한국산학기술학회논문지
Vol. 9, No. 5, pp. 1249-1254, 2008

변분 근사화 분포의 유도 및 변분 베이지안 가우시안 혼합 모델의 구현

이기성[1*]

Implementation of Variational Bayes for Gaussian Mixture Models and Derivation of Factorial Variational Approximation

Gi-Sung Lee[1*]

요 약 그래프 모델에서 가장 중요한 부분은 관찰 데이터가 주어진 상황에서은닉 변수와 더불어 파라미터의 사후 확률 분포의 계산이다. 이 논문에서는 가우시안 혼합 모델에 대한 변분 베이지안 방법의 구현과 변분 근사화 분포의 분해 유도를 제안한다. 이 방법은 정보 검색이나 데이터 시각화와 같은 데이터 분석 등에 적용이 가능하다.

Abstract The crucial part of graphical model is to compute the posterior distribution of parameters plus with the hidden variables given the observed data. In this paper, implementation of variational Bayes method for Gaussian mixture model and derivation of factorial variational approximation have been proposed. This result can be used for data analysis tasks like information retrieval or data visualization.

Key Words : Gaussian mixture, variational Bayes, EM algorithm

1. 서론

그래프 모델(graphical model) 학습 기법 중 가장 대표적인 방법이 최대 우도(maximum likelihood) 방법이다[1]. 학습 데이터가 주어진 경우 최대 우도 방법은 주어진 그래프 모델에 대한 모델 파라미터를 학습 데이터를 사용하여 추정하게 된다. 최대 우도 방법의 단점은 데이터에 지나친 overfitting을 하게 되어 복잡한 모델이 결과로서 얻어진다는 점이다.

혼합 모델(mixture model)은 주어진 데이터의 분류(classification) 또는 확률 밀도 추정과 같은

unsupervised learning에 많이 사용되는 통계 기법이다. 특히 정보 검색이나 컴퓨터 비전과 같은 분야에서 단어의 분류 또는 장면에서 특정 객체의 추출과 같은 응용 분야에 사용되고 있다. 혼합 모델을 사용함에 있어 주어진 데이터에 특정 분포(모델)를 가정하고 그 분포의 파라미터를 추정해 나가는 방법으로 혼합 모델을 사용하게 되는데 이 때 최대 우도 방법 또는 베이지안 방법을 사용하게 된다.

베이지안 방법은 특정 모델에 집중하기보다는 여러 모델에 대한 데이터 조건부 사후 확률을 구한 뒤 이 확률을 일종의 가중치로 사용하여 각 모델이 데이터를 얼마나 잘 반영하는지 가중치로 사용하게 된다[2]. 따라서 베이지안 방법을 사용하면 최대 우도 방법이 가지고 있는 단점을 어느 정도 해결할 수 있다. 베이지안 방법은 이와 같은 장점에도 불구하고 아주 간단한 모델의 경우에도 수학적인 계산량이 많아 사용에 제약이 따르고 있다[3]. [4]에서는 변분 베이지안 방법을 제안했는데 이 방법은 통계 물리학 분야에서 사용하는 mean-field 이론에 바탕을 두고 있다. [5,6]는 베이지안 파라미터 추정과 모델 비교에 변분법을 사용하였다.

2. 변분 베이지안 방법론

2.1 변분법

17세기 스위스의 수학자 베르누이가 최단강하선 문제

이 논문은 2008년 호원대학교 교내연구비의 지원에 의하여 연구되었음.
[1]호원대학교 컴퓨터게임학부
*교신저자: 이기성(ygslee@howon.ac.kr)
접수일 08년 07월 25일 수정일 08년 09월 30일 게재확정일 08년 10월 16일

(brachistochrone problem)를 제시하면서, 변분법에 대한 연구가 시작되었다. 최단강하선 문제는 위 아래로 떨어진 두 점 사이를 마찰 없이 중력에 의해 가장 빨리 이동할 수 있는 경로는 무엇인가의 문제였으며 직선이 아닌 사이클로이드 곡선을 따라 움직이는 경로가 해답으로 알려져 있다. 일반적으로 변분법에 관련된 문제는 주어진 적분을 최대 또는 최소화하는 특정 함수를 찾는 문제의 형태이다.

x-y 평면의 두 점 $(x_1,y_1),(x_2,y_2)$을 가장 최단 거리로 연결하는 방법은 직선이라는 것 또한 변분법을 통해 증명할 수 있다. 먼저 두 점 $(x_1,0)$과 $(x_2,0)$을 연결하는 임의의 곡선(이차 미분이 가능)을 $\eta(x)$라고 정의한다. 두 점 (x_1,y_1)과 (x_2,y_2)을 최단 거리로 연결하는 곡선을 $y(x)$라고 하면 함수 $Y(x)$를 다음과 같이 정의한다.

$$Y(x)=y(x)+\epsilon\eta(x) \tag{1}$$

ϵ는 파라미터로 정의되었으며 $Y(x)$는 두 점 (x_1,y_1)과 (x_2,y_2)을 잇는 임의의 곡선이 된다. 따라서 다음 수식을 최소화하는 $Y(x)$을 찾는 것이 주어진 문제이다.

$$I=\int_{x_1}^{x_2}\sqrt{1+Y'}\,dx \tag{2}$$

I는 ϵ의 함수이며($\epsilon=0$일 때 $Y=y(x)$) $\epsilon=0$인 경우 I가 최소값을 가져야 하므로 다음을 만족한다.

$$\frac{dI}{d\epsilon}=0, when\,\epsilon=0 \tag{3}$$

식 을 ϵ에 대해 미분하면 다음 결과를 얻는다.

$$\frac{dI}{d\epsilon}=\int_{x_1}^{x_2}\frac{1}{2}\frac{1}{\sqrt{1+Y'}}2Y'\left(\frac{dY'}{d\epsilon}\right)dx \tag{4}$$

식 을 x에 대해 미분하면 다음 결과를 얻는다.

$$Y'(x)=y'(x)+\epsilon\eta'(x)\Rightarrow\frac{dY'}{d\epsilon}=\eta'(x) \tag{5}$$

위의 수식을 그 위의 수식에 대입한 뒤 $\epsilon=0$일 때 $dI/d\epsilon=0$을 풀면 다음의 수식을 얻는다.

$$\left(\frac{dI}{d\epsilon}\right)_{\epsilon=0}=\int_{x_1}^{x_2}\frac{y'(x)\eta'(x)}{\sqrt{1+y'^2}}dx=0 \tag{6}$$

부분 적분 공식을 사용하여 위 수식을 적분해서 풀면 다음 결과를 얻는다.

$$\frac{d}{dx}\left(\frac{y'}{\sqrt{1+y'^2}}\right)=0 \tag{7}$$

위 결과에서 $y'/\sqrt{1+y'^2}$가 상수임을 알 수 있으며 이는 $y(x)$는 직선이라는 것을 알 수 있다.

일반적으로 변분법은 최대값 또는 최소값 을 찾고자하는 적분 형태를 파악하고 다음의 오일러 방정식을 사용한 미분방정식을 푸는 방식으로 진행된다.

$$I=\int_{x_1}^{x_2}F(x,y,y')dx \tag{8}$$

$$\frac{d}{dx}\frac{\partial F}{\partial y'}-\frac{\partial F}{\partial y}=0 \tag{9}$$

2.2 가우시안 혼합 모델

가우시안(Gaussian) 혼합 모델(mixture model)은 확률밀도 측정, 클러스터(cluster) 분석 등에 널리 사용되는 통계 기법이다. 하나의 가우시안 분포만을 사용한 데이터 모델링이 비현실적인 경우, 가우시안 혼합 모델을 도입하면 좀 더 유연한 모델링이 가능하다. 가우시안 혼합 모델은 다음 수식과 같이 K개의 서로 다른 가우시안 밀도 함수를 선형적으로 결합한 형태이다.

$$p(X)=\sum_{k=1}^{K}\pi_k N(X|\mu_k,\Sigma_k) \tag{10}$$

가우시안 밀도 함수 $N(X|\mu_k,\Sigma_k)$는 파라미터가 μ_k,Σ_k인 혼합 모델의 k번째 성분이 된다. π_k는 $\pi_k=p(z_k=1)$로 정의하는데, z_k는 차원이 K인 벡터 Z의 k번째 성분에 해당된다. 벡터 Z의 각 성분은 0 또는 1의 값을 취하게 되며 다음을 반드시 만족해야 한다.

$$\sum_{k}^{K}z_k=1 \tag{11}$$

위의 식에서 벡터 Z의 K개의 성분 중에서 하나의 성분만 1이 됨을 알 수 있다. 따라서 π_k는 다음을 만족해야

한다.

$$\sum_{k=1}^{K}\pi_k = 1 \quad 0 \le \pi_k \le 1 \tag{12}$$

벡터 Z의 확률을 다음과 같이 정의할 수 있다.

$$p(Z) = \prod_{k=1}^{K}\pi_k^{z_k} \tag{13}$$

벡터 Z의 k번째 성분이 1이라는 의미는 현재 데이터가 k번째 가우시안 밀도에 의해 생성되었다는 것을 의미하며 따라서 다음과 같은 조건부 확률을 정의할 수 있다.

$$p(X|z_k = 1) = N(X|\mu_k, \Sigma_k) \tag{14}$$

또는

$$p(X|Z) = \prod_{k=1}^{K}N(X|\mu_k, \Sigma_k)^{z_k} \tag{15}$$

따라서 $p(X)$는 다음과 같이 가우시안 혼합 모델의 형태로 정의가 된다. 즉, 가우시안 혼합 모델은 새로운 변수 Z의 도입으로 유도할 수 있다.

$$\begin{aligned} p(X) &= \sum_{Z} p(Z)p(X|Z) \\ &= \sum_{k=1}^{K}\pi_k N(X|\mu_k, \Sigma_k) \end{aligned} \tag{16}$$

N개의 데이터 $\{X_1,...,X_N\}$의 경우, 각각의 데이터 포인트에 대해 N개의 K차원 벡터 $Z_1,...,Z_N$가 있으며, 이 경우 n번째 데이터 X_n에 대한 벡터 Z_n의 성분은 $z_{n1}, z_{n2},..., z_{nK}$가 된다. 특정 데이터 X가 k번째 가우시안 밀도 함수에 포함될 확률(또는 k번째 가우시안 밀도 함수가 X를 생성했을 확률)을 다음과 같이 정의한다.

$$p(z_k = 1|X) = \frac{p(z_k = 1)p(X|z_k = 1)}{\sum_{j=1}^{K}p(z_j = 1)p(X|z_j = 1)} \tag{17}$$

N개의 데이터 포인트 $X_1,...,X_N$ 관찰 결과가 주어진 경우, 가우시안 혼합 모델을 사용하여 각각의 데이터가 어떤 가우시안 성분에 해당되는지 추정하는 과정을 통해 클러스터링과 같은 분석을 하게 된다. 이를 위해 maximum likelihood, Bayesian 분석 등과 같은 기법을 사용하게 된다.

파라미터 집합 $\pi = \{\pi_1,...,\pi_K\}$, $\mu = \{\mu_1,...,\mu_K\}$, $\Sigma = \{\Sigma_1,...,\Sigma_K\}$과 N개의 데이터 집합 $\{X_1,...,X_N\}$이 주어진 경우 로그 우도(likelihood) 함수는 다음과 같이 정의 된다. 아래 수식에서 각 데이터 포인트는 i.i.d. 샘플로 가정한다.

$$\log p(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N}\log\left\{\sum_{k=1}^{K}\pi_k N(X_n|\mu_k, \Sigma_k)\right\} \tag{18}$$

2.3 변분 근사화 분포의 유도

변분 근사화 분포의 유도 과정은 기존의 논문이나 책에서 자세히 기술되지 않은 것이 사실이다. 특히 분해 변분 근사화가 실제로 어떻게 사용되는지에 대한 과정이 자세히 기술되지 않았다. 이 절에서는 변분 근사화 분포의 유도 과정을 제시하고 추정할 파라미터를 유도한다.

데이터 집합 $X = \{X_1,...,X_N\}$, 이진 벡터 집합 $Z = \{z_1,...,z_N\}$의 n번째 벡터 z_n의 k번째 성분 z_{nk} $(n = 1,...,N, k = 1,...,K)$, 혼합 분포 가중치 집합 $\pi = \{\pi_1,...,\pi_K\}$이 주어진 경우 조건부 확률 $p(Z|\pi)$은 다음과 같다.

$$p(Z|\pi) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}} \tag{19}$$

한편 파라미터와 Z변수가 주어진 데이터 X의 조건부 확률 $p(X|Z,\mu,\Lambda)$는 다음과 같다.

$$\begin{aligned} &p(X|Z,\mu,\Lambda) = \prod_{n=1}^{N}\prod_{k=1}^{K}N(X_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}, \\ &\Lambda = \{\Lambda_k\},\ \Lambda_k^{-1} \equiv \Sigma_k \end{aligned} \tag{20}$$

파라미터 μ,Λ,π의 prior는 다음과 같이 정의한다.

$$p(\pi) = Dir(\pi|\alpha_0) = C(\alpha_0)\prod_{k=1}^{K}\pi_k^{\alpha_0 - 1} \tag{21}$$

각각의 가우시안 성분에 대한 파라미터 μ,Λ에 대한 prior는 다음과 같이 정의한다.

$$p(\mu,\Lambda)=p(\mu|\Lambda)p(\Lambda) \tag{22}$$
$$=\prod_{k=1}^{N}N(\mu_k|m_0,(\beta_0\Lambda_k)^{-1})\,W(\Lambda_k|W_0,\nu_0)$$

$W(\bullet|a,b)$는 파라미터가 a,b인 Wishart분포이다. m_0의 값은 대칭성을 위해 **0**으로 설정된다.

모델 파라미터에 대한 prior가 정의 되었으므로 모델 내의 모든 확률 변수(파라미터, 은닉 변수 포함)에 대한 결합 확률을 변수 간의 의존 구조를 고려해 표시하면 다음과 같다.

$$p(X,Z,\pi,\mu,\Lambda)=p(X|Z,\mu,\Lambda)p(\pi)p(\mu|\Lambda)p(\Lambda) \tag{23}$$

위 결합 확률 분포는 수학적인 계산 또는 처리에 어려움이 있으므로 근사 분포 q를 사용하여 다음과 같이 분해(factorization)한 다음 로그 우도(likelihood) 함수를 최적화 하는 방법을 사용한다.

$$q(Z,\pi,\mu,\Lambda)=q(Z)q(\pi,\mu,\Lambda) \tag{24}$$

이와 같은 방법을 "분해 변분 근사화(factorial variational approximation)"라고 한다. 파라미터 θ에 임의의 분포를 적용하면 로그 증거(evidence)는 Jensen's Inequality 법칙에 의해 다음과 같은 lower bound를 가지게 된다.

$$\begin{aligned}\log p(X)&=\log\int p(X,\theta)d\theta \\ &=\log\int q(\theta)\frac{p(X,\theta)}{q(\theta)}d\theta \\ &\geq\int q(\theta)\log\frac{p(X,\theta)}{q(\theta)}d\theta \\ &=F(q)\end{aligned} \tag{25}$$

예를 들어 $q(\theta)=\prod_{i}^{3}q(\theta_i)$인 경우를 살펴보면 $F(\theta)$는 다음과 같다.

$$F(\theta)=\int q_1q_2q_3\log\frac{p(X,\theta)}{q_1q_2q_3}d\theta_1d\theta_2d\theta_3 \tag{26}$$

식 (26)을 최대화하는 것은 제약 조건 $\int q(\theta)d\theta=1$을 만족시켜야 하는 최적화 문제이며 변분법을 사용하게 되는 부분이 되기도 한다. 먼저 새로운 함수 $g(\theta)$을 다음과 같이 정의한다.

$$g(\theta)=\int q_1(\theta'_1)q_2(\theta'_2)q_3(\theta'_3)d\theta'_1d\theta'_2d\theta'_3 \tag{27}$$

Newton 표기법 $\dot{g}$을 써서 다음 결과를 얻을 수 있다.

$$\dot{g}-q_1q_2q_3=0 \tag{28}$$

식 (26)의 $q_1q_2q_3\log\frac{p(X,\theta)}{q_1q_2q_3}$를 $h(q_1,q_2,q_3,\theta)$로 정의하고 Lagrange 승수 λ를 써서 최대화 시키려는 함수를 다음과 같이 정의한다.

$$\begin{aligned}&h_a(q_1,q_2,q_3,\theta,g,\lambda)= \\ &q_1q_2q_3\log\frac{p(X,\theta)}{q_1q_2q_3}+\lambda(\dot{z}-q_1q_2q_3)\end{aligned} \tag{29}$$

$F(q)$를 분해된 분포 q_1,q_2,q_3에 대해 최대화시키기 위해서 다음의 오일러 방정식을 사용한다.

$$\begin{aligned}\frac{\partial g_a}{\partial q_i}-\frac{d}{d\theta}\left(\frac{\partial g_a}{\partial\dot{q}_i}\right)=0 \\ \frac{\partial g_a}{\partial g}-\frac{d}{d\theta}\left(\frac{\partial g_a}{\partial\dot{g}}\right)=0\end{aligned} \tag{30}$$

위의 식에서 $\dot{q}\equiv dq_i/d\theta$를 사용하였다. 위의 식을 q_i에 대해 풀면 다음과 같은 결과를 얻는다. 아래 수식에서 상수 값은 정규화 상수 값에 해당된다.

$$\log q_i=E_{j\neq i}[\log p(X,\theta)]+Const \tag{31}$$

이 결과를 사용하여 수식 (24)에서 분해된 두 개의 분포 $q(Z)q(\pi,\mu,\Lambda)$를 다음과 같이 표현할 수 있다.

$$\begin{aligned}&q^*(Z)=\prod_{n=1}^{N}\prod_{k=1}^{K}r_{nk}^{z_{nk}}, \\ &r_{n_k}=\frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}, \\ &\ln\rho_{nk}=E[\ln\pi_k]+E[\ln|\Lambda_k|]-\frac{D}{2}ln(2\pi) \\ &-\frac{1}{2}E_{\mu_k,\Lambda_k}[(X_n-\mu_k)^T\Lambda_k(X_n-\mu_k)]\end{aligned} \tag{32}$$

$$q(\pi,\mu,\Lambda)=q(\pi)\prod_{k=1}^{K}q(\mu_k,\Lambda_k) \tag{33}$$

식 33의 최적화 해는 다음과 같이 구할 수 있다.

$$q^*(\pi)=Dir(\pi|\alpha),\alpha_k=\alpha_0+N_k \tag{34}$$

$$q^*(\mu_k,\Lambda_k)=N(\mu_k|m_k,(\beta_k\Lambda_k)^{-1})\,W(\Lambda_k|W_k,\nu_k) \tag{35}$$

위 수식에서 사용된 파라미터는 다음과 같다[7].

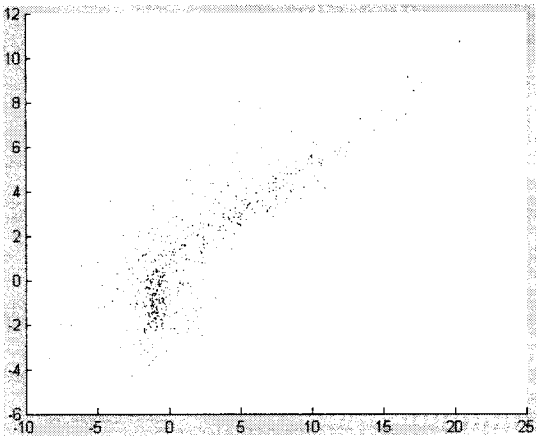
$$\begin{aligned}
&\beta_k=\beta_0+N_k, \qquad (36)\\
&m_k=\frac{1}{\beta_k}(\beta_0 m_0+N_k\overline{X_k}),\\
&W_k^{-1}=W_0^{-1}+N_kS_k+\\
&\frac{\beta_0 N_k}{\beta_0+N_k}(\overline{X_k}-m_0)(\overline{X_k}-m_0)^T\\
&\nu_k=\nu_0+N_k\\
&N_k=\sum_{n=1}^{N}r_{nk}\\
&\overline{X_k}=\frac{1}{N_k}\sum_{n=1}^{N}r_{nk}X_n\\
&S_k=\frac{1}{N_k}\sum_{n=1}^{N}r_{nk}(X_n-\overline{X_k})(X_n-\overline{X_k})^T
\end{aligned}$$

4. 실험

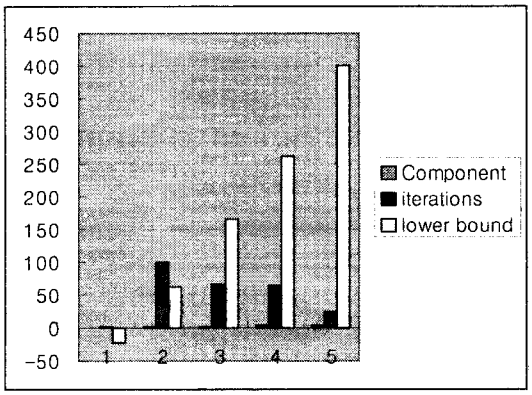
본 논문에서는 기존의 EM 알고리즘과 달리 변분 베이지안 EM 알고리즘을 2차원 데이터에 적용하였다. 2.3절에서 유도한 파라미터 업데이트 방법을 적용한 변분 베이지안 EM 알고리즘을 적용하였는데 구성 성분의 개수를 달리 하면서 로그 우도 값을 비교하였다. 그림 1은 실험에 사용된 데이터를 출력한 것이다. 데이터는 2차원의 인위적인(artificial) 데이터이며 각 데이터는 두 개의 클래스 중의 하나의 클래스에 속하게 된다.

실험은 가우시안 성분의 개수를 1,2,3,4,5인 경우에 대해 로그 우도 값의 lower bound를 측정하는 방식으로 진행했다. 실험의 결과 성분의 개수는 5개가 주어진 데이터를 가장 잘 설명하는 것으로 나타났다.

BIC(Bayesian Information Criterion)의 최소화 원칙에 의해서 적용했을 때에도 5개의 성분 개수가 가장 데이터를 잘 설명하는 것으로 나타났다. 그림 2에서 x축은 성분의 개수를 나타내며 iterations이라고 표기된 부분은 로그 우도 값이 수렴하기 까지 필요한 반복의 개수를 나타낸다. Lower bound는 변분 베이지안 방법론에서 최대화 하려는 목표 값에 해당한다.



[그림 1] 실험에 사용된 데이터



[그림 2] 성분 개수 및 lower bound

5. 결론

본 논문에서는 변분법을 이용한 변분 근사화 분포의 유도를 제시하였으며 변분 가우시안 혼합 모델을 구현하여 2차원 데이터에 대한 학습에 적용하였다. 혼합 모델은 가장 많이 사용하는 것이 가우시안 혼합 모델이지만 이외에도 포아송 혼합 모델, 감마 혼합 모델 등이 존재한다. 각각의 어플리케이션에 따라 어떤 혼합 모델이 가장 적합한 모델인지 선택하는 문제와 정보검색이나 컴퓨터 비전과 같은 분야에 혼합 모델을 적용함에 있어 앞으로 더 다양한 혼합 모델의 연구 및 개발이 필요하다고 할 것이다.

참고문헌

[1] I. J. Myung, Tutorial on maximum likelihood estimation, Journal of Mathematical Psychology, Vol. 47(1): 90-100, 2003.

[2] Peter M. Lee, *Bayesian Statistics: An Introduction*, Arnold Publication, 2004.

[3] H. Attias, Independent Factor Analysis, *Neural Computations*, 11: 803-851, 1999.

[4] H. Attias, "Inferring Parameters and Structure of Latent Variable Models by Variational Bayes," In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

[5] D.J.C. Mackay, Ensemble Learning and Evidence Maximization, Technical report, Cavendish Laboratory, University of Cambridge, 1995.

[6] D.J.C. Mackay, Ensemble Learning for Hidden Markov Models, Technical report, Cavendish Laboratory, University of Cambridge, 1998.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

이 기 성(Gi-Sung Lee) **[종신회원]**

- 1993년 2월 : 숭실대학교 컴퓨터학과 (공학사)
- 1996년 2월 : 숭실대학교 컴퓨터학과 (공학석사)
- 2001년 8월 : 숭실대학교 컴퓨터학과 (공학박사)
- 2001년 9월 ~ 현재 : 호원대학교 컴퓨터게임학부 교수

<관심분야>
이동통신, 멀티미디어 통신, 네트워크 보안, 정보 검색