan, K.C. (2005). Mining changes in association rules: a fuzzy approach. Fuzzy Sets and Systems, 149(1): 87-104.

[2] Bay, S. D. and Pazzani, M. J. (1999). 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.

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

[4] Bay, S. D. and Pazzani, M. J. (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/MLResoesitory.html].

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

[7] Cho, Y. B., Cho, Y. H. and Kim, S. H. (2005). Mining changes in customer buying behavior for collaborative recommendations. Expert Systems with Applications, 28(2): 359-369.

[8] Cong, G. and Liu, B. (2002). Speed-up Iterative Frequent Itemset Mining with Constraint Changes. In: Proceedings of the International Conference on Data Mining (ICDM 2002), pp 107-114.

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

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

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

[12] Li, H. F., Lee, S. Y. and Shan, M. K. (2005). Online Mining Changes of Items over Continuous Append-only and Dynamic Data Streams. Journal of Universal Computer Science, 11(8): 1411-1425.

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

[14] Liu, B., Hsu, W., Han, H. S. and Xia, Y. (2002). Mining Changes for Real-Life Applications. DaWaK 2000, pp337-346.

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

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

[17] Owen, A. (2001). Empirical likelihood. Chapman & Hall, New York.

[18] Rao, J. (1996). On variance estimation with imputed survey data. J. Amer. Statist. Assoc., 91: 499-520.

[19] Wang, K., Zhou, S. Q., Fu, A. W. C. and Yu, X. J. (2003). Mining Changes of Classification by Correspondence Tracing. In: SIAMDM'03, SIAM International Conference on Data Mining, May 1-3, San Francisco.

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

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

[22] Webb, G. I., Butler, S.M. and Newlands, D.A. (2003). On detecting differences between groups. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'03, pp.256-265.

[23] Ying, A. T., Murphy, G. C., Raymond, T. N. and Mark, C. C. (2004). Predicting Source Code Changes by Mining Change History. IEEE Trans. Software Eng., 30(9): 574-586.