당뇨병의 예측을 위한 분류기 앙상블의 BKS 결합

박한샘° 조성배
연세대학교 컴퓨터과학과
sammy@sclab.yonsei.ac.kr° sbcho@cs.yonsei.ac.kr

BKS Fusion of Classifier Ensemble for Prediction of Diabetes

Han-Saem Park° and Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

경제 여건의 향상 및 생활양식의 변화로 최근 우리나라에서도 당뇨병 환자가 늘어남에 따라 당뇨병의 예측 및 치료가 중요한 관심사가 되고 있다. 본 논문은 1993년과 1995년 두 차례에 걸쳐 경기도 연천 지역 주민들의 여러 가지 신체 지수 등을 조사한 데이터를 대상으로, 1차 년도의 데이터로부터 동일한 환자가 2차 년도에 정상상태를 유지하는지 혹은 당뇨병으로 진행이 되는지를 예측하는 문제를 다룬다. 혈당량, 허리둘레 등의 수치가 당뇨병의 발병에 영향을 끼치는 것은 알려진 사실이므로, 현재의 데이터로부터 앞으로의 발병 가능성을 예측하는 것이 가능하며, 이는 환자에게 보다 정확한 정보를 알려줄 수 있으므로 의미가 있는 일이다. 예측을 위해 본 논문에서는 분류기를 사용하며, 예측율을 높이기 위해 여러 분류기를 BKS로 결합하였다. BKS (behavior knowledge space) 결합 방법은 분류기간의 독립 가정이 필요 없으며, 데이터 크기가 크고 전형적인 경우에 좋은 결과를 낼 수 있는 방법이다. BKS 결합 방법을 통해 실험을 해 본 결과 단일 분류기로 실험을 한 결과보다 향상된 성능을 얻을 수 있었으며, 투표 결합 방법과 비교하여 더 좋은 성능을 보였다.

1. 서 론

경제적 여건의 향상 및 생활양식의 변화로 최근 우리나라에서도 당뇨병 환자가 늘어나고 있으며, 이에 따라 당뇨병의 예방 및 치료가 중요한 관심사가 되고 있다. 당뇨병은 심장 질환, 고혈압, 신장 질환, 각종 신경계 질환 등 여러 가지 다른 병의 원인이 될 수 있으며, 90년대 이후 당뇨병 발병률 뿐 아니라 당뇨병에 의한 사망률도 꾸준히 증가하고 있다. 표 1은 주요 사망 원인별 사망자 수를 나타내는 것으로 당뇨병이 주요 사망 원인 중 하나가 되었음을 보여준다[1].

표 1. 주요 사망 원인 별 사망자 수

연도	1998	1999	2000	2001	2002
뇌혈관 질환	34,355	34,410	34,817	35,354	37,134
폐암	9,583	10,417	11,606	11,971	12,587
당뇨병	9,791	10,296	10,746	11,403	12,090
위암	11,102	11,309	11,550	11,483	11,771
간암	9,302	9,747	10,118	10,215	11,115
고혈압성 질환	3,899	3,568	4,238	4,875	5,125
결핵	3,478	3,297	3,413	3,221	3,352
계	240,254	246,539	247,346	242,730	246,515

본 논문은 경기도 연천 지역 주민들을 대상으로 1993년, 1995년 2차에 걸쳐 조사된 여러 가지 신체 지수 등의 데이터를 분석한다. 이 데이터는 1차 년도에는 정상이었다가 2차 년도에는 정상과 당뇨병 환자로 나뉘어 진 연천지역 주민들의 여러 가지 정보를 담고 있는데, 이 데이터에서 1차년도의 데이터로부터 2차 년도에 정상 상태를 유지하는지, 아니면 당뇨병 환자가 되는지를 예측해 내는 것이 본 논문에서 다루고자 하는 문제이다. 이제까지 하지만 혈당량, 허리둘레, 체지방 지수 등 여러 가지 신체적 데이터가 당뇨병의 진행에 영향을 끼친다고 알려져 왔으므로, 과거의 데이터로부터 앞으로의 발병 가능성을 예측함으로써 환자에게 보다 정확한 정보를 알려 줄 수 있을 것이다[2].

예측을 위해서 패턴인식 분야에서 사용되는 분류 기법을 사용하였는데, 단일 분류기를 사용한 분류율의 향상은 한계가 있기 때문에 보다 높은 성능을 얻기 위해 여러 개의 분류기를 BKS (behavior knowledge space) 결합 방법을 통해 결합하였다. BKS 결합 방법은 다른 방법과는 달리 사용한 개별 분류기들이 서로 독립이라는 가정이 필요 없으며, 데이터의 크기가 충분히 크고, 전형적인(representative) 경우에 좋은 결과를 낼 수 있는 방법이다[3].

2. 관련 연구

당뇨병과 관련한 의학 분야의 관련 연구가 활발히 진행되어 왔는데, 이들은 우리 사회에서의 당뇨병 발생률이 예상보다 훨씬 높다고 보고하고 있다[4, 5].

표 2. 관련 연구

저자	방법	데이터
Kim *et al.*	evolutionary neural network	colon cancer data
Park *et al.*	sequential MLP	HRA (health risk appraisal)
Xu *et al.*	Dempster-Shafer and bayesian formalism	handwriting data
Valentini *et al.*	bagged ensemble of SVM	colon cancer, lymphoma data

표 2는 질병을 예측하거나 질병의 클래스를 분류하기 위해 분류기를 사용했던 관련 연구들을 보여준다. Kim 등은 대장암을 예측하기 위해 진화 신경망(evolutionary neural network)을 사용했으며[6], Park 등은 당뇨병의 예측 모델을 만들기 위해

SMLP (sequential MLP)를 사용하였다[7].

패턴 분류를 위해서 하나의 분류기만을 사용하기도 하지만, 단일 분류기의 개선을 통한 분류율의 향상은 한계가 있기 때문에 여러 분류기를 결합하여 사용하기도 한다. 이런 맥락에서 분류기의 결합에 대해서도 많은 연구가 이루어져 왔으며, 분류기를 결합해서 사용했을 때의 결과가 단일 분류기를 사용한 것 보다 우수한 성능을 보임이 여러 연구들을 통해 확인되었다[8, 9]. 따라서 이런 분류기 결합은 질병을 예측하거나 분류하는 목적으로 사용되기도 한다. Valentini 등은 대장암, 백혈병 등의 데이터를 분류하기 위해 분류기 결합 방법 중 bagging으로 SVM (support vector machines)을 결합하여 사용하였다[9].

3. BKS 결합 앙상블 분류기

앞에서 언급했듯이 단일 분류기를 사용한 인식률의 향상은 한계가 있기 때문에, 여러 연구팀들이 성능의 향상을 위해 분류기를 결합하여 사용하였다[3, 8, 9, 10].

본 논문에서는 당뇨병의 예측을 위해 BKS 결합 방법으로 분류기를 결합하였다. 개별 분류기로는 K-최근접 이웃(*k*-nearest neighbour, KNN), 다층 신경망(multi-layered perceptron, MLP), 그리고 SVM(support vector machines)이 사용되었다. SVM은 서로 다른 두 가지의 커널 함수를 사용하였는데, 아래 그림 1에서의 SVM(L)과 SVM(R)은 각각 선형(linear) 커널 함수와 RBF (radial basis function) 커널 함수를 사용한 SVM을 뜻한다.

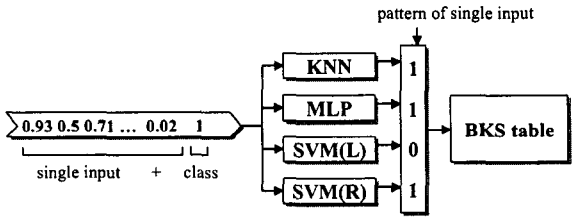

그림 1. BKS 테이블을 위한 하나의 입력 패턴

그림 1은 입력 데이터로부터 만들어 진 하나의 학습 패턴이 BKS 테이블에 저장되는 과정을 보여준다. 하나의 학습 데이터는 네 개의 개별 분류기를 거쳐 그림 1에서 보이는 4-비트 이진 코드의 형태로 BKS 테이블에 저장될 하나의 입력 패턴이 되고, 이러한 패턴은 입력 데이터의 클래스와 함께 BKS 테이블에 저장되게 된다. 모든 학습 데이터(750개)에 대하여 이러한 과정이 반복되고 나면 표 3와 같은 BKS 테이블이 만들어진다.

표 3. BKS 테이블

pattern	label	
	0	1
1 1 1 1	43	638
1 1 1 0	10	3
1 1 0 1	10	16
1 1 0 0	4	2
0 1 1 0	8	0
0 1 0 0	13	1
0 0 0 0	2	0

표 3에서 볼 수 있듯이 BKS 테이블에서는 각 패턴과 패턴에 따른 여러 레이블의 수를 기억한다. 본 논문에서 사용한 당뇨병 데이터는 2 클래스 데이터 이고, 결합을 위해서 4개의 단일 분류기가 사용되었으므로 BKS 테이블을 위해 필요한 패턴의 수는 2^4이 된다. 표 3은 16개의 패턴 중 7개에 대한 레이블만을 보여주는데, 여기에 나타나지 않은 패턴에 대해서는 학습 데이터가 없었기 때문이다.

학습 데이터를 통해 구성된 BKS 테이블은 테스트를 위해 사용된다. 각 테스트 데이터는 학습 데이터와 마찬가지로 그림 1에서 처럼 네 개의 분류기를 거쳐 패턴이 생성되고, 생성된 패턴을 BKS 테이블의 내의 패턴에서 검색하여 레이블을 결정한다. 이 때 레이블은 학습 데이터가 더 많은 쪽을 선택한다. 네 개의 단일 분류기를 거친 테스트 데이터의 패턴이 BKS 테이블에 없다면, 그 데이터는 거부(reject)된다. 이 경우 본 논문에서는 전체 비율에 따른 확률이 높은 레이블을 선택하였다. 마지막으로 모든 테스트 데이터에 대해 앞의 과정을 반복하여 분류율을 계산한다. 이와 같이 BKS 결합 방법은 해당 데이터에 대한 레이블만을 이용해 분류기를 결합하는 추상 레벨(abstract level) 결합 방법이다[3].

앞에서도 설명했듯이 BKS 결합 방법은 분류기 간의 독립 가정이 필요 없으며, 데이터 크기가 크고 전형적인 경우 높은 성능을 내는 방법이다. 본 논문의 당뇨병 데이터는 데이터의 크기가 충분히 크지는 않지만, 클래스가 2개뿐이고 또한 분류기를 4개밖에 사용하지 않았으므로 데이터 수가 조금 부족한 점이 보완될 수 있다. 또한 BKS 테이블 내 학습 데이터의 패턴이, 테스트 데이터의 패턴과 거의 일치하는 전형적인 데이터이기 때문에 BKS가 결합 방법으로서 적합하다. BKS 결합 방법의 단점으로 클래스와 분류기의 수가 커질 경우 필요한 저장 공간이 기하학적으로 늘어난다는 점과, BKS 테이블의 클래스 간 레이블 수가 비슷할 때는 분류율이 떨어지는 점을 들 수 있다[3]. 본 논문에서는 2 클래스 문제의 해결을 위해 4개의 분류기를 결합하였으므로 무리한 저장 공간을 필요로 하지 않으며, 표 3에서 볼 수 있듯이 한 패턴에 대해 레이블 수가 비슷한 경우 또한 거의 없다. 이런 여러 가지 사항을 고려할 때, 본 논문의 문제 해결을 위한 방법으로 BKS 결합이 적합하다고 할 수 있다.

4. 실험 및 결과

4.1. 실험 데이터

실험에 사용된 데이터는 본래 지역 사회를 기반으로 한 역학 연구를 위해 조사되었다. 1993년에 경기도 연천군에 거주하는 2520명 중 당뇨병이 없던 2293명을 대상으로 1차 조사를 하였으며, 1995년 2차 조사 시에는 그 중 1193명을 조사하였다[4]. 이 중 1040명이 정상 상태였고, 153명은 당뇨병으로 진행이 되었다. 즉, 정상과 당뇨를 분류해야 하는 2 클래스 문제를 위한 데이터이다. 본 논문은 두 차례의 조사에 모두 참여한 1193명의 데이터 중 많은 결측값을 포함하고 있던 일부의 샘플을 제외한 1111명의 데이터를 대상으로 하였다. 각 데이터는 연령, 성별 등의 외형적인 정보부터 혈당, 인슐린, 체질량 지수 등과 같은 신체 지수까지 여러 속성을 포함하고 있으며, 본 논문에서는 여러 속성들 가운데 분류에 유용한 20개의 속성을 사용하였다. 표 4는 각 데이터의 속성을 보여준다.

표 4. 데이터의 속성

연령	성별
신장	혈압
체중	허리 둘레
수축기 혈압	엉덩이 둘레
고혈압 여부	이완기 혈압
간수치 (GOT)	간수치 (GPT)
공복 혈당	중성 지방
HDL 콜레스테롤	콜레스테롤
허리, 엉덩이 둘레 비	체질량 지수
식사 후 두 시간째 혈당	당뇨병 여부

4.2. 실험 결과

가. 실험 환경

전체 1111개의 샘플을 학습 샘플 750 (정상:660/당뇨:90), 테스트 샘플 361 (정상:317/당뇨:44)로 나누어 실험하였다. 먼저 MLP, KNN의 실험 환경은 다음과 같다. MLP의 경우 (20-10-2)개의 (입력-히든-출력)노드 수를 사용했으며, 학습률은 0.01, 최대 반복회수는 200회로 설정하였다. KNN의 k는 예비 실험에서 가장 좋은 결과를 보인 5로 설정하였다.

나. 실험 결과 및 분석

표 5는 파라미터를 튜닝하여 얻은 각 단일 분류기의 분류율과 BKS 결합 방법을 통해 얻은 분류율을 보여준다. 단일 분류기의 분류결과는 MLP의 결과가 91.7%로 가장 높았으며 BKS 결합 방법을 사용했을 때는 92.8%의 분류률을 얻어 결합으로 인한 분류율의 향상을 확인했다. 테스트 샘플의 패턴이 BKS 테이블에 없어서 거부(reject)된 경우가 하나 있었으나 전체 분류율에 큰 영향을 주지는 못하였다. 이 비율이 커질 경우 BKS 결합결과에 안좋은 영향을 끼치게 되지만, 전형적인 데이터를 대상으로 할 경우에는 큰 문제가 되지 않는다.

표 5. 단일 분류기와 BKS 결합 방법의 분류율 비교

분류기	MLP	KNN	SVM(L)	SVM(R)	BKS 결합
분류율(%)	91.7	88.6	91.1	87.8	92.8

표 6은 간단하면서도 많이 쓰이는 분류기 결합 방법인 투표 결합 방법과 가중 투표 결합 방법으로 얻은 분류율을 BKS 결합 결과와 비교한 것이다. 가중 투표 결합의 가중치는 단일 분류기의 인식률을 사용하였으며, 공정한 비교를 위해 BKS 결합 방법과 동일한 학습 샘플로 개별 분류기를 학습하고, 역시 동일한 테스트 샘플로 테스트를 하였다. 투표 결합과 가중 투표 결합의 결과는 예상 외로 단일 분류기 중 분류율이 가장 좋은 MLP의 결과보다도 낮은 분류율을 보였다.

표 6. 결합 방법간의 분류율 비교

결합 방법	투표	가중투표	BKS
분류율(%)	91.4	90.8	92.8

표 7은 실험 결과 가운데 위와 같은 결과가 나오는 원인이 되는 패턴을 분석한 것이다. 투표 결합 방법은 다수결의 원리에 의한 것이므로 과반수이상의 분류기가 분류를 잘못하게 되면 결합 결과도 틀릴 수밖에 없다. 하지만 BKS 결합 방법은 여러 단일 분류기의 결과를 패턴으로 기억하는 동시에 실제 클래스도 기억을 하게 되므로, 다수의 분류기가 분류를 잘못했다고 해도 정답을 맞출 수 있게 된다. 대표적인 예가 표 7의 첫 번째 패턴인 「1110」이다. 투표 결합을 했을 때에는 이 패턴을 클래스 1로 분류하게 되므로, 실제로는 클래스 0에 속하는 다수의 샘플을 잘못 분류하게 된다. 반면에 BKS 결합 방법의 경우는 학습 샘플을 통해 만들어진 표 3의 BKS 테이블을 참고하므로, 이 패턴을 클래스 1로 분류한다. 「1100」, 「0110」과 같이 0과 1의 수가 동일한 패턴도 투표 결합 방법과 비교해 BKS의 분류율이 더 높게 나오는 원인 중 하나이다.

표 7. BKS 결합에 유리한 패턴

pattern	실제 클래스	샘플 수
1 1 1 0	0	17
	1	5
1 1 0 0	0	6
	1	3
0 1 1 0	0	11
	1	0

5. 결론 및 향후 연구

본 논문에서는 연천 지역의 주민들로부터 얻은 1993년의 데이터로부터 동일한 사람들이 1995년에 당뇨병으로 진행이 되는지를 예측하기 위해 패턴 인식 방법인 분류 기법을 사용하였다. 또한 예측율의 향상을 위해 여러 개의 분류기를 BKS 결합 방법을 통해 결합하였다. 실험을 통하여 단일 분류기를 사용했을 경우에는 87.8%~91.7%의 분류율을 얻었으며, BKS 결합 방법을 통해서는 92.8%의 향상된 분류율을 얻었다. 마지막으로 BKS로 결합한 결과가 투표 결합 방법과 비교하여 더 높은 성능을 보임을 확인하였다.

본 논문에서 사용한 분류기를 통한 예측은 여러 속성 값의 패턴을 통해 학습과정을 거쳐 분류에 사용하는 것이므로, 이 때의 분류기는 블랙 박스로서 역할을 수행하여 결과만을 알려준다. 하지만 베이지안 네트웍 등 속성들 간의 관계를 분석해 볼 수 있는 방법을 사용한다면 직접적인 분석이 가능해지므로 또한 의미 있는 연구가 될 것이다.

감사의 글

본 논문은 보건복지부의 보건의료기술진흥사업의 지원에 의하여 이루어진 것임.

참고 문헌

[1] 통계청: 사망원인통계연보. 2002.

[2] E.-J. Kim, *et al.*, "Relationship between serum total lipids level and sex, age, blood sugar level, body weight and vascular complications in Korean Diabetics," *Journal of Korean Diabetes*, vol. 2, pp. 13-17, 1974.

[3] S. Raudys and F. Roli, "The behavior knowledge space fusion method: Analysis of generalization error and strategies for performance improvement", *4th Int. Workshop on Multiple Classifier Systems*, Springer, LNCS 2709, pp. 55-64, 2003.

[4] H.-K. Lee, *et al.*, "Prevalence of diabetes and IGT in Yonchon County, South Korea," *Diabetes Care*, vol. 18, pp. 545-548, 1995.

[5] C.-S. Shin, *et al.*, "Risk factors for the development of NIDDM in Yonchon county, Korea," *Diabetes Care*, vol. 20, pp. 1842-1846, 1997.

[6] K.-J. Kim, *et al.*, "Prediction of colon cancer using an evolutionary neural network," *Neurocomputing*, 2004. (in press)

[7] J. Park, *et al.*, "A sequential neural network model for diabetes prediction," *Artificial Intelligence in Medicine*, vol. 23, pp. 277-293, 2001.

[8] L. Xu, *et al.*, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, pp. 418-435 1992.

[9] G. Valentini, *et al.*, "Cancer recognition with bagged ensembles of support vector machines," *Neurocomputing*, vol. 56, pp. 461-466, 2004.

[10] C. Park, *et al.*, "Searching for optimal ensemble of feature-classifier pairs in gene expression profile using genetic algorithm," *Journal of KISS*, vol. 31, pp. 525-536, 2004.