Journal of Korea Multimedia Society Vol. 13, No. 5, May 2010(pp. 722-729)

VRIFA: LRBF 커널과 Nomogram을 이용한 예측 및 비선형 SVM 시각화도구

김성철[†], 유환조[††]

요 약

예측 문제를 해결하기 위한 데이타마이닝 기법은 다양한 분야에서 주목받고 있다. 이것에 대한 한 예로 컴퓨터-기반의 질병의 예측 혹은 진단은 CDSS(Clinical Decision support System)에서 가장 중요한 요소이기도 하다. 이러한 예측 문제를 해결하기 위해서 RBF커널 같은 비선형 커널을 사용한 SVM이 가장 널리 사용되고 있는데, 이는 비선형 SVM이 어떠한 다른 분류기법보다 정확한 성능을 보이기 때문이다. 하지만 비선형 SVM을 사용한 경우에는 모델내부를 시각화하는 일이 어려워서 예측결과에 대한 직관적인 이해가 힘들고, 의학 전문가들은 이러한 비선형 SVM의 사용을 기피하고 있는 실정이다. Nomogram은 SVM을 시각화하기 위해 제안된 기법이다. 하지만 이는 선형 SVM의 경우에만 사용이 가능하고. 이 문제를 해결하기 위해서 LRBF 커널이 제안된 바 있다. LRBF 커널은 기존의 RBF 커널을 사용한 SVM과 대등한 결과를 보이면서도 예측결과의 선형적 분석도 가능하게 한다. 본 논문에서는 노모그램(Nomogram)과 LRBF 커널을 사용한 SVM이 통합되어 있는 예측 툴 VRIFA를 제안한다. 이 툴은 사용자와 상호작용하며 비선형 SVM 모델의 내부구조를 데이타의 각 속성별로 보여주는 방법으로 사용자가 예측결과를 직관적으로 이해하도록 도와준다. VRIFA는 Nomogram 기반의 피쳐선택(feature selection) 기능도 포함하고 있는데, 이 기능은 예측결과에 부정적인 영향을 끼치거나 중복된 연관성을 보이는 속성을 제거함으로써 모델의 정확도를 높이는 데 기여한다. 그리고 데이터에 포함된 클래스의 비율이 한 쪽으로 치우쳐져 있는 경우에는 ROC 곡선 넓이(AUC)를 예측결과를 평가하기 위한 측도로 사용할 수 있다. 이 툴은 컴퓨터-기반의 질병 예측 혹은 질병의 위험 요소 분석에 대해 연구하는 연구자들에게 유용하게 사용될 것으로 전망하는 바이다.

VRIFA: A Prediction and Nonlinear SVM Visualization Tool using LRBF kernel and Nomogram

Sung-Chul Kim[†], Hwan-jo Yu[††]

ABSTRACT

Prediction problems are widely used in medical domains. For example, computer aided diagnosis or prognosis is a key component in a CDSS (Clinical Decision Support System). SVMs with nonlinear kernels like RBF kernels, have shown superior accuracy in prediction problems. However, they are not preferred by physicians for medical prediction problems because nonlinear SVMs are difficult to visualize, thus it is hard to provide intuitive interpretation of prediction results to physicians. Nomogram was proposed to visualize SVM classification models. However, it cannot visualize nonlinear SVM models. Localized Radial Basis Function (LRBF) was proposed which shows comparable accuracy as the RBF kernel while

※ 교신저자(Corresponding Author) : 유환조, 주소 : 경북 포항시 남구 효자동 산 31번지 포항공과대학교 공학2동 데이터마이닝연구실(412-140), 전화 : 054)279-2941, FAX : 054)279-2299, E-mail : hwanjoyu@postech.ac.kr
접수일 : 2009년 12월 1일, 수정일 : 2009년 12월 15일
완료일 : 2010년 1월 20일

[†] 정회원, POSTECH 컴퓨터공학과 통합과정
(E-mail : subright@postech.ac.kr)
[††] 유환조, POSTECH 컴퓨터공학과 교수
(E-mail : hwanjoyu@postech.ac.kr)
※ 본 연구는 '구조형 웹 데이터 분석을 통한 커널 기반 실시간 추천 시스템 개발.' 연구(4.0004376.02)지원으로 수행되었음.

the LRBF kernel is easier to interpret since it can be linearly decomposed. This paper presents a new tool named VRIFA, which integrates the nomogram and LRBF kernel to provide users with an interactive visualization of nonlinear SVM models, VRIFA visualizes the internal structure of nonlinear SVM models showing the effect of each feature, the magnitude of the effect, and the change at the prediction output. VRIFA also performs nomogram-based feature selection while training a model in order to remove noise or redundant features and improve the prediction accuracy. The area under the ROC curve (AUC) can be used to evaluate the prediction result when the data set is highly imbalanced. The tool can be used by biomedical researchers for computer-aided diagnosis and risk factor analysis for diseases.

Key words: Decision support systems(의사 결정 지원 시스템), feature selection(피쳐셀랙션), localized radial basis function(LRBF) kernel(커널), nomograms(노모그램), support vector machines (SVMs)(서포트 벡터 머신), visualization(시각화)

1. 서 론

최근 데이터 마이닝과 기계 학습 분야에서는 컴퓨터 기반의 의학 문제 해결을 위해 많은 노력을 기울이고 있다. 이것에 대한 한 예로 컴퓨터-기반의 질병의 예측 혹은 진단은 CDSS(Clinical Decision support System)에서 가장 중요한 요소이기도 하다. SVM[1], [2]은 기계 학습 분야에서 가장 활발히 사용되고 있는 분류 기법으로써 수많은 의학 문제에 적용되어 왔다[4,11- 15]. 예를 들면, SVM은 축적된 환자의 데이터로부터 분류기를 학습하여 새로운 환자의 데이터를 가지고 특정 질병의 발병 여부를 예측하는 작업 등을 할 수 있다. 특별히 비선형 커널을 사용한 SVM의 경우에는 어떠한 다른 분류기법보다 높은 정확도를 보인다. 하지만 비선형 SVM을 사용한 경우에는 모델내부를 시각화하는 일이 어려워서 예측결과에 대한 직관적인 이해가 힘들다. 이 때문에 의학 전문가들은 이러한 비선형 SVM은 사용하지 않고 여전히 로지스틱 회귀분석과 같은 선형 모델에 의지하고 있다. 선형 모델은 정확도는 비선형 모델보다 뒤처지지만, 결과를 분석하는 일이 쉽기 때문이다.

비선형 모델의 또 다른 단점은 피쳐선택(Feature Selection)이 어렵다는 것이다. 피쳐선택은 분류 결과에 가장 주요한 영향을 준 피쳐를 찾거나 순위를 정하기 위한 중요한 작업이다. 실제 환자를 진찰하는 상황에서 의사는 특정 질병의 주요 위험 인자가 무엇인지를 가장 알고 싶어 할 것이다. 즉, 그 인자의 값이 변함에 따라서 실제 해당 질병의 발병 가능성이 어떻게 변화하는 지 또한 알고 싶어 할 것이다. 피쳐선택기법도 활발히 연구되어 왔지만 대부분이 선형 모델에 적합하고 비선형 모델에 적용 가능한 모델은 전무하다.

따라서 비선형 모델을 의학 분야에 적용하기 위해서는 모델 시각화와 피쳐선택이 필수적이라고 할 수 있다. 본 논문에서는 예측 문제에서 높은 정확도를 유지하면서도 비선형 SVM의 결과를 시각화 할 수 있는 예측 및 비선형 SVM 시각화툴 VRIFA (visualization for risk factor analysis)를 제안한다. VRIFA는 축적된 환자데이터로부터 비선형 SVM모델을 학습하여 진단/예측 모델을 생성하고, 모델을 이용하여 새로운 환자에 대해 질병 발병 확률을 예측하여 보여준다. 또한 각 피쳐가 환자의 질병 발병 여부에 어떠한 영향을 주는지를 그래프로 시각화하여 보여줌으로써 의학 전문가들이 쉽게 환자의 질병 발병 요인과 분석을 가능하게 한다. VRIFA는 비선형 SVM모델을 시각화하여 보여주는 최초의 툴로써 예측의 정확도 뿐 아니라 원인 분석이 중요한 의학 분야에 유용하게 이용될 수 있다.

본 논문의 구성은 다음과 같다. 섹션 II에서는 노모그램을 적용하여 비선형 SVM을 시각화하는 방법에 대해 서술하고, 섹션 III에서는 LRBF커널에 대해 소개하고, 섹션 IV에서는 적용 가능한 피쳐선택에 기법에 대해 알아본다. 그 이후 UCI 데이터셋을 사용한 VRIFA의 구현정보와 데모 그리고 결과와 결론을 마지막으로 본 논문이 마무리된다.

2. 노모그램을 활용한 SVM 시각화

노모그램 분류기(classifier)를 시각화하기 위한 도구이다. 이번 섹션에서는 노모그램의 개념에 대해 그리고 노모그램을 어떻게 SVM에 적용할 수 있는지 설명할 것이다.

2.1 노모그램

그림 1은 신종플루 환자의 사망여부를 예측하는

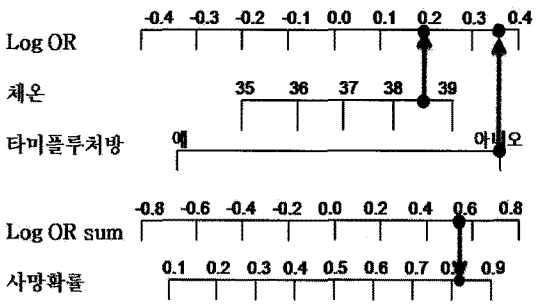

그림 1. 두 개의 피쳐(체온, 타미플루처방여부)를 가지는 환자 데이터로부터 신종플루 사망확률을 예측하기 위해 적용된 노모그램예제

분류기를 노모그램을 사용하여 시각화한 예제이다. 이해를 돕기 위해서 환자의 피쳐는 체온과 타미플루 처방여부 두 가지로 제한하였다. 노모그램을 예측에 사용하기 위해서는 우선 각 Log OR(log odds ratios) 수치로 표현된 각 피쳐의 값을 더하고, 그 합을 사망확률을 예측하는 데에 사용한다. 위 예제에 따르면 각 피쳐의 값을 가장 위에 위치한 Log OR 선으로 측정하면, 체온의 Log OR값은 약 0.19이고, 타미플루 처방여부의 Log OR값은 약 0.35이 된다. 두 피쳐의 Log OR 값의 합은 0.19+0.36 = 0.55가 되고 이를 가장 아래에 위치한 사망확률 확률선에 빗대어보면 약 0.82에 해당한다.

또한 노모그램에서는 Log OR 선의 길이를 보고 각 피쳐가 최종 확률이나 Log OR 합에 얼마나 큰 영향을 주는지 알아낼 수 있다. 예를 들어 체온의 최대값은 39인데, 이는 Log OR에서 0.25 그리고 최종 확률로는 약 0.67에 해당하는 값이다. 이를 통해서 노모그램에서는 각 피쳐선의 길이가 길수록 더 넓은 범위의 Log OR값을 가질 수 있다는 것을 알 수 있다. 이것은 그만큼 최종확률에 더 큰 영향을 준다는 것을 의미한다. 그림 1에서는 타미플루처방여부선의 길이가 체온선의 길이보다 길다. 이는 사망확률이 체온보다 타미플루 처방여부의 영향을 더 많이 받는 다는 것을 의미한다. 노모그램의 이러한 특징은 노모그램이 피쳐선택(feature selection)에도 활용될 수 있음을 보여준다.

2.2 노모그램을 활용한 SVM 시각화

SVM의 경우에도 노모그램을 사용하여 SVM의 모델 내부를 시각화할 수 있고, OR값의 평균을 그리는 방법(회귀분석 등을 활용)으로 각 피쳐의 효과도 나타낼 수 있다 [8]. SVM의 경우를 살펴보면, SVM에서는 데이터 샘플 (x, y)와 하이퍼플랜과의 거리는 독립변수이며 $\delta(x)$로 표현한다. 그리고 x와 y의 유사도를 돌려주는 커널함수 $k(x, y)$가 주어졌을 때, 그 거리는 결정함수(decision function)를 사용하여 아래와 같이 표현할 수 있다.

$$\delta(x) \simeq b + \sum_{j=1}^{N} y_j \alpha_j K(x, z_j)$$

이 때, b는 바이어스이고, α는 SVM에서 서포트벡터 z의 계수이고, N은 서포트벡터의 개수이다. 커널이 각 피쳐에 대해 선형적으로 분해할 수 있는 형태이면 그 거리는 다음과 같이 표현할 수 있다.

$$\delta(x) \simeq b + \sum_{k=1}^{M} [w]_k$$

$$[w]_k = \sum_{j=1}^{N} y_j \alpha_j K(x_k, z_{j,k})$$

이 때, M은 총 피쳐의 개수이고, x_k는 데이터벡터 x에서 k번째 피쳐의 값이고, $z_{j,k}$는 j번째 서포트벡터의 k번째 피쳐이다. 이 때 샘플 x가 파지티브클래스에 포함될 확률은 다음과 같이 구한다.

$$P(y=1 \mid x) = \frac{1}{1+e^{-(A+B\times\delta(x))}}$$

위 식의 파라미터 A와 B는 회귀 분석 모델을 사용하여 log linklihood를 최적화함으로써 얻을 수 있고, 또한 cross validation을 적용함으로써 overfitting을 예방하였다.

파라미터 A와 B를 구한 이후에 위 식은 아래와 같이 표현될 수 있다.

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0 + \sum_{k=1}^{M} [\beta]_k)}}.$$

위 식에서 $\beta_0 = A + B\times b$, $[\beta]_k = B\times[w]_k\beta_0$ 이다. β_0는 피쳐값이 없을 때 해당 피쳐의 prior 확률로 동작하는 상수 intercept이다. $[\beta]_k$는 스코어 벡터로써 k번째 피쳐값을 이 벡터를 사용하여 Log OR선에 매핑함으로써 k번째 피쳐의 점수를 구한다. 이 점수가 노모그램에서 사용되는 해당 피쳐의 Log OR값이 된다.

그러나 polynomial커널이나 RBF커널과 같은 비선형 커널을 사용하는 SVM의 경우에는 위와 같은 접근법을 사용할 수가 없다. (RBF커널은 높은 유연성과 정확도를 보이면서 가장 활발하게 사용되고 있는

커널이다[3,5]). 따라서 위에서 설명한 방법은 선형 모델에만 적용이 가능하고, 이를 해결하기 위해서 LRBF 커널이 제안되었다.

3. LRBF 커널

LRBF 커널은 비선형 커널을 사용한 경우에 따르는 부작용을 해소하기 위하여 제안된 방법으로, LRBF 커널을 사용하여 학습된 모델은 기존의 RBF 커널을 사용한 경우[3]와 대등한 결과를 보이면서도 예측결과의 선형적 분석도 가능하다는 장점을 지닌다[9].

3.1 LRBF 커널의 정의

RBF 커널은 높은 유연성을 이유로 SVM에 적용되어 활발하게 사용되고 있다. RBF커널은 다음과 같다.

$$K(x,z)=e^{(-\gamma\|x-z\|^2)}=\prod_{k=1}^{M}e^{(-\gamma(x_k-z_k)^2)}.$$

위 식과 같이 RBF커널은 두 벡터의 유사도를 지수의 거듭제곱으로 표현되는 각 피쳐유사도의 곱으로 표현된다. RBF커널이 SVM에서 사용되는 경우, 커널은 데이터벡터 x와 서포트벡터 z의 곱으로 유사도를 계산한다. 즉, SVM에서 사용된 RBF커널 피쳐공간 상에서 데이터벡터와 서포트벡터의 거리를 기준으로 해당 데이터벡터의 클래스를 결정한다.

이렇게 곱으로 표현된 RBF의 특성상 이것을 선형적으로 분해할 수가 없다. 즉, 노모그램을 사용한 시각화가 불가능하다는 것이다. LRBF커널은 이 문제를 해결하기 위해 제안되었으며, LRBF 커널은 노모그램을 사용해 시각화 할 수 있다. 식은 다음과 같이 표현한다.

$$K(x,z)=\sum_{k=1}^{M}e^{(-\gamma(x_k-z_k)^2)}.$$

LRBF 커널은 입력 데이터에서 각 피쳐의 값이 커널의 결과값에 어느 정도의 영향을 주는 지 확인할 수 있다는 점에서 RBF커널과는 차별화된다. RBF커널은 각 피쳐값의 유사도의 곱으로 표현되는 반면에, LRBF 커널은 각 피쳐의 유사도의 합으로 구해진다. 즉, LRBF커널은 선형적으로 분해되어 노모그램을 사용해 시각화 할 수 있다.

3.2 노모그램으로 LRBF커널 시각화하기

SVM에 선형적으로 분해가 가능한 LRBF커널을 사용하면, 노모그램도 적용이 가능하다. 기존의 선형 SVM을 노모그램에 적용한 경우와는 다르게 LRBF 커널을 사용한 SVM을 시각화하는 경우에 노모그램에서의 이팩트벡터는 각 피쳐마다 비선형적 특징을 가지는 Log OR 커브를 생성한다.

그림 2에서 보듯이 VRIFA를 사용해서 학습된 UCI heart 데이터의 분류기와 그 모델의 내부를 피쳐별로 확인할 수 있다. 오른쪽 상단의 그래프는 피쳐값에 따른 Log OR값에 대한 비선형 관계를 볼 수 있다(본 예제에서는 12번 피쳐의 그래프를 보여주고 있다.). 파란점은 현재 환자의 피쳐값(x축)과 거기에 해당하는 Log OR값(y축)을 나타내고 있다. 여기서 그래프 하단에 위치한 컨트롤도구를 통해서 피쳐값이 변함에 따라서 Log OR값이 비선형 그래프를 따라 어떻게 변하는지 확인할 수 있고, 그 변화가 최종확률에는 어떠한 영향을 주는지 왼쪽 하단의 그래프에서 확인할 수 있다. 최종확률은 모든 피쳐의 Log OR값의 합에 따라 변하므로 위 예제에서 피쳐값을 왼쪽으로 움직이면 해당 피쳐의 Log OR값이 증가할 것이고 Log OR합도 증가하여 최종 확률 또한 증가할 것이다.

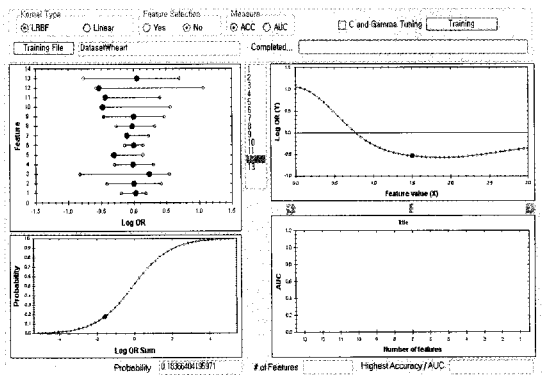

그림 2. VRIFA툴 예시: 피쳐선택을 적용하지 않은 경우, 정확도를 기반으로 UCI heart 데이터의 분류기를 학습한 결과.

4. 피쳐선택

피쳐선택(Feature selection)은 데이터셋에서 연관성이 적거나 중복된 정보를 포함하고 있는 피쳐를 삭제함으로써 학습모델의 성능을 개선하고자 하는 방법이다. 피쳐선택은 크게 필터, 랩퍼, 임베디드 접근법 이렇게 세 가지로 구분할 수 있는데[7], 필터방법은 모델을 학습하기 전에 각 피쳐를 통계학적으로

분석하여 상위에 랭크된 피쳐를 선택하는 방법이고, 랩퍼 방법은 학습된 모델의 결과로 피쳐를 평가하여 상위에 랭크된 피쳐를 선택하는 방법이고, 임베디드 방법은 모델을 학습하는 과정에서 중간 과정의 학습 모델과 상호작용하면서 성능을 저해하는 피쳐를 골라내는 방법이다. 랩퍼방법은 앞에 언급한 방법에 비해 더 효율적이라고 여겨지지만 임베디드 방법이 보통 가장 좋은 결과를 낸다.

4.1 노모그램을 활용한 피쳐선택

RFE(Recursive Feature Elimination) 은 대표적인 임베디드 피쳐선택방법으로 분류기를 반복적으로 학습하면서 각 반복에서 학습된 분류기 내부를 분석하여 적은 가중치를 가지는 피쳐를 삭제하면서 가장 좋은 결과를 보이는 피쳐부분집합을 선택하는 방법으로 SVM에 적용된 바 있다[6,10]. 하지만 RFE를 적용하기 위해서는 분류기에 사용되는 모델이 선형적이어야 한다는 조건이 있기 때문에 일반적인 비선형 커널 SVM에는 적용할 수가 없다.

본 논문에서는 LRBF커널과 노모그램을 활용하여 비선형 SVM에도 RFE기법을 적용하였다. 노모그램에 따르면 더 높은 가중치를 가지는 피쳐가 더 중요한 피쳐이므로, 노모그램에서 더 긴 선을 가지는 피쳐가 더 중요한 피쳐가 될 것이다. (그림 1 참고) 따라서 비선형으로 표현된 각 피쳐커브의 길이를 계산하기만 하면, 노모그램을 그리는 방법으로 각 피쳐의 중요도를 확인할 수 있다. 일반적인 RFE에서는 반복적으로 분류기를 학습한 이후 선형적으로 표시된 모델에서 각 피쳐에 해당하는 가중치만 비교하면 되었는데, 노모그램-RFE에서는 비선형적인 모델을 사용하기 때문에 추가적으로 노모그램상에서 각 피쳐를 표현하는 커브와 그 길이를 계산해야한다. 이 과정 이후에는 RFE와 같은 방법으로 각 프로세스마다 가장 낮은 가중치를 가지는 피쳐(가장 짧은 길이의 선을 가지는 피쳐)를 하나씩 제거함으로써 피쳐선택을 수행한다. 이 과정을 반복하면서 높은 정확도를 가지는 피쳐부분집합을 찾을 수 있다. 그림 3에서 오른쪽 상단에 위치한 그래프가 피쳐의 비선형적 특징을

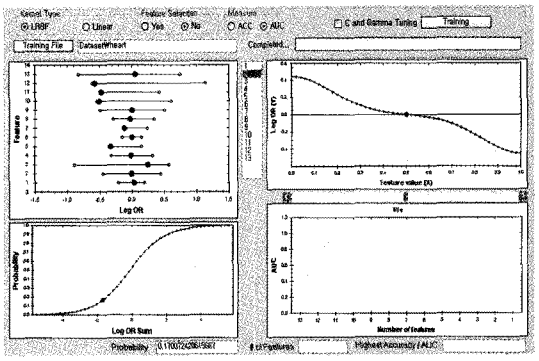

그림 3. VRIFA툴 예시: 피쳐선택을 적용한 경우, 정확도를 기반으로 UCI heart 데이터의 분류기를 학습한 결과

표 1. SVM-RFE 알고리즘

Line	Code
Input	Training instances X_0 and their class labels y
1	Initialize subset of surviving features $[1, 2, ..., n]$
2	Initialize feature ranking list $r = []$
3	while $(s \neq [])$
4	Restrict training instances to the subset of surviving features $X = X_0(:, s)$
5	Train the SVM with the restricted instances and their class labels Compute the weight vector of the SVM classifier
6	$w = \sum_{l}^{sv} y_i \alpha_i x_i$
7	Compute the ranking criteria $C_k = (w_k)^2$, for all k
8	Update the feature ranking list according to the criteria
9	Eliminate features with the lowest ranking
10	end while
outputs	Feature ranking list r

보여주는 그래프이고, 오른쪽 하단에 위치한 그래프가 각 피쳐 커브의 길이에 따라서 피쳐선택을 수행한 후 학습된 모델의 정확도를 각 피쳐부분집합의 개수에 따라 보여주는 그래프이다.

5. VRIFA 구현 및 데모

VRIFA(visualization for risk factor analysis)는 위에서 언급한 노모그램과 LRBF커널을 사용한 SVM을 통합되어 있는 예측 및 SVM모델 시각화도구로써 사용자는 대화형의 인터페이스를 활용하여 환자 데이터에 대한 시각적인 정보를 얻을 수 있다. 이는 의사들이 환자를 진단할 경우에 주요 위험 요인 판별에 유용하게 사용될 것으로 보인다. VRIFA는 LRBF커널을 사용하여 SVM모델을 학습하고, 그 모델을 분해하여 최종 확률에 대한 피쳐별 영향력을 사용자에게 제공한다. VRIFA는 또한 노모그램기반의 피쳐선택기능을 포함하고 있어 중요한 피쳐를 자동으로 선택함으로써 정확도를 높일 수 있다.

VRIFA는 Visual C# 2005와 .NET 버전의 LIBSVM을 사용하여 구현하였다. VRIFA를 사용하기 위해서는 Microsoft .NET Framework를 설치해야만 한다. VRIFA는 http://dm.postech.ac.kr/vrifa에서 다운로드 받을 수 있다.

5.1 데모

VRIFA의 테스트를 위하여 UCI Repository에서 가져온 심장병 데이터셋을 사용하였다. 이 데이터는 270개의 데이터를 가지고 있고(양성데이터 120개, 음성데이터 150개), 레이블을 포함한 총 피쳐는 13개이다. VRIFA의 실행을 위해서는 우선 'Training File'을 클릭하여 분류기를 학습하고자 하는 데이터를 선택한 이후(단, 데이터는 libsvm 포맷을 따라야한다.) 'Kernel Type'에서 LRBF커널을 사용할 것인지 Linear 커널을 사용할 것인지를 선택하고, 'Feature Selection'에서 피쳐선택의 적용여부를 선택하고, 'Measure'에서 어떤 척도를 기준으로 분류기를 학습할 것이지를 선택하고, 마지막으로 파라미터 튜닝여부를 선택한 후, 'Training'버튼을 클릭하면 된다.

그림 2과 그림 3는 정확도를 기반으로 하여 UCI heart 데이터셋의 분류기를 학습한 결과이고 그림 4과 그림 5는 AUC를 기반으로 하여 같은 데이터셋의

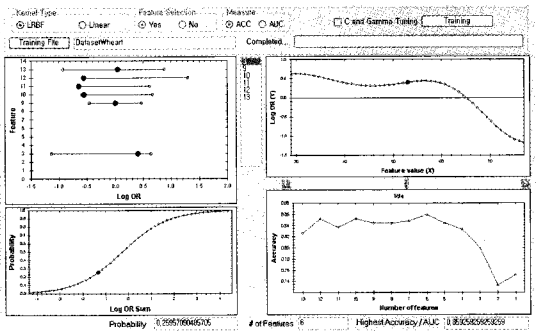

그림 4. VRIFA툴 예시 : 피쳐선택을 적용하지 않은 경우, AUC를 기반으로 UCI heart 데이터의 분류기를 학습한 결과

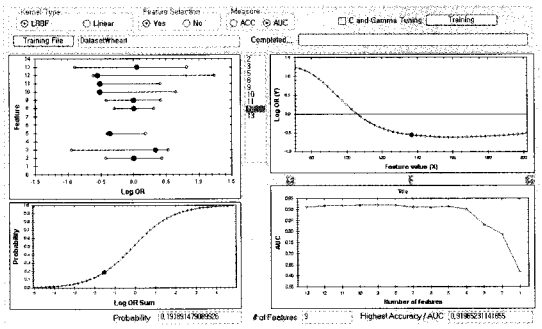

그림 5. VRIFA툴 예시: 피쳐선택을 적용한 경우, AUC를 기반으로 UCI heart 데이터의 분류기를 학습한 결과

분류기를 학습한 결과이다. 위 그림들에서 왼쪽 상단에 위치한 그래프는 각 피쳐가 결과값에 영향을 주는 정도(선의 길이)와 각 피쳐값이 가질 수 있는 전체 범위(선) 그리고 특정 환자의 각 피쳐값(빨간점)을 보여준다. 그 그래프 바로 옆에 위치한 박스에서 피쳐 넘버를 선택하면, 오른쪽 상단에 각 피쳐값의 Log OR값을 보여준다. 그리고 각 값을 조정하면서 해당 피쳐값이 변화함에 따라서 최종 확률이 어떻게 변하는 지를 왼쪽 하단의 그래프에서 확인할 수 있다. 왼쪽 하단의 그래프는 모든 피쳐의 Log OR값의 합과 그것을 기반으로 한 최종 확률을 확인할 수 있다. 마지막으로 오른쪽 하단의 그래프는 선택된 피쳐의 부분집합의 성능을 보여준다. 최종 결과는 이 중에서 가장 좋은 결과를 보이는 피쳐의 부분집합을 사용한 결과이다.

5.2 한계점

VRIFA의 한계점은 커널을 분해하여 피쳐별로 그

중요성을 판별할 수 있는 반면에, 피쳐들 간의 관계에 대한 정보는 주지 못한다는 것이다. 그리고 최적의 결과를 얻기 위해서는 파라미터 튜닝을 해야 하는데 이 작업은 다양한 경우의 수의 파라미터를 모두 사용하여 학습함으로써 파라미터를 찾아내야 하기 때문에 많은 시간을 필요로 하기도 하고, 특정 파라미터의 경우에는 결과가 수렴하기까지도 오랜 시간이 걸린다. 효율적인 파라미터 탐색은 지금도 활발히 연구되고 있는 분야이다.

6. 결 론

본 논문에서는 LRBF커널을 사용하여 SVM으로 예측문제를 해결하고 그 결과를 시각화하는 도구인 VRIFA를 제안하였다. LRBF 커널은 RBF 커널과 대등한 결과를 보이면서도 커널을 사용한 모델의 내부를 분해할 수 있어서 모델을 시각화할 수 있었다. VRIFA는 사용자에게 SVM의 결과를 피쳐별로 보여주고 최종 확률 또한 제시함으로써 사용자가 학습결과를 직관적으로 이해하고 그 결과를 활용할 수 있는데 도움을 주었다. 이 정보는 메디컬 데이터를 분석하는 의사들에게 해당 데이터에서 중요한 역할을 하는 요인을 분석하게 하여 진단작업을 도와줄 수 있을 것으로 보인다. 또한 VRIFA에 피쳐선택기능도 구현하여 결과에 유사한 관계를 가지는 중복 피쳐나 전혀 관계가 없는 피쳐들을 삭제함으로써 분류기의 성능을 개선할 수 있었다.

7. ACKNOWLEDGEMENT

이 논문은 2009년 정부의 재원으로 두뇌한국 21 사업과 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2008-331-D00483).

참 고 문 헌

[1] C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, Vol.20, pp. 273-297, 1995.

[2] V. Vapnik, *The Nature of Statistical Learning Theory*, New York: Springer-Verlag, 1995.

[3] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Discovery*, Vol.2, pp. 121-167, 1998.

[4] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, Vol.16, pp. 906-914, 2000.

[5] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines [Online]," Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[6] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, Vol.46, pp. 389-422, 2002.

[7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *the Journal of Machine Learning Research*, Vol.3, pp. 1157-1182, 2003.

[8] A. Jakulin, M. Mozina, J. Demsar, I. Bratko, and B. Zupan, "Nomograms for visualizing support vector machines," *Knowledge Discovery and Data Mining*, 2005.

[9] B. Cho, H. Yu, J. Lee, Y. Chee, and I. Kim, "Nonlinear support vector machine visualization for risk factor analysis using nomograms and localized radial basis function kernels," *the Institute of Electrical and Electronics Engineers Transactions on Information Technology in Biomedicine*, 2005.

[10] J. H. OH, J. Gao, A. nandi. P. Gurnani, L. Knowles, J. Schorge, and K. P. Rosenblatt. "Multicategory classification using extended SVM-RFE and markov blanket on SELDI-TOF mass spectrometry data," *the Institute of Electrical and Electronics Engineers Symposium. Computational Intelligence in Bioinformatics and Computational Biology*, 2005.

[11] K. Takeuchi and N. Collier, "Bio-medical entity extraction using support vector machines,"

Artificial Intelligence in Medicine, Vol.33, pp. 125-137, 2005.

[12] T.Arodz, M. Kurdziel, E. O. D. Sevre, and D. A. Yuen, "Pattern recognition techniques for automatic detection of suspicious-looking anomalies in mammograms," *Computer Methods and Programs in Biomedicine*, Vol.79, pp. 135-149, 2005.

[13] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler, "Learning from imbalanced data in surveillance of nosocomial infection," *Artificial Intelligence in Medicine*, Vol.37, pp. 7-18, 2006.

[14] M. E. Mavroforakix, H. V. Georgiou, N. Dimitropuoulox, D. Cavoura, and S. Theodoridis, "Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers," *Artificial Intelligence in Medicine*, Vol.37, pp. 145-162, 2006.

[15] L. Ramirez, N. G. Durdle, V. J. Raso, and D. L. Hill, "A support vector machines classifier to assess the severity of idiopathic scoliosis from surface topology," *the Institute of Electrical and Electronics Engineers Transactions on Information Technology in Biomedicine*, Vol.10, No.1, pp. 84-91, 2006.

김 성 철

2008년 POSTECH 컴퓨터공학과 (학사)

2008년~현재, POSTECH 컴퓨터공학과 통합과정

유 환 조

1998년 중앙대학교 컴퓨터공학과 (학사)

2004년 Univ. of Illinois at Urbana-Champaign(PhD)

2004년~2008년, Univ. of Iowa, Assistant professor

2008년~현재 POSTECH 컴퓨터공학과 교수