# 行政院國家科學委員會專題研究計畫 成果報告

## 利用分解及重組兩階段方式搜尋未排比核糖核酸之共通結構元

## 研究成果報告(精簡版)

計畫類別：個別型
計畫編號：NSC 99-2221-E-009-142-
執行期間：99年08月01日至100年07月31日
執行單位：國立交通大學資訊工程學系（所）

計畫主持人：胡毓志

計畫參與人員：碩士班研究生-兼任助理人員：莫士逸
碩士班研究生-兼任助理人員：陳彥嘉
碩士班研究生-兼任助理人員：黃韻潔
碩士班研究生-兼任助理人員：蘇莆仁
碩士班研究生-兼任助理人員：郭思廷

報告附件：出席國際會議研究心得報告及發表論文

處理方式：本計畫可公開查詢

中華民國 100年08月04日

# 行政院國家科學委員會補助專題研究計畫 ■成果報告 □期中進度報告

## 利用分解與重組兩階段方式搜尋未排比核糖核酸之共同結構元

計畫類別：■個別型計畫　□整合型計畫
計畫編號：NSC　99 - 2221 - E - 009 - 142 -
執行期間：99 年 08 月 01 日至 100 年 07 月 31 日

執行機構及系所：交通大學資工系

計畫主持人：胡毓志
共同主持人：無
計畫參與人員：陳彥嘉，黃韻潔，莫士逸，郭思婷，蘇莆仁

成果報告類型(依經費核定清單規定繳交)：■精簡報告　□完整報告

本計畫除繳交成果報告外，另須繳交以下出國心得報告：
□赴國外出差或研習心得報告
□赴大陸地區出差或研習心得報告
■出席國際學術會議心得報告
□國際合作研究計畫國外研究報告

處理方式：除列管計畫及下列情形者外，得立即公開查
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

中　華　民　國　100　年　8　月　3　日

# 行政院國家科學委員會專題研究計畫成果報告

## 國科會專題研究計畫成果報告撰寫格式說明

## Preparation of NSC Project Reports

計畫編號：NSC 99-2221-E-009 -142-
執行期限：99 年 8 月 1 日至 100 年 7 月 31 日

主持人：胡毓志　交通大學資訊工程系
計畫參與人員：陳彥嘉，黃韻潔，莫士逸，郭思婷，蘇莆仁
交通大學資訊工程系

一、中文摘要

許多功能性的核糖核酸俱有在演化過程中所保留的二級結構藉以維持它們在細胞中的所扮演的角色。已有許多針對單一核糖核酸結構預測之研究，然而隨著已知核糖核酸數量的增加，如何發掘核糖核酸家族的共通特性已日形重要。與過去方法不同的是，我們提出一種將發掘共通結構元與核糖核酸排比分開處理的演算法。我們先利用核糖核酸摺疊預測工具將核糖核酸的二級結構做整理，之後再採行類似 Gibbs Sampling 的方法搜尋核糖核酸中的共通結構元。我們使用多個在 Rfam 核糖核酸家族檢測此方法的可行性，並與其他現行方法做比較。

關鍵詞: 核糖核酸，二級結構，結構元

**Abstract**

*Many functional RNAs have evolutionarily conserved secondary structures in order to fulfill their roles in a cell. A lot of works have been done for single RNA structure prediction; however, as more RNA sequence data have been produced, finding characteristic structure motifs within RNA families becomes very important. Unlike some methods that find consensus structures from a multiple sequence alignment if available or others that align sequences and structures simultaneously, our approach separates consensus motif finding from sequence folding. After applying RNA folding algorithms to each sequence of given RNAs as a preprocess, we then combine structure decomposition and Gibbs sampling techniques to identify common structure motifs in unaligned RNA sequences. To demonstrate the performance, we tested it on several RNA families in Rfam. The experimental results show our new approach is competitive with other current prediction systems.*

Keywords: RNA, secondary structure, structure element

**Introduction**

Most of the current structural bioinformatics research is focused on proteins, and yet thousands of genes produce transcripts exerting their functions without ever producing protein products [1-3]. We can easily argue that the comprehensive understanding of the biology of a cell requires, besides proteins, the knowledge of the identities of all functional RNAs (both noncoding and protein-coding) and their molecular structures.

A fundamental principle of biology is that a stable 3D structure is essential for biological functions. Many functional RNAs have evolutionarily conserved structures in order to fulfill their roles in a cell. Some of the functions can be presented by functional motifs, such as several well-understood structurally conserved RNA motifs in viral RNAs, e.g. the TAR and RRE structures in HIV and the IRES regions in Picornaviridae [4]. Although experimental assays for basepairing in RNAs constitute the most reliable method for secondary structure determination, yet it is often difficult and expensive to acquire the 3D spectrum data of RNA molecules [5].

Various computational methods for the prediction of RNA secondary structures have been developed. According to the search strategies applied and the structure representations used, they can be roughly classified into the following categories: (1) free energy minimization [6-8] (2) comparative sequence analysis [9-10] (3) stochastic context-free grammars [11-13] (4) heuristics [14-16] (5) graph theoretical approach [17-18] and (6) hybrid [19-22].

We developed a new method for identifying consensus secondary structure motifs in a set of unaligned RNA sequences. It operates in two stages that distinguish folding RNAs from finding consensus structure motifs. In the first stage, each of the given sequences is fed to some folding algorithm, e.g. Mfold [6] or RNAfold [8], to get a predicted secondary structure. In the second stage, it takes as the input the predicted secondary structures from the folding algorithm to identify common motifs. There are several important features in our method. First, it is applicable to unaligned RNA sequences with long flanking regions and low sequence similarity. Second, it has flexibility in incorporating new tools for single RNA global structure prediction in the first stage. Third, secondary structures predicted in the first stage are transformed to an abstract form that helps constrain the search space of consensus motifs. We tested our method on 7 Rfam [23] families respectively to evaluate the accuracy of consensus structure motif prediction for each single family.

## Methods

Given functionally related RNA sequences, there are currently three main approaches to the finding of common secondary structures [24]. The first approach aligns sequences using standard multiple sequence alignment tools, e.g. ClustalW [25], and then detects consensus secondary structures based on mutual information, free energy or sequence covariance [9-10][26] etc. However this approach strongly depends on a reliable multiple sequence alignment. It is not suitable for RNAs with low sequence similarity. An alternative approach is to fold sequences and align structures at the same time. Though this approach can be applied in the case of unavailability of multiple sequence alignment, its high computational complexity restricts its practical use [27-29]. If there is no enough sequence conservation, and the complexity of structure motifs exceeds the pragmatic limit of the above approaches, we may take the last approach. It first predicts the secondary structure for each sequence, and then aligns the structures directly [20-21][30].

Our proposed method adopts the last approach, but the objective of our system is to find consensus structure motifs within a set of RNAs rather than a multiple global structure alignment. We define a legal structure motif for an RNA family as a commonly shared structure: (1) that is folded from continuous nucleotides, (2) that begins with a 5'segment of a stem, and ends with a 3'segment which may be the half of another stem or paired with the first 5'segment, and (3) that has no unpaired 5' or 3'segment between the first 5'segment and the last 3'segment. For each predicted structure from a folding algorithm, we exhaustively decompose it and enumerate all the possible substructures that comply with the three constraints above. A legal structure motif can be further defined as a component motif if it satisfies all three constraints above, and cannot be broken into smaller legal structure motifs.

Gibbs sampling is one of the MCMC (Markov Chain Monte Carlo) algorithms. In Gibbs sampling, we iteratively sample each variable conditioned on the most recent values of the other variables. Starting with a set of initial values of all the variables, we cycle through the sampling process for each single variable in any order until the values of all the variables converge to a stable state. Under this framework, the given RNAs are the variables, and the motif patterns are treated as their values. Our goal here is, by Gibbs sampling, to find the appropriate value (i.e. motif pattern) for each variable (i.e. RNA) such that the values can reach a stable distribution which corresponds to a consensus structure motif.

We define the similarity between two structure patterns, $s_i$ and $s_j$, as the following:

$$\text{if } RLD(s_i,s_j) > \theta_L, sim(s_i,s_j) = 0; \quad (1)$$

$$\text{otherwise } sim(s_i,s_j) = w_1 \cdot seqalign(s_i,s_j) + w_2 structalign(s_i,s_j). \quad (2)$$

$$RLD(s_i,s_j) = | len(s_i) - len(s_j) | / \max(len(s_i), len(s_j)). \quad (3)$$

$len(s_k)$ is the sequence length of structure $s_k$.

where $RLD(s_i,s_j)$ is the relative length difference between $s_i$ and $s_j$, $seqalign(s_i,s_j)$ is the sequence alignment score based on the Needleman-Wunsch algorithm [31], assuming match=1, mismatch=0 and gap=-1, and $structalign(s_i,s_j)$ is the structure alignment score computed by RSmatch [32]. Both *seqalign*($s_i$,$s_j$) and *structalign*($s_i$,$s_j$) are normalized between zero and one. Note that we assign zero to *sim*($s_i$,$s_j$) directly when *RLD*($s_i$,$s_j$) is greater than $\theta_L$ to save the time for the computation-intensive alignment procedures. The motivation behind this is our observation of most families in Rfam shows that the relative length difference between family members is usually insignificant, which makes it an effective filter. In eq.(2), *sim*($s_i$,$s_j$) is computed as the weighted sum of the sequence and structure alignment scores, where $w_1$+$w_2$=1.

The Gibbs sampling process in our system starts with an initial state of a consensus motif represented by a set of seeds, *SEED*, each of which is a possible occurrence of the motif in a particular RNA sequence. In each iteration, we sample the motif patterns for one RNA, e.g. *R*, conditioned on the currently selected motif occurrences in the others, and a structure pattern $p_i \in R$ will be chosen as a new seed (i.e. a new motif occurrence) if it satisfies either of the following conditions.

If *R* does not currently have a seed in *SEED*, then

$$p_i = \operatorname{argmax}_{p_j} 1/|SEED| \cdot \sum_{s_k \in SEED} sim(p_j, s_k) \qquad (4)$$

under the constraint that

$$1/|SEED| \cdot \sum_{s_k \in SEED} sim(p_i, s_k) > \theta_s$$

If *R* already has a seed in *SEED*, then

$$p_i = \operatorname{argmax}_{p_j} 1/(|SEED| - 1) \cdot \sum_{s_k \in SEED, s_k \notin R} sim(p_j, s_k) \qquad (5)$$

under the constraint that

$$1/(|SEED| - 1) \cdot \sum_{s_k \in SEED, s_k \notin R} sim(p_i, s_k) > \theta_s$$

As we iterate over every RNA, we can either add new patterns as new motif occurrences when the above condition is satisfied, or delete old seeds from the seed set if they no longer meet the constraint. We update the set *SEED* with the aim to increase the total pairwise pattern similarity $sim_{total}$(*SEED*) defined below. We repeat the same sampling process until no change of motif occurrences can be made to improve $sim_{total}$(*SEED*).

$$sim_{total}(SEED) = \sum_{i \neq j} sim(s_i, s_j) \quad \forall s_i, s_j \in SEED \qquad (6)$$

The initial seeds determine where and how fast Gibbs sampling converges, and the size of the initial seed set does not need to be equal to the total number of the RNAs given. Since we can start Gibbs sampling with different initial seeds, it can terminate at various sets of final seeds. When Gibbs sampling stops after it converges, the size of converged *SEED* will ideally be equal or approximate to that of the given RNA family, and the seeds *per se* are the predicted occurrences of a consensus motif. According to $sim_{total}$, we rank all the motifs to which Gibbs sampling converges, and report them in a sorted list after the user specifies the number of top-ranked motifs required in the output.

## Experimental Results

Several recent tools were selected for comparison, including MARNA [21], CMfinder [22], and RNAshapes [33]. As these algorithms were derived from different design philosophies, we followed Yao *et al.* [22] to test each algorithm on the same input data using default parameter settings to conduct a reasonably fair and consistent comparative study.

We picked 7 families of different sizes from the Rfam database as the test data. The seed alignment for each family is considered the consensus motif, whose number of hairpins varies from one to three among different families. Unlike Yao *et al.* [22], who included a fixed number of genomic sequence bases (e.g. 200 bases), we instead included genomic sequence flanking the motif such that the ratio of the motif length to the sequence total length is set between 0.1 and 0.6 at random, to reflect the reality that motif positions are usually unknown. The smaller the ratio, the larger the length difference between motifs and sequences. The average flanking genomic sequence length can then vary from 50 to more than 250 bases for different families. As the length of genomic flanking sequences has a larger deviation by our setting, the test data are more challenging than Yao *et al.'s*, and these test datasets are summarized in Table 1.

The performance was measured at the base pair level relative to the Rfam annotation. We compared the predicted motif against the annotated seed alignment provided in Rfam. Let $P_t$ (true positive) denote the number of base pairs that exist in annotated seed alignments and are correctly predicted, $P_f$ (false positives) denote the number of base pairs that do not exist in annotated alignments but are predicted, and $N_f$ (false negatives) denote the number of base pairs that exist in seed alignments but are not predicted. The overall accuracy of a prediction is computed as the MCC (Matthews Correlation Coefficient) approximated by the geometric mean of sensitivity and positive predictive value [34].

$$MCC \approx \sqrt{\frac{P_t}{P_t + N_f} \cdot \frac{P_t}{P_t + P_f}} \qquad (7)$$

As MARNA has a lower limit on input size, for those families larger than 20 RNAs, we randomly picked 20 sequences for testing. The MCC for each method is presented in Table 2. For a complete comparison, we tested all the methods, except MARNA, on the full set of seed sequences in each family, and summarized the results in Table 3. According to Table 2 and 3, our approach outperformed RNAshapes and MARNA in most of the tests, and was comparable to CMfinder.

To further compare our system with CMfinder in robustness, we added noise to the datasets by putting in 15 random non-family RNA sequences. We present the results in Table 4, and it shows no significant difference in all test datasets except two families, lin-4 and IRE. Note that the family size of lin-4 is much smaller than that of the others. It contains only nine seed sequences. After we added 15 noise sequences, the low signal/noise ratio affected CMfinder more significantly than our approach. On the other hand, though compared with the others IRE is not a small family (34 RNAs), yet the IRE motif is relatively small. Its size is only 28 nts, which could be easily clouded by noise. Table 4 indicates that our system was more robust than CMfinder in these tests.

**Table 1.** Summary of Rfam families for testing

| | Max/Min/Avg Seq Length | Avg Motif Length | No. of Hairpins in Motif | Total Sequences |
|---|---|---|---|---|
| ctRNA_pGA1 | 303/299/300 | 62 | 2 | 17 |
| Entero_CRE | 312/212/231 | 39 | 1 | 56 |
| HepC_CRE | 202/152/170 | 48 | 2 | 47 |
| IRE | 181/81/140 | 28 | 1 | 34 |
| lin-4 | 322/320/321 | 68 | 1 | 9 |
| Purine | 201/99/190 | 72 | 3 | 35 |
| s2m | 164/160/163 | 41 | 1 | 38 |

**Table 2.** Summary of prediction accuracies (MCC) for partial Rfam families

| | MARNA | RNAshapes | CMfinder | Ours |
|---|---|---|---|---|
| ctRNA_pGA1 | 0.890 | 0.873 | 0.950 | 0.959 |
| Entero_CRE | 0.765 | 0.844 | 0.954 | 0.936 |
| HepC_CRE | 0.659 | 0.911 | 0.998 | 0.987 |
| IRE | 0.499 | 0.569 | 0.899 | 0.847 |
| lin-4 | 0.793 | 0.797 | 0.795 | 0.711 |
| Purine | 0.749 | 0.558 | 0.900 | 0.864 |
| s2m | 0.282 | 0.241 | 0.855 | 0.899 |

**Table 3.** Summary of prediction accuracies (MCC) for complete Rfam seed sets

| | RNAshapes | CMfinder | Ours |
|---|---|---|---|
| ctRNA_pGA1 | 0.790 | 0.950 | 0.959 |
| Entero_CRE | 0.816 | 0.913 | 0.934 |
| HepC_CRE | 0.805 | 0.999 | 0.976 |
| IRE | 0.502 | 0.970 | 0.902 |
| lin-4 | 0.796 | 0.795 | 0.711 |
| Purine | 0.749 | 0.923 | 0.903 |
| s2m | 0.160 | 0.897 | 0.923 |

**Table 4.** Robustness comparison

| | CMfinder | Ours |
|---|---|---|
| ctRNA_pGA1 | 0.950 | 0.959 |
| Entero_CRE | 0.913 | 0.934 |
| HepC_CRE | 0.999 | 0.976 |
| IRE | 0.862 | 0.871 |
| lin-4 | 0.478 | 0.660 |
| Purine | 0.923 | 0.903 |
| s2m | 0.897 | 0.923 |

## Conclusion

Given a set of unaligned RNA sequences, the goal is to find the consensus structure motifs in these RNAs. In this work, we proposed a two-stage approach by separating motif finding from sequence folding. Within this framework, not only can new folding tools be easily added to increase reliability, but other optimization techniques than Gibbs can also be applied to improve accuracy. The competitive performance of the new approach was demonstrated by testing it on various Rfam families.

In the future work we plan to extend the approach in two directions. First, we will increase its applicability to finding characteristic structure motifs in mixed unaligned RNAs from multiple families. Second, we will develop an adaptive mechanism for parameter tuning of *RLD* and *sim* thresholds so the system can adjust the threshold automatically.

## References

[1] Storz, G. (2002) An expanding universe of noncoding RNAs. *Science*, 296, 1260-1263.

[2] Lai, E.C. (2003) RNA sensors and riboswitches: self-regulating messages. *Current Biology*, 13, R285-R291.

[3] Nudler, E. and Mironov, A.X. (2004) The riboswitch control bacterial metabolism. *Trends Biol. Sci.*, 29, 11-17.

[4] Hofacker, I.L., Priwitzer, B. and Stadler, P.F. (2004) "Prediction of locally stable RNA secondary structures for enome-wide surveys", *Bioinformatics*, 20, 186-190.

[5] Furtig, B., Richter, C., Wohnert, J. and Schwalbe, H. (2003) NMR spectroscopy of RNA. *Chembiochem.*, 4, 936-962.

[6] Zuker, M and Stiegler, P. (1981) Optimal computer folding of larger RNA sequences using therdynamics and auxiliary information. *Nucleic Acids Res.*, 9, 133-148.

[7] Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, 244, 48-52.

[8] Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M. And Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie*. 125, 167-188.

[9] Chiu, D.K. and Kolodziejczak, T. (1991) Inferring consensus structure from nucleic acid sequences. *Compu Appl Biosci*, 7, 347-352.

[10] Gutell, R.R. (1993) Evolutionary characteristics of RNA: inferring higher-order structure from patterns of sequence variation. *Curr Opin. Struct. Biol.*, 3, 313-322.

[11] Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, 22, 2079-2088.

[12] Sakakibara, Y., Brown, M., Hughey, R., Mian, I., Sjolander, K., Underwood, R. and Haussler, D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, 22, 5112-5120.

[13] Knudsen, B. and Hein, J. (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.*, 31, 3423-3428.

[14] Batenburg,F.H.D.van, A.P. Gultyaev & C.W.A. Pleij (1995) An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. theor. Biol*, 174, 269-280.

[15] Gultyaev, A.P., F.H.D.van Batenburg and C.W.A. Pleij (1995) The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, 250, 37-51.

[16] Hu, Y. (2002) Prediction of consensus structural motifs in a family of coregulated RNA sequences. *Nucleic Acids Research*, 30, 3886-3893.

[17] Ji, Y., Xu, X and Stormo, G.D. (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, 20, 1591-1602.

[18] Hamada, M., Tsuda, K., Kudo, T., Kin, T. and Asai, K. (2006) Mining frequent stem patterns from unaligned RNA sequences. *Bioinformatics*, 22, 2480-2487.

[19] Juan, V. and Wilson, C. (1999) RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol.*, 289, 935-947.

[20] Hochsmann, M., Toller, T., Giegerich, R. and Kurtz, S. (2003) Local similarity of RNA secondary structures. *Proc of the IEEE Bioinformatics Conference*. 59-68.

[21] Siebert, S. and Backofen, R. (2003) MARNA: A server for multiple alignment of RNAs. Proc of the German Conference on Bioinformatics, 35-40.

[22] Yao, Z., Weinberg, Z and Ruzzo, W. (2006) CMfinder: a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22, 445-452.

[23] Griffiths-Jones, S. *et al.* (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33, 121-124.

[24] Gardner, P.P. and Giegerich, R. (2004) A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5:140.

[25] Thompson, J., Higgins, D. and Gibson, T (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap

penalties and weight matrix choice. *Nucleic Acids Research*, 22, 4673-4680.

[26] Hofacker, I.L., Fekete, M. and Stadler, P. (2002) Secondary structure prediction for aligned RNA sequences. *J. Molecular Biology*, 319, 1059-1066.

[27] Sankoff, D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Applied Math*, 45, 810-825.

[28] Gorodkin, J., Heyer, L. and Stormo, G. (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Research*, 25, 3724-3732.

[29] Touzet, H. and Perriquet, O. (2004) CARNAC: folding families of related RNAs. *Nucleic Acids Research*, 32, W142-145.

[30] Giegerich, R., Voβ, B. and Rehmsmeier, M. (2004) Abstract shapes of RNA. *Nucleic Acids Research*, 32, 4843-4851.

[31] Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Molecular Biology*, 48, 443-453.

[32] Liu, J., Wang, J., Hu, J. and Tian, B. (2005) A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics*, 6:89.

[33] Steffen, P., VoB, B., Rehmsmeier, M., Reeder, J. and Giegerich, R. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22, 500-503.

[34] Gorodkin, J. *et al*. (2001) Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Research*, 29, 2135-2144.

# 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）

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）

目前國內已有許多研究單位的多位學者及專家致力於核糖核酸之相關研究，其中包含 RNAi 的研究居多，成果也大多局限於資料庫的開發。國內學術研究的重點多以論文發表為主軸，除了強調引用次數以提升國內在國際的學術地位外，與國外研究方向相較，鮮少有實際的應用，或許政府單位可以扮演橋樑角色，協助較多產學合作的機會，雖然政府協助產學合作已有不少成績，但大多以硬體生產為主，如可能，應朝其他更高階的產業發展。我們希望藉由與其他相關研究之合作，拓展我國在核糖核酸相關領域之研究發展，資料庫與分析預測工具的結合應可加速核糖核酸研究的應用。

# 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：100 年 8 月 3 日

| | | | |
|---|---|---|---|
| 計畫編號 | NSC 99 - 2221 - E - 009 - 142 - | | |
| 計畫名稱 | | | |
| 出國人員姓名 | 胡毓志 | 服務機構及職稱 | 交大資工 |
| 會議時間 | 100 年 7 月 24 日<br>至<br>100 年 7 月 26 日 | 會議地點 | Rome, Italy |
| 會議名稱 | The 2011 European Conference on Data Mining | | |
| 發表論文題目 | Applications of Data Mining to Postoperative Pain Management | | |

## 一、 參加會議經過

The first impression on Rome was not so good as expected after my arrival in Rome on Jul 23rd. The public transportation system, including Metro and Bus, was on strike. I heard of the economic situation of Italy before the trip, but never thought I'd encounter something like labor on strike. As the conference would be held in some hotel not located in the central Rome, it caused me lots of troubles to get there. The final resort was to take a taxi.

I registered at the front desk of ECDM the following morning. To my surprise, I ran into a colleague of mine at a different school. After a chat, we attended the opening remarks, and the keynote. Since some other related conferences had been held in the same hotel, most of the participants continued to join the follow-up conferences. This certainly drew more audience to the events. I also noticed that a significant number of researchers were from China, which suggests that China has been quite active in this field too. Since my presentation was scheduled right after the keynote, I reported to the session chair immediately afterwards. Each speaker was allowed for 25min followed by a 5min Q&A.

There were several interesting questions regarding our work raised after my talk. Gladly, they were all answered properly, I think. Two issues that most interested me were: (1) the feasibility of applying logistic regression to PCA patient attribute analysis, and (2) the potential improvement by applying feature selection to our 280 attributes. We already tested feature selection before the conference, including wrapper and filter, but neither of them showed any significant improvement. Thus, I "humbly" replied to the audience what we had found earlier. As for the first comment, I think it is worth a try.

On the very last day, on my way home, it was certainly not a decent experience in Rome either. The Custom

Officials were on strike. The line of people waiting for custom declaration just got longer and longer, but the officials simply ignored the inconvenience they caused. I could not imagine what if this happened in Taiwan. I kept hearing the complaints from tourists all over the world, e.g. Thai, Australia, USA, etc. I can come up with all kinds of excuses for these Italians for their strikes, but the damage has certainly been done, not only on themselves, but also the international tourists, including me.

## 二、 與會心得

Though most of the works presented were related to the conventional topics in data mining/machine learning, e.g. classification, clustering, etc., yet several novel applications really caught my eyes. I summarized some of these works as below.

1. Mobile Mind: A fully mobile platform based machine learning applications
2. Mining for chest pain diagnosis for the elderly in an emergency department
3. Who TWEETS: detecting user types and TREET quality using supervised classification
4. Extracting maximally connected sub-graphs used for community generation

Among them, two topics are related to my current studies, i.e. medical informatics and network community analysis. Given this opportunity, I have a chance to appreciate what other researchers are working on, and what the potential applications of data mining/machine learning may be.

Data mining/machine learning has been a very active topic among the main streams of research. Numerous significant real-world applications have been reported in literature. Judging from my experience of attending international conferences, I cannot help but notice that those advanced countries have put tremendous efforts and resources to this field. Not only the novelty of their projects, but also the scale of the budgets really opened my eyes. Being a nobody, I may be in no position to judge the current status of the research on data mining/machine learning in Taiwan, and yet I indeed sincerely and humbly hope that things may change a bit in the right direction in the near future, so that we can carry out more flexible research projects of larger scales if possible. The aims of the projects will no longer be just getting more publications, or filing more patents. Instead, perhaps we can expand the aims a bit, or adjust the aims a bit. How to integrate the efforts of academia and industry should be a higher priority. We have witnessed a significant progress in collaboration of academia and computer H/W corps; nevertheless, collaboration in more high-end industry has been limited, especially in intelligence-based S/W development. It is true that it is also the responsibility of the academia to seek its opportunity in industry, and in fact, this is how it works in some other more advanced countries. However, we have not reached that stage yet. Government can definitely play an important and effective role in setting up the bridge between the academia and the private corps.

## 三、 考察參觀活動(無是項活動者略)

The conference provided a city tour around Rome, but it required extra charge. Due to my limited NSC travel fund, I had to pass.

## 四、 建議

Though being just a nobody with words of no significance, if I may, I humbly suggest that government support more researchers to attend international conferences, provide sufficient funds to encourage all to continue what they are doing, build a bridge through the academia can communicate well with the private corps. Across the Straight, China has been extremely active in hosting international conferences. In contrast, we lack the opportunities. With the full support of government, I believe there are a sufficient number of academia leaders who are willing to organize world-known high-quality international conferences. We all have seen the benefits a conference can ever bring along, in addition, by far, it is one of the most effective ways to show Taiwan to the World.

## 五、 攜回資料名稱及內容

The Proceedings of ECDM 2011 (CD ROM)

## 六、其他

# APPLICATIONS OF DATA MINING TO POSTOPERATIVE PAIN MANAGEMENT

Yuh-Jyh Hu, Rong-Hong Jan, Kuochen Wang, Yu-Chee Tseng

*Department of Computer Science, National Chiao Tung University*

*Hsinchu, Taiwan*

Tien-Hsiung Ku, Shu-Fen Yang

*Department of Anesthesia, Changhua Christian Hospital*

*Changhua, Taiwan*

### ABSTRACT

Appropriate postoperative pain management contributes to earlier mobilization, shorter hospitalization, and reduced cost. Undertreatment of pain may impede short-term recovery, and may even have a detrimental long-tern effect on health. Despite the advancement in postoperative pain management, pain relief and patient satisfaction still does not meet some patients' requirement. By applying data mining techniques, this study aimed to identify the predictive factors for anesthetic dosage and PCA (Patient Controlled Analgesia) demands. With the assistance of Changhua Christian Hospital, we collected 1655 PCA patient records. We analyzed patient PCA usage profiles. We concentrated on two prediction tasks in this study: (a) postoperative analgesic consumption, and (2) PCA setting readjustment.

### KEYWORDS

Intramuscular (IM), Patient Controlled Analgesia (PCA), decision trees, bagging, clustering

## 1. INTRODUCTION

Pain is one of the most commonly reported postoperative symptoms (Chung, et al., 1996). It is a highly personal experience influenced by multiple factors, including sensitivity to pain, age, genetics, physical status, and psychological factors (Turk and Okifuji, 1999; Bisgaard, et al., 2001). Along with the progress of medical science, people gradually become aware of the importance of pain management.

Intramuscular (IM) opioid injection is the most commonly used treatment for postoperative pain relief. However, different surgeries cause different degrees of pain, and pain endurance varies among people. IM opioid injection does not take effect until several tens of minutes, and it is hard to know beforehand the correct analgesic dosage to meet the patient's need. On the other hand, PCA (Patient Controlled Analgesia) is a delivery system for pain medication that makes effective and flexible pain treatments possible by allowing patients to adjust the dosage of anesthetics themselves. From previous research (Walder, et al., 2001; Dolin, et al., 2002), PCA has become one of the most effective techniques for postoperative analgesia. It is widely used in hospitals for the management of postoperative pain, especially for major surgeries. In this paper, we concentrate on the study of PCA.

There are two objectives in this paper. First, we investigate the correlation between PCA demand profiles and patient physical states. The motivation is that with better knowledge of the correlations, anesthesiologists can prescribe appropriate PCA settings for a patient, e.g. PCA dose, lockout, 4-hour limit, etc. Second, to lead to tailored methods to reduce severe postoperative pain and improve acute and chronic outcomes, we propose applying computational tools to predict: (1) postoperative analgesic requirement after a period of time, and (2) whether PCA initial controls require readjustment based on the patient's physical states and the early-stage PCA usage, e.g. the first 24 hours.

## 2. MATERIALS AND METHODS

To extend our previous study (Hu, et al., 2010), with the assistance of Acute Pain Service at Changhwa Christian Hospital (CCH), we collected and processed more patient records (from 1182 to 1655), each containing more than 200 attributes, which we divided into three categories: (a) patient physical states, e.g. systolic and diastolic blood pressures, (b) operation-related attributes, e.g. surgical type and duration, and (c) PCA settings and demand after surgery, e.g. PCA dose and lockout. Their values are either symbolic or numeric.

### 2.1 PCA Demand Analysis

Given the PCA demand frequency within each unit time, we can represent each patient's PCA demand profile as a time course, as illustrated in Figure. 1. The X-axis represents the 24-hr time scale, and the Y-axis indicates the PCA demand frequency. We applied the *k*-means clustering algorithm to the PCA demand profiles. Our objective was to discover from the time courses the significant patterns that characterized PCA demand. For example, we could identify three demand patterns in Figure 1, each including two patients.

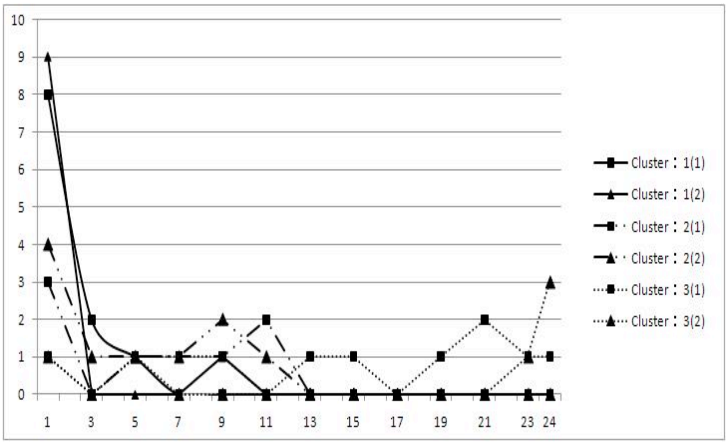


Figure 1. Example of PCA Demand Profiles. Each curve represents the distribution of PCA demands of a patient in 24 hours.

After clustering, we assigned each PCA usage record its cluster membership, e.g. 1, 2 or 3 when *k*=3. Though we know which patients have similar PCA demand profiles by their memberships, yet we do not understand which patient attributes contribute to the similar profiles. One common approach is to find the individual attributes significantly correlated with cluster membership, e.g. Pearson correlation coefficient and Spearman correlation. However, these methods only address linear relationships, and do not take into account attribute interactions that exist in real-world problems, e.g. BMI (Body Mass Index) relates to both weight and height. The relationship between cluster membership and an individual patient attribute may be too weak to recognize. Nevertheless, there are still significant associations between cluster membership and a set of attributes when considered together. Instead of evaluating the correlation between cluster membership and each single attribute, we measured the associations between cluster membership and a set of attributes based on mutual information, as defined below.

$$
\begin{aligned}
I(C;A) &= H(C) - H(C \mid A) = -\sum_{C_i \in C} p(C_i)\lg p(C_i) - \sum_{A_i \in A} p(A_i) H(C \mid A_i) \\
&= -\sum_{C_i \in C} p(C_i)\lg p(C_i) + \sum_{A_i \in A} p(A_i) \sum_{C_i \in C} p(C_i \mid A_i)\lg p(C_i \mid A_i) \\
&= -\sum_{C_i \in C} p(C_i)\lg p(C_i) + \sum_{A_i \in A, C_i \in C} p(C_i, A_i)\lg p(C_i \mid A_i)
\end{aligned}
$$

where *I*(*C;A*) is the mutual information used to measure the association between the cluster membership *C* and the attribute set *A*. *H*(*C*) is the marginal entropy of *C*, and *H*(*C*|*A*) is the entropy of *C* conditional on *A*. The cluster membership presents all the legal clusters, i.e. $\sum p(C_i) = 1$, and the attribute set specifies the attributes selected to reduce uncertainty, which satisfies $\sum p(A_i) = 1$, and each $A_i$ represents a possible attribute value combination. Mutual information, which is a symmetric measure to quantify the statistical information shared between two distributions (Cover and Thomas, 1991), provides a sound indication of the shared information between a pair of variables. The motivation is to remove the uncertainty about cluster membership as much as possible by knowing about some subset of attributes. We normalized mutual information to avoid the natural bias in mutual information that favors attributes with many values (Quinlan, 1993), or a set of attributes with many value combinations. Several normalizations are possible based on the observation that *I*(*C;A*) ≤ min(*H*(*C*)*;H*(*A*)). They include normalization using the arithmetic or geometric mean of *H*(*C*) and *H*(*A*). Since *H*(*X*) = *I*(*X;X*), we adopted the geometric mean because of the analogy with a normalized inner product in Hilbert space (Strehl and Ghosh, 2002). We define the normalized mutual information as:

$$
I(C;A)_{norm} = \frac{I(C;A)}{\sqrt{H(A)H(C)}}
$$

We consider a set of patient attributes with higher normalized mutual information is more related to the PCA demand profiles, and thus worth further medical analysis.

Since *A* is any subset, except the empty set, of the total attributes **A**, there are $2^{|\mathbf{A}|}$-1 possibilities of *A*. Currently we have only 15 attributes other than PCA setting and demand related attributes. They can produce $2^{15}$-1 different attribute subsets. We computed the normalized mutual information of every attribute subset to evaluate the correlation between the patient's behavior in using PCA and his (or her) physical states or surgery-related attributes.

## 2.2 PCA Dosage and Control Prediction

Given the physical states of patients and their early-hour PCA demand profiles, we first wish to predict the total anesthetic dose taken in later hours. In this work, we focused on longer early-hour PCA usage profiles than those in (Hu, et al., 2010). As in previous works, we tried to predict the symbolic value of anesthetic dose as we noticed that the dose could generally fall in several distinct ranges. We divided the numeric value of dosage into several intervals. This was done by an iterative optimization procedure that identified the cut-points such that the dose deviation in each interval was minimal. We then discretized the numeric value of anesthetic dose into a number of symbolic values, e.g. "low", "medium" and "high."

We can compare and evaluate prediction methods according to their accuracy and comprehensibility. The accuracy of a predictor refers to its ability to correctly predict the target value, e.g. total anesthetic dosage, for previously unseen data; the comprehensibility of a predictor refers to the level of ease with which humans can interpret the predictions. A predictor that is able to make both accurate and comprehensible predictions is most desirable, but unfortunately, for any prediction method available, there exist some trade-offs between these two criteria due to inductive bias. Our aim was to develop a prediction tool that is able to make predictions about PCA control with high accuracy as well as acceptable comprehensibility for the anesthesiologists.

Decision tree learning is among the most widely used and practical methods for inductive inference (Quinlan, 1993). It is a method of approximating the function for the target attribute by learning a decision tree from the examples. Compared with other inductive learning methods, e.g. artificial neural networks (Rumelhart, et al., 1994), support vector machines (Vapnik, 1998), genetic algorithms (Goldberg, 1989), etc., decision tree learning is more human interpretable because a decision tree is a pictorial representation, and can be easily translated into a set of if-then-else rules. To maintain sufficient comprehensibility in predictions, we built our PCA predictor based on decision tree learning. Though decision tree learning has proved useful in many real-world applications, e.g. SKICAT (Fayyad, et al., 1993), and text classification (Lehnert, et al., 1995), further studies have shown that an ensemble of decision trees are often more accurate than any single tree in the ensemble (Bauer and Kohavi, 1999; Dietterich, et al., 1996; Breiman, 1996a; Freund, 1995). Two popular methods for creating accurate ensembles are Bagging (Breiman, 1996b) and Boosting (Schapire, 1990). Both methods rely on “re-sampling" techniques to obtain different training sets for each predictor in the ensemble, but the analyses indicate that Boosting is more prone to overfitting the training data than Bagging (Freund and Schapire, 1996; Opitz and Maclin, 1999). As a result, the performance of Boosting may decrease more than Bagging in the presence of noise. We applied bagging and boosting to the decision learning algorithms, and compared their performance on the PCA datasets.

Besides predicting the total anesthetic dosage in later hours, we also wish to predict if any PCA control, e.g. lockout time or PCA dosage, should be readjusted later to improve patient's satisfaction. We consider this as an anomaly detection problem (Chandola, et al., 2009) as only few patients require readjustment of PCA settings. Learning from imbalanced datasets, where the number of examples of one (minority) class is much smaller than the other (majority), presents an important challenge to the machine learning community. Traditional machine learning algorithms are typically biased towards the majority class, and produce poor predictive accuracy over the minority class. There have been various approaches proposed for coping with imbalanced datasets. Kubat and Matwin under-sampled examples of the majority class (Kubat and Matwin, 1997); Ling and Li over-sampled examples of the minority class (Ling and Li, 1998); Chawla *et al.* over-sampled the minority class and under-sampled the majority class (Chawla, et al., 2002); Cardie and Howe weighted examples in an effort to bias the learning toward the minority class (Cardie and Howe, 1997); Joshi *et al.* evaluated boosting algorithms to classify rare classes (Joshi, et al., 2002); Guo and Viktor combined boosting and synthetic data to improve the prediction of the minority class (Guo and Viktor, 2004).

Instead of applying sampling techniques, or generating artificial data, we identified the examples that could bias the learning algorithms, and removed them from the data to balance the classes. Motivated by the nearest-neighbor approach for outlier detection (Knorr, et al., 2000; Angiulli and Pizzuti, 2002; Eskin, et al., 2002), we identified in a neighborhood the candidate examples for removal. Unlike those techniques that used the distance of a data instance to its *k*th nearest neighbor as the anomaly score, for each example of the minority class, we first identified its *k* nearest neighbors, and from them we marked the majority-class neighbors as “dirty.” After examining each example in the minority class and its neighbors, we removed those “dirty” examples. The rationale behind this is that these nearest majority-class neighbors of a minority-class member are likely to mislead the learning algorithm, and without them, learning algorithms can identify the boundary of the minority class more easily.

## 3. EXPERIMENTAL RESULTS

### 3.1 PCA Demand Profile Analysis Results

We ran the k-means algorithm on the 1655 PCA demand profiles in the patient records collected from CCH. We discovered three profile clusters sized 510, 511, and 644 respectively. We calculated the mean demand frequency as the representative demand pattern, and presented them in Figure 2. For clarity, we also presented an enlarged version of the representative pattern in clusters two and three in Figure 3.

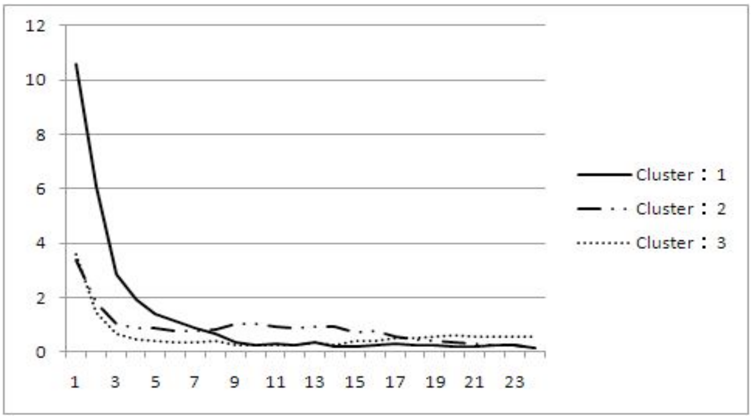


Figure 2. Representative PCA Demand Patterns in three clusters

For all possible attribute subsets of the 15 patient physical states, and surgery-related factors, we calculated the normalized mutual information to evaluate their association with the PCA demand patterns. The top two attribute subsets were listed in Table 1, and a random small attribute subset was included for comparison. The results showed that more than half of the attributes related to PCA demand, and they interacted. Also note that the normalized mutual information is small as $H(A)$ in the denominator is large due to the large number of attribute value combinations. To verify if the difference in normalized mutual information between the top and the second attribute subsets is merely by chance, we performed a bootstrap test for estimating the confidence interval. It showed that the difference was not random but rather significant at the level of $p$-val < 0.05. Despite the low normalized mutual information, our experimental results demonstrated that these 15 attributes contributed to a patient's PCA demand behavior differently. It encouraged us to identify more factors and different interactions to better characterize PCA demand behaviors.

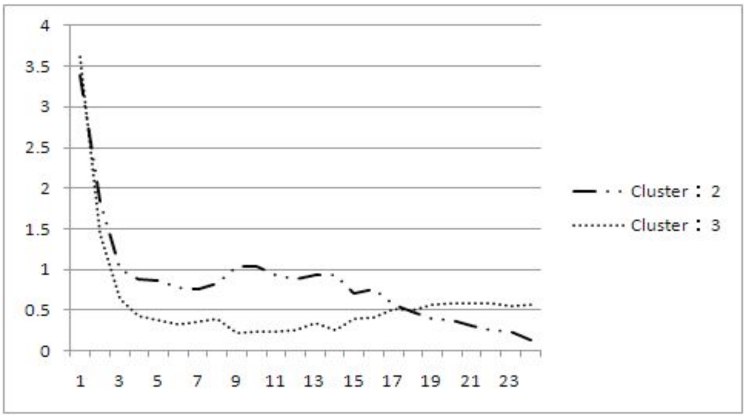


Figure 3. Representative PCA Demand Patterns in Clusters 2 and 3 (enlarged version)

Table 1. Attribute Subsets and Normalized Mutual Information

| Attribute Subset | Normalized MI |
|---|---|
| **bp_high, pulse, weight, OP_CLASS, op_time, ASA_TYPE1, M1_ANS_WAY, age, AMI** | 0.047 |
| **bp_high, bp_low, pulse, OP_CLASS, op_time, ASA_TYPE1, M1_ANS_WAY, age** | 0.045 |

| | |
|---|---|
| **OP_CLASS, DM, HT** | 0.004 |

## 3.2 PCA Dosage and Control Prediction Results

### 3.2.1 Anesthetic Dosage Prediction

Instead of making predictions based on short-term PCA usage, e.g. 3~12hrs in (Hu, et al., 2010), we predicted the total 72hour anesthetic dose of a patient from the first 24hour PCA usage data as well as the patient's physical states, and surgery-related attributes. We discretized the numeric value of anesthetic dose into three symbolic values, "low", "medium" and "high." The 1099 out of total 1655 patients that used PCA more than 72 hours were divided into three classes according to the symbolic values, and the class size was 399 (low), 551 (medium), and 149 (high) respectively. We compared the performance of C4.5 (Quinlan, 1993) with bagging and boosting in a stratified 10-fold cross validation (cv), and measured the performance in terms of recall and precision of each class, and overall accuracy in all classes. In each run of the cross validation experiment, we generated 200 bootstrap samples from the training data, and created an ensemble of 200 decision trees. For a test example, we made the prediction by taking a majority vote from the bagging trees. To implement boosting, we adopted the AdaBoost ensemble algorithm (Freund and Schapire, 1996). For all three methods, the prediction performance in "high dose" was poorer than the other two classes. A possible reason is that the number of patients that required high anesthetic dosage was smaller than the other two. For reference, we also included ANN (artificial neural network) and SVM (support vector machine) in comparison. We conducted a paired t-test between bagging and the other methods. An asterisk denoted bagging performed significantly better (*p*-val < 0.05), e.g., bagging's overall accuracy was significantly better than the others. Our experiments confirmed that boosting is more prone to overfitting than bagging (Freund and Schapire, 1996; Opitz and Maclin, 1999); as in the presence of noticeable noise in the PCA data, bagging outperformed the other methods. See Table 2 for details.

### 3.2.2 PCA Setting Readjustment Prediction

A second prediction task was to predict whether any readjustment of PCA settings was needed in later hours, which is a binary-class classification problem. Like PCA control prediction, PCA setting readjustment prediction was also based on the first 24-hr PCA usage data as well as patient's physical states, and surgery-related attributes. As the number of patients that needed PCA readjustment was much smaller than that of patients who did not, it caused class imbalance. In our PCA data, the ratio is 81% (negative class) to 19% (positive class). When classes are imbalanced, conventional learning algorithms often produce classifiers that do little more than predict the most common class.

Table 2. Results of Anesthetic Dosage Prediction

| Anesthetic Dosage | C4.5 | C4.5 | C4.5 | ANN | SVM |
|---|---|---|---|---|---|
| Prediction (%) | bagging | AdaBoost | | | |
| **"low dose" Recall** | 84.3 | 79.2* | 77.4* | 69.8* | 7.9* |
| **"medium dose" Recall** | 83.5 | 75.8* | 72.8* | 79.6* | 96.1 |
| **"high dose" Recall** | 62.4 | 61.2 | 60.6* | 21.6* | 0.0* |
| **"low dose" Precision** | 84.3 | 78.8* | 76.3* | 80.2* | 59.4* |
| **"medium dose" Precision** | 83.5 | 75.3* | 73.5* | 66.1* | 50.6* |
| **"high dose" Precision** | 62.4 | 66.0 | 62.8 | 56.9* | 0.0* |
| **Overall Accuracy** | 80.9 | 75.1* | 72.8* | 68.5* | 50.7* |

We first compared bagging with AdaBoost when classes were imbalanced, and then applied either under-sampling or over-sampling to balance classes. The goal of PCA readjustment prediction is to predict correctly the patient whose initial PCA settings require modification. Appropriate performance measures proposed include the F-score (Lewis and Gale, 1994) of positive examples, and the geometric mean of positive accuracy and negative accuracy (Kubat and Matwin, 1997). The other measures, e.g. overall accuracy, were included for reference. We showed the results in Table 3. All the values were averaged over the 10 runs in a 10-fold cv. We found that without removing "dirty" data, Bagging combined with under-sampling performed the best in both measures of positive F-score and geometric accuracy. Nevertheless, without under-sampling to balance the classes, bagging had very poor performance, even worse than the baseline original C4.5. These experimental results indicated that class imbalance had a stronger impact on bagging than on boosting. Class imbalance in our PCA data misled bagging toward the majority class. Like in Table 2, we used an asterisk to indicate the significantly better performance of bagging combined with under-sampling technique in a paired t-test, e.g. the positive F-score of bagging with under-sampling outperformed all the others significantly. The results also showed that ANN and SVM were severely biased by the majority class.

Table 3. Comparison of Bagging and Boosting (before data cleaning)

| Method | C4.5 | C4.5 | C4.5 | C4.5 | C4.5 | C4.5 | C4.5 | ANN | SVM |
|---|---|---|---|---|---|---|---|---|---|
| Prediction (%) | bagging | bagging | bagging | AdaBoost | AdaBoost | AdaBoost | | | |
| | | over-sampling | under-sampling | | over-sampling | under-sampling | | | |
| **Pos Recall** | 4$^*$ | 16$^*$ | 42 | 20$^*$ | 33$^*$ | 47 | 26$^*$ | 0$^*$ | 0$^*$ |
| **Pos Precision** | 39 | 28 | 28 | 28 | 25$^*$ | 23$^*$ | 25.$^*$ | 0$^*$ | 1$^*$ |
| **Pos F-score** | ***8***$^*$ | ***20***$^*$ | ***33*** | ***23***$^*$ | ***28***$^*$ | ***31***$^*$ | ***25***$^*$ | **0**$^*$ | **0.1**$^*$ |
| **Neg Recall** | 98.7 | 90 | 74 | 88 | 77 | 63$^*$ | 83 | 100 | 100 |
| **Neg Precision** | 82$^*$ | 82$^*$ | 85 | 83$^*$ | 83$^*$ | 84 | 83$^*$ | 81$^*$ | 82$^*$ |
| **Neg F-score** | 89 | 86 | 79 | 85 | 80 | 72$^*$ | 83 | 89 | 90 |
| **Geometric Accu** | ***20***$^*$ | ***38***$^*$ | ***56*** | ***41***$^*$ | ***50***$^*$ | ***55*** | ***46***$^*$ | **0**$^*$ | **0.2**$^*$ |
| **Overall Accu** | 81 | 76 | 68 | 75 | 68 | 60 | 72 | 81 | 81 |

We next tested our strategy for data cleaning and class balancing. After removing "dirty" negative examples, we reduced the ratio of negative to positive from 81:19 to 65:35. We presented the results in Table 4. Compared against Table 3, our data cleaning strategy improved all the classifiers in positive F-score and geometric accuracy except ANN and AdaBoost with undersampling. The classifier that gained the most from data cleaning was bagging.

## 4. CONCLUSION

We introduced a real-world application of data mining to anesthesiology. To increase patient satisfaction, we have analyzed PCA patient data, and run several experiments to evaluate the potential of applying data mining algorithms to assist anesthesiologists in PCA control. The results demonstrated the feasibility of our combinatorial approach to medical applications.

Previous research (Walder, et al., 2001; Dolin, et al., 2002) has shown PCA is one of the most effective techniques for postoperative analgesia and widely used in hospitals for the management of postoperative pain. Though PCA provides the medical staff with a convenient way to control pain, it requires constant attention, e.g. manually collecting each patient's PCA data, printing out analgesia usage data, entering readings into appropriate databases, etc.

As the advance of information technology and wireless networks, the objective has also shifted from increasing hardware performance alone to providing better services and wider applicability. Among many potential applications of information network technology is medical care. The automation of data collection and maintenance could significantly reduce the labor work to increase efficiency. We will extend the Zigbee sensor network and the IEEE 802.11 network by developing a 3G-gateway module to collect and transmit the PCA-related data to databases for pain management. We can connect more medical devices than PCA with Zigbee nodes to collect a wider variety of patient vital signals, e.g. $SpO_2$, HRV, etc. We can further explore these signals to characterize PCA demand behaviors better, and predict PCA dosage and control more accurately. Based on the same concept, we plan to expand this framework to homecare applications.

Table 4. Comparison of Bagging and Boosting (after data cleaning)

| Method | C4.5 | C4.5 | C4.5 | C4.5 | C4.5 | C4.5 | C4.5 | ANN | SVM |
|---|---|---|---|---|---|---|---|---|---|
| Prediction (%) | bagging | bagging | bagging | AdaBoost | AdaBoost | AdaBoost | | | |
| | | over-sampling | under-sampling | | over-sampling | under-sampling | | | |
| **Pos Recall** | 41 | 51 | 54 | 42 | 55 | 54 | 44 | 0 | 4 |
| **Pos Precision** | 29 | 26 | 26 | 24 | 24 | 22 | 23 | 0 | 28 |
| **Pos F-score** | ***34*** | ***34*** | ***35*** | ***31*** | ***33*** | ***31*** | ***30*** | **0** | **6** |
| **Neg Recall** | 77 | 66 | 64 | 69 | 58 | 56 | 66 | 100 | 98 |
| **Neg Precision** | 85 | 85 | 86 | 84 | 85 | 84 | 83 | 81 | 81 |
| **Neg F-score** | 81 | 74 | 73 | 76 | 69 | 67 | 74 | 90 | 89 |
| **Geometric Accu** | ***56*** | ***58*** | ***59*** | ***54*** | ***57*** | ***55*** | ***54*** | **0** | **13** |
| **Overall Accu** | 71 | 63 | 62 | 64 | 57 | 56 | 62 | 81 | 80 |

# ACKNOWLEDGEMENT

This work was partially supported by National Science Council (NSC99-2218-E-009-002, NSC99-2221-E-009-142).

# REFERENCES

Angiulli, F. and Pizzuti, C., 2002. Fast outlier detection in high dimensional spaces. in *Proc. 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pp.15-26.

Bisgaard, T. et al., 2001. Characteristics and prediction of early pain after laparoscopic cholecystectomy. *Pain,* 90, pp.261-269.

Breiman, L., 1996a. Stacked regressions. *Machine Learning*, 24, pp.49-64.

Breiman, L., 1996b. Bagging predictors. *Machine Learning*, 24, pp.123-140.

Bauer, E. and Kohavi, R., 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36, pp.105-139.

Cardie, C. and Howe, N., 1997. Improving minority class prediction using case-specific feature weights. in *Proc. 14th International Conference on Machine Learning,* pp.57-65.

Chandola, V. et al., 2009. Anomaly Detection: A Survey. *ACM Computing Surveys*, 41(3), doi: 10.1145/1541880.1541882.

Chawla, N. et al., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artificial Intelligence Research*, 16, pp.321-357.

Chung, V. et al., 1996. Postoperative symptoms 24 hours after ambulatory anaesthesia. *Can J Anaesth*, 43, pp.1121-1127.

Cover, T.M. and Thomas, A.T., 1991. *Elements of Information Theory*. Wiley.

Dietterich, T. et al., 1996. Applying the weak learning framework to understand and improve C4.5. in *Proc. 13th International Conference on Machine Learning*, pp.96-104.

Dolin, S.J. et al., 2002. Effectiveness of acute postoperative pain management: evidence from published data. *Br J Anaesth*, 89(3), pp.409–423.

Eskin, E. et al., 2002. A geometric frame-work for unsupervised anomaly detection. in *Proc. Applications of Data Mining in Computer Security*, pp.78-100.

Fayyad, U.M. et al., 1993. SKICAT: A machine learning system for automated cataloging of large scale sky survey, in *Proc. 10th International Conference on Machine Learning*, pp. 112-119.

Freund, Y., 1995. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2), pp.256–285.

Freund, Y. and Schapire, R., 1996. Experiments with a new boosting algorithm. in *Proc. 13th International Conference on Machine Learning*, pp.148-156.

Goldberg, D., 1989. *Genetic algorithms in serach, optimization, and machine learning*. Reading, MA: Addison-Wesley.

Guo, H and Viktor, H.L., 2004. Boosting with data generation: Improving the Classification of Hard to Learn Examples. in *Proc. 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp.17-20.

Hu, Y., Jan, R.H., Wang, K., Tseng, Y.C., Ku, T.H., Yang, S.F., 2010. An Application of Sensor Networks with Data Mining to Patient Controlled Analgesia. IEEE HealthCom, Lyon, France.

Joshi, M.V. et al., 2002. Predicting rare classes: can boosting make any weak learner strong? in *Proc. 8th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp.297-306.

Knorr, E.M. et al., 2000. Distance-based outliers: algorithms and applications. *The VLDB Journal, 8*, pp.237-253.

Kubat, M. and Matwin, S., 1997. Addressing the curse of imbalanced training sets: One-sided selection. in *Proc. 14th International Conference on Machine Learning*, pp.179-186.

Lehnert, W. et al., 1995. Inductive Text Classification for Medical Applications. *J Experimental and Theoretical Artificial Intelligence,* 7(1), pp.271-302.

Lewis, D.D. and Gale, W.A., 1994. A sequential algorithm for training text classifiers. in *Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, pp.3–12.

Ling, C. and Li, C., 1998. Data Mining for Direct Marketing Problems and Solutions. in *Proc. of 4th International Conference on Knowledge Discovery and Data Mining*, pp.73-79.

Opitz, D. and Maclin, R., 1999. Popular Ensemble Methods: An Empirical Study. *J. Artificial Intelligence Res*., 11, pp. 169-198.

Quinlan, J.R., 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco, USA.

Rumelhart, D. et al., 1994. The basic ideas in neural networks. *Communications of ACM*, 37, pp.87-92.

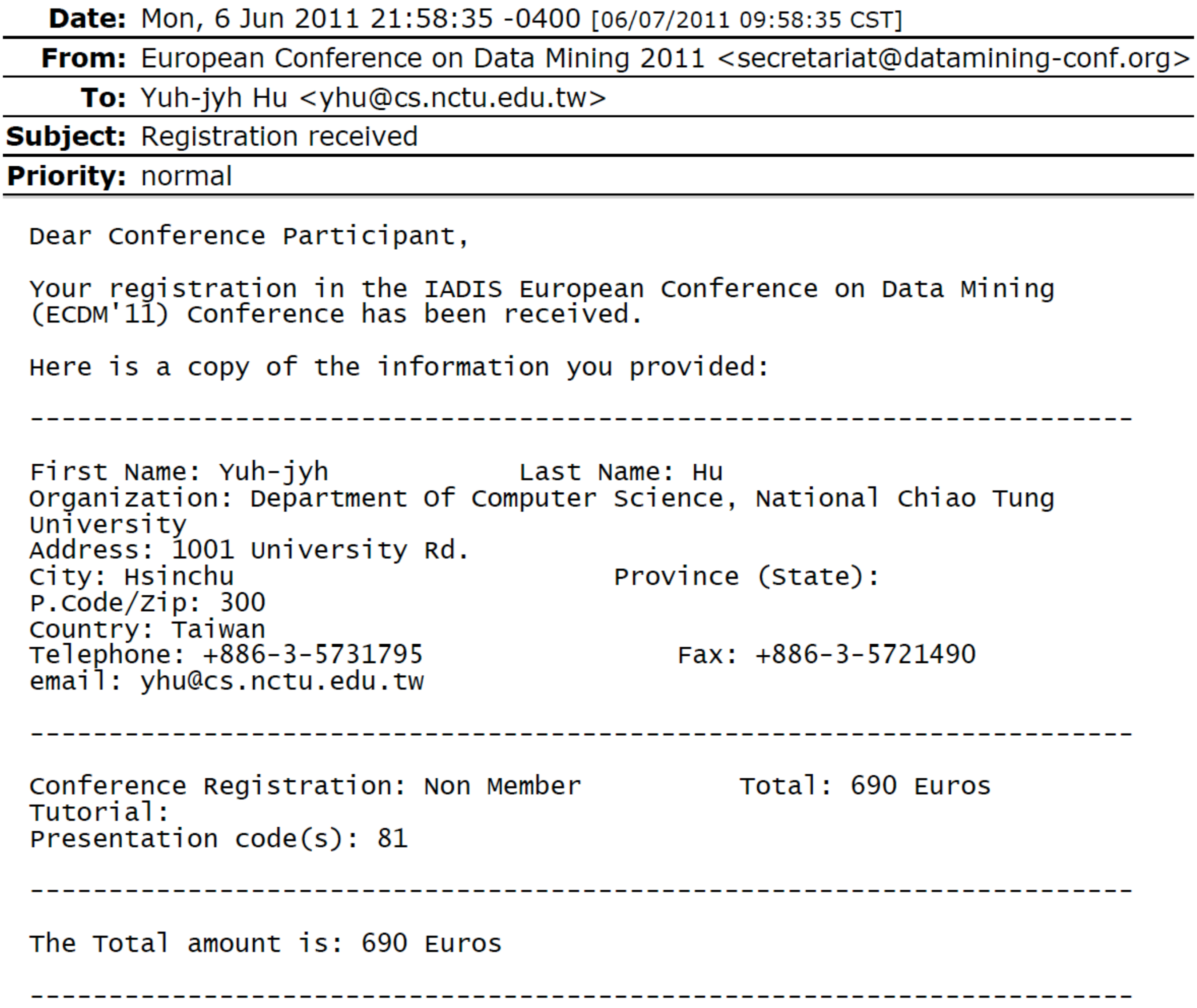
Schapire, R., 1990. The strength of weak learnability. *Machine Learning*, 5 (2), pp.197-227.

Strehl, A and Ghosh, J., 2002. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *J. Machine Learning Research*, 3, pp.583-617.

Turk, D.C. and Okifuji, A., 1999. Assessment of patients' reporting of pain: An integrated perspective. *Lancet*, 352, pp.1784-1788.

Vapnik, V.N., 1998. *Statistical Learning Theory*. John Wiley&Sons.

Walder, B., et al., 2001. Efficacy and safety of patient-controlled opioid analgesia for acute postoperative pain. *Acta Anaesthesiol Scand*, 45(7), pp.795-804.



**Date:** Mon, 6 Jun 2011 21:58:35 -0400 [06/07/2011 09:58:35 CST]

**From:** European Conference on Data Mining 2011 <secretariat@datamining-conf.org>

**To:** Yuh-jyh Hu <yhu@cs.nctu.edu.tw>

**Subject:** Registration received

**Priority:** normal

```
Dear Conference Participant,

Your registration in the IADIS European Conference on Data Mining
(ECDM'11) Conference has been received.

Here is a copy of the information you provided:

----------------------------------------------------------------------

First Name: Yuh-jyh            Last Name: Hu
Organization: Department Of Computer Science, National Chiao Tung
University
Address: 1001 University Rd.
City: Hsinchu                        Province (State):
P.Code/Zip: 300
Country: Taiwan
Telephone: +886-3-5731795                Fax: +886-3-5721490
email: yhu@cs.nctu.edu.tw

----------------------------------------------------------------------

Conference Registration: Non Member          Total: 690 Euros
Tutorial:
Presentation code(s): 81

----------------------------------------------------------------------

The Total amount is: 690 Euros

----------------------------------------------------------------------
```

# 國科會補助計畫衍生研發成果推廣資料表

日期:2011/08/03

| | |
|---|---|
| 國科會補助計畫 | 計畫名稱: 利用分解及重組兩階段方式搜尋未排比核糖核酸之共通結構元 |
| | 計畫主持人: 胡毓志 |
| | 計畫編號: 99-2221-E-009-142- 學門領域: 生物資訊 |

無研發成果推廣資料

# 99 年度專題研究計畫研究成果彙整表

| | |
|---|---|
| 計畫主持人：胡毓志 | 計畫編號：99-2221-E-009-142- |
| 計畫名稱：利用分解及重組兩階段方式搜尋未排比核糖核酸之共通結構元 | |

| 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 | 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% | | |
| 國外 | 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 | 1 | 1 | 100% | | Given the fund, our lab was able to get a paper (in a different topic) published in European Conference on Data Mining |
| | | 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 | 5 | 5 | 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<br>(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）

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）

目前國內已有許多研究單位的多位學者及專家致力於核糖核酸之相關研究,其中包含 RNAi 的研究居多，成果也大多局限於資料庫的開發。國內學術研究的重點多以論文發表為主軸，除了強調引用次數以提升國內在國際的學術地位外，與國外研究方向相較，鮮少有實際的應用，或許政府單位可以扮演橋樑角色，協助較多產學合作的機會，雖然政府協助產學合作已有不少成績，但大多以硬體生產為主，如可能，應朝其他更高階的產業發展。

我們希望藉由與其他相關研究之合作，拓展我國在核糖核酸相關領域之研究發展，資料庫與分析預測工具的結合應可加速核糖核酸研究的應用。