*The Korean Journal of Applied Statistics* (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.3.477
**28**(3), 477–483

# A Note on the Decision of Sample Size by Relative Standard Error in Successive Occasions

GeunShik Han[a,1] · Gi-Sung Lee[b]

[a]Department of Computer Engineering, Hanshin University
[b]Department of Children Welfare, Woosuk University

(Received March 2, 2015; Revised March 31, 2015; Accepted March 31, 2015)

**Abstract**

This study deals with the decision problem of sample size by the relative standard error of estimates derived from survey results in successive occasions. The population of the construction in business survey results is used to calculate quartile of the relative standard error of the 1,000 sample obtained from simple or stratified random sampling. The sample size at time $t$ with a relative standard error of the point $(t-1)$ in the successive occasions were calculated according to the sampling method. As a result, in terms of the sample size according to the size of the relative standard error of the $(t-1)$, simple random sampling differs significantly from stratified sampling. In addition, we could see differences in sample size (depending on how the population is stratified) and that careful attention is required in the problem of sample size by the relative standard error of estimates derived from survey results in successive occasions.

Keywords: successive occasions, relative standard error, sample size, simple random sampling, stratified sampling

## 1. 서론

표본설계시 고려해야하는 문제 중의 하나가 표본크기를 결정하는 것이다. 표본크기가 적정 수 이상일 경우 고비용의 문제가 발생하며 반대로 표본크기가 작을 경우 추정치의 정도가 낮아지는 문제가 발생한다. 따라서 표본크기는 사전에 결정된 목표정도를 만족시키는 범위 내에서 가능한 한 작은 것이 적절하다. 표본크기를 결정하기 전에 연구자들은 사전에 목표정도를 결정하게 되는데 이때 활용하는 목표정도는 허용오차를 응용한 것으로 모수 $\theta$와 추정값 $\hat{\theta}$과의 차이를 이용하는 절대오차 $|\hat{\theta}-\theta| = d$, 상대오차 $|(\hat{\theta}-\theta)/\theta| = d_0$, 상대표준오차 $C_{\hat{\theta}} = (S_{\hat{\theta}}/\hat{\theta}) * 100(\%)$ 등이 주로 활용된다. 이때 표본크기는 다음 식을 만족하는 $n$을 구하는 것과 같다.

$$\Pr\left(\left|\hat{\theta}-\theta\right| \geq d\right) = \alpha.$$

만약 모평균 $\mu_y$ 추정을 위한 상대오차를 $|(\bar{y}-\mu_y)/\mu_y| = d_0$라 하면 표본평균 $\bar{y}$가 정규분포를 따른다는

This research was supported by Hanshin University Research Funding.
[1]Corresponding author: Department of Computer Engineering, Hanshin University, 137, Hanshindae-gil, Osan-si, Gyeonggi-do 142-791, Korea. E-mail: gshan@hs.ac.kr

가정 하에서 $|(\bar{y}-\mu_y)/\mu_y| = d_0 = t\sqrt{(N-n)/N}\,(C_y/\sqrt{n})$을 $n$에 대해 정리하면 다음과 같은 표본크기 결정 식을 얻는다.

$$n = \frac{\left(\frac{tC_y}{d_0}\right)^2}{1+\frac{1}{N}\left(\frac{tC_y}{d_0}\right)^2}, \tag{1.1}$$

여기서 $C_y$는 모변동계수이고, $t$는 신뢰계수이다.

대부분의 경우 표본크기는 식 (1.1)을 응용하여 결정하게 된다. 그러나 통계청을 비롯한 일부기관에서 매년 시행되는 계속조사의 경우에는 Park (1989)이 제안한 과거 시점의 추정량의 변동계수와 현 시점의 추정량의 목표오차의 변동을 이용하여 다음과 같이 표본의 크기를 구하고 있다.

$$n_t = n_{t-1}\left(\frac{C_{t-1}}{C_t}\right)^2, \tag{1.2}$$

여기서 $n_t$는 구하려는 시점 $t$의 표본크기이며, $n_{t-1}$는 과거 조사 시점 $(t-1)$의 표본크기이다. $C_{t-1}$는 과거 조사 시점 $(t-1)$의 결과에서 얻은 추정값의 상대오차이며, $C_t$는 구하려는 시점 $t$의 목표오차이다.

최근 지난 시점의 자료가 존재할 때 표본의 크기에 대한 연구로는 Kim (2012)과 Park과 Na (2014)가 있다. Kim (2012)은 현 시점의 추정량의 목표오차와 과거 시점의 추정량의 변동계수와 모집단의 크기 변동을 사용한 표본의 크기 구하는 공식을 제시하였다. Park과 Na (2014)는 모집단의 변동계수와 모집단의 크기의 변동과 추정량의 목표오차, 그리고 과거 시점의 추정량의 변동계수를 사용한 표본크기 문제를 연구하였다. Kim (2012)은 표본크기를 구하는데 모집단의 크기 변동만을 반영하였으나, Park과 Na (2014)는 모집단의 산포변동도 추가로 반영하는 새로운 표본크기 공식을 연구하였다. 또한, Yoo와 Shin (2011)은 패널조사에서 비율과 총계 추정에 표본의 크기가 미치는 영향을 연구하기도 하였다.

본 연구에서는 계속조사에서 과거의 조사결과로 얻은 추정값의 상대표준오차를 이용한 표본크기 결정 문제에 대하여 실제 사업체 조사자료를 활용하여 살펴보고자 한다. 사업체 조사결과 중 건설업을 모집단으로 이용하여 표본크기를 500개에서 3,000개까지 500개씩 증가시켜가면서 표본을 1,000개씩 단순임의추출 또는 층화추출하여 추출된 각 표본으로부터 상대표준오차들을 계산한다. 그리고 이들 값들을 토대로 계속조사에서 시점 $(t-1)$에서의 상대표준오차를 이용한 시점 $t$에서의 표본크기를 추출법에 따라 구하여 비교해 봄으로써 과거의 조사결과에서 얻은 추정값의 상대표준오차를 이용한 표본크기 결정 식의 활용에 대한 제언을 하고자 한다.

## 2. 추출법에 따른 상대표준오차와 표본크기

이 절에서는 사업체 조사결과(2012년 결과) 중 건설업을 조사대상으로 선정하여, 계속조사에서 과거의 조사결과에서 얻은 추정값의 상대오차를 이용한 표본크기 결정 문제에 대하여 다루어 보고자 한다. 추정하고자 하는 모수가 매출액일 경우 매출액을 이용하여 표본크기를 결정하는 것이 바람직하지만 대개의 경우 매출액에 대한 정보가 부족하여 사업체의 종사자 수를 이용하여 표본크기를 결정하게 된다. 통계청의 2012년 사업체 조사결과, 건설업 사업체의 수는 109,201개이고, 종사자 수 1,000명 이상인 사업체는 46개로 나타났다. 종사자 규모 1,000명 이상인 사업체들의 종사자 수에 대한 변동량은 매우 크므로 전수조사를 하고 이들을 제외한 사업체 109,155개를 조사모집단으로 정의하여 표본크기 결정 문제를 다루고자 한다.

**Table 2.1.** Quartile of the relative standard error due to changes in sample size

| Quartile | Sample size($n$) | | | | | |
|---|---|---|---|---|---|---|
| | 500 | 1,000 | 1,500 | 2,000 | 2,500 | 3,000 |
| Q1 | 10.1 | 8.4 | 7.1 | 6.3 | 5.8 | 5.4 |
| median | 12.4 | 9.7 | 8.2 | 7.2 | 6.5 | 5.9 |
| Q3 | 15.6 | 11.7 | 9.2 | 8.0 | 7.1 | 6.4 |

**Table 2.2.** The sample size at the point of time $t$ by using the relative standard error in the point of time $(t-1)$ (simple random sampling)

| Sample size | The sample size at the point of time $t$ by using the relative standard error in the point of time $(t-1)$ | | |
|---|---|---|---|
| | Q1 | median | Q3 |
| 500 | 294 | 443 | 702 |
| 1,000 | 705 | 940 | 1,368 |
| 1,500 | 1,128 | 1,505 | 1,894 |
| 2,000 | 1,497 | 1,955 | 2,414 |
| 2,500 | 2,012 | 2,526 | 3,014 |
| 3,000 | 2,498 | 2,982 | 3,509 |

### 2.1. 단순임의추출에서의 표본크기

식 (1.2)의 표본크기 결정 식을 활용하기 위하여 건설업 모집단으로부터 추출되는 표본크기는 500, 1,000, 1,500, 2,000, 2,500, 3,000개로 500개씩 크기를 증가시켜가면서 단순임의 추출하였다. 크기가 $n$인 표본을 1,000개씩 추출하였으며 추출된 각 표본으로부터 상대표준오차를 계산하였다. 다음 Table 2.1은 크기가 $n$인 1,000개의 표본들로부터 계산된 상대표준오차들의 사분위수를 나타낸다.

Table 2.1에서 Q1은 제1사분위수, Q3는 제3사분위수를 나타낸다. 크기가 500인 표본을 단순임의 추출하는 경우, 상대표준오차들의 중위수는 12.4이며, 사분위수 범위는 5.5이고, 범위는 19.93으로 나타났다. 그리고 표본크기가 커질수록 상대표준오차들의 사분위수가 작아짐을 알 수 있다.

시점 $t$에서의 표본크기를 계산하기 위하여 앞에서 추출한 크기가 $n$인 1,000개의 표본들로부터 구한 상대표준오차들을 시점 $(t-1)$의 상대표준오차로 가정하였고, 시점 $t$에서 연구자가 원하는 상대표준오차는 1,000개 상대표준오차 값들의 중위수를 이용하였다.

표본크기 결정 식 (1.2)를 이용하여 크기가 $n$인 시점 $(t-1)$에서 1,000개의 표본으로부터 계산한 상대표준오차를 이용하여 구한 시점 $t$에서의 표본크기를 각 표본크기에 따른 사분위수 별로 정리하면 다음 Table 2.2와 같다.

계속조사를 시행하며 과거 시점 $(t-1)$의 모집단과 현시점 $t$에서의 모집단에 변화가 없다고 가정하자. 만약 과거 시점 $(t-1)$ 조사에서 표본크기가 500개였으며 상대표준오차가 12.4였다고 하면, 현 시점 $t$에서의 표본크기는 식 (1.2)에 의해 500개가 된다. 그러나 과거 시점 $(t-1)$에서의 상대표준오차가 제1사분위수에 해당하는 10.1이었다면 현 시점 $t$에서의 표본크기는 294개가 되며, 제3사분위수에 해당하는 15.6이었다면 현 시점 $t$에서의 표본크기는 702개가 된다. 즉, 크기가 500개인 표본을 위 모집단으로부터 단순임의 추출하는 경우 $\binom{N}{n} = \binom{109,155}{500}$개의 가능한 표본 중 어느 표본을 활용하여 실사를 하느냐에 따라 조사모집단의 변동이 전혀 없는 상태에서도 표본크기는 매우 큰 변동을 보이고 있다.

표본크기의 변화에 따른 상대표준오차를 상자와 수염 그림으로 나타내 보면 다음 Figure 2.1과 같다.

Table 2.1과 Figure 2.1의 상자와 수염 그림에서처럼 같이 표본크기가 500에서 3,000개로 증가하면서

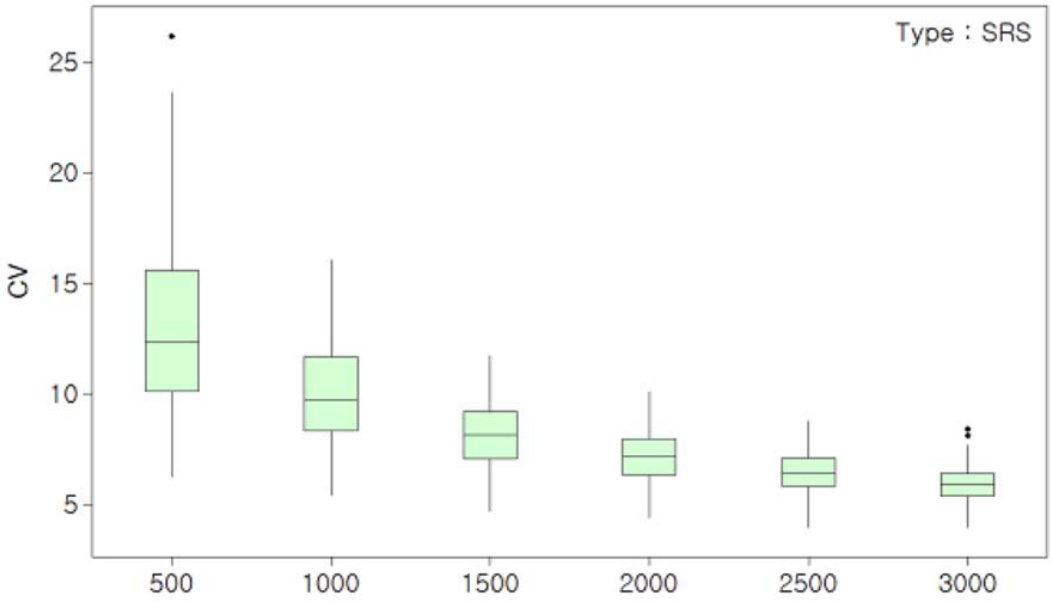


**Figure 2.1.** Box plot of the relative standard error due to sampling size change (simple random sampling).

**Table 2.3.** Population distribution according to the size of construction workers

| workers size category | $N$ | mean | std |
|---|---|---|---|
| 0∼4 | 67,121 | 2.029 | 1.05 |
| 5∼9 | 22,273 | 6.504 | 1.35 |
| 10∼49 | 16,934 | 18.755 | 9.34 |
| 50∼249 | 2,541 | 94.184 | 45.12 |
| 250∼999 | 286 | 438.584 | 188.99 |
| 1000∼ | 46 | - | - |

상대표준오차의 범위가 좁아지는 것을 볼 수 있으나 표본크기의 변동은 적지 않다는 것을 볼 수 있다.

### 2.2. 층화추출에서의 표본크기

앞서 단순임의 추출에서 사용한 건설업 사업체 109,201개를 종사자 규모별로 층화하면 다음 Table 2.3과 같다. 사업체 수는 종사자 규모 5명 미만이 67,121개로 가장 많으며 1,000명 이상 사업체는 46개인 것으로 나타났다. 종사자 규모 1,000명 이상인 층의 사업체들의 종사자수에 대한 변동량은 매우 크므로 여기에서는 이들을 전수층으로 처리하는 것으로 가정하고 표본 추출에서 제외하기로 하였다.

앞서 단순임의 추출에서 사용한 조사모집단을 종사자 규모별로 층화 후 네이만 할당한 표본크기는 다음 Table 2.4와 같다. 네이만 할당의 경우 표본크기가 2,500개 이상부터 종사자 규모 250명 이상에서 전수층이 형성되었다.

식 (1.2)의 표본크기 결정 식을 활용하기 위하여 건설업 모집단으로부터 추출되는 표본크기는 500, 1,000, 1,500, 2,000, 2,500, 3,000개로 500개씩 크기를 증가시켜가면서 층화추출하였다. 크기가 $n$인 표본을 1,000개씩 추출하였으며 추출된 각 층의 표본으로부터 상대표준오차를 계산하였다. 다음 Table

**Table 2.4.** Sample size according to the size of construction workers

| Workers size category | Sample size($n$) | | | | | |
|---|---|---|---|---|---|---|
| | 500 | 1,000 | 1,500 | 2,000 | 2,500 | 3,000 |
| 0∼4 | 83 | 165 | 248 | 330 | 419 | 513 |
| 5∼9 | 35 | 71 | 106 | 141 | 179 | 219 |
| 10∼49 | 185 | 370 | 555 | 740 | 937 | 1,149 |
| 50∼249 | 134 | 268 | 402 | 536 | 679 | 833 |
| 250∼999 | 63 | 126 | 189 | 253 | 286 | 286 |

**Table 2.5.** Range and quartile range of the relative standard error according to the sample size

| | Sample size | | | | | |
|---|---|---|---|---|---|---|
| | 500 | 1,000 | 1,500 | 2,000 | 2,500 | 3,000 |
| Range | 1.387 | 0.691 | 0.463 | 0.355 | 0.378 | 0.306 |
| Quartile range | 0.314 | 0.167 | 0.113 | 0.076 | 0.068 | 0.068 |

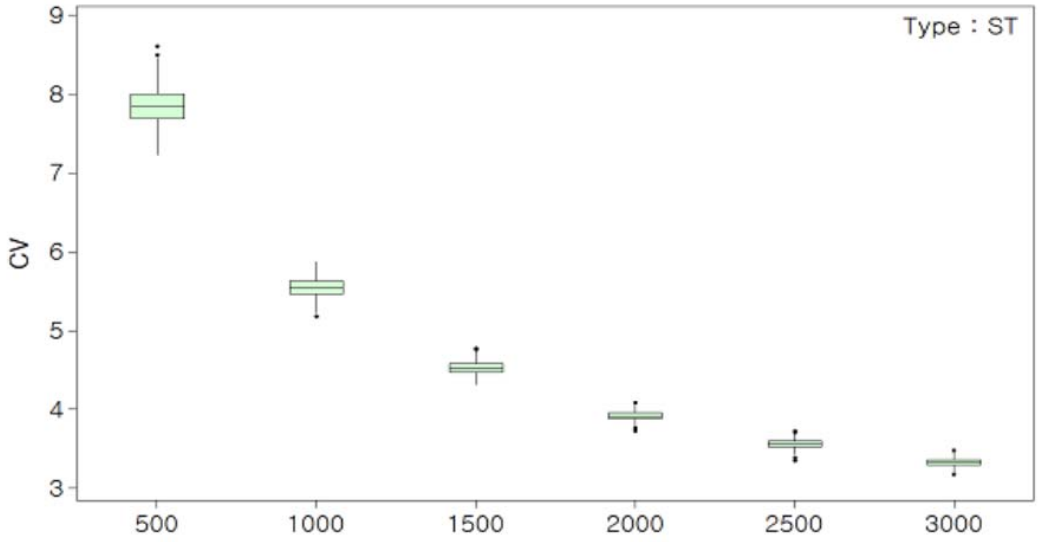


**Figure 2.2.** Box plot of the relative standard error due to sampling size change (stratified sampling).

2.5는 크기가 $n$인 1,000개의 표본들로부터 계산된 상대표준오차들의 범위 및 사분위수 범위를 나타낸다. 크기가 500인 표본을 층화추출하는 경우, 상대표준오차들의 범위는 1.387이며, 사분위수 범위는 0.314로 나타났다. 그리고 표본크기가 커질수록 상대표준오차들의 범위와 사분위수 범위가 작아짐을 알 수 있다.

표본크기의 변화에 따른 상대표준오차를 상자와 수염 그림으로 나타내 보면 다음 Figure 2.2와 같다.

표본크기 결정 식 (1.2)를 이용하여 크기가 $n$인 시점 $(t-1)$에서 1,000개의 표본으로부터 계산한 상대표준오차를 이용하여 구한 시점 $t$에서의 표본크기를 각 표본크기에 따른 사분위수 별로 정리하면 다음 Table 2.6과 같다. Table 2.6에서 보는바와 같이 층화추출의 경우, 조사모집단의 변화가 없다는 가정 하에서 식 (1.2)를 이용하여 시점 $t$에서의 표본크기를 결정하였을 때 시점 $(t-1)$에서 실사에 활용할 수 있는 가능한 표본들로부터 계산된 상대표준오차의 변동량이 작아 추정된 표본크기에도 변동량이 작다는

**Table 2.6.** The sample size at the point of time $t$ by using the relative standard error in the point of time $(t-1)$ (stratified sampling)

| Sample size | The sample size at the point of time $t$ by using the relative standard error in the point of time $(t-1)$ | | |
|---|---|---|---|
| | Q1 | median | Q3 |
| 500 | 481 | 500 | 521 |
| 1,000 | 970 | 1,000 | 1,030 |
| 1,500 | 1,462 | 1,505 | 1,537 |
| 2,000 | 1,962 | 1,997 | 2,040 |
| 2,500 | 2,453 | 2,502 | 2,548 |
| 3,000 | 2,941 | 3,004 | 3,063 |

것을 볼 수 있다.

식 (1.2)를 이용하여 표본크기를 결정하였을 때 시점 $t$에 활용할 표본크기가 시점 $(t-1)$에서의 표본크기와의 차이가 $\pm$50개 정도인 경우 실제 표본크기를 증가(감소)시켜야 할지는 고민할 필요가 있다.

## 3. 결론 및 제언

본 연구에서는 계속조사에서 과거의 조사결과에서 얻은 추정값의 상대표준오차를 이용한 표본크기 결정 문제에 대하여 실제 사업체 조사자료를 활용하여 살펴보았다. 사업체 조사결과 중 건설업을 모집단으로 이용하여 표본크기를 500개에서 3,000개까지 500개씩 증가시켜가면서 표본을 1,000개씩 단순임의추출 또는 층화추출하여 추출된 각 표본으로부터 상대표준오차들의 사분위수를 계산하였다. 그리고 이들 값들을 토대로 계속조사에서 과거 시점 $(t-1)$에서의 상대표준오차를 이용한 시점 $t$에서의 표본크기를 추출법에 따라 제시하였다. 그 결과 단순임의추출의 경우는 과거 시점 $(t-1)$에서의 상대표준오차의 크기에 따라 표본크기가 매우 크게 차이가 나타남을 알 수 있었으며, 층화추출의 경우에는 단순임의추출보다는 크게 차이가 나타나지는 않았다. 하지만 이 두 추출법에 대한 비교는 조사모집단의 변화가 없다는 가정 하에 이루어진 결과이므로 매년 조사모집단의 특성에 변화가 심하고 이를 표본에 반영하기 위해서는 표본재설계가 시행될 필요가 있다.

본 연구를 통해 계속조사에서 과거의 조사결과에서 얻은 추정값의 상대표준오차를 이용한 표본크기 식을 활용하는데 있어서 주의를 기울일 필요가 있음을 알 수 있었다. 특히, 단순임의추출의 경우 과거의 조사결과에서 얻은 추정값의 상대표준오차를 이용한 표본크기가 매우 큰 차이가 있고, 층화추출의 경우도 어떻게 층화를 하느냐에 따라 표본크기에 차이가 있을 수 있으므로 표본크기 식 활용에 세심한 주의가 필요하다.

## References

Cochran, W. G. (1977). *Sampling Techniques*, 3rd Ed., John Wiley and Sons, New York.

Park, H. A. and Na, S. R. (2014). Decision of sample size on successive occasions, *The Korean Journal of Applied Statistics*, **27**, 513–521

Kim, K. S. (2012). Sample size determination in repeated surveys with varying population sizes, *Survey Research*, **13**, 159–174.

Park, H. N. (1989). *Statistical Survey (2nd Edition)*, Youngji Publishers, Seoul.

Yoo, Y. and Shin, K. L. (2011). A study on the decision of sample size for panel survey design, *The Korean Journal of Applied Statistics*, **24**, 25–34.

# 계속조사에서 상대표준오차를 이용한 표본크기 결정에 관한 고찰

한근식[a,1] · 이기성[b]

[a]한신대학교 컴퓨터공학부, [b]우석대학교 아동복지학과

(2015년 3월 2일 접수, 2015년 3월 31일 수정, 2015년 3월 31일 채택)

**요 약**

본 연구에서는 계속조사에서 과거의 조사결과에서 얻은 추정값의 상대표준오차를 이용한 표본크기 결정 문제에 대하여 실제 사업체 조사자료를 활용하여 살펴보았다. 통계청 사업체 조사결과 중 건설업을 모집단으로 이용하여 표본크기를 500에서 3,000까지 500씩 증가시켜가면서 표본을 1,000개씩 단순임의추출 또는 층화추출하여 추출된 각 표본으로부터 상대표준오차들의 사분위수를 계산하였다. 그리고 이들 값들을 토대로 계속조사에서 시점 $(t-1)$에서의 상대표준오차를 이용한 시점 $t$에서의 표본크기를 추출법에 따라 구하였다. 그 결과 단순임의추출의 경우는 층화추출의 경우보다 시점 $(t-1)$에서의 상대표준오차들의 크기에 따라 표본크기가 매우 크게 차이가 나타남을 알 수 있었으며, 층화추출의 경우도 어떻게 층화를 하느냐에 따라 표본크기에 차이가 있을 수 있음을 알 수 있었다. 따라서 계속조사에서 과거의 조사결과에서 얻은 추정값의 상대표준오차를 이용한 표본크기 식을 활용하는데 있어서 세심한 주의가 필요함을 확인할 수 있었다.

주요용어: 계속조사, 상대표준오차, 표본크기, 단순임의추출, 층화추출

이 논문은 한신대학교 학술연구비 지원에 의하여 연구되었음.

[1]교신저자: (142-791) 경기도 오산시 한신대길 137, 한신대학교 컴퓨터공학부. E-mail: gshan@hs.ac.kr