K-최근접 이웃 알고리즘을 활용한 심장병 진단 및 예측

박평우*, 이석원**
*아주대학교 컴퓨터공학과
**아주대학교 소프트웨어학과
e-mail : *bryan2056@ajou.ac.kr, **leesw@ajou.ac.kr

Classification of Heart Disease Using K-Nearest Neighbor Imputation

Pyoung-Woo Park *, Seok-Won Lee**
*Dept. of Computer Engineering, Ajou University
**Dept. of Software and Computer Engineering, Ajou University

요 약

본 논문은 심장질환 도메인에 데이터 마이닝 기법을 적용한 연구로, 기존 환자의 정보에 대하여 K-최근접 이웃 알고리즘을 통해 결측 값을 대체하고, 대표적인 예측 분류기인 나이브 베이지안, 서포트 벡터 머신, 그리고 다층 퍼셉트론을 적용하여 각각 결과를 비교 및 분석한다. 본 연구의 실험은 K 최적화 과정을 포함하고 10-겹 교차 검증 방식으로 수행되었으며, 비교 및 분석은 정확도와 카파 통계치를 통해 판별한다.

1. 서론

1.1 연구배경

최근 인공지능을 통해, 우리 현세대의 인간은 그 본연의 능력을 넘어서는 수준의 데이터를 활용할 수 있게 되었으며, 보다 정확하고 효율적인 의사결정이 가능하게 되었다. 특히, 의료 분야에서 인공지능은 그 정의와 기술력을 바탕으로 막대한 효과를 거둘 수 있다.

현재, 의료 지식 및 데이터의 축적이 가속화되면서 의료 전문가와 의사는 이러한 모든 정보를 활용하고 완벽한 처방과 진단을 내리기는 어려운 것이 현실이다. 또한, 의사마다 다른 관점과 특정 상황에 따라 같은 환자에 대하여 서로 상이한 진단을 내릴 수 있으며, 그 정확도와 신뢰도도 객관적으로 가늠하기 어렵다. 따라서 의사의 진단과 처방에 대하여 보조할 수 있는 시스템의 필요성이 대두되어왔고, 그에 대한 연구가 60 년 전부터 진행되어오고 있다.

본 논문에서 제안하는 심장질환 진단을 위한 예측 방법의 필요성은 다음과 같다. 첫 번째로, 의료 진단 및 처방의 정확도를 향상시키기 위함이다. 이는 체계적이고 결함이 없으며, 복합적인 데이터로부터 통합적인 결정을 완벽하게 내릴 수 있다면, 의사의 진단과 처방의 정확도는 높아질 수 있다. 두 번째, 신뢰도 측면에서 의료 의사결정 시스템은 효과적이다. 지속적으로 축적된 실제 데이터에 대하여 검증과 테스트를 거쳐, 상호적으로 시스템과 의사의 판단에 대한 신뢰도를 점차 높일 수 있다. 마지막 세 번째로, 시간과 비용이라는 한정된 자원 안에서 효율적이다. 조금 더 빠르고 정확하며 신뢰할 수 있는 의사결정을 내릴 수 있다면, 의사와 환자 모두에게 도움될 수 있기 때문이다.

본 논문에서의 타깃 도메인은 허혈성 심장질환으로, 인체의 심장에 혈액을 공급해주는 혈관인 관상동맥이 좁아지게 되면서, 심장근육 일부에 혈액 공급이 부족하게 되어 발생하는 질환이다. 이는 협심증과 심근경색, 심장 돌연사 등을 유발한다. 또한, 심장 근육 세포는 혈액이 30 분 이상 공급되지 못하면 죽게 되고, 심장 근육 세포가 죽은 부위는 제대로 기능을 할 수 없게 되어 심부전으로 이어질 수 있다. 그리고 증상이 나타난 지 한 시간 이내에 사망하는 심장 돌연사로도 진행될 수 있기 때문에 빠르고 정확하며 신뢰할 수 있는 의사의 진단과 처방이 필요하다. 더욱이, 심장질환은 우리나라의 주요 사망원인 중 2 위로 집계되었고, 국제연합은 21 세기 보건정책의 목표를 심혈관 질환, 만성 질환 관리로 전환하였으며, 세계 각국은 고령화 사회로 진전되고 질병 구조가 만성병 구조로 바뀌고 있음을 인지하여 범국가적 차원에서 만성질환 예방 사업의 중요성을 재인식하고 있다.

1.2 연구목표

본 연구의 목표는 다음 두 가지 목표를 두고 진행한다. 첫 번째 목표는 기존의 데이터를 이용하여 의료 전문가와 의사를 보조하는 것이다. 이는, 심장질환 의심 환자의 임상 정보와 검사 정보 등을 바탕으로 기존의 데이터와 비교 및 분석하여 진단함으로써 목적을 달성할 수 있다. 이를 통하여, 질환의 인과 관계

를 설명하는데 도움을 줄 수 있고, 불필요한 검사를 최소화하여 시간 및 비용 효율성을 높일 수 있다. 두 번째 목표는 심장질환 관련 의료인을 위한 교육 및 학습용 모델로 활용하는 것이다. 이는 일반적이거나 예외적인 상황까지 도출 및 예측된 결과를 통해, 자신의 경험과 지식을 효율적으로 극대화하는 데에 그 의의가 있다.

이에 따라, 본 논문에서는 인공지능의 한 분야인, 데이터 마이닝 기술 및 도구를 활용하여 심장질환에 대한 데이터를 분석하고 처리하며, 진단을 돕는 예측 방식을 제시한다.

2. 데이터 세트

2.1 데이터 세트 설명 및 분석

본 논문에서는 어바인, 캘리포니아 대학교에서 공개적으로 제공하는 데이터 세트를 활용한다. 해당 데이터 세트는 주로 심장질환 연구에 사용되는 것으로, 심장 이상의 진단 및 예측 시스템에서도 유용하게 쓰이고 있다 [1].

해당 데이터 세트는 클리블랜드, 헝가리, 스위스, 롱 비치의 병원에서 제공한 것으로, 결측 데이터는 물음표 또는 -9 값으로 표현되어 있으며 각 환자의 클래스 라벨은 0 부터 4 까지의 정수 값으로 환자의 심장병 진단 및 질환의 유무를 표현한다. 그러나 클리블랜드를 제외한 나머지 병원의 데이터들은 속성 데이터에 결측 값이 상당히 많기 때문에 정확한 진단의 지표가 되기 어렵다. 다음의 <표 1>에서는 각 병원의 클래스 분포와 결측 값을 가진 데이터의 수를 정리한다.

<표 1> 클래스 분포 및 결측 값 정리

데이터 세트	클래스 분포						결측 데이터 수
	0	1	2	3	4	전체 데이터 수	
클리블랜드	164	55	36	35	13	303	6
헝가리	188	37	26	28	15	294	293
스위스	8	48	32	30	5	123	123
롱 비치	51	56	41	42	10	200	199
전체 데이터	411	196	135	135	43	920	621

네 개의 모든 데이터 세트는 다음 <표 2>와 같이, 숫자 형식 13 개의 속성과 심장병 유무를 나타내는 마지막 1 개의 클래스 라벨로 구성되어 있다.

<표 2> 심장질환 데이터 세트 구성

이름	설명	데이터 형식
나이 (age)	환자 나이	수치 자료형
성별 (sex)	환자 성별	1=남성/0=여성
흉통 유형 (cp)	환자 흉통 유형	1=전형적 협심증/ 2=비전형적 협심증/ 3=비협심증성 흉통/ 4=무통
안정 혈압 (trestbps)	입원 시 안정 혈압 (mmHg)	수치 자료형
혈청 콜레스테롤 (chol)	환자 혈청 콜레스테롤 (mg/dl)	수치 자료형
공복 혈당 (fbs)	공복 혈당량 120mg/dl 기준 판별	1=120mg/dl 초과/ 0=120mg/dl 이하
심전계 결과 (restecg)	환자 안정 시 심전계 결과	0=정상/ 1=비정상 ST-T 파동/ 2=좌심실 비대증
최대 심박수 (thalach)	환자 최대 심박수	수치 자료형
운동 유발 협심증 (exang)	운동 시 협심증 유발 유무	1=유/0=무
ST 분절 하강 (oldpeak)	운동 시 ST 분절 하강 정도	수치 자료형
ST 분절 경사 (slope)	최고 ST 분절 기울기	1=오르막/2=편형/ 3=내리막
주요 혈관 수 (ca)	형광 투시법 시 혈관 수	0=0 개/1=1 개/2=2 개/ 3=3 개
결손 (thal)	심장 결손	3=정상/6=고정 결손/ 7=가역 결손
심장질환 진단 (num)	혈관 조영검사 결과	0=정상/1~4=질병

2.2 데이터 세트 처리

본 논문에서는 클리블랜드, 헝가리, 스위스, 롱 비치 네 개의 데이터 세트를 모두 통합하여 하나의 데이터 세트로 만들고, 클래스 라벨 값을 0 과 1 만으로 수정한다. 즉, 원본 데이터 세트에서 0 값을 갖는다면 그대로 0 값으로 두고, 1 부터 4 까지의 값을 갖는다면 1 의 값으로 수정한다. 이는 1 이상의 값을 가질 때 모두 심장질환이 있다고 판단할 수 있기 때문이다. 또한, -9 값으로 표현된 결측 값에 대하여 모두 동일하게 물음표 표기로 대체한다.

3. 전처리 과정

3.1 결측 데이터

연구 데이터 세트에서 결측 데이터는 하나 이상의 결손 값을 가지고 있는 경우의 데이터 수로, 전체 데이터 920 개 중 621 개, 약 67.5%를 차지한다. 따라서 데이터 마이닝 기술을 적용하였을 경우, 정확도와 신뢰도를 보장하기 어렵다. 다른 논문에서는 해당 데이터 세트에 대한 결측 데이터 처리 작업을 수행하지 않았고 [2], 평균 값을 사용하였으며 [3], 이를 무시하고 가정하거나 그대로 삽입하여 정확도를 측정하였다 [4].

3.2 K-최근접 이웃 알고리즘

본 연구에서는 결측 데이터를 처리하는 방안으로 K-최근접 이웃 알고리즘을 사용한다. K-최근접 이웃 대체법은 결측 값이 발생한 개체와 가장 가까운 거리에 있는 K 개의 이웃 개체와 거리를 계산하여 다수결로 결측 값을 대체하는 방법이다 [5].

이 때, 최근접 이웃 알고리즘의 거리 계산은 고어 거리 계산 방식을 활용하여 수행하는데, 이는 이진 데이터 형식, 범주형 자료 형식, 순서형 자료 형식, 연속 자료 형식 등 다양한 자료 형식을 가진 데이터 세트에서 연산을 지원한다. 고어 거리는 계산 시, 각 변수 값에 무게 값을 곱하여 평균으로 연산하는 방식으로, 해당 수식은 다음과 같다.

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \delta_{i,j,k}}{\sum_{k=1}^{p} w_k}$$

이후, 고어 거리 계산을 통해 가장 가까운 K 개의 이웃 개체를 선정하고 비슷한 값끼리 다수결로 판별하여 해당 결측 값을 채운다.

4. 분류 알고리즘

4.1 나이브 베이지안

나이브 베이지안은 베이즈 정리에 기반한 알고리즘으로, 다른 분류기에 비해 간단한 접근 방식과 명료한 의미 표현, 확률적인 지식의 학습을 지원한다. 계산 시 가정하는 조건은 크게 두 가지로, 각 예측에 대한 속성 값이 서로 조건부 독립이라는 것, 드러나지 않거나 잠재적인 속성이 예측 과정에 영향을 미치지 않는다는 것이다 [6].

4.2 서포트 벡터 머신

서포트 벡터 머신은 패턴 인식, 자료 분석을 위한 지도 학습 모델로, 주로 분류와 회귀 분석으로 사용된다. 두 클래스 중 어느 하나에 속한 데이터 세트가 주어졌을 때, 해당 알고리즘은 주어진 데이터 세트를

바탕으로 새로운 데이터가 어느 클래스에 분류되는지 판단하는 모델을 생성하고, 이 때 생성된 분류 모델은 비확률적 이진 선형 분류 형식으로 사상 공간에서 경계를 구분 짓는다.

본 논문에서는 기존의 모델에 순차적 최소 최적화 방식이 적용된 서포트 벡터 머신 알고리즘을 사용한다. 이는 이차적 문제에서 가능한 작은 시리즈로 나누어 처리하는 방식으로 상대적으로 계산이 빠르고 선형 분류 정확도가 높다 [7].

4.3 다층 퍼셉트론

다층 퍼셉트론은 단층 퍼셉트론에 은닉층이 더해진 형태로, 입력층에서 전달되는 출력 값이 은닉층으로 전달되고 은닉층의 출력 값이 출력 층에 전달되는 구조를 가진다. 또한, 다층 퍼셉트론에서는 은닉층의 목표 값을 정의할 수 없기 때문에 역전파 기법을 활용하여 오차를 통해 가중치를 수정한다. 본 연구에서는 해당 네트워크에 시그모이드 함수를 적용하여 수행한다.

5. 실험 결과 및 분석

5.1 전처리 실험

전처리 수행 과정에서는 오픈소스로, 통계 및 데이터 마이닝 분야에서 많이 활용되고 있는 프로그램인 R 을 활용한다. 결측 데이터를 처리한 후, 전체 데이터 세트는 결측 값이 없는 완성된 데이터 세트로 변환된다.

결측 데이터 대체 시, K 값의 범위는 1 부터 13 까지 과반수 값을 획득할 수 있는 홀수 값만으로 설정하여, K-최근접 이웃 알고리즘에 적용한다. K 값 최적화 과정은 직접 1, 3, 5, 7, 9, 11, 13 에 대하여 대체된 데이터 세트를 대표적인 세 분류 알고리즘, 나이브 베이지안, 서포트 벡터 머신, 그리고 다층 퍼셉트론에 적용하여 정확도 및 카파 통계치를 근거로 하여 판단할 수 있다. 다음의 <표 3>은 K 값에 따른 세 알고리즘의 정확도를 표현하였고, <표 4>는 K 값에 따른 세 알고리즘의 카파 통계치를 표현한다. 알고리즘의 적용은 자바 기반의 오픈소스 소프트웨어인 WEKA 를 사용하였고, 테스트 및 검증 방식으로 10-겹 교차 검증 방식을 이용한다.

10-겹 교차 검증 방식은 과적합 현상을 피하기 위하여 데이터 세트를 10 개의 서브 세트로 분할하여 사용한다. 또한, 본 연구에서는 데이터의 양이 충분하지 않고, 분류기 성능의 통계적 신뢰도를 높이기 위하여 사용하였다.

5.2 K-최근접 이웃 적용 결과 및 분석

먼저, K 값에 1 부터 13 까지 홀수만을 대입하여 전처리 과정을 완료한 후, 각 K 값에 따라 생성된 데이터 세트를 알고리즘에 적용하였다. 그 정확도와 카파 통계치는 아래 <표 3>과 <표 4>와 같다.

<표 3> K 값에 따른 알고리즘의 정확도

K	나이브 베이지안 (%)	서포트 벡터 머신 (%)	다층 퍼셉트론 (%)	평균 (%)
1	86.7391	87.5000	88.2609	87.5000
3	88.0435	89.4565	90.9783	89.4928
5	88.4783	88.9130	88.5870	88.6594
7	89.0217	88.9130	89.2391	89.0579
9	89.1304	89.2391	90.0000	89.4565
11	89.1304	89.7826	88.9130	89.2753
13	89.8913	90.0000	89.1304	89.6739
평균 (%)	88.6335	89.1149	89.3012	

<표 4> K 값에 따른 알고리즘의 카파 통계치

K	나이브 베이지안	서포트 벡터 머신	다층 퍼셉트론	평균
1	0.7315	0.7472	0.7631	0.7473
3	0.7576	0.7864	0.8179	0.7873
5	0.7665	0.7752	0.7690	0.7702
7	0.7774	0.7756	0.7825	0.7785
9	0.7798	0.7826	0.7981	0.7868
11	0.7802	0.7938	0.7756	0.7832
13	0.7956	0.7985	0.7715	0.7885
평균	0.7698	0.7799	0.7825	

정확도는 전체 테스트 세트에서 학습한 데이터를 기반으로 예측하였을 때, 클래스 값과 얼마나 동일한가를 알아보는 것으로, K 값이 3, 7, 11, 13 인 경우 대체적으로 높은 정확도를 보였다. 제일 높은 정확도는 K 값이 3 이고 다층 퍼셉트론 알고리즘을 적용하였을 때 나타났음을 알 수 있다.

카파 통계치는 코헨의 카파로 불리는 통계량으로, 두 명의 평가자가 있다고 가정하였을 때 이 일치도를 확인한 수치 값이다 [8]. 평가자 간 예측 및 분류 결과가 다르면 0 에 가깝고 일치하면 1 에 가까운 값을 갖는다. 일반적으로, 코헨의 카파 값이 0.6 이상일 경우 타당하다고 판단할 수 있는데, 연구 결과 중 K 값이 대체로 3, 9, 11, 13 일 때 좋은 결과를 보인다. 가장 타당한 결과는 정확도와 마찬가지로 K 값이 3 이고 다층 퍼셉트론 알고리즘을 적용한 경우이다.

5.3 실험 결과 비교

K-최근접 이웃 알고리즘을 결측 값 대체로 처리한 결과, 최적화 시 K 의 값은 3 으로 다층 퍼셉트론이 예측 알고리즘에 적합하다는 것을 알 수 있다. 마지막으로, 원본 데이터의 나이브 베이지안과 서포트 벡터 머신 그리고 다층 퍼셉트론 예측 결과와 비교 및 분석하며 실험을 종료한다. 아래 <표 5>와 <표 6>은 각각 원본 데이터 세트와 3-최근접 이웃 알고리즘을 적용한 데이터 세트 정확도 및 카파 통계치를 나타낸 결과이다.

<표 5> 원본 및 전처리 데이터 세트 정확도

데이터 세트	나이브 베이지안 (%)	서포트 벡터 머신 (%)	다층 퍼셉트론 (%)	평균 (%)
원본	83.2609	81.5217	77.6087	80.7971
전처리	88.0435	89.4565	90.9783	89.4928
우위	전처리	전처리	전처리	전처리
향상 값	4.7826	7.9348	13.3696	8.6957

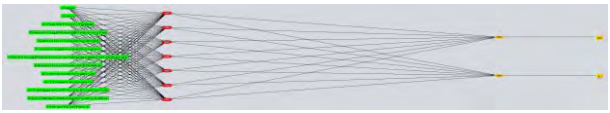
<표 6> 원본 및 전처리 데이터 세트 카파 통계치

데이터 세트	나이브 베이지안	서포트 벡터 머신	다층 퍼셉트론	평균
원본	0.6606	0.6248	0.5487	0.6114
전처리	0.7576	0.7864	0.8179	0.7873
우위	전처리	전처리	전처리	전처리

결과를 살펴보면, 예측 정확도 측면에서 나이브 베이지안의 경우 4.7826%, 서포트 벡터 머신은 7.9348%, 그리고 다층 퍼셉트론은 13.3696% 향상되었고, 카파 통계치 값은 모두 상승하였다. 따라서 3-최근접 이웃 알고리즘을 적용하여 결측 값을 대체한 데이터 세트는 기존의 데이터 세트와 비교하여 모든 면에서 향상되었다고 할 수 있다.

5.4 실험 결과 분석

구체적으로, 본 실험에서 적용한 다층 퍼셉트론은 각 교차 검증마다 500 번 학습을 반복하였고, 학습 속도는 0.3, 그리고 네트워크는 다음 (그림 1)과 같이 구성하였다.



(그림 1) 다층 퍼셉트론 네트워크 구성도

3-최근접 이웃 알고리즘을 적용한 결과, 예측 정확도와 카파 통계치 값이 원본 데이터 세트보다 효율적임을 수행 결과를 통해 알 수 있었고, 다층 퍼셉트론에 대한 예측 분류가 가장 타당한 분류 방법이라는 것을 알 수 있었다.

6. 결론

6.1 연구결론

본 연구에서는 우리나라의 주요 사망원인 2 위로 집계되는 심장질환을 연구 도메인으로 설정하고, 인공지능의 데이터 마이닝 기술을 접목하여 의사와 환자 모두에게 이익이 되는 효율적인 접근 방식을 제안한다. 연구의 궁극적 목적은 심장질환 의심 환자의 누적된 정보를 바탕으로, 의료 전문가 및 의사를 보조하는 것과 의료인을 위한 교육 및 학습용 모델로 활용되도록 하는 것이다.

타깃 데이터 세트는 데이터 마이닝 심장질환 연구에 대표적으로 사용되는 캘리포니아 대학교 공개 데이터 세트를 기반으로 하였다. 그러나 해당 원본 데이터 세트를 분석한 결과, 상당히 많은 결측 데이터를 가지고 있음을 알 수 있었고, 이에 따라 진단 및 처방 예측을 목적으로 수학적 알고리즘을 적용하는 데에 정확도와 신뢰도 측면에서 보장할 수 없다는 판단을 하게 되었다. 본 논문에서는 해당 데이터 세트의 결측 값을 대체하기 위한 방법으로 K-최근접 이웃 알고리즘을 사용하였고, 대표적인 예측 분류기인 나이브 베이지안, 서포트 벡터 머신, 그리고 다층 퍼셉트론을 적용하였다. 실험 시, 정확도와 카파 통계치를 통해 그 성능을 판별하였으며, 3-최근접 이웃 알고리즘을 적용한 데이터 세트의 다층 퍼셉트론 예측 방식이 가장 뛰어나다는 결과를 도출하였다. 즉, 최종적으로 본 실험 및 연구의 결과에 따라, 3-최근접 이웃 결측 값 대체 알고리즘과 다층 퍼셉트론 예측 방식이 심장질환 판단 및 예측에 유용하다고 결론지을 수 있었다.

6.2 한계점

본 연구의 한계점은 크게 두 가지로 정리할 수 있다. 첫 번째, 타깃 데이터 세트로 설정한 캘리포니아 대학교 공개 데이터 세트만으로는 보편타당하고 완벽한 진단을 내리기 어렵다는 것이다. 이는 데이터 세트 자체의 환자 수와 속성 수가 다른 질병 연구에 비해 상대적으로 부족하다고 판단될 수 있기 때문이다. 두 번째는 본 연구에서 수행한 결측 값 대체 방식이 다른 관련 연구에서도 항상 옳을 수만은 없다는 것이다. 즉, 연구실험에서 수행한 K-최근접 이웃 알고리즘은 타깃 데이터 세트에 한하여 높은 정확도와 신뢰도를 보장하였지만, 해당 데이터 세트가 아닌 다른 개인 데이터나 타 심장질환 연구 데이터에서는 불안정한 데이터 세트를 만들 수도 있다는 것이다.

6.3 향후 연구

위에서 다룬 본 연구 한계점을 극복하기 위하여, 타 연구 기관 및 병원의 데이터 세트를 기존의 데이터 세트에 추가하여 연구를 진행할 수 있다. 필요하다면 전체 속성을 수정한 후, 결측 값은 K-최근접 이웃 알고리즘으로 다시 적용하여 완벽한 데이터 세트를 만들 수 있다. 이를 통해, 환자의 수와 속성의 수와 관련된 문제를 해결한다. 또한, K-최근접 이웃 알고리즘 외에도 다른 결측 값 대체 알고리즘을 사용하여 연구를 진행할 수 있다. 이 때에는 결측 값 대체 방식과 수학적 예측 분류기를 교체해가며 예측 결과를 비교 및 분석하는 방향으로 진행한다. 이 외에도 데이터 세트의 속성에 대하여 유전 알고리즘, 또는 대칭적 불확실성 특성 선택 알고리즘을 적용해볼 수 있다.

Acknowledgment

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW 중심대학지원사업의 연구결과로 수행되었음 (20150009080031001)

이 논문은 2017 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.2017R1D1A1B03034279)

참고문헌

[1] Noreen Kausar, and others (2015). “Review of Data Mining Approaches for Extraction and Classification of Clinical Data in Diagnosis of Coronary Artery Disease”, *ARPN Journal of Engineering and Applied Sciences*, Vol. 10, No. 15, pp. 6679-6685.

[2] Shruti Ratnakar, K. Rajeswari and Rose Jacob (2013). “Prediction of Heart Disease Using Genetic Algorithm for Selection of Optimal Reduced Set of Attributes”, *International Journal of Advanced Computational Engineering and Networking*, Vol. 1, Issue 2, pp. 51-55.

[3] Atul Kumar Pandey, Prabhat Pandey and K. L. Jaiswal (2014). “Classification Model for the Heart Disease Diagnosis”, *Global Journal of Medical Research*, Vol. 14, Issue 1, pp. 9-14.

[4] Xiaoyong Liu and Hui Fu (2014). “PSO-Based Support Vector Machine with Cuckoo Search Technique for Clinical Disease Diagnoses”, *The Scientific World Journal*, Vol. 2014, Article, 548483.

[5] Serena G Liao, and others (2014). “Missing Value Imputation in High-Dimensional Phenomic Data: Imputable or Not, and How?”, *BMC Bioinformatics*, 15:346.

[6] George H. John and Pat Langley (1995). “Estimating Distributions in Bayesian Classifiers”, *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, San Mateo, USA, pp. 338-345.

[7] John C. Platt. “Fast Training of Support Vector Machines Using Sequential Minimal Optimization”, *Microsoft Research*, Redmond, USA, pp. 41-65.

[8] Mikel Aickin. “Maximum Likelihood Estimation of Agreement in the Constant Predictive Probability Model, and Its Relation to Cohen’s Kappa”, *International Biometric Society*, Vol. 46, No. 2, pp. 293-302.