퍼지 클러스터 분석 기반 특징 선택 방법

이 현 숙†

요 약

특징선택은 문제 영역에서 관찰된 다차원데이터로부터 데이터가 묘사하는 구조를 잘 반영하는 속성을 선택하여 효과적인 실험 데이터를 구성하는 데이터 준비과정이다. 이 과정은 문서분류, 영상인식, 유전자 선택 분야에서와 같은 분류시스템의 성능향상에 중요한 구성요소로서 상관관계 기법, 차원축소 및 상호 정보 처리 등의 통계학이나 정보이론의 접근방법을 중심으로 연구되어왔다. 이와 같은 특징 선택 분야의 연구는 다루는 데이터의 양이 방대해지고 복잡해지면서 더욱 중요시 되고 있다.

본 논문에서는 데이터가 가지는 특성을 반영하면서 새로운 데이터에 대하여 일반화 할 수 있는 특징선택 방법을 제안하고자한다. 준비된 데이터의 각 속성 데이터에 대하여 퍼지 클러스터 분석에 의하여 최적의 클러스터 정보를 얻고 이를 바탕으로 군집성과 분리성의 정도를 측정하여 그 값에 따라 특징을 선택하는 메카니즘을 제공한다. 제안된 방법을 실세계의 컴퓨터 바이러스 분류에 적용하여 기존의 대비에 의한 휴리스틱 방법에 의해 선택된 데이터를 가지고 분류한 것과 비교하고자 한다. 이를 통하여 주어진 특징에 서열을 부여할 수 있고 효과적으로 특징을 선택하여 시스템의 성능을 향상 시킬 수 있음을 확인한다.

키워드 : 특징 선택, 퍼지 클러스터 분석, 성능 측정자, 정보이론

A Feature Selection Method Based on Fuzzy Cluster Analysis

Rhee, Hyunsook†

ABSTRACT

Feature selection is a preprocessing technique commonly used on high dimensional data. Feature selection studies how to select a subset or list of attributes that are used to construct models describing data. Feature selection methods attempt to explore data's intrinsic properties by employing statistics or information theory. The recent developments have involved approaches like correlation method, dimensionality reduction and mutual information technique. This feature selection have become the focus of much research in areas of applications with massive and complex data sets.

In this paper, we provide a feature selection method considering data characteristics and generalization capability. It provides a computational approach for feature selection based on fuzzy cluster analysis of its attribute values and its performance measures. And we apply it to the system for classifying computer virus and compared with heuristic method using the contrast concept. Experimental result shows the proposed approach can give a feature ranking, select the effective features, and improve the system performance.

Key Words : Feature Selection, Fuzzy Cluster Analysis, Performance Measure, Information Theory

1. 서 론

패턴인식, 로보틱스, 전문가시스템 등의 지능형시스템은 공통적으로 system identification (SI)과정을 기반으로 설계되어 있다. SI는 주어진 입력데이터로부터 출력데이터로의 사상(mapping) 방법을 중심으로 연구되어왔다. 사상방법에는 회귀기법과 같은 표면적인 사상방법과 최근 많은 연구결과를 만들어 낸 신경망과 퍼지 시스템 등의 소프트 컴퓨팅에 의한 암시적 사상방법으로 나눌 수 있다. 신경망의 연구를 중심으로 하는 암시적 사상방법은 black-box 모델로서 그 처리과정을 설명하기는 어렵지만 병렬처리에 의하여 데이터를 학습해 가는 견고한 최적 메카니즘으로 여러 분야에서 타당한 연구결과를 도출하였다. 특히 역 전파 신경망 Backpropagation Neural Networks)과 자기조직화 지도(Self Organizing Map) 학습을 통하여 분류해 내는 대표적인 신경망 모델은 음성인식, 영상인식, 자동화 시스템, 여러 종류의 예측시스템 등의 분야에 적용되어 실용화 단계의 안정된 성능을 나타내었다[1].

이와 같은 SI 모델의 가장 핵심적인 기능은 목적으로 하는 출력공간안의 여러 상태로의 분류기법이다. 또한 분류기능은 새로운 정보가 계속해서 발생하고 비슷한 상황에 적절히 대처해야하는 대부분의 지능형 시스템에서 학습을 통하

† 정 회 원 : 동양공업전문대학 전산정보학부 부교수
논문접수 : 2007년 2월 22일, 심사완료 : 2007년 4월 2일

여 구현되고 있다. 이러한 분류시스템은 크게 데이터 준비 과정, 학습 처리 과정, 의사 결정 과정으로 나눌 수 있으며 지금까지의 연구는 학습 처리 과정과 의사 결정 과정에 초점을 맞추어 진행되어왔다[2]. 분류에서 다루는 데이터는 문서분류, 영상인식, 유전자인식, 컴퓨터 파일 분류와 같이 그 양이 방대하며 실세계에서 관찰되는 데이터를 그대로 시스템의 입력으로 사용될 수 없기 때문에 데이터 준비 과정은 시스템의 성능을 좌우하는 중요한 과정으로 알려져 있다. 그러나 이를 위하여 시스템 구축 시 준비된 데이터의 속성이나 분포에 대한 정보를 알고 있는 전문가의 개입이 필요하며 이는 시스템 설계를 일반화시키기 어렵게 만드는 부분이므로 분리하여 처리해왔다. 최근 이와 같은 데이터 준비 과정의 중요성을 인식하고 분류 시스템 안에 통합시키려는 연구가 진행되고 있다[3].

특히 특징선택은 문제 영역에서 관찰된 다차원데이터로부터 데이터가 묘사하는 구조를 잘 반영하는 속성을 선택하여 효과적인 실험데이터를 구성하는 데이터 준비과정의 핵심부분이다. 이 과정은 문서분류, 영상인식, 유전자 선택 분야에서와 같은 분류시스템의 성능향상에 중요한 구성요소로서 상관관계 기법, 차원축소 및 상호 정보 처리 등의 정보이론이나 통계학의 접근방법을 중심으로 연구되어왔다. 이와 같은 특징 선택 분야의 연구는 다루는 데이터의 양이 방대해지고 복잡해지면서 더욱 중요시 되어 인공지능 분야의 휴리스틱 방법이 적용되기도 하였다. 이는 전문가 시스템 분야의 지식획득의 어려움(Knowledge Acquisition Bottleneck)과 유사하며 앞으로 계속 연구되어야 할 분야로 각광받고 있다[4,5].

본 논문에서는 데이터가 가지는 특성을 반영하면서 새로운 데이터에 대하여 일반화할 수 있는 특징 선택 방법을 제안하고자한다. 준비된 데이터의 각 속성 데이터에 대하여 퍼지 클러스터 분석에 의하여 최적의 클러스터 정보를 얻고 이를 바탕으로 군집성과 분리성의 정도를 측정하여 그 값에 따라 특징을 선택하는 일반적인 메카니즘을 제공한다. 제안된 방법을 실세계의 컴퓨터 바이러스 분류에 적용하여 기존의 대비에 의한 휴리스틱 방법에 의해 마련된 데이터[6,7]를 가지고 분류한 것과 비교하고자 한다. 이를 통하여 주어진 특징에 서열을 부여할 수 있고 효과적으로 특징을 선택하여 시스템의 성능을 향상 시킬 수 있음을 확인한다.

2장에서는 본 논문을 위하여 미리 연구 발표된 퍼지 클러스터링을 위한 목적기반 퍼지 신경망, FNN-B(Fuzzy Neural Networks-Batch Learning Version)와 본 논문에서 적용하고자 하는 컴퓨터 바이러스 분류 분야에서 사용해 온 특징선택 방법에 대하여 기술한다. 3장에서는 논문의 중심부분으로 제안된 클러스터 분석 기반 특징 선택 방법과 이에 사용된 클러스터 성능 측정자에 대하여 기술한다. 4장에서는 실세계의 파일로부터 추출한 바이러스 분류를 위한 데이터에 적용하여 제안된 방법에 의하여 특징을 추출한 후 분류에 적용하는 과정을 기술하고 기존의 대비에 의한 휴리스틱 방법[6,7]과 비교한다. 5장은 결론으로서 제안한 방법을 요약하고 앞으로의 발전 방향을 기술한다.

2. 관련 연구

2장에서는 본 논문의 기초가 되는 관련 연구로서 퍼지 클러스터링을 위한 목적 함수 기반 퍼지 신경망 FNN-B를 고찰해 보고 적용하고자 하는 컴퓨터 바이러스 분류 분야에서 사용해 온 특징추출 방법에 대하여 기술한다[8,9].

2.1 목적 함수 기반 퍼지 신경망

FCM(Fuzzy c-Means) 알고리즘 목적함수 J_m을 비교사 학습신경망에 결합시켜(그림 1)와 같은 퍼지신경망, FNN-B 구성하였다[8]. 이렇게 구성된 신경망에서 다음의 알고리즘을 통하여 입력 층에 제공된 데이터 $X = \{\boldsymbol{x}_1,...,\boldsymbol{x}_n\}$는 대표정보인 클러스터의 중심점 $(v_1,v_2,...,v_c)$을 학습해 간다. 이러한 학습을 통해 형성된 클러스터 층은 데이터 $\boldsymbol{x}_j$ 의 클러스터 i에 속하는 소속 값, u_{ij} 을 포함하는 정보 사이의 관계를 표현하는 값인 $(\alpha_1,\alpha_2,...,\alpha_c)$를 계산하여 그 결과를 다음 학습에 활용한다. 이러한 학습 알고리즘은 클러스터링의 결과가 만들어 내는 오류 값을 요약하는 퍼지 함수를 설정한 후 그 값이 최소가 되도록 학습의 방향을 유도하는 메카니즘에 의해 진행된다. 또한 제안된 방법은 입력과 출력 사이의 관계를 기술하기 어려운 경우도 쉽게 처리하는 비교사 학습신경망의 장점도 함께 가지고 있다.

단계 1 : c, m, ϵ의 값을 설정하고 입력데이터 셋을 준비한다. c 는 클러스터의 수, m 은 FCM 알고리즘의 weighting exponent 이다.

단계 2 : 초기 가중치 벡터 $V=(v_1,v_2,...,v_c)$와 퍼지 C 분할 U를 0과 1 사이의 난수로 초기화한다. $t = 0$.

단계 3 : 다음 식을 이용하여 $(v_1,v_2,...,v_c)$ 를 계산하고

$$\eta_i = \frac{1}{\sum_{j=1}^{n}\alpha_{ij}} \text{ 이라고 하자.}$$

$$\alpha_{ij} = \frac{2m}{m-1}\left\{\sum_{l=1}^{c}(u_{lj})^{m+1}\left(\frac{\|x_j - v_l\|^2}{\|x_j - v_i\|^2}\right)\right\}^{\frac{m}{m-1}}$$

단계 4 : 다음 식을 이용하여 가중치 벡터를 수정한다.

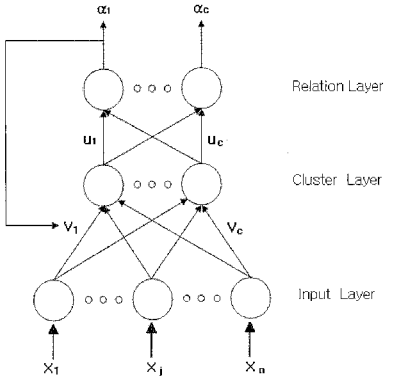
$$\Delta v_i = \eta_i \sum_{j=1}^{n}\alpha_{ij}(x_j - v_{ij})$$

단계 5 : 다음 식을 이용하여 퍼지 C 분할 U를 계산한다.

$$u_{ij} = \frac{1}{\sum_{s=1}^{c}\left\{\frac{\|x_j - v_i\|}{\|x_j - v_s\|}\right\}^{2/(m-1)}}$$

단계 6 : $\boldsymbol{v}_{i,t}$ 는 현재 가중치 벡터이고 $\boldsymbol{v}_{i,t+1}$는 단계4 에 의하여 수정된 가중치 벡터일때

$diff = \sum_{i=1}^{c}\|v_{i,t+1} - v_{i,t}\|^2 < e$ 이면 알고리즘을 끝내고 그렇지 않으면 단계 3으로 가서 처리를 계속한다.



(그림 1) 목적함수기반 퍼지 신경망(FNN-B)

2.2 컴퓨터 바이러스 분류를 위한 특징선택 방법

1986년 컴퓨터 바이러스가 처음 출현한 이래 매년 수많은 새로운 바이러스가 중요한 정보를 저장하고 있는 컴퓨터를 위협하고 있으므로 이는 점점 현실적인 문제로 대두되고 있다. 이에 바이러스 파일을 분석하여 시스템으로부터 이를 탐지하여 대처하기 위한 연구들이 계속되고 있다. 이러한 연구의 대부분은 탐지를 위한 중요한 정보로서 코드의 특정부분 문자열을 시그내처로 추출하기 위한 방법론에 초점을 맞추고 있다. 최근에 그동안 문서분류와 음성인식 등의 자연언어처리 분야에 활용되어 온 n-gram 분석 기법이 컴퓨터바이러스 탐지를 위한 시그내처 추출에 활용되고, Computer immune system 의 설계를 위하여 제안되기도 하였다[10]. 여기서 시그내처는 바이러스 탐지 시스템의 특징이 된다.

Abou-Assaleh 등[11]은 Commom N-Gram(CNG) 방법을 제안하여 알려지지 않은 새로운 파일을 진단하는데 활용하였다. 악성코드와 정상코드로 구성된 데이터베이스로부터 자주 출현하는 n-gram을 시그내처 로서 추출하여 저장한다. 이렇게 추출된 n-gram은 특정파일의 구조를 반영하는 정보를 함축하고 있으며 바이러스 침입자가 쉽게 예측하기 어려운 것으로 알려져 있다. 분석하고자하는 코드에 대하여 이미 저장된 시스내처로부터 k-nearest 알고리즘을 적용하여 정상코드인지 악성코드인지 분류 하게 된다. 이 방법은 파라메타 n과 추출된 시스내처의 수 L에 따라 민감하게 그 성능이 좌우되는 것으로 보고 되었으나 알려지지 않은 새로운 파일을 대상으로 하는 초기연구로서 가치가 있다. 또한 Kolter 등[12]도 비슷한 방법으로 접근하고 있으나 정보공학의 기법을 적용하였다. 특징으로 추출된 n-gram 들이 준비된 각 파일에 존재여부를 지시하는 이진데이터를 모아 평균상호 정보(average mutual information)를 계산하여 그 값이 큰 500개의 데이터를 선택하여 WEKA[13]에서 구현한 학습방법-Instance-based Learner, TFIDF, Naive Bayes, a support vector machines, a decision tree and a booted classifier-에 적용하였다. 준비된 데이터의 분류정확도에 의하여 분류하지 않고 ROC(receiver operating characteristics)에 의하여 평가하였다.

이와 같이 n-gram 기법들은 적용 가능 하기는 하나 부분문자열의 크기 n과 특징패턴의 개수 L과 같은 파라메타에 종속적인 결과를 가져오므로 일반적인 접근방법으로 발전시키기는 어렵다. n-gram 분석방법이 분류에 공헌하는 점을 관찰하여 이진실행파일을 역어셈블 하여 명령어를 구성하는 연산코드로부터 마이닝 기법을 활용하여 instruction sequence를 특징패턴으로 선택하는 기법이 제안되기도 하였다. 문서분류나 영상인식 분야와 마찬가지로 컴퓨터바이러스 분류의 대상이 되는 파일로부터 관찰되는 연산코드의 종류가 다양하다. 그러므로 바이러스 분류에 적합한 연산코드를 선택하여 활용할 수 있도록 정제하여 실험데이터를 구성하는 특징선택과정은 중요한 연구 분야이다. 그러나 이러한 특징선택은 시스템의 목적이나, 응용 영역에 따라 상이한 처리과정이 필요하므로 그 동안 분류를 위한 연구에서 전문가의 휴리스틱을 적용하는 것 이상의 접근방법을 다루지 못했다[6,7]. 최근 특징추출과정의 시스템 성능에 미치는 영향이 대두되면서 일반적인 접근을 위한 연구들이 진행되고 있다[3].

3. 클러스터 분석 기반 특징선택 방법

분류시스템을 통하여 시스템이 가진 문제를 해결하기 위하여 가장 중요한 과정은 입력데이터의 준비과정이다. 특징선택은 문제 영역에서 관찰된 다차원데이터로부터 데이터가 묘사하는 구조를 잘 반영하는 속성을 선택하여 효과적인 실험데이터를 구성하는 데이터 준비과정의 핵심부분이다. 이 과정은 문서분류, 영상인식, 유전자 선택 분야에서와 같은 분류시스템의 성능향상에 중요한 구성요소로서 상관 관계 기법, 차원축소 및 상호 정보 처리 등의 정보이론이나 통계학의 접근방법을 중심으로 연구되어왔다. 이와 같은 특징선택 분야의 연구는 다루는 데이터의 양이 방대해지고 복잡해지면서 더욱 중요시 되어 인공지능 분야의 휴리스틱 방법이 적용되기도 하였다.

본 논문에서는 퍼지 클러스터링과 그 결과 형성된 클러스터의 타당성을 측정하는 계산적 접근 방법에 의한 특징선택의 일반적 접근방법을 제시하고자한다. 실세계에서 준비된 데이터의 각 속성 데이터를 퍼지 클러스터 분석에 의하여 그룹핑하고 각 속성의 클러스터에 대한 군집성과 분리성의 정도를 측정하여 그 값에 따라 특징을 선택하는 일반적인 메카니즘을 제안한다.

m개의 속성을 가진 n개의 수집된 데이터는 $n \times m$ 데이터 집합으로 표현할 수 있다. 이때 데이터의 특징속성을 추출하기 위하여 n개의 각 속성 데이터를 그룹핑하여 클러스터를 만들고 c개의 중심 값과 $n \times c$의 퍼지 c 분할 정보를 얻어낸다. 이를 위하여 미리 연구하여 그 성능을 테스트한 퍼지 학습 신경망 FNN-B[8]를 적용한 후 형성된 클러스터에 대한 주어진 속성 데이터의 밀집성과 분리성의 정도를 측정하는 값을 계산한다. (정의1)에서는 군집성 측정자(compactness index) C를 정의하고 (정의2)에서는 분리성 측정자(separation index) D를 정의하고 있다[8]. 즉 C의

값이 작을수록, D의 값이 클수록 속성 데이터가 분류에 적합한 데이터를 가지고 있음을 알 수 있다.

정의1 : 퍼지 분할의 밀집성 C는 같은 클러스터 안에서의 임의의 두 데이터 사이의 거리의 평균으로서 전체 데이터 $a_1, \ldots, a_n$에 대하여 다음의 식 (1)과 같이 정의된다. 이때 ω_1은 임의의 두 데이터가 같은 클러스터에 속하는 소속정도를 나타낸다.

$$C = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \sum_{i=1}^{c} |a_j - a_k| \omega_1 \quad (1)$$

,where $\omega_1 = \min\{u_{ij}, u_{ik}\}$

정의2 : 퍼지분할의 분리성 D는 서로 다른 클러스터에 속하는 임의의 서로 다른 두 데이터 사이의 거리로서 전체 데이터 $a_1, \ldots, a_n$에 대하여 다음의 식 (2)와 같이 정의된다. 이때 ω_2는 임의의 두 데이터가 각각 서로 다른 두 클러스터에 속하는 소속정도를 나타낸다.

$$D = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} |a_j - a_k| \omega_2 \quad (2)$$

,where $\omega_2 = \min\{u_{i_1 j}, u_{i_2 k}\}$

이때 데이터 a_j의 가장 큰 소속 값을 가지는 클러스터를 i_1 이라하고 i_1이 아닌 클러스터 중 다른 데이터 a_k의 가장 큰 소속 값을 가지는 클러스터는 i_2라 한다.

이러한 퍼지 클러스터 분석 방법을 바탕으로 제안한 특징추출 과정을 요약하여 자세히 기술하면 다음과 같다.

단계 1 : $n \times m$의 데이터 $x(j,p)$를 준비한다. 클러스터의 수 c와 선택할 특징의 수 L을 정의한다.

단계 2 : $p = 1, \ldots, m$ 에 대하여

단계 2.1 : p 번째 속성데이터를 가지고 FNN-B에 의하여 퍼지 클러스터링을 수행한 후 c개의 중심 값, $v(1), \ldots, v(c)$와 $n \times c$의 퍼지 분할 정보 $u_p(j,i)$를 얻는다.

단계 2.2 : 2.1.의 정보를 이용하여 주어진 속성 데이터의 군집성 기준 C와 분리성 기준 D를 계산한다.

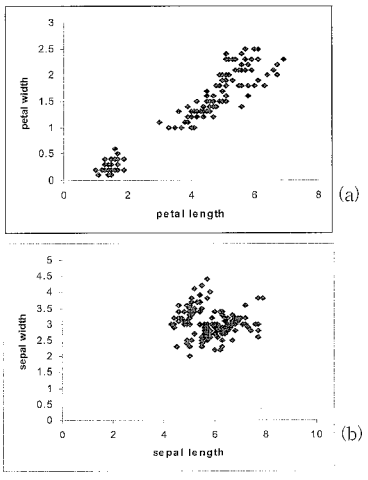
단계 2.3 : C와 D를 가지고 다음과 같이 특징을 선택한다. (경우 1) $C \neq 0$ 이면 $I(p) = D/C$ 로 계산하여 저장한다. (경우 2) $C = 0$ 이고 $D = 0$인 경우 속성 p는 특징으로 선택될 수 없다. (경우 3) $C = 0$ 이고 $D \neq 0$인 경우 속성 p는 특징으로 선택한다.

단계 3 : $I(p)$의 값이 큰 순서로 속성 p를 특징으로 선택하여 최적의 L개의 특징을 선택하여 $n \times L$ 의 정제된 데이터 집합을 얻는다.

$I(p)$는 밀집성과 분리성의 비율로 나타내므로 데이터 분포와 클러스터 사이의 관계를 반영한 값을 산출하며 데이터 집합 사이의 상대적인 비교가 가능하게 하였다. 보통 성능 측정지가 가지는 c가 커짐에 따라 단조 감소하는 경향이 적으며 클러스터 대표 값 보다는 실제 데이터를 기준으로 하기 때문에 (단계 2)와 같이 분석할 수 있다. (단계 2)에서의 (경우 2)는 모든 데이터가 같은 값을 가지는 경우이므로 특징 데이터는 분류에 도움을 주지 못하므로 선택될 수 없다. 또한 (경우 3)은 같은 그룹에 속해 있는 데이터는 모두 같은 값을 가지고 있는 hard partitioning 된 경우로 특징데이터로 선택될 수 있다. 그러나 (경우 2)와 (경우 3)은 예외적인 경우이며 대부분의 데이터는 모든 속성에 대하여 얻은 측정값 $I(p)$를 기준으로 그 값이 큰 순서로 특징을 서열화(feature ranking) 할 수 있으며 정해진 L 개의 특징을 추출하게 된다.

제안된 방법의 타당성을 확인하기 위하여 널리 알려진 벤치마크 데이터인 iris 데이터를 사용하였다. iris 데이터는 3가지 종류의 iris 식물로부터 sepal length, sepal width, petal length, petal width의 4가지 속성을 측정한 150개의 데이터이다. 이 데이터는 한 가지 종류는 다른 것으로부터 선형분리 가능하고 나머지 두 종류는 선형분리 가능하지 않는 것으로 알려져 있다. 이 데이터를 가지고 위의 알고리즘에 의하여 실행한 결과, $I(p)$의 값이 큰 두 가지 속성은 petal length와 petal width로서 특징으로 선택되었다. (그림 2)에서 알 수 있듯이 특징으로 선택된 petal length와 petal width는 추출한 iris data의 분포를 반영하고 있으며 선택되지 않은 sepal length와 sepal width는 그 분포를 반영하지 않고 있음을 알 수 있다.

이와 같이 제안된 방법은 데이터 집합이 가지고 있는 속성이 독립적이며 보통 전문가에 의해 부여되는 속성의 정성적인 가중치를 고려하지 않았다. 이러한 가정 하에서 속성 상호간의 관계정보나 중복성과 같은 복잡한 요소를 고려하지 않고 각 속성이 주어진 클러스터 형성에 공헌하는 정도를 $I(p)$의 값으로 요약하여 특징선택의 기준으로 마련하였다. 실세계로부터 무작위로 관찰된 다차원 데이터의 경우 $I(p)$의 값을 기준으로 관찰된 속성에 대하여 군집성과 분



(그림 2) iris data 분포 (a) petal length-petal width (b) sepal length-sepal width

리성을 기준으로 서열(ranking)을 정하는 것은 데이터 활용을 위한 정보로 사용할 수 있다.

4. 실험 및 결과

본 장에서는 3장에서 제안한 클러스터 분석 기반 특징선택을 컴퓨터 바이러스 분류에 적용하여 그 타당성을 검증해 보고자한다. 컴퓨터바이러스 분류를 위한 파일을 준비하면 이를 분석하여 전 처리 과정을 거쳐 속성이 될만한 변수를 규정하고 이에 따라 입력데이터를 만든다. 제안된 방법을 컴퓨터 바이러스 분류시스템에 적용하기 위하여 VX heaven[14] 으로부터 200개의 바이러스 파일과 윈도우시스템 실행 파일로부터 200개의 정상파일을 수집하여 다음과 같은 전 처리 과정을 통하여 입력 데이터를 준비하였다. 우선 수집한 이진실행파일(binary executables)을 IDA Pro[15]를 사용하여 역어셈블 한다. 이 과정에서 대부분의 실행파일은 완전하게 역어셈블 되지 않기 때문에 경우에 따라 휴리스틱을 사용하여 추출하기도 한다. 이제 역어셈블 된 코드를 블록으로 나누고 각 블록에 대하여 블록이름과 그 안에 있는 명령어의 연산 코드(instruction operation code)로 구성된 중간 파일을 만든다. 이 중간 파일로부터 각 명령어의 출현빈도수를 구하고 이를 바탕으로 특징 연산코드를 추출하게 된다. 이러한 처리과정은 악성코드의 탐지는 정상코드와 구별되는 특징패턴으로부터 알 수 있고 그런 정보는 파일을 구성하는 명령어로부터 얻어질 수 있다는 자연스러운 아이디어에서 출발하고 있다. 특히 명령어는 연산코드와 피연산자 부분으로 되어 있는데 연산코드만으로도 파일의 내용을 모두 표현할 수 있고 특징패턴을 추출할 수 있다. 이때 피연산자 부분은 고려하지 않으므로 중간파일의 크기를 상당히 줄일 수 있고 입력데이터의 준비를 단순화 시킬 수 있다. 지난 연구에서 사용된 데이터에서는 출현빈도수로부터 50개의 특징연산코드를 추출하였다.

이제 이렇게 마련된 입력데이터의 속성을 분석하여 분류시스템에 필요한 속성을 선택하여 실험데이터로 적용해야한다. 제안된 방법을 비교하기 위하여 바이러스 파일에 자주 나오는 명령어 열은 정상파일에는 잘 나오지 않는다는 휴리스틱과 전문가에 의한 파일분석을 통하여 26개의 명령어 열을 특징패턴으로 선정하였다[6,7]. 이는 각 명령어 열의 각 클래스 안에서의 정규화된 출현횟수에 기초한 것으로 학습하고자하는 파일이 변경되거나 새로운 파일 정보를 학습에 추가시키기 위해서는 처음부터 다시 분석해야하는 단점을 가지고 있다. 이를 편의상 대비에 의한 휴리스틱 방법이라 하고 각 파일에 대하여 선정된 26개 명령어 패턴의 정규화된 출현횟수를 구하여 분류에 사용될 400×26의 입력 데이터 Hdata를 마련하여 본 논문에서 제안된 방법과 성능 비교를 하고자 준비하였다.

본 논문에서는 준비된 파일의 중간파일로부터 얻은 50개의 연산 코드 패턴의 파일 안에서의 정규화 된(normalized) 출현횟수를 구하여 400×50의 데이터를 마련한다. 여기에 대비에 의한 휴리스틱 방법을 적용하는 대신 본 논문의 3장에서 기술한 특징선택방법의 입력으로 사용하여 26개의 특징패턴을 준비하고 400×26의 Cdata를 준비한다. 여기서 Hdata와의 비교를 용이하게 하기 위하여 L을 26으로 하였다. 분류율을 비교하기 위하여 특징패턴의 선정에 참여했던 50개의 정상파일과 50개의 바이러스 파일로부터 Hdata 구성에 참여한 26개 패턴에 대한 정규화된 출현횟수를 구하여 100×26의 데이터 HtestI을 구성하고 Cdata 구성에 참여한 패턴에 대하여 CtestI를 구성한다. 다음으로 특징선택에 참여하지 않은 파일을 처리하는지 테스트하기 위하여 VX heaven으로부터 50개의 바이러스 파일과 윈도우시스템 실행파일로부터 50개의 정상 파일을 선정하여 마찬가지로 Hdata 패턴에 대한 HtestII, Cdata 패턴에 대한 CtestII를 준비하였다.

준비된 테스트 데이터를 가지고 잘 알려진 분류 알고리즘[13], K-nearest neighbor(KNN) 알고리즘, Support Vector Machine(SVM), Fuzzy Neural Network(FNN-B)에 적용한 결과는 <표 1>과 같다. <표 1>의 결과로부터 이미 학습한 데이터를 가지고 분류한 경우 대비에 의한 휴리스틱 방법에 의해 준비된 HtestI과 본 논문에서 제안된 방법에 의해 준비된 CtestI을 가지고 실험한 결과 유사한 결과를 얻을 수 있었다. 그러나 학습에 참여하지 않은 임의의 데이터로부터 준비한 HtestII와 CtestII를 가지고 분류한 경우는 본 논문에서 제안한 방법이 더 나은 분류율을 나타내었다. 또한 특징선택과정에서 사용한 클러스터 분석방법을 포함한 분류알고리즘 FNN-B의 경우 가장 나은 분류율을 보여 주고 있음을 알 수 있다.

제안된 방법을 좀 더 고찰하기 위하여 L의 값을 35, 26, 22, 15, 10과 같이 변경하여 논문에서 제안한 방법에 의해 준비된 데이터 CtestI과 CtestII을 퍼지 신경망 FNN-B의 입력으로 사용하여 분류한 결과 분류율은 <표 2>와 같다. L의 값은 시스템의 성능을 좌우하는 가장 중요한 매개변수로 알려져 있으며 데이터에 종속적이기 때문에 데이터 준비과정에서 실험을 통하여 적절한 값을 제시해야한다. 본 실험에서 사용한 L의 값 26은 [6,7]의 연구로부터 실험을 통하여 적절한 값으로 정해졌다. <표 2>로부터 L의 크기가 26인 경우와 22인 경우는 크게 분류율이 다르지 않음을 알 수 있

〈표 1〉 서로 다른 특징선택에 의해 마련된 데이터의 분류율 비교

	KNN	SVM	FNN-B
HtestI	88.2%	91.7%	93.5%
CtestI	87.5%	91.5%	94.2%
HtestII	75.2%	82.6%	82.8%
CtestII	78.3%	84.7%	85.3%

〈표 2〉 L의 크기에 따른 제안된 방법의 분류율 비교

	L=35	L=26	L=22	L=15	L=10
CtestI	91.7%	94.2%	94.5%	85.7%	81.8%
CtestII	82.6%	85.3%	84.8%	76.5%	70.4%

다. 이때 L이 22인 경우 CtestI을 가지고 실험한 경우와는 달리 CtestII의 경우 성능이 떨어진 것을 볼 수 있는데 일반화를 위하여 좀 더 많은 특징이 필요함을 확인할 수 있다. 그러나 L이 35인 경우와 26인 경우를 통하여 L의 크기가 크다고 좋은 성능을 나타내지 않는다는 것을 알 수 있다.

5. 결 론

문서분류, 영상인식, 유전자 선택 분야에서와 같은 분류시스템에서는 관찰된 다차원의 데이터로부터 효과적으로 분류를 수행할 수 있는 실험 데이터를 준비해야 한다. 특징선택은 이러한 준비과정의 핵심적인 부분으로 시스템의 성능향상에 중요한 구성요소로서 상관관계 기법, 차원축소 및 상호 정보 처리 등의 통계학이나 정보이론의 접근방법을 중심으로 연구되어왔다. 또한 특징선택분야의 연구는 다루는 데이터의 양이 방대해지고 복잡해지면서 더욱 중요시 되어 인공지능분야의 휴리스틱 방법이 적용되기도 하였다. 이러한 특징선택을 통하여 데이터의 차원을 축소시키거나 데이터의 양을 줄일 수 있으므로 시스템의 성능을 향상 시킬 수 있고, 관련 없거나 중복된 데이터를 찾아내어 시스템의 정확도를 높일 수 있다. 그러나 데이터는 문제에 종속적인 성질을 가지고 있기 때문에 일반적인 접근방법이 어렵고 데이터를 수집한 전문가의 분석과 경험에 의해 준비되어왔다.

본 논문에서는 데이터가 가지는 특성을 반영하면서 새로운 데이터에 대하여 일반화하여 처리할 수 있는 특징선택 방법을 제안하였다. 준비된 데이터의 각 속성 데이터에 대하여 퍼지 클러스터 분석에 의하여 최적의 클러스터 정보를 얻고 이를 바탕으로 군집성과 분리성의 정도를 측정하여 그 값에 따라 특징을 추출하는 메카니즘을 제안하였다. 또한 제안된 방법을 실세계의 컴퓨터 바이러스 분류에 적용하여 기존의 휴리스틱에 의해 마련된 데이터를 가지고 분류한 것과 비교, 고찰하여 그 타당성을 입증하였다. 특히 학습에 참여하지 않은 테스트 데이터에 대하여 상대적으로 더 좋은 분류율을 볼 수 있는데 이로부터 제안된 특징선택방법이 더 일반화가 뛰어남을 확인할 수 있다.

본 논문에서 사용한 방법은 데이터 집합이 가지고 있는 속성이 독립적이며 보통 전문가에 의해 부여되는 속성의 정성적인 가중치를 고려하지 않았다. 또한 제안된 방법은 주어진 속성 값이 분류할 c개의 클래스에 잘 분포되어 있는 정도를 측정하는 평균적인 계산방법에 의하여 처리되므로 예외적인 데이터의 영향을 받을 수 있다는 단점이 있다. 앞으로 다양한 데이터에 대하여 L의 값의 변화를 기준으로 그 결과를 분석하고 또한 새로운 데이터를 점증적으로 학습하는 방법을 고안하여 이미 학습한 방대한 양의 파일 정보를 재처리할 필요 없이 처리할 수 있는 적응성이 향상된 체계적인 방법으로 더욱 발전시켜야겠다.

참 고 문 헌

[1] Gupta, M. M., Jin, L., and Homma, N., Static and Dynamic Neural Networks : From Fundamentals to Advanced Theory, Wiley-IEEE Press, April 2004.

[2] Chin-Teng Lin, Chang-Mao Yeh, Shen-Fu Liang, Jen-Feng Chung and Nimit Kumar, "Support-Vector-Based Fuzzy Neural Network for Pattern Classification", IEEE Trans. on Fuzzy System, Vol. 14, No. 1, Feb. 2006.

[3] Debrup Chakraborty and Nikhil R. Pal, "Integrated Feature Analysis and Fuzzy Rule-Based System Identification in a Neuro-Fuzzy Paradigm", IEEE Trans. on System, Man and Cybernetics, Vol. 31, No. 3, June 2001.

[4] Isabelle Guyon and Andre Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3, 2003.

[5] Huan Liu, "Evolving Feature Selection", IEEE Intelligent Systems and Their Applications Vol. 20, Issue 4 Nov.-Dec. 2005.

[6] Jianyong Dai, Joohan Lee and Morgan C. Wang, "Detecting Unknown Computer Virus Using Data Mining Techniques", Business Intelligent Symposium, poster presentation, April, 2006.

[7] Jianyong Dai, Muazzam Siddiqui, Joohan Lee and Morgan C. Wang, "Detecting Computer Viruses Mining Instruction Sequences", Submitted to IEEE Trans. on Dependable and Secure Computing, Jan. 2007.

[8] 이현숙, "퍼지 성능 측정자를 이용한 적응 데이터 마이닝 모델", 정보처리학회 논문지, 제13-B권 5호, 2006.

[9] 이현숙, "점증적 학습 퍼지 신경망을 이용한 적응 분류 모델", 퍼지 및 지능시스템 학회 논문지, Vol. 16, No. 6, 2006.

[10] J. O. Kephart, "A Biologically Inspired Immune System for Computers.", Proceedings of the 4th Workshop on Synthesis and Simulation of Living Systems, pp.130-139, 1994.

[11] Abou-Assaleh, Nick Cercone, Vlado Keselj, and Ray Sweidan, "Detection of New Malicious Code Using N-grams Signatures, Proceedings of the Second Annual Conference on Privacy, Security and Trust (PST'04), pp. 193-196, 2004.

[12] Kolter, J.Z., and Maloof, M. A., "Learning to detect malicious executables in the wild", In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 470-478. New York, NY, 2004.

[13] I. Witten and E. Frank, "Data mining: Practical machine learning tools and techniques with java implementations", Morgan Kaufmann, San francisco, CA, 2000.

[14] VX Heaven : http://vx.netlux.org

[15] http://www.datarescue.com

이 현 숙

email : hsrhee@dongyang.ac.kr

1989년 서강대학교 전자계산학과(학사)

1991년 포항공과대학교 컴퓨터공학과(석사)

1997년 서강대학교 컴퓨터학과(박사)

1991년~1997년 한국전자통신연구소 (ETRI) 연구원

1997년~현재 동양공업전문대학 전산정보학부 부교수

관심분야 : 소프트컴퓨팅, 패턴인식, 데이터마이닝