行政院國家科學委員會專題研究計畫 成果報告

運用隨機森林分類方法在檢定基因集的顯著性
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 100-2118-M-004-004-
執 行 期 間 ：100 年 08 月 01 日至 101 年 07 月 31 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人：薛慧敏
共 同 主 持 人：蔡政安
計畫參與人員：博士班研究生-兼任助理人員：許嫚荏

報 告 附 件 ：出席國際會議研究心得報告及發表論文

公 開 資 訊 ：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中 華 民 國 101 年 08 月 24 日

中 文 摘 要 ： 近年來在基因微陣列(microarray)實驗中，越來越多的研究人員將研究目的由檢定個別基因與外顯表現變數(phenotype)的相關性，擴展到檢定特定基因集合(gene-set)的顯著性。研究人員依據基因之生物功能將基因歸類，目前已有多個公開資料庫提供基因組相關資訊。基因集合的顯著性檢定可分為兩類，第一類稱為競爭性檢定(competitive test)，主要目的為檢定一特定基因集合在相較於其他的基因集合下，有特別顯著的表現。第二類則稱為自足的檢定(self-contained test)，主要在檢定此特定基因集合是否有顯著表現。在這個研究中，我們將建立依據基因集合的分類器，並以此分類器的預測誤差率來評估此集合與外顯變數的相關性，我們將利用隨機森林(random forest)來建立分類器。由於此二個檢定的虛無假設不同，故其虛無分配也不同，我們在研究中也將探討各檢定的P值的計算方式。本方法將被應用在實際資料上以與其他方法作比較，另外也透過電腦模擬實驗來驗證本方法的有效性。

中文關鍵詞： 基因集合分析，競爭性檢定，自足性檢定，隨機森林，顯著值

英 文 摘 要 ： In DNA microarray studies, a gene-set analysis (GSA) is used to evaluate the association between the expression of biological pathways, or a priori defined gene sets, and a particular phenotype. Two types of differentially expressed testing are of research interest: the competitive testing and the self-contained testing. The competitive test is to determine whether the specific gene set is relatively differentially expressed when compared to other gene sets. The self-contained test is interested in finding whether the gene set alone is differentially expressed. The two tests involve different null distributions. To take consideration on the interaction or correlation within the gene set, we consider assessing the significance of the gene set by the performance of a classifier developed upon the gene set. In this study, the Random Forest classification is applied. For each of the two tests, the corresponding empirical P-value of an observed out-of-bag (OOB) error rate of the classifier is introduced by using adequate resampling method. Several real examples are analyzed for comparison. A

simulation study is conducted for verification.

英文關鍵詞： Gene set analysis, competitive test, self-contained test, random forest, p-value

Assessing the Significance of a Gene Set

Huey-Miin Hsueh[1], Chen-An Tsai[2] and Da-Wei Zhou[1]

[1]Department of Statistics, National Chengchi University, Taiwan

[2]Graduate Institute of Biostatistics & Biostatistics Center. China Medical University, Taiwan

Abstract

In DNA microarray studies, a gene-set analysis (GSA) is used to evaluate the association between the expression of biological pathways, or a priori defined gene sets, and a particular phenotype. Two types of differentially expressed testing are of research interest: the competitive testing and the self-contained testing. The competitive test is to determine whether the specific gene set is relatively differentially expressed when compared to other gene sets. The self-contained test is interested in finding whether the gene set alone is differentially expressed. The two tests involve different null distributions. To take consideration on the interaction or correlation within the gene set, we consider assessing the significance of the gene set by the performance of a classifier developed upon the gene set. In this study, the Random Forest classification is applied. For each of the two tests, the corresponding empirical P-value of an observed out-of-bag (OOB) error rate of the classifier is introduced by using adequate resampling method. Several real examples are analyzed for comparison. A simulation study is conducted for verification.

1. Introduction

In DNA microarray studies, a gene-set analysis (GSA) is used to evaluate the association between the expression of biological pathways, or a priori defined gene sets, and a particular phenotype. Genes that serves a molecular function, a biological process, or a cellular component are annotated to the same term and grouped together into sets. The annotation terms can be obtained from public-domain web-libraries such as Gene Ontology (GO) or KEGG, BioCarta and Broad Institute. See Pang et al. (2006) and Delongchamp (2006).

Many statistical methods are proposed for the gene-set data analysis in literatures. These

existing approaches not only are distinct in the test statistic of use, but also are found different in the null hypothesis and hence the problem of research interest. Tian et al. (2005) and Goeman and Bühlmann (2007) summarize the methods into two types of methods: the competitive and self-contained tests. The null hypothesis of a competitive test is the specific gene set is not differentially expressed when compared to other gene sets. The method involves not only the gene set of research interest but also the full data set. The sampling unit in construction of the null distribution for calculating a P-value is gene. A positive finding is obtained only when the gene set has more significant association with the phenol-type variable than other gene sets. On the other hand, the self-contained test is interested in determining whether the gene set is differentially expressed. The analysis is taken with respect to the specific gene set data alone. To obtain a P-value, the re-sampling unit is the sample as the conventional approach. A thorough review can be found in Goeman and Bühlmann (2007). In this study, the emphasis will be placed on the self-contained problem. A new statistical testing procedure will be proposed. Later more discussions will be provided in the article.

So far, many existing gene-set analyses are based on summarizing the results of testing the significance of individual genes. However, it's difficult to take consideration on the interaction or correlation between genes by using this kind of single gene-based method. The other multiple gene-based methods utilize mostly the information of mean difference. Pang et al. (2006) proposed to build a classifier based on the gene set and use the performance of the classifier to assess the importance of a gene set. They applied the Random Forest classification in their article and show that this method is adequate in terms of producing a better test-set error rate estimate. Nevertheless, their approach only allows investigator to rank gene sets, but does not provide conclusions on the significance of a specific gene set. Recently, Ojala and Garriga (2010) introduce the permutation tests for evaluating a classifier. In this paper, we will apply the idea of Ojala and Garriga (2010) to determine the significance of the Random Forest classification build based on a gene-set.

In this paper, we will develop the self-contained statistical testing procedures to test whether the gene set is differentially expressed. The proposed methods will be applied on five real

examples from Pang et al. (2006). For comparisons with existing approaches, a simulation will be designed and conducted to verify the validity of the proposed method as well.

2. Method

Consider a microarray study of m genes of size n with k phenotypes. Assume there are I gene sets of interest, $S_1,\ldots,S_I$. The set S_i includes m_i genes. There are two types of hypotheses in the gene set analysis: competitive test and self-contained test. While the null hypothesis of a self-contained test of the gene set S_i is

$$H_0^S: \textit{The gene set } S_i \textit{ is not differentially expressed;}$$

the null hypothesis of a competitive test is

$$H_0^C: \textit{The gene set } S_i \textit{ is at most as differentially expressed as other gene sets}.$$

See .Basically, a self-contained test reveals a marginal association of the gene set on the phenotype, but a competitive test seeks for the conclusion with some baseline adjustment. However, the definition of a competitive test is not clear. The null hypothesis given above indicates that there exists at least one other gene set that is more differentially expressed than the specific gene set. Consequently, one will not obtain a significant result unless the specific gene set has the strongest association with the phenotype. In other words, the aim is to determine whether the specific gene set is the best among all possible sets. It is a too stringent requirement. A more reasonable alternative hypothesis may be that the specific gene set is belonging to the top group, which includes the gene sets with the highest association with the phenotype. However, since there is an enormous amount of possible gene sets (2^k-1), the hypothesis testing is a difficult task. This study only focuses on the self-contained test.

To construct a testing procedure, the determination of a testing procedure is essential and important. The existing methods commonly use conventional statistical statistics, such as two-samples-t-test, or F-test in an ANOVA table. Only the mean differences across different pheno-type groups are taken into account. Here to fully utilize the information of multiple genes of a gene set, a complex classifier is build and its testing error rate is of interest. The lower the error, the more the evidence shows to support the significance of the gene set. Hence the testing

error can serve as a test statistic and the correspondent p-value can be used for a statistical conclusion.

We consider using the Random Forests for classification. The Random Forest is constituted by many classification trees, which every one of them is constructed by a bootstrap subset of the original dataset. The samples, which are not selected, are used for calculating a classification error rate of this tree. Further, this partial data set is called the out-of-bag (OOB) data. An overall OOB error rate is given when the specified numbers of trees are added to the forest. The OOB error rate is used to evaluate a gene-set. A gene-set with lower OOB error rate is regarded to have a better predicting power to the phenotype variable and hence has greater significance. Since the Random Forests takes numerous bootstrappings in building a forest, the cross-validation, usually required for a test-set error rate, is unnecessary. The OOB error rate provides an unbiased estimate of the test set error rate. On the other hand, applying classification trees makes the method time-efficient.

Given an observed error rate e_0 of a classifier, a permutation-based P-value can be obtained as following,

$$\text{P} - \text{value} = \frac{\sum_{k=1}^{N} I\{e^{(k)} \le e_0\}}{N}, \qquad (1)$$

where $e^{(k)}$ is the error rate of the classifier based on the k-th permutation null sample and N is the number of permutations. See Ojala and Garriga (2010). One draws a significant conclusion if p-value is less than or equal to the predetermined significance level α. The way of permutation depends on the null hypothesis to test. For a self-contained test of the gene set S_i, the null distribution is produced by shuffling the phenotype labels of the n samples in each run. Along with the partial data set, which consists of only the gene set, the randomized labels are used for a random forest classification and an OOB error rate. Repeat the process for N runs, the P-value against H_0^S can be found by (1) and the statistical conclusion can be drawn.

3. Results

3.1 Real examples

The proposed methods are applied on five examples from

http://bioinformatics.med.yale.edu/pathway-analysis/rf.htm, Breast dataset (Farmer *et al*., 2005) and P53 data set. 49 patients of Breast dataset were classified into three tumor types: 6 of them were apocrine type, 16 were basal and 27 were luminal. Over twenty thousands of gene expressions are provided. In P53 example, the data are regarding 50 NCI-60 cells, where 17 of them were P53+ and 33 were P53 mutant. See Table 1 for the data summary. The pathways are determined either by KEGG, BioCarta or manually. In applying our proposed tests, the tree size of the random forest is 50,000 and the permutation size for the null distribution is 2,000. The results of the top 10 significant pathways of the two examples are presented in Table 2-5.

Table 1. Data summary

Dataset	n	m	Pathway	Phenotype
Breast	49	22215	444	Three tumor
P53	50	12625	440	p53+/p53 mutant

Table 2. The results of the top 10 pathways in testing self-contained hypothesis in Breast data.

Pathway	Gene set size	Observed Error rate	Self-contained	
			Null Error rate mean (std)	p-value (α=0.05)
BC-Regulation of BAD phosphorylation	24	0.0816	0.5098(0.0582)	<0.0005
BC-GATA3 participate in activating the Th2 cytokine genes expression	21	0.0816	0.5147(0.0622)	<0.0005
BC-CARM1 and Regulation of the Estrogen Receptor	24	0.0816	0.5088(0.0584)	<0.0005
Jak-STAT signaling pathway	71	0.0816	0.5035(0.0519	<0.0005
Fructose and mannose metabolism	39	0.0816	0.5052(0.0572)	<0.0005
Glycolysis-Gluconeogenesis	68	0.0816	0.5018(0.0556)	<0.0005
Carbon fixation	25	0.1020	0.5175(0.0628)	<0.0005
Estrogen-response	10	0.1020	0.5244(0.0669)	<0.0005
Downregulated of MTA-3 in ER-negative Breast Tumors	19	0.1020	0.5203(0.0639)	<0.0005
Pentose phosphate pathway	22	0.1020	0.5118(0.0621	<0.0005

Table 3. The results of the top 20 pathways in testing competitive hypothesis in Breast data

Pathway	Gene set size	Observed Error rate	Competitive	
			Null Error rate mean(std)	p-value (α=0.05)
BC-Regulation of BAD phosphorylation	24	0.0816	0.5084(0.0578)	<0.0005
BC-GATA3 participate in activating the Th2 cytokine genes expression	21	0.0816	0.2471(0.0735)	0.0040
BC-CARM1 and Regulation of the Estrogen Receptor	24	0.0816	0.2385(0.0708)	0.0045
Jak-STAT signaling pathway	71	0.0816	0.1702(0.0391	0.0070
Fructose and mannose metabolism	39	0.0816	0.1973(0.0531)	0.0125
Glycolysis-Gluconeogenesis	68	0.0816	0.1695(0.0405)	0.0145
Carbon fixation	25	0.1020	0.2332(0.0678)	0.0045
Estrogen-response	10	0.1020	0.3185(0.0928)	0.0060
Downregulated of MTA-3 in ER-negative Breast Tumors	19	0.1020	0.2532(0.0730)	0.0070
Pentose phosphate pathway	22	0.1020	0.2411(0.0717	0.0120

Table 4. The results of the top 10 pathways in testing self-contained hypothesis in P53 data.

Pathway	Genes Set size	Observed Error rate	Self-contained	
			Null Error rate mean (std)	p-value (α=0.05)
SA_G1_AND_S_PHASES	24	0.12	0.3894(0.0468)	<0.0005
g2 Pathway	44	0.14	0.3815(0.0425)	<0.0005
SA PROGRAMMED CELL DEATH	24	0.14	0.3883(0.0470)	<0.0005
Mitochondria pathway	33	0.16	0.3821(0.0436)	<0.0005
DNA_DAMAGE_SIGNALLING	49	0.16	0.3698(0.0359)	<0.0005
p53hypoxiaPathway	40	0.16	0.3814(0.0441)	<0.0005
bad pathway	41	0.18	0.3785(0.0412)	<0.0005
bcl2family and reg. network	59	0.18	0.3777(0.0426)	0.0005
p53Pathway	40	0.18	0.3843(0.0459)	0.0005
chemical pathway	44	0.20	0.3800(0.0423)	<0.0005

Table 5. The results of the top 10 pathways in testing competitive hypothesis in P53 data.

Pathway	Gene Set size	Observed Error rate	競爭型	
			Null Error rate mean(std)	p-value (α=0.05)
SA_G1_AND_S_PHASES	24	0.12	0.3538(0.0483)	<0.0005
g2 Pathway	44	0.14	0.3413(0.0424)	<0.0005
SA PROGRAMMED CELL DEATH	24	0.14	0.3551(0.0490)	<0.0005
Mitochondria pathway	33	0.16	0.3489(0.0359)	<0.0005
DNA_DAMAGE_SIGNALLING	49	0.16	0.3244(0.0320)	<0.0005
p53hypoxiaPathway	40	0.16	0.3427(0.0449)	0.0005
bad pathway	41	0.18	0.3438(0.0432)	0.0020
bcl2family and reg. network	59	0.18	0.3354(0.0381)	0.0010
p53Pathway	40	0.18	0.3431(0.0446)	0.0020
P53_signalling	153	0.20	0.3242(0.0327)	<0.0005

3.2 Simulation

The simulation study refers to Liu et al. (2007). Consider m=100 gene sets and one binary phenotype, says normal and diseased. In each phenotype group, there are 10 samples. The gene expressions of a sample are generated from a multivariate normal distribution. First, the means of the 100 genes in the normal group are iid generated from uniform (0,10). After that, the means of the 100 genes in the diseased group are consecutively determined by adding 2r in the first 20 genes; subtracting 2r in the next 20 genes, and staying the same in the remaining 60 genes, where r ranges from zero to 1.2. The two groups have the same covariance matrix, where the variances of the 100 genes are iid from uniform (0.1,10). The pairwise correlation coefficient between the 100 genes has the following form:

$$\rho_{jl} = \begin{cases} \rho, 1 \le j \ne l \le 20 \\ \rho, 21 \le j \ne l \le 40 \\ 0, 41 \le j \ne l \le 100 \end{cases},$$

where ρ=0, 0.3, 0.5 and 0.9. In the following, the random forest tree size is 50, 000 and the permutation size is 1000. The simulation size is 1000. See Table 6 for the empirical type I error rates of 8 proposed self-contained methods and Figure 1 for the empirical powers at various ρ. From these results, we

conclude that our method outperforms the existing procedures.

Table 6. The type I error rate of the self-contained test.

Method	ρ=0	ρ=0.3	ρ=0.5	ρ=0.9
Hotelling's T^2	0.050	0.039	0.038	0.050
PCA	0.053	0.042	0.052	0.062
SAM-GS	0.046	0.042	0.038	0.055
ANCOVA	0.042	0.038	0.034	0.052
Global	0.001	0.009	0.016	0.034
GSEA	0.059	0.058	0.052	0.048
MaxMean	0.093	0.094	0.107	0.098
Random Forests	0.040	0.034	0.027	0.036

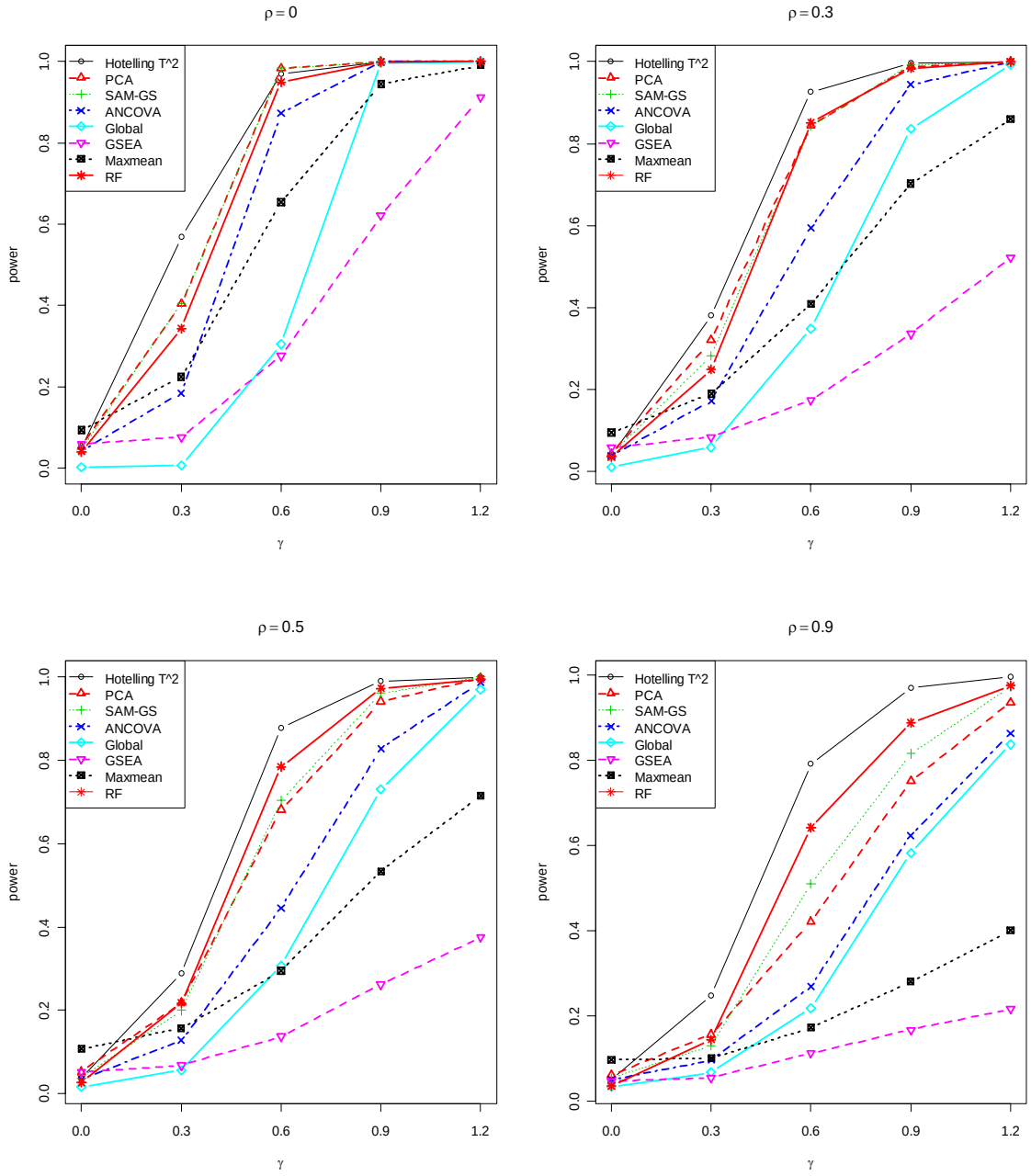

Figure 1. The power function of the 8 methods at various ρ.

4. Conclusion

We find that the proposed methods have satisfactory performance in controlling the type I error rate and in power via simulation. Moreover, real data applications show that the proposed methods are feasible in identifying the significance of gene set. However, the methods request a moderate amount of computations, which becomes less burdensome due to the great advance in recent technology.

REFERENCE

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., Floyd, E. and Zhao, H. (2006) Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**, 2028-2036.

Delongchamp, R., Lee, T., and Velasco, C. (2006) A method for computing the overall statistical significance of a treatment effect among a group of genes. *BMC Bioinformatics*, **7**: S11.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S. and Park, P. J. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci*., **102**, 13544-13549.

Goeman, J. J. and Bühlmann, P. (2007) Analyzing gene expression data in terms of gene sets methodological issues. *Bioinformatics*, **23**, 980-987.

Ojala, M. and Garriga, G. C.(2010) Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, **11**, 1833-1863.

Chen, J. J., Lee, T., Delongchamp, R.R, Chen, T. and Tsai, C. A. (2007) Significance analysis of groups of genes in expression profiling studies. *Bioinformatics*, **23**, 2104-2112.

國科會補助專題研究計畫項下出席國際學術會議心得報告

日期：101 年 2 月 28 日

計畫編號	NSC 100－2118－M－004－004－		
計畫名稱	運用隨機森林分類方法在檢定基因集的顯著性		
出國人員姓名	薛慧敏	服務機構及職稱	政治大學統計系
會議時間	2012 年 2 月 2 日 至 2012 年 2 月 3 日	會議地點	韓國首爾市國立首爾大學
會議名稱	(中文)2012 年東亞區域生物計量會議 (英文) East Asia Regional Biometric Conference 2012		
發表論文題目	(中文)使得 ROC 曲線部份線下面積最大化之生物標記最大線性組合 (英文)The linear combinations of biomarkers maximizing the partial AUC		

一、參加會議經過

此為一小型國際會議，主要與會人員為來自中國、台灣、日本、韓國與印度的學者。本屆大會由國立首爾大學主辦。在兩天的研討會議中分為四個場次，同時至多有兩個演講廳進行論文發表與討論。此外在會議廳外，也有海報發表區。本人在2/3上午發表論文,在其他時間則聆聽與會學者的文章發表,學習目前最新學術發展，並適時參與討論。

二、與會心得

三、考察參觀活動(無是項活動者略)

無。

四、建議

五、攜回資料名稱及內容

包括書面議程與摘要資料一冊。

六、其他

EAR-BC 2012

International Biometric Society Korean Region

East Asia Regional Biometric Conference 2012
February 2(Thu)-3(Fri), 2012
Seoul National University
Hoam Convention Center, Seoul, Korea

Jan 16, 2012

Dear Prof. Huey-Miin Hsueh,

The IBS (International Biometric Society) Korean Region have the pleasure of inviting you to the 3rd East Asia Regional Biometric Conference to be held in Seoul, Korea between 2nd and 3rd February 2012.The conference will be held at the Hoam Convention Center in Seoul National University.

On behalf of the Program Committee for EAR-BC 2012 we are pleased to inform you that your paper has been accepted for inclusion and presentation in the oral program.

This year our theme for conference is Biometric Applications in East Asia. Participants from East Asia countries as well as different parts of the will would exchange ideas and establish long-term collaborative relationships.

These sessions included prominent representatives from a range of different disciplines in these fields. Two-day conference includes invited sessions, contributed sessions and poster presentations. We also will be offered free attendance at the conference banquet and cultural program.

This official invitation does include accommodation expenses from 1st ~4th February in Hoam Faculty House (http://www.hoam.ac.kr/english). We will send accommodation confirmation letter with you soon.

We look forward to seeing you in Seoul.

Sincerely yours

Ho Kim

Ho Kim, Ph.D.
the IBS(International Biometric Society) Korean Region President

If you have questions, please reply to this e-mail.
Email : earbc2012@gmail.com
Tel : (82)2-880-2829
Department of Biostatistics, School of Public Health,
Seoul National University, 599 Kwanak-Gu Kwanak-ro, Seoul 151-742, Korea

The linear combinations of markers which maximize the partial area under the ROC curves

Man-Jen Hsu, Huey-Miin Hsueh

Department of Statistics, National ChengChi University, Taipei, Taiwan

As biotechnology has made remarkable progress nowadays, there has also been a great improvement on data collection with lower cost and higher quality outcomes. More often than not investigators can obtain the measurements of many disease-related features simultaneously. When multiple potential markers are available for constructing a diagnostic tool of a disease, an effective approach is to combine these markers to build one single indicator. For continuous-scaled variables, the use of linear combinations is popular due to its easy interpretation. Su and Liu (1993) derived the best linear combination under the criterion of the area under the receiver operating characteristic (ROC) curve, when the joint normality of markers is assumed. However, in many investigations, the emphases are placed only on a limited extent, instead of the whole ROC curve. The goal of this study is to investigate the linear combination that maximizes the partial area under a ROC curve (pAUC) for a given specificity range. We find that the pAUC maximizer may not be unique and local maximizers sometimes do exist, in contrast to the AUC maximization. In this talk, computational issues will be discussed, an estimated optimal linear combination will be introduced, and the asymptotic property of the proposed estimator will be given. Numerical studies on both synthetic and real data sets will be performed for validations.

國科會補助專題研究計畫項下出席國際學術會議心得報告

日期：101 年 8 月 10 日

計畫編號	NSC 100－2118－M－004－004－		
計畫名稱	運用隨機森林分類方法在檢定基因集的顯著性		
出國人員姓名	薛慧敏	服務機構及職稱	政治大學統計系
會議時間	2012 年 6 月 23 日 至 2012 年 6 月 26 日	會議地點	美國波士頓 Westin 飯店
會議名稱	(中文)「國際泛華統計協會之應用統計研討會」 (英文) ICSA 2012 APPLIED STATISTICS SYMPOSIUM		
發表論文題目	(中文)醫學診斷之生物標記選取 (英文)Biomarker Selection in Medical Diagnosis		

一、參加會議經過

第一天為報到與短期課程，後三天則為研討會議。每一天包括一場專題演講與二至三個場次的演講,每場次有多個演講廳同時進行會議。本人在 6/25 下午發表論文，在其他時間則至各演講廳，聆聽與會學者的文章發表，學習目前最新學術發展，並適時參與討論。

二、與會心得

本屆與會人員多數為生物統計、醫學統計領域之專家與學者。參加人員除了來自學界外,也包括美國食品藥物管理局(FDA)或美國衛生研究院(NIH)以及製藥產業等。此會議提供很好的機會讓產業界、學術界以及政府機構的人員能夠就近期統計理論、方法與實際運用上交流與分享。本人在此次會議上有豐富收穫。

三、考察參觀活動(無是項活動者略)

無。

四、建議

近年來，由於經濟因素，至歐美參加會議之旅費與生活費逐年高漲，國科會補助通常不敷使用，建議能因應客觀環境，適當提高補助經費。

五、攜回資料名稱及內容

包括書面議程與摘要資料一冊。

六、其他

無。

hsueh

寄件者: Shi, Hongliang [Hongliang.Shi@MPI.com]
寄件日期: 2012年5月19日星期六 下午 9:23
收件者: hsueh@nccu.edu.tw
副本: tcai.hsph
主旨: Your contributed abstract to ICSA 2012 applied statistics symposium

Dear Huey-Miin,

We would like to inform you that your submitted abstract has been accepted by our programming committee. Instead of presenting at a contributed session, your talk 'Biomarker selection in medical diagnosis' will be about 25 mins and be included in the invited session s02 (6/25 15:00-16:50).

Please plan accordingly and let me know if you have any question!

Best,
Hongliang

Biomarker Selection in Medical Diagnosis

Man-Jen Hsu[1], Yuan-Chin Ivan Chang[1,2], Huey-Miin Hsueh[1]

[1]Department of Statistics, National ChengChi University, Taipei, Taiwan
[2]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

As biotechnology has made remarkable progress nowadays, there has been a great improvement on data collection with lower cost and higher quality outcomes. More often than not investigators can obtain the measurements of many disease-related features simultaneously. The multiple continuous-scaled biomarkers are combined by a linear combination for a medical diagnosis. This study aims to construct a procedure to select a subset of important biomarkers. The criterion under investigation is the partial area under the ROC curve (pAUC) over a pre-determined specificity range. The importance of an individual biomarker is assessed upon its contribution to the best linear combination maximizing the pAUC. The testing procedures to identify either the significance of a set of biomarkers or the significance of a single biomarker are proposed. Furthermore, two biomarker selection approaches, embedding the proposed statistical tests, are developed. Numerical studies on both synthetic and real data sets are performed for validations.

無研發成果推廣資料

100 年度專題研究計畫研究成果彙整表

計畫主持人：薛慧敏	計畫編號：100-2118-M-004-004-
計畫名稱：運用隨機森林分類方法在檢定基因集的顯著性	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	0	2	100%	人次	
		博士生	1	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	1	100%	篇	除了參加國際研討會議發表論文，也致力於期刊投稿。
		研究報告/技術報告	0	0	100%		
		研討會論文	0	1	100%		已投稿至 Genome Informatics Workshop 2012 大會，目前正審理中。
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 ■未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

本計畫之研究論文已投稿至 Genome Informatics Workshop 2012，目前正在審核中。

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限）

此成果在學術研究方面，對於基因集資料，提供適當統計檢定方法，可進一步應用在後續生物或醫學上，有利於生醫相關產業發展。