퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 5, pp. 707-711

# Finding Informative Genes From Microarray Gene Expression Data Using FIGER-test

Kyoung Oak Choi[1], Hwan Mook Chung[2]

[1] Computer Engineering, Catholic University of Daegu, Korea
E-mail: okajaa@cu.ac.kr
[2] Computer Engineering, Catholic University of Daegu, Korea
E-mail: hmchung@cu.ac.kr

## Abstract

Microarray gene expression data is believed to show the functions of living organism through the gene expression values. We have studied a method to get the informative genes from the microarray gene expression data. There are several ways for this. In recent researches to get more sophisticated and detailed results, it has used the intelligence information theory like fuzzy theory. Some methods are to add fudge factors to the significance test for more refined results.

In this paper, we suggest a method to get informative genes from microarray gene expression data. We combined the difference of means between two groups and the fuzzy membership degree which reflects the variance of the gene expression data. We have called our significance test the Fuzzy Information method for Gene ExpRession data(FIGER). The FIGER calculates FIGER variation ratio and FIGER membership degree to show how strongly each object belongs to the each group and then it results in the significance degree of each gene. The FIGER is focused on the variation and distribution of the data set to adjust the significance level. Our simulation shows that the FIGER-test is an effective and useful significance test.

Key Words : Fuzzy logic, Fuzzy membership function, Significance Test, Microarray gene expression

## 1. 서 론

To identify the functions of genes in the living organisms, the traditional method has worked based on "one gene in one experiment", however a new technology, called DNA microarray gene chip, has made it possible to see the whole picture of the genome[1].

DNA microarray technology can disgorge the expression data, however to get the informative genes we need to know the significance level of each gene. The popular ways for the significance test are based on traditional statistical method. In recent researches, there are several ways to use fuzzy theory or to add empirical factors[2,3].

In this paper, we suggests the Fuzzy Information method for Gene ExpRession data(FIGER). The FIGER has two steps, one is the FV(FIGER variation ration) and the other is the FMD(Fuzzy Membership Degree). The FV is the ratio of standard deviation to mean. The FV shows the distribution of data set of gene expression data. The FMD is about how much close the each object is to the mean of the group. For the simulation of FIGER, we downloaded a gene expression data set from the public repository of microarray gene expression data and computed the results with R-project that is a statistical language and a tool. The FIGER has showed the reasonable results for the analysis of microarray gene expression data. In this paper we showed the FIGER is a useful method for the microarray gene expression data analysis.

접수일자 : 2007년 4월 14일
완료일자 : 2007년 7월 30일

## 2. Related Researches

Since the DNA microarray technology has been developed by Pat Brown laboratory, it has become a way to identify the kinds and the amounts of gene(mRNA molecule) transcription which tells how the genes respond to its circumstances and/or needs[4]. There are several ways to get the significant genes. Some are based on statistics method, some are on empirical fudge factors, and some has used intelligent information theory like fuzzy theory[2,3].

One of the most commonly used method of significance test is T-test. To cover the some problems of T-test, the SAM method was suggested. The SAM method is also commonly used method since it has been developed in 2001[3]. Another approach has been developed in recent researches that used fuzzy membership degree[2]. We described about those methods.

### 2.1 Statistics significance test

There are commonly used methods for the significance test; T-test, Wilcoxon rank sum test, etc. The T-test compares the difference between the means of two groups in relation to the variation which expressed as the standard deviation of the two groups. The Wilcoxon rank sum test is an alternative method to the two-sample t-test in nonparametric group, that is based on the order of observations[2].

The t-statistics is defined as

$$t(S_1,S_2)=\frac{|\mu_{S_1}-\mu_{S_2}|}{\sqrt{\frac{\sigma_{S_1}^2}{|S_1|}+\frac{\sigma_{S_2}^2}{|S_2|}}} \quad (1)$$

where $\mu_s$ is mean, $\sigma_s$ is standard deviation of sample $S$.

The T-test is a method to check the mean difference, so a large mean difference can make the large t-statistics. However the small variance can also generate the large t-statistics. Besides this, the T-test considers only the mean of a set but not much considers the distribution of the data set. There are several attempts to amend these problems.

### 2.2 Significance test based on empirical fudge factor

There is a method which adds a fudge factor to maximize the statistics value. The SAM(Significance Analysis of Microarrays) is a method to modify the problems of T-test. The difference between the SAM and the T-test is the score of $s_0$ which is a positive constant added to the denominator of equation $d(i)$. The value $s_0$ is used to minimize the coefficient of variation[3].

$$d(i)=\frac{x_I(i)-x_U(i)}{s(i)+s_0} \quad (2)$$

where $x_I(i)$ and $x_U(i)$ are the average of $gene(i)$ in group $I$ and $U$. The standard deviation $s(i)$ is defined as

$$s(i)=\sqrt{a\{\sum_m[x_m(i)-x_I(i)]^2+\sum_n[x_n(i)-x_U(i)]^2\}} \quad (3)$$

where $\sum_m$ and $\sum_n$ are summations of the expression values of the group $I$ and $U$.

$a=(1/n_1+1/n_2)/(n_1+n_2-2)$, $n_1$ and $n_2$ are the numbers of the group $I$ and $U$.

### 2.3 Significance test based on fuzzy theory

One another approach for the significance test is to use intelligence information theory. The FM-test(Fuzzy set theory based Method test) calculates the FM c-value (FM convergence degree) which reflects how much the object belongs to the other group not to its own group. This is a different approach compared to the SAM or the T-test. The FM-test used the exponent function to get the fuzzy membership degree of each gene expression data. With the FM c-value, FM-test takes the FM d-value (FM divergence value). The p-value of FM-test is based on the empirical p-value.

$$c(S_1,S_2)=\frac{\sum_{i=1}^{n_1+n_2}(I_{S_1}(x_i)f_{F(S_2)}(x_i)+I_{S_2}(x_i)f_{F(S_1)}(x_i))}{n_1+n_2} \quad (4)$$

$$d(S_1,S_2)=1-c(S_1,S_2) \quad (5)$$

where the $f_{F(S_1)}$ and $f_{F(S_2)}$ are the membership function for each group,

$S=S_1\cup S_2=\{x_i,i=1,...,n_1+n_2\}$, $n_1=|S_1|$, $n_2=|S_2|$, and $I_{S_i}(x)=1$

if $x\in S_i$ and 0 otherwise for $i=1,2$ 이다[1].

## 3. FIGER Significance Test

### 3.1 FIGER weighting factors

To complement the problems of existing significance test, recent researches shows the weighting factors based on fuzzy theory or empirical values[2,3]. Our FIGER focuses on the variation and the distribution of the data set.

We defined the FIGER variation ratio (FV) which uses the exponent function and the ratio of standard deviation to mean. Our FV (FIGER variation ratio) is represented by the equation (8).

$$f_{FV_{S_1}}=e^{-\left(\frac{\delta_{S_1}}{\mu_{S_1}}\right)}\times\alpha \quad (6)$$

$$f_{FV_{S_2}}=e^{-\left(\frac{\delta_{S_2}}{\mu_{S_2}}\right)}\times\alpha \quad (7)$$

$$FV(S_1,S_2)=\frac{f_{FV_{S_1}}+f_{FV_{S_2}}}{2} \quad (8)$$

where $\mu_{S_1}$ and $\mu_{S_2}$ are means of group $S_1$ and $S_2$, $\delta_{S_1}$ and $\delta_{S_2}$ are standard deviations of group $S_1$ and $S_2$. The $\alpha$ is the constant to adjust the FV. In this paper we set 1.5 for the constant $\alpha$.

The standard deviation is a statistic that tells that how the data set are close to the mean. If a data set is a normal distribution, it means that most of the data are close to the average. In statistics, the coefficient of variation(CV) is a percent of the standard deviation to the mean. In our FV definition, we computed how the standard deviation is close to the mean based on the CV. For the fuzzification, we used exponent function.

We also defined the FMD (Fuzzy Membership Degree)

which measures how much close the each object is to the mean of the group. When an object is between one standard deviation, the object has heigh membership degree.

$$FMD(S_1) = \frac{\sum_{1}^{n} fmd(x_{S_1})}{n} \tag{9}$$

$$fmd(x_{S_1}) = \begin{cases} \mu_{S_1} - \delta_{S_1} \le x_{S_1} \le \mu_{S_1} + \delta_{S_1} & then \quad 1 \\ \mu_{S_1} - 2\delta_{S_1} \le x_{S_1} \le \mu_{S_1} + 2\delta_{S_1} & then \quad 0.7 \\ \mu_{S_1} - 3\delta_{S_1} \le x_{S_1} \le \mu_{S_1} + 3\delta_{S_1} & then \quad 0.4 \\ others & then \quad 0.1 \end{cases} \tag{10}$$

$$FMD(S_2) = \frac{\sum_{1}^{m} fmd(x_{S_2})}{m} \tag{11}$$

$$fmd(x_{S_1}) = \begin{cases} \mu_{S_1} - \delta_{S_1} \le x_{S_1} \le \mu_{S_1} + \delta_{S_1} & then \quad 1 \\ \mu_{S_1} - 2\delta_{S_1} \le x_{S_1} \le \mu_{S_1} + 2\delta_{S_1} & then \quad 0.7 \\ \mu_{S_1} - 3\delta_{S_1} \le x_{S_1} \le \mu_{S_1} + 3\delta_{S_1} & then \quad 0.4 \\ others & then \quad 0.1 \end{cases} \tag{12}$$

$$FMD(S_1, S_2) = \frac{FMD(S_1) + FMD(S_2)}{2} \tag{13}$$

where $\mu_{S_1}$ and $\mu_{S_2}$ are menas of group $S_1$ and $S_2$, $\delta_{S_1}$ and $\delta_{S_2}$ are standard deviations of group $S_1$ and $S_2$. $S_1 = \{x_{S_{1i}}, i = 1, 2, .., n\}$,

$$n_1 = |S_1|, \; S_2 = \{x_{S_{2i}}, i = 1, 2, .., m\}, \; m = |S_2|.$$

In normal distribution, one standard deviation is away from the mean in either direction, that accounts for 68 percent of data set. We started the definition of the FMD values from this statistics.

### 3.2 FIGER Significance Test

Our approach was based on analysis of a statistics significance and FIGER value. The FIGER is consists of FV and FMD. The FV is the ratio of standard deviation to mean and the FMD is the membership degree of each object.

Equation (14) and (15) shows the FIGER test.

$$t_{FIGER}(S_1, S_2) = \frac{|\mu_{S_1} - \mu_{S_2}|}{\sqrt{\frac{\sigma_{S_1}^2}{|S_1|} + \frac{\sigma_{S_2}^2}{|S_2|}}} \times figer(S_1, S_2) \tag{14}$$

$$figer(S_1, S_2) = \frac{FV(S_1, S_2) + FMD(S_1 + S_2)}{2} \tag{15}$$

where $\mu_{S_1}$ and $\mu_{S_2}$ are means of group $S_1$ and $S_2$, $\delta_{S_1}$ and $\delta_{S_2}$ are standard deviations of group $S_1$ and $S_2$. $|S_1|$ and $|S_2|$ are the numbers of group $S_1$ and $S_2$.

## 4. Simulation Method and Results

### 4.1 Method

We downloaded a gene expression data set from the GEO(Gene Expression Omnibus)[3]. The GEO is a public repository for a high-throughput experimental data. The GEO has microarray-based experiments measuring mRNA, miRNA, genomic DNA, etc[4].

The downloaded data is consisted with two groups of 7,129 genes each. One group has five insulin resistant samples and the other has five insulin sensitive samples[5]. We used 1518 genes which have no null values and the expression values are greater than 100.

We used R-project which is a tool and a language for statistics computing and graphics[6]. The significance test for T-test and our FIGER-test has computed with R-project.

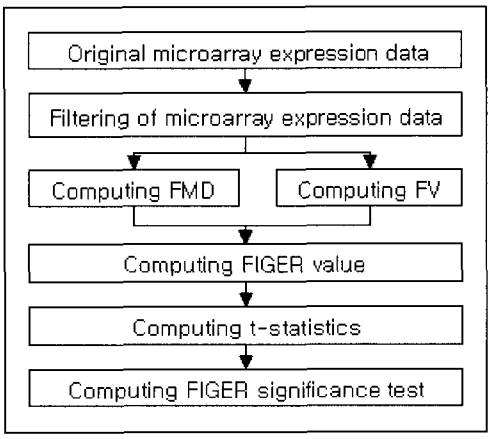


Fig. 1. The work flow of FIGER significance test

To get the FIGER value, there are two steps. One is to get the FMD value, and the other is to get the FV value. After getting the t-statistics, the FIGER is used for the variance factor for the significance test. Fig.1 is the work flow of the FIGER significance test.

### 4.2 Results

After computing the significance test of the T-test and the FIGER-test, We compared the results between FIGER-test results and T-test results. The Fig.2 shows the distribution of the p-values calculated by FIGER-test and T-test.

From the Fig.2 we can see the difference of p-values between the FIGER-test results and the T-test results. The results distribution is similar in the good data set whose p-value is <= 0.05 and the bad data set whose p-value is close to 1.0. It tells that the data variation of those regions is smaller than the data variation of other region.

Table 1. Best-ranked genes of significance test

| | a1 | a2 | a3 | a4 | a5 | b1 | b2 | b3 | b4 | b5 | ttest-p-val | figer-p-val | FIGER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M60858_rna1_at | 2134 | 2266 | 1979 | 1802 | 2207 | 2573 | 2380 | 2983 | 2719 | 2756 | 0.002 | 0.002 | 0.978 |
| L07033_at | 750 | 843 | 758 | 535 | 578 | 516 | 492 | 386 | 448 | 441 | 0.005 | 0.005 | 1.027 |
| L07648_at | 664 | 488 | 637 | 568 | 627 | 719 | 881 | 850 | 714 | 728 | 0.005 | 0.005 | 1.022 |
| M95610_at | 492 | 400 | 609 | 384 | 384 | 291 | 326 | 265 | 230 | 306 | 0.007 | 0.006 | 1.026 |
| X81003_at | 243 | 252 | 325 | 331 | 302 | 418 | 342 | 339 | 457 | 433 | 0.008 | 0.009 | 0.973 |
| X57959_at | 4859 | 3522 | 4137 | 4075 | 3488 | 5983 | 4592 | 5130 | 4855 | 5171 | 0.011 | 0.009 | 1.03 |
| Z26491_s_at | 804 | 655 | 647 | 984 | 1200 | 1127 | 1395 | 1250 | 1347 | 1126 | 0.011 | 0.011 | 1.005 |
| U68111_at | 131 | 152 | 222 | 238 | 261 | 269 | 231 | 349 | 359 | 301 | 0.02 | 0.022 | 0.977 |
| M74089_at | 541 | 650 | 529 | 368 | 435 | 220 | 430 | 316 | 316 | 391 | 0.022 | 0.023 | 0.992 |
| M32598_at | 4225 | 4095 | 6688 | 4636 | 5497 | 7589 | 6464 | 6536 | 6023 | 6166 | 0.025 | 0.02 | 1.054 |
| U43944_at | 313 | 513 | 451 | 408 | 433 | 197 | 122 | 335 | 297 | 391 | 0.029 | 0.041 | 0.916 |
| M37238_s_at | 602 | 722 | 480 | 339 | 392 | 249 | 390 | 280 | 335 | 309 | 0.03 | 0.03 | 1.001 |
| M37435_at | 607 | 776 | 779 | 384 | 333 | 416 | 148 | 304 | 348 | 287 | 0.03 | 0.038 | 0.94 |
| Z29481_at | 515 | 483 | 314 | 309 | 275 | 245 | 186 | 238 | 281 | 260 | 0.03 | 0.027 | 1.029 |
| L42611_f_at | 1563 | 1644 | 1418 | 491 | 539 | 683 | 572 | 114 | 460 | 239 | 0.031 | 0.054 | 0.865 |
| S62539_at | 270 | 358 | 280 | 339 | 246 | 419 | 350 | 756 | 748 | 412 | 0.031 | 0.049 | 0.884 |
| D28118_at | 328 | 237 | 381 | 332 | 339 | 343 | 517 | 375 | 435 | 441 | 0.032 | 0.032 | 1.003 |
| U53445_at | 604 | 649 | 716 | 557 | 523 | 440 | 547 | 562 | 501 | 374 | 0.033 | 0.03 | 1.03 |
| HG2815-HT4023_s_at | 5307 | 3930 | 4643 | 3882 | 4562 | 4098 | 3902 | 3714 | 3526 | 3270 | 0.034 | 0.023 | 1.098 |
| U36221_at | 646 | 961 | 528 | 457 | 248 | 312 | 225 | 323 | 277 | 149 | 0.034 | 0.043 | 0.945 |
| L27559_s_at | 391 | 379 | 268 | 323 | 380 | 774 | 506 | 416 | 468 | 449 | 0.035 | 0.043 | 0.942 |
| U47054_at | 881 | 674 | 939 | 1349 | 1184 | 1343 | 1110 | 1808 | 1760 | 1287 | 0.036 | 0.043 | 0.957 |
| X59834_at | 1319 | 1047 | 965 | 1552 | 1604 | 1804 | 1180 | 2245 | 1946 | 2127 | 0.038 | 0.046 | 0.95 |
| M80482_at | 160 | 116 | 158 | 166 | 154 | 222 | 198 | 233 | 327 | 149 | 0.039 | 0.051 | 0.93 |
| U90907_at | 1443 | 1520 | 1676 | 518 | 764 | 507 | 910 | 551 | 494 | 503 | 0.04 | 0.049 | 0.946 |
| D31884_at | 828 | 706 | 790 | 306 | 386 | 427 | 207 | 311 | 343 | 360 | 0.042 | 0.047 | 0.97 |
| U35048_at | 1025 | 1131 | 775 | 832 | 914 | 705 | 580 | 905 | 814 | 563 | 0.043 | 0.04 | 1.021 |
| S81003_at | 753 | 1055 | 732 | 515 | 621 | 568 | 363 | 553 | 498 | 539 | 0.046 | 0.039 | 1.043 |
| J03278_at | 395 | 727 | 506 | 344 | 354 | 109 | 377 | 296 | 230 | 318 | 0.047 | 0.057 | 0.946 |
| M35252_at | 572 | 446 | 631 | 801 | 771 | 855 | 563 | 1132 | 1065 | 1011 | 0.048 | 0.059 | 0.942 |
| M83667_rna1_s_at | 111 | 452 | 266 | 355 | 344 | 298 | 549 | 584 | 601 | 429 | 0.049 | 0.063 | 0.932 |
| S90469_at | 387 | 349 | 370 | 154 | 294 | 414 | 449 | 400 | 369 | 395 | 0.065 | 0.049 | 1.081 |

Table 2. Different significance level between FIGER-test and T-test. The a and the b are the sample groups.

| | ttest-p-val | figer-p-val | FV | std/mean a | std/mean b | FMD | FIGER |
|---|---|---|---|---|---|---|---|
| M37238_s_at | 0.03 | 0.03 | 1.183 | 0.307825 | 0.172368 | 0.82 | 1.001 |
| M37435_at | 0.03 | 0.038 | 1.06 | 0.366445 | 0.328682 | 0.82 | 0.94 |
| Z29481_at | 0.03 | 0.027 | 1.208 | 0.292629 | 0.146184 | 0.85 | 1.029 |
| HG2815-HT4023_s_at | 0.034 | 0.023 | 1.345 | 0.131372 | 0.086962 | 0.85 | 1.098 |
| U36221_at | 0.034 | 0.043 | 1.04 | 0.463177 | 0.278128 | 0.85 | 0.945 |
| S90469_at | 0.065 | 0.049 | 1.251 | 0.303704 | 0.072308 | 0.91 | 1.081 |

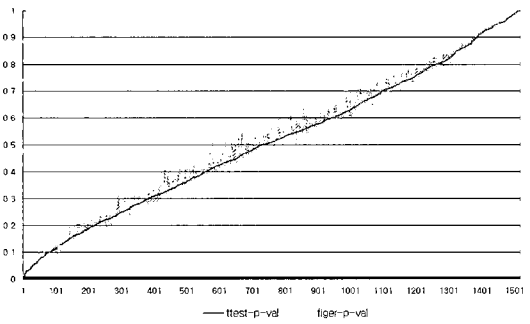


Fig. 2. FIGER-test and T-test results

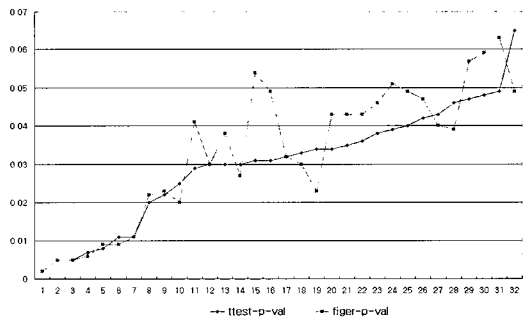


Fig. 3. 5% level of significance test

There are best-ranked genes by FIGER-test and T-test in Table 1. As shown in Table 1, most of the FIGER-test results are as good as the T-test results. The T-test results in 31 gene whose p-value is <= 0.05, the FIGER-test results in 27 genes whose p-value is <= 0.05. The number of genes whose p-values are <= 0.05 in both test is 26, the p-value of 5 genes are <= 0.05 only in the T-test, 1 gene is <= 0.05 only in the FIGER test. We showed the distribution of p-values of 5% significant level in Fig.3.

In Table 2 we showed some genes whose p-value is same in the T-test but different in the FIGER-test. The gene M37238_s_at, gene M37435_at and gene Z29481_at have p-value of 0.03 at T-test but 0.03, 0.038, 0.027 at FIGER-test. It is hard to say that the p-value of 0.027 is much more significant than the p-value of 0.03, however in the case that the p-value is almost close to 0.05 but it is larger than 0.05, the gene could not be included in significant genes group. The FIGER test can give more sophisticated results in this case based on the data

distribution. We searched papers to get the information about the gene that is not indicated as a significant gene by T-test but indicated as a significant gene by FIGER-test. The gene S90469 is p-value 0.05 in the T-test but 0.049 in the FIGER test. The gene title of S90469_at is "P450 (cytochrome) oxidoreductase" and whose location is 7q11.2. We found a paper that shows the candidate genes for insulin resistance syndrome(IRS) are located at the chromosomal region 7q11.2[7]. This tells that the FIGER can get the resonable results compared to the T-test. The FIGER can get the 26 genes out of 32 genes in common with the T-test and also the FIGER identified a candidate gene that can not be identified by the T-test.

## 5. Conclusion

We proposed a new approach, the FIGER-test, that estimates the variance of observed data set and the membership degree of each observations. The FIGER is based on the FIGER value which is combined with FV and FMD. The FV value of the FIGER-test estimates the ratio between standard deviation and mean, the FMD value of FIGER-test estimates the membership degree of each observation.

For the simulation of the FIGER test, we downloaded the gene expression data from a public repository and also we computed the result by the T-test. The original gene set are 7,129, after filtering we got 1518 gene. Form the 1518 genes, the T-test identified 32 significant genes and the FIGER-test identified 27 significant genes. Within 27 genes, 26 genes are also identified as significant gene by T-test and one gene is identified only by the FIGER-test. However we confirmed that the gene could be a candidate gene for the significant gene based on biological paper.

In this paper we showed that the different significant level can be calculated not only by the values of mean differences but also by the value of data distribution. Based on this research, we are going to apply the FIGER method to other gene expression analysis. The FIGER can be a method to regulate a weight values of data set or to find the clustering groups depending on the data distribution. The FIGER can be a useful and effective method for the gene expression data analysis.

## 6. References


[1] http://www.gene-chips.com/

[2] L.R.Liang, S.Lu, X.Wang, Y.Lu, V. Mandal, D.Patacsil, D. Kumar, "FM-test: a fuzzy-set-theory-based approach to differential gene expression data analysis," BMC Bioinformatics, Vol. 7, Suppl No. 4, S7, 2006.

[3] V.G.Tusher, R.Tibshirani, G.Chu, "Significance analysis of microarrays applied to the ionizing radiation response" PNAS, Vol.98, No.9, pp 5116 - 5121, 2001.

[4] http://www.ncbi.nlm.nih.gov/

[5] ftp://ftp.ncbi.nih.gov/pub/geo/

[6] http://www.r-project.org/

[7] R. Arya, J. Blangero, K. Williams, L. Almasy, T.D. Dyer, R.J.Leach, P.O'Connell, M.P.Stern, R.Duggirala, "Factors of Insulin Resistance Syndrome-Related Phenotypes Are Linked to Genetic Locations on Chromosomes 6 and 7 in Nondiabetic Mexican-Americans" DIABETES, Vol. 51, No. 3, pp 841-847, 2002.

[8] Sholom M. Weiss, Nitin Indurkhya, Predictive Data Mining A Practical Guide, Morgan Kaufmann Publishers, 1998.

[9] Leonard kaufman, Peter J. Rousseeuw, Finding Groups in Data, Wiley series in probability and statistics, 2005.

[10] Yuanchen He, Yuchun Tang, Yan-Qing Zhang, Raishekar Sunderraman, "Fuzzy- Granular Gene Selection from Microarray Expression Data",Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), 2006.


## 저 자 소 개


**최경욱(Kyoung Oak Choi)**

1997년 : 대구가톨릭대학교 전자계산학과 졸업(이학사)

1999년 : 대구가톨릭대학교 전산통계학과 전자계산전공 석사졸업(이학석사)

2002년 : 대구가톨릭대학교 전산통계학과 전자계산전공 박사수료

2002년~2006년 : 한국생명공학연구원 연구원

관심분야 : 퍼지이론, 생물정보, 데이터마이닝

E-mail : ok0822@chol.com

**정환묵(Hwan Mook Chung)**

제16권 제6호 참조