Journal of the Korean Institute of Industrial Engineers
Vol. 26, No. 4, pp. 354-362, December 2000.

데이터융합, 앙상블과 클러스터링을 이용한 교통사고 심각도 분류분석†

손소영 · 이성호

연세대학교 공과대학 컴퓨터과학 · 산업시스템공학과

Data Fusion, Ensemble and Clustering for the Severity Classification of Road Traffic Accident in Korea

Soyoung Sohn · Sung-Ho Lee

Increasing amount of road traffic in 90's has drawn much attention in Korea due to its influence on safety problems. Various types of data analyses are done in order to analyze the relationship between the severity of road traffic accident and driving conditions based on traffic accident records. Accurate results of such accident data analysis can provide crucial information for road accident prevention policy. In this paper, we apply several data fusion, ensemble and clustering algorithms in an effort to increase the accuracy of individual classifiers for the accident severity. An empirical study results indicated that clustering works best for road traffic accident classification in Korea.

1. 서 론

우리 나라에서는 산업발달과 국민소득의 증가로 생활수준이 향상됨에 따라 자동차 이용도 급격하게 증가하였으며 운전면허소지자 또한 꾸준한 증가추세를 보이고 있다. 이와 같이 계속적인 증가 추세를 보이고 있는 교통량은 교통사고 및 환경문제를 발생시키고 있다. 특히, 인적피해사고는 부상자와 사망자를 포함하기 때문에 사회문제시되고 있으며, 이중 교통부상사고는 후유 장애인으로 그 피해가 계속될 수 있다. 교통사고 발생건수를 살펴보면 95년 24만8천8백65건에 비해 96년도에는 26만5천52건으로 약 1만6천건이 증가했다. 그 중 부상자수는 95년은 33만1천7백47명, 96년에는 35만5천9백62명, 97년에는 34만3천1백59명이고, 사망자수는 95년은 1만3백32명, 96년에는 1만2천6백53명, 97년에는 1만1천6백3명으로 연간 1만 명이 넘고 있으며 물적피해 역시 상당량으로 집계되고 있다(통계청, 1999). 따라서, 이를 감소시키기 위한 노력이 시급해지고 있다. 매년 집계된 사고자료를 바탕으로 교통사고의 발생과 관련된 인적, 도로환경적, 차량특성을 조사 분석하여 이들을 바탕으로 사고심각도 예측모형이 수립되면 교통사고 예방을 위한 적절한 조치를 취하는 등 여러 가지 정책개발에 효과적으로 활용될 수 있을 것이다.

이를 위하여 기존의 다양한 도로교통사고 연구들이 시도되었다. Lupton & Bolsdon (1999)는 교통사고 국제데이터베이스(IRTAD)를 사용하여 16개국을 비교한 결과 사고 수는 감소했지만, 1993년 전체 교통사고사망자당 보행자의 교통사고율이 뉴질랜드, 네덜란드, 프랑스에서는 12%에서부터 영국, 아일랜드에서는 30% 등 변하고 있음을 보여준다. 이들 나라들 사이의 차이는 도시밀도, 도로기반, 이동시 걷는 횟수, 알코올과 속도에 관한 법규 등 다양한 요인들에 기인한다고 보았다. 이일병, 임병현(1990)의 연구에서는 우리 나라의 82~89년의 전국 교통사고를 단위로 전체 인구수, 자동차 보유 대수, 운전면허 소지자수, 도로 연장거리, 교통 경찰관 수, 국민 총생산 등의 자료를 회귀분석에 이용하여 교통사고 예측을 하였다. 오윤석, 고양선(1992)은 대형 교통사고의 요인들 중 인적 요인에 속하는 운전자의 법규 위반 유형별과 가해 운전자의 사망여부를 인적, 차량적, 그리고 도로 환경적 요인들에 대한 판별함수를 이용해 분석하였다. 김미영(1993)은 교통사고 통계원표의 항

† 본 연구는 한국과학재단 특정기초연구(1999-1-303-005-3)로 수행되었음.

목 중 인적 요인으로 음주, 과속운전 여부, 보행자의 무단 횡단에 따른 사망정도가 차이가 있는지를 대수선형 분석(log-linear analysis)으로 비교하였다. 도로교통사고 자료처리의 일환으로 Sohn & Shin(2000)은 교통사고 심각도 분류분석을 함에 있어 전형적인 데이터 마이닝 기법인 신경망(neural network)과 Decision-Tree, 로지스틱 회귀분석(logistic regression)을 이용하였다. 이들은 교통사고통계원표에 기록된 여러 가지 범주형 설명변수들을 고려하여 사고심각도를 3가지 범주(치명적 상해, 경미한 상해, 물적피해)와 2가지 범주(신체상해, 물적피해)로 분류하고 기법간의 분류정확도를 비교, 분석하였다. 그러나 Sohn & Shin(2000)의 교통사고 심각도 분석의 분류정확도가 분류기별로 2범주의 경우 약 72~73%, 3범주의 경우 50~54% 정도로 그다지 높지 않은 것으로 나타났다. 본 논문에서는 이러한 단일 분류기의 분류정확도를 높이고자 다수의 분류기 결과를 활용하여 데이터융합기법 및 앙상블기법 그리고 클러스터링 알고리듬을 이용하여 사고예방 정책에 기여하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 데이터융합기법, 앙상블기법의 문헌고찰을 하였다. 3절에서는 본 논문에서 다루고자 하는 클러스터링방법, 데이터융합기법, 앙상블기법을 제시하였다. 4절에서는 실제 사례를 이용하여 본 논문에서 다루고 있는 분류기법들의 성능을 비교하였다. 5절에서는 결론 및 향후 연구방향을 제시하였다.

2. 문헌고찰

데이터 융합(Data Fusion)은 하나의 센서에 의해 성취될 수 있는 것 이상의 추론과 개선된 정확성을 얻기 위해 여러 개의 센서로부터 감지된 데이터를 조합하는 기술을 말한다. 즉, 각 수집체계에 대한 물리적 사건(event), 활동(activity), 또는 상황(situation)에 관한 추론을 하기 위해 다양한 수집자료들을 적절히 활용하고 이들의 영향도를 평가하여 최적의 결과 값을 산출하는 것이다. 센서와 컴퓨터 하드웨어의 발달과 더불어 실시간 데이터 융합에 대한 연구가 활발해졌으며 관련기법으로는 인공지능, 패턴인식, 통계적 추정 등 여러 영역을 포함하고 있다. 데이터융합의 응용분야로는 자동타겟인식, 전장감시, 전략적 경보방어시스템 등의 군사분야 및 복잡한 기계의 모니터, 의학진단, CBM(condition-based maintenance) 등의 비군사분야를 포함한다.

현재 여러 응용분야에서 데이터융합을 위해 다중센서가 주로 쓰이고 있으며 센서로부터 감지된 데이터는 다양한 수준에서 조합되고 융합된다. 데이터퓨전은 크게 연관(association), 추정(estimaiton), 정체규정(identity declaration)으로 나눌 수 있다. 목표와의 관련성을 나타내는 연관분야에서는 관찰치와 예측된 관찰치에 대한 거리, 상관계수 등의 척도가 쓰이며, 모델의 모수, 관심특성치를 예측하는 추정분야에서는 최우추정법(MLE), 최소자승법(LS) 등의 최적기준을 이용하며, 칼만필터(kalman filter) 등과 같은 순차적 기법도 사용하고 있다. 또, 목표에 대한 정체를 구별하는 정체규정분야에서는 융합되는 수준에 따라 특징수준융합(feature-level fusion), 결정수준융합(decision-level fusion), 데이터수준융합(data-level fusion)으로 나눌 수 있다. 먼저 특징수준융합은 감지된 데이터로부터 대표되는 특징을 추출해내는 과정이다. 상태수준융합이라고도 일컬어지는데, 이 수준에서 특징들은 다중센서로 관찰된 데이터를 시그널 프로세싱 기법(웹트럴분석, 시간-빈도분석, 비선형동적기법, 상위하위순위분석)을 사용하여 추출한다. 그리고 추출된 데이터는 신경망, 패턴인식, 퍼지로직, 클러스터링 알고리듬, 템플리트방법을 기반으로 한 통합된 특징으로 분류된다. 결정수준융합은 각각의 센서가 감지하는 객체의 속성, 위치 등에 대한 정보를 융합하는 과정이다. 관련기법으로 가중결정방법, 전형적 추론방법, 베이지안추론(Bayesian), 뎀스터-쉐퍼방법(Dempster-shafer), 로지스틱융합방법(logistic fusion) 등이 사용되고 있다. 데이터수준융합은 센서로부터 직접 수집된 데이터를 융합하고 그후 특징추출, 정체규정을 하는 과정으로서 특징 및 결정수준의 기법들이 모두 쓰이고 있다(Hall, 1997; Kam, 1997).

본 논문에서 사용하고자 하는 결정수준융합의 응용사례를 몇 가지 살펴보면 다음과 같다. Buede *et al.*(1997)은 대공전장에서 사용될 수 있는 ESM(Electronic Support Measure), IFF (Identification Friend or Foe), 레이더의 세 가지 센서를 이용하여 비행기의 형태를 규정하고자 하였다. 이를 위해 베이지안방법 및 뎀스터-쉐퍼방법의 데이터 융합기법을 이용하여 시뮬레이션을 실시하였다. 센서데이터가 입력되지 않은 경우를 고려하여 분류확률값이 일정수준에 수렴하기 위한 수렴시간을 두 가지 방법에 대해 비교한 시뮬레이션 결과로서 베이지안방법이 우수함을 보였다. Xufeng *et al.*(1997)은 자동차 동력장치에 사용되는 두 종류의 부품의 고장분류정확도를 높이기 위해 결정수준에서 데이터를 융합하였다. 첫 번째 부품에서는 실험을 통하여 시간이 지남에 따라 발생하는 동력장치의 진동신호 특성치(평균, 분산 등)들을 웨이브릿변환(wavelet transform)하여 추출된 패턴데이터로 고장여부를 알아내고자 했다. 이와 더불어 변환한 웨이브릿신호와 고장나지 않은 상태의 표준신호의 거리차이를 이용하여 고장형태를 더 자세히 분류하였다. 이러한 분류작업을 위하여 4개의 다중센서를 이용하였으며 각 센서에서 실험데이터에 의해 학습된 신경망을 통하여 얻어진 분류결과의 신뢰도를 높이기 위해 데이터융합기법인 뎀스터-쉐퍼방법을 이용하였다. 또 다른 부품에서는 고장여부를 분류하기 위해 표준신호와 입력된 신호의 상관정도를 보는 그레이 시스템(Grey system)을 이용하여 나온 상관관계(correlation) 값들을 뎀스터-쉐퍼방법을 이용하여 융합하였다. Dar and Vachtsevanos (1989)는 꽃잎의 길이, 두께 등의 4가지의 특성을 이용하여 붓꽃의 하위 종을 3가지 클래스(class)로 분류하고자 하였다. 개개의 특성치에 퍼지이론을 적용하여 각 경우가 세 개 중 하나의 클래스로 분류될 확률값을 구한 후 뎀스터 쉐퍼방법을 이용하여 결정수준에서 융합하였다. 또, 특성치들의 3가지 순서변화

를 주어 같은 융합된 확률값의 수준을 얻기 위해 가장 좋은 특성치의 순서를 추출하였다. Shaw and Garvey (1992)는 5개 수준의 이산신호에서의 융합을 위하여 이산시간에 따라 두 개의 센서를 통해서 입력되는 신호에 대해서 감지된 신호 수준에 확률값을 할당하였다. 뎀스터 쉐퍼방법을 이용하여 각 센서에 의해 감지된 신호수준의 확률값을 데이터 수준에서 융합하고 융합된 확률값의 상한(Plausibility)과 하한(Support)을 계산하였다. 기아정보시스템(1997)에서는 CCTV, 검지기, Probe vehicle 등 세 개의 센서에서 추출된 통행시간을 바탕으로 융합된 통행시간을 예측하는 데 가중평균방법을 이용하였다. Blanco *et al.* (1999)은 개인이나 조직에 대해 얻어진 여러 모델의 신용도 결과들을 종합하기 위해 선형융합, 로지스틱융합, 베이지안 융합모델을 제시하였다.

다음은 하나의 데이터로부터 Bootstrap resampling 방법에 의해 여러 번 추출된 자료를 바탕으로 추정된 분류결과를 앙상블(ensemble)하는 arcing, bagging 기법과 데이터의 특성을 고려하여 몇 개의 군집으로 나눈 뒤 각 군집별로 적절한 분류모형을 적용하는 클러스터링 기법에 대해 자세히 살펴보고자 한다. 분류기 앙상블(classifier ensemble)은 서로 다른 분류기(예:인공신경망, DT)들의 분류결과를 융합한다기보다는 하나의 분류기도 training 자료의 성격에 따라 결과가 다르게 나올 수 있다는 점을 감안하여 여러 개의 Bootstrap resample에 근거한 분류기들의 결과를 하나의 결과로 모아주는 것이다. 지금까지 보편적으로 알려져 있는 분류기 앙상블에 대한 기법으로는 Bagging(Bootstrap AGGregatING), Arcing(Adaptive Resampling and Combining)을 들 수 있다(Breiman, 1994; Breiman, 1996). Arcing은 분류관련 문제에서만 독보적으로 다루어진 기법으로 Schapire (1990)에 의해 Boosting이라는 기법으로 처음으로 개발되었으나, Breiman(1996)이 이것을 Arcing이라 다시 이름지었다. 이 기법들은 하나의 분류기를 사용함으로써 나타나는 불안정성의 단점을 보완하고자 개발되었으며, 분류확률값을 융합하는 데이터 융합기법과는 다르게 Bootstrap resample을 이용하여 여러 개의 데이터 집합을 생성함으로써 각각 Bootstrap 집합에 대하여 분류기를 구성한다. 그리고 임의로 생성된 몇 개의 Bootstrap 집합으로 Training된 분류기의 분류결과를 융합하여 분류하는 것을 bagging이라 한다. 그리고 각각의 분류기에 대한 분류결과만을 적용하는 것이 아니라 분류기의 정확도를 나타내는 가중치를 적용함으로써 융합하는 것을 arcing이라 한다(Breiman, 1996; Opitz *et al.*, 1997).

앙상블의 응용사례를 살펴보면 다음과 같다. Optiz *et al.* (1997)은 14개의 서로 다른 데이터집합에 한 개의 분류기, 데이터를 모두 사용하여 학습시킨 단순 앙상블, 그리고 데이터를 재추출하여 학습시킨 bagging, arcing을 이용하였다. 분류기로는 신경망과 Decision-Tree를 사용하였으며 각 기법별 시험데이터에서의 오분류 확률값을 비교한 결과 대부분의 경우 bagging, arcing 기법의 분류에러율이 낮음을 보였다. Quinlan (1996)은 예측력을 개선하기 위해 Decision Tree(C4.5)를 사용하여 bagging, arcing(boosting)을 실시하였다. 27개의 데이터집합을 이용하여 C4.5, bagging, arcing의 오분률을 비교한 결과 arcing이 우수함을 보였다. Merz *et al.*(1999)은 여러 개의 분류기의 결과를 융합하기 위해서 주성분회귀분석을 이용하였다. 각 분류기의 예측결과 행렬로부터 주성분벡터를 만들어 실제결과에 대한 주성분 회귀분석모형을 제시하였다.

한편 군집분석을 위해 자주 쓰이는 k-평균 클러스터링 알고리듬을 살펴보면 다음과 같다. 데이터집합에서 임의의 k개의 데이터를 클러스터의 중심으로 놓는다. 그리고 그 나머지 데이터($n-k$) 중에서 임의의 데이터 하나를 추출하여 거리척도면에서 k개의 중심점과 가까운 클러스터에 데이터를 할당한다. 그리고 할당된 클러스터의 중심은 기존의 중심점과 새로 할당된 점의 평균으로 수정된다. 그후, 위의 과정을 $n-k$개의 데이터를 k개의 클러스터에 모두할당하고 중심점을 구할 때까지 수행하는 알고리듬이다(Gose, 1996). k-평균 클러스터링의 응용예로 Xu *et al.*(1999)는 웹상에서 주제별 정보검색을 위하여 주제의 수를 클러스터로 놓고 k-평균 클러스터링 방법을 이용하여 문서들을 클러스터링하였다. 여기서는 거리척도로서 Kullback-Leibler divergence를 이용하였다. Lee(1995)는 숫자인식을 위하여 클러스터 신경망을 제시하였다. 16×16 픽셀이미지를 5개의 4×4의 특징으로 변환하고 5개의 클러스터를 구성하여 각 특징을 입력으로 5개의 신경망을 구성하였다. 출력 단에서는 10개의 노드를 구성하여 0부터 9까지의 숫자를 분류하였다.

3. 개선모형

본 장에서는 Sohn and Shin(2000)의 교통사고 심각도 분류분석에 사용된 신경망과 Decision-Tree 기법의 분류능력을 개선하기 위해 사용할 데이터융합, 앙상블, 클러스터링방법을 소개하고자 한다. 저자들은 차도폭, 차체형상, 사고유형, 사고직전속도, 난폭운전, 보호장구를 설명변수로 사고의 심각도를 물적피해 또는 신체상해 두 가지 중 하나로 분류하였다.

개선된 모형을 위한 데이터 융합기법으로 본 논문에서는 뎀스터-쉐퍼, 베이지안, 로지스틱방법을 중심으로 사용하였다. 먼저 뎀스터-쉐퍼방법에 대해 살펴보면 다음과 같다. 두 개의 센서(신경망, Decision-Tree)로 결정될 수 있는 기본제안을 각각 A_1, A_2(교통사고의 결과 A_1 : 신체상해, A_2 : 물적피해)로, 기법(신경망, Decision-Tree)별 분류결과 제안에 대한 믿음의 함수를 m_N, m_D로 정의하자. 이때 가능한 제안의 수 l은 공집합을 포함해서 $2^2=4$ 이 된다. 그리고 이에 따르는 융합된 제안은 u_l (u_1 : 신체상해, u_2 : 물적피해, u_3 : 물적피해 또는 신체상해, u_4 : 공집합)로, 믿음의 정도는 식 (1)과 같이 $m(u_l)$로 정의된다(Hall, 1992; Choi *et al.*, 1998).

$$m(u_l)=\frac{\sum_{A_i,A_j=u_l} m_N(A_i)m_D(A_j)}{1-C} \tag{1}$$

표 1. 뎀스터-쉐퍼 방법에 의한 융합규칙(물피 : 물적피해, 신체 : 신체상해)

신경망 / Decision-Tree	m_N(신체)	m_N(물피)	m_N(물피∪신체)
m_D(신체)	m_N(신체)×m_D(신체)	m_N(물피)×m_D(신체)	m_D(신체)×m_N(물피∪신체)
m_D(물피)	m_N(신체)×m_D(물피)	m_N(물피)×m_D(물피)	m_D(물피)×m_N(물피∪신체)
m_D(물피∪신체)	m_N(신체)×m_D(물피∪신체)	m_N(물피)×m_D(물피∪신체)	m_N(물피∪신체)×m_D(물피∪신체)

$$C=\sum_{A_k A_m=\varnothing} m_N(A_k)m_D(A_m)$$

이러한 뎀스터-쉐퍼융합방법이 교통사고분류 분석에 사용된 예를 들면, Decision-Tree가 물적피해라 분류하고, 신경망이 물적피해라 선언한 경우, 그 사건이 물적피해일 믿음의 정도는 식 (2)와 같이 얻을 수 있다(<표 1>, <그림 1>참조).

$$m(\text{물피})=\frac{m_N(\text{물피})m_D(\text{물피})+m_N(\text{물피})m_D(\text{물피}\cup\text{신체})+m_D(\text{물피})m_N(\text{물피}\cup\text{신체})}{1-C} \quad (2)$$

$$C=m_N(\text{물피})m_D(\text{신체})+m_N(\text{신체})m_D(\text{물피})$$

m(신체), m(물피∪신체)도 같은 요령으로 구할 수 있으며 m(물피)가 다른 제안의 믿음의 정도보다 클 때 그 사건을 물적피해사건이라 분류한다. 또, m_N(물피)보다 m_D(물피)가 클 때 신경망의 분별력이 Decision-Tree의 분별력보다 m_N(물피)−m_D(물피)만큼 크다고 한다. 이렇게 뎀스터-쉐퍼방법에 의해 얻어진 m(물피), m(신체)를 베이지안 방법의 사전확률값으로 적용될 수 있다.

두 번째 융합방법으로 베이지안방법에 대해서 살펴보면, 뎀스터-쉐퍼방법과는 달리 제안들의 상호배반이라는 가정이 첨가된다. 분류기법의 결과를 바탕으로 데이터 융합에서 사용되는 사후확률(융합된 신뢰도)을 정의하면 식 (3)과 같다(Hall, 1992).

$$P(u_R\,|I_N, I_D)=\frac{P(u_R)P(I_N\,|u_R)P(I_D\,|u_R)}{\sum_{j=1}^{n}P(u_j)P(I_N\,|u_j)P(I_D\,|u_j)} \quad (3)$$

n : 제안의 수

R : 선택된 제안(R=1,2)

I : 신경망(N), Decision-Tree(D)가 분류한 A_1 또는 A_2

$P(u_R)$: 사전에 융합된 제안 u_R일 것이라는 믿음의 정도($m(u_R)$)

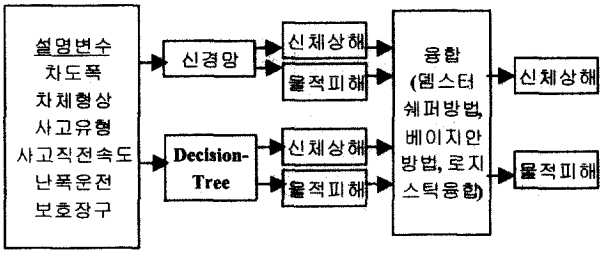

그림 1. 교통사고 심각도 분류분석을 위한 데이터융합.

$P(I_N|u_R)$: 제안이 u_R이었을 때 그 사건을 신경망이 I라고 판단됐던 것일 믿음의 정도

식 (3)을 교통사고 심각도 분류분석에 적용하면, 예를 들어 신경망이 물적피해라 분류하고 Decision-Tree가 신체상해라 판단한 사건이 실제로 물적피해 사건일 확률 P(물피|물피$_N$, 신체$_D$)는 다음과 같이 얻어진다.

$$\frac{P(\text{물피})P(\text{물피}_N|\text{물피})P(\text{신체}_D\,|\text{물피})}{P(\text{물피})P(\text{물피}_N|\text{물피})P(\text{신체}_D\,|\text{물피})+P(\text{신체})P(\text{물피}_N\,|\text{신체}).P(\text{신체}_D\,|\text{신체})} \quad (4)$$

여기서, 각 사고당 P(물피), P(신체)는 각각 뎀스터-쉐퍼방법에 의해 얻어진 m(물피), m(신체)로 대체될 수 있으며 P (신체$_N$|신체)는 실제사고가 신체상해였을 때 신경망이 신체상해로 분류할 확률값을 의미한다. 이러한 조건부 확률은 기존의 사고자료로부터 구해질 수 있다. 데이터융합 후 P(물피|물피$_N$, 신체$_D$)는 P(신체|물피$_N$, 신체$_D$)와 비교되어 결정을 내리는 과정에서는 더 큰 신뢰도의 제안을 선택하게 된다(<그림 1>참조).

세 번째 융합방법으로 로지스틱융합방법을 살펴보면 다음과 같다. 신경망과 Decision-Tree에서 얻은 분류확률값을 각각 $P(I_N)$, $P(I_D)$라 할 때 로지스틱 회귀분석모형을 사용한 융합모형은 식 (5)와 같다(Blanco *et al.*,1999).

$$P(A_1\,|I_N, I_D)=\frac{1}{1+\exp(-\beta_0-\beta_1P(I_N)-\beta_2P(I_D))} \quad (5)$$

이러한 로지스틱회귀모형을 바탕으로 신경망이 물적피해라 분류하고 Decision-Tree가 신체상해라 판단한 사건이 실제로 신체상해 사건일 확률 P(신체|물피$_N$, 신체$_D$)는 실제 데이터를 이용하여 얻어진 모수를 통해 다음과 같이 얻어진다. 실제 데이터를 적용한 예는 4절에서 더 자세히 다루도록 하겠다.

$$P(\text{신체}|\text{물피}_N,\ \text{신체}_D)=\frac{1}{1+\exp(-\beta_0-\beta_1P(\text{물피}_N)-\beta_2P(\text{신체}_D))} \quad (6)$$

이 외에도 여러 가지 융합기법이 존재하는데 분류기가 두 개인 경우 $\max(0, P(I_N)+P(I_D)-1)$, $P(I_N)P(I_D)$, $\min(P(I_N), P(I_D))$, $2P(I_N)P(I_D)/(P(I_N)+P(I_D))$, $\text{root}(P(I_N)P(I_D))$, $(P(I_N)+P(I_D))/2$, $\max(P(I_N), P(I_D))$, $P(I_N)+P(I_D)-P(I_N)P(I_D)$와 같은 규칙을 적용할 수 있다(Madanl *et al.*, 1998).

여러 개의 분류기를 융합하여 분류결과를 산출하는 앙상블 기법으로는 bagging, arcing을 이용하고자 한다. 먼저 bagging에 대해 살펴보면 여러 개의 Bootstrap resample을 추출하고 각 표본별 관측치에 신경망이 신체상해라 분류한 경우 1, 물적피해라 분류한 경우 0을 할당하여 가장 많은 분류결과를 해당사고 심각도로 선언한다. 이러한 bagging모형을 자세히 살펴보면 다음과 같다(Breiman, 1996; 최대우 외, 1999).

① 본래의 데이터와 동일한 크기를 갖는 Bootstrap resample을 B개 만들어 원래의 데이터를 대체하여 하나의 분류기를 적용하여 resample 수만큼 각각 학습시킨다.
② 입력 x 와 출력 y 로 구성된 Bootstrap resample b에 대한 분류기 C_b의 가능한 예측 범주를 신체상해, 물적피해라 한다. 가능한 분류결과값 $C_b(x)$로는 신체상해와 물적피해에 1과 0을 각각 할당한다.
③ 이러한 분류를 B개의 Bootstrap resample의 개수만큼 되풀이하여 식 (7)과 같이 C_{bag}를 구한 후 C_{bag}가 0.5 이상이면 해당 관찰치를 신체상해라 하고 0.5 이하이면 물적피해라 분류한다(<그림 2>참조).

$$C_{bag} = \frac{1}{B}\sum_{b=1}^{B} C_b(x) \tag{7}$$

Arcing의 기본 아이디어는 Bagging과 같으나 샘플링할 때의 분류 정확성을 높이는 것으로 나타난 관측치가 샘플링될 확률이 높도록 Bootstrap resampling한다. Arcing의 진행 순서는 다음과 같다.

① 전체 데이터 N개에서 각 관측치가 추출될 확률을 $P(i)=1/N$ 로 같은 값을 적용하여 Bootstrap resample을 실시한다 $(i=1,\cdots, N)$.
② N개의 관측치를 가지는 b번째 Bootstrap resample로 학습된 분류기 C_b를 형성한다.
③ i 번째 경우에 대해서 C_b를 이용하여 분류한 결과가 오분류되었을 때는 1의 값을, 정분류되었을 때는 0의 값을 갖는 오분류를 판별하기 위한 더미변수 $d(i)$를 정의한다.
④ 분류기 C_b의 오분류율 ε_b와 $P(i)$를 갱신하는 데 필요한 β_b를 계산한다.

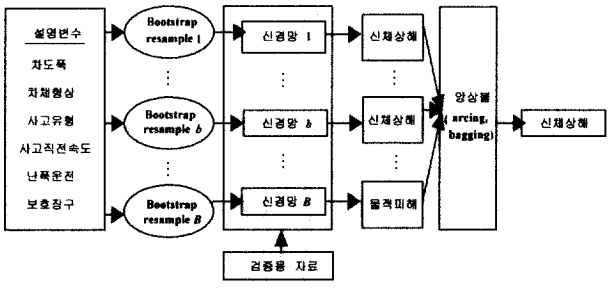

그림 2. 교통사고 심각도 분류분석을 위한 앙상블.

$$\varepsilon_b = \sum_{i=1}^{N} P(i)d(i), \quad \beta_b = \frac{(1-\varepsilon_b)}{\varepsilon_b} \tag{8}$$

⑤ β_b를 바탕으로 $b+1$번째 Bootstrap resample에서 관측치 i가 샘플링될 확률을 다음과 같이 갱신한다.

$$P_{b+1}(i) = \frac{P_b(i)\beta_b^{d(i)}}{\sum P_b(i)\beta_b^{d(i)}} \tag{9}$$

⑥ 이와 같은 과정을 B번 반복한 후, 입력변수 x를 갖는 경우 사고심각도를 신체상해 또는 물적피해(1 또는 0)로 분류한 결과에 각 분류기마다 $\log(\beta_b)$의 가중치를 주어 가중 평균한 값을 취하게 된다.

$$\frac{C_{\text{arc}} = \sum_{b=1}^{B} w_b C_b(x)}{\sum_{b=1}^{B} w_b} \tag{10}$$

$w_b = \log(\beta_b)$, 만약 $\varepsilon_b < 1/2$ 면
$w_b = 0$, 만약, $\varepsilon_b > 1/2$ 면

여기서 w_b는 ε 값이 1/2 이상이 되면 0의 값을 가진다. 이는 하나의 분류기가 데이터의 반 이상을 오분류하면 의미가 없어지기 때문이다. 앙상블 결과 C_{arc} 값이 0.5 이상이면 신체상해라 분류한다.

bagging과 arcing은 각 분류기를 통하여 나온 분류결과를 융합하여 입력 데이터의 분산으로 인한 결과의 편차를 줄여 보고자 하는 데 의의가 있다. 입력자료의 분산이 매우 큰 경우에는 분류결과의 앙상블보다는 입력자료 자체를 군집으로 구분하여 군집별 분류분석을 해주는 것이 나을 수 있다. 본 논문에서는 이런 점을 고려하여 클러스터링방법을 제시하였다. 그 과정을 살펴보면, k-평균클러스터링 방법을 사용하여 Training 자료를 적당한 크기로 클러스터링한 후 각 클러스터 개수만큼 분류기 C_k를 구성한다. 그리고 클러스터링된 데이터집합을 바탕으로 각 분류기를 Training시킨다. 그 후 Test 데이터가 입력되면 그 데이터가 속하는 클러스터의 분류기를 통하여 분류결과 $C_k(x)$를 산출하게 된다. 따라서, 입력데이터의 패턴을 고려하여 분류결과를 산출하기 때문에 높은 정확도가 예측된다(<그림 3>참조).

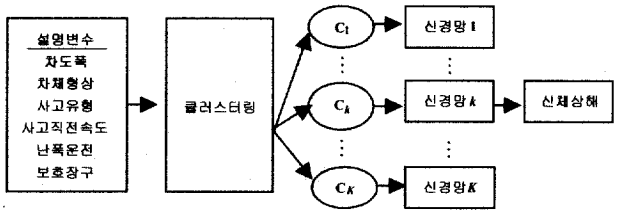

그림 3. 교통사고 심각도 분류분석을 위한 클러스터링.

$$C_{\text{cluster}} = \sum_{k=1}^{K} q_k C_k(x) \quad (11)$$

K : 클러스터의 개수($k=1 \sim K$)

$q = 1$ ...만약 입력데이터가 k 클러스터에 속한다면,

0 ...그렇지 않으면

식 (11)과 같이 C_{cluster}를 구하여 그 결과가 0.5 이상이면 신체상해라 분류하고 0.5보다 적은 경우는 물적피해라 분류한다.

4. 사례 응용

교통사고 자료는 그 양이 매우 방대하며 자료간에 복잡한 상관관계가 있어 분석을 하는 데 많은 비용과 시간이 소요된다. Sohn and Shin(2000)의 연구에서는 신경망, Decision Tree, 로지스틱 회귀분석을 이용하여 교통사고 심각도 분류분석을 하였다. 이 연구에서는 모형수립을 위하여 얻어진 11564건의 교통사고 자료 중 60%를 Training, 40%를 Validation 자료로 각각 할당하고, 79개 교통사고 관련 항목 중 Decision Tree로 선택된 6개 입력변수(<표 2> 참조)를 선택하여 2개 범주(신체상해, 물적피해)를 분류하였다. 그러나 신경망, Decision Tree, 로지스틱 회귀분석을 이용한 분석결과, Test자료의 분류정확성은 낮았으며 기법별 평균 분류정확도의 유의한 성능 차이는 거의 없는 것으로 나타났다. 따라서 본 연구에서는 로지스틱회귀분석은 신경망의 부분집합으로 보고 이것을 제외한 후 Decision-Tree 및 신경망 분류기법들을 센서로 생각하고 각각의 분류결과, 즉 신체상해와 물적피해를 가능한 제안으로 생각하였다. 또, 분류정확성 향상을 도모하고자 앞장에서 제시한 데이터융합기법(뎀스터-쉐퍼방법, 베이지안방법, 로지스틱융합방법 및 일반적인 산술평균, max값 등), 앙상블기법(arcing, bagging), 클러스터링방법을 이용하였고 분류능력은 분별력과 분류정확도 관점에서 평가하였다. 이를 위하여 Sohn and Shin(2000)의 연구와 동일한 자료를 데이터융합을 위해서는 Training, Validation, Test 자료로 각각 40%, 30%, 30%를 할당하였고 앙상블, 클러스터링방법을 위하여 Training(validation 포함), Test 자료로 각각 70%, 30%를 할당하였다. 분류방법들의 성능은 분별력과 분류정확성관점에서 비교되었다.

분별력은 한 사건에 대해 오분류를 할 경우 손실이 큰 경우에 데이터를 융합함으로써 분류의 신뢰도를 한층 높여 손실을 줄일 수 있는 이점을 제공한다. 먼저 이러한 이점을 주는 분별력 관점에서 각 기법의 성능을 비교하고자 한다. 분별력은 분류를 함에 있어서 임의의 관찰치에 대해 하나의 분류기법이 다른 기법보다 분류신뢰도가 높을 때 발생하는 신뢰도의 차이를 말하고 각 기법이 실제사건을 제대로 분류할 확률값으로 판단할 수 있다. 예를 들어, 3469개의 Test자료에서 임의의 10개의 자료를 이용하여 각 기법이 물적피해라고 분류할 확률을 구해보면 <그림 4>와 같고, 예를 들어 한 자료의 분별력을 구해 보면, <그림 4>에서 각 기법이 물적피해라 분류할 때 1번 데이터의 경우 뎀스터-쉐퍼방법(62%)이 신경망(60%)에 비해 2%, Decision-Tree(52%)에 비해 10%, 베이지안방법(96%)의 경우는 신경망에 비해 36%, Decision-Tree에 비해 44% 분별력이 증가하는 것으로 나타난다. 3번 데이터의 경우 분별력은 뎀스터-쉐퍼방법의 경우 신경망에 비해 10%, Decision-Tree에 비해 7%, 베이지안방법의 경우는 신경망에 비해 10%, Decision-Tree에 비해 7% 증가하는 결과로 나타난다. 분별력분석 결과 3469개 Test 자료에서 베이지안방법이 뎀스터-쉐퍼방법보다 평균 17.67%, 신경망보다 평균 13.81%, Decison-Tree보다 평균 22.56%, 로지스틱융합방법보다 평균 22.52% 정도 분별력이 높은 결과로 나타났다(<표 3> 참조). <그림 4>에서 보면 신경망, Decision-Tree에서 산출된 분류확률값이 0.5 이상일 때는 뎀스터-쉐퍼방법과 베이지안방법, 로지스틱융합방법의 확률값은 이들 기법들보다 위에 위치하고, 0.5 이하일 때는 아래에 위치하고 있다. 그렇지만 각 기법의 확률값이 상충될 때, 즉 신경망이 물적피해라 할 확률값이 0.5 이상이고 Decision-Tree에 의한 것이 0.5 이하일 때는 뎀스터-쉐퍼방법과 베이지안방법, 로지스틱융합방법의 확률값은 두 기법(신경망, Decision-Tree)의 확률값 사이에 있게 되고 분별력이 떨어지는 것을 알 수 있다. 앙상블과 클러스터링방법은 각 분류기의 분류확률값을 융합하는 것이 아니라 분류결과값, 즉 1 또는 0 값을 융합하는 것이기 때문에 분별력관점에서는 비교가 불가능하므로 제외하기로 한다.

분류정확도관점에서 자세히 살펴보면 다음과 같다. 각 기법에 할당된 2개의 범주형 분류자료들의 분류확률을 입력자료

표 2. Decision tree 가 선택한 변수들

종속변수(2개의 범주)	
설명변수명	내 용
X34	차도폭
X43	차체형상
X49	사고유형
X50	사고직전속도
X56	난폭운전
X81	보호장구

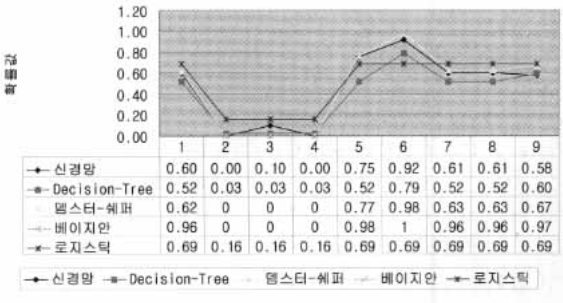

그림 4. 각 분류기법에 근거한 물적피해의 확률값.

표 3. 단일기법과 데이터융합기법의 분별력 비교

자료1 : 설명변수 6개		
		분별력(베이지안대비)
단일 분류기법	Decision Tree	< 22.56%
	신경망	< 13.81%
융합 기법	뎀스터-쉐퍼	< 17.67%
	베이지안	0
	로지스틱	< 22.52%

로 이용하여 뎀스퍼-쉐퍼방법 및 베이지안, 로지스틱융합방법을 이용하여 융합하였다. 각각 기법의 결정단계에서는 두 개의 범주형 자료 중 큰 신뢰도가 있는 쪽을 선택하였다. 3469개의 Test자료를 이용하여 분류정확도를 산출한 결과 뎀스터-쉐퍼방법에 의한 융합된 결과는 72.79%, 베이지안방법은 71.23%, 로지스틱융합방법은 72.30%의 분류정확도를 보였다. Decision-Tree와 신경망보다는 뎀스터-쉐퍼방법과 로지스틱융합방법이 분류정확도의 향상을 보였으나 큰 차이는 없는 것으로 나타났다(<표 4> 참조).

베이지안 방법을 적용하기 위하여 교통사고자료 중 4625개의 Training자료로부터 <표 5>와 같이 조건부 확률을 구하였다. 예를 들어, P(물적피해), P(신체상해)는 뎀스터-쉐퍼방법에 의해 나온 m(물적피해), m(신체상해)의 신뢰도를 그대로 사용하였다. $P(\text{물피}_N|\text{물피})$는 실제자료가 물적피해였을 때 신경망이 물적피해라고 분류한 경우를 누적시켜 실제 물적 자료의

표 4. 기법별 분류정확도 결과

자료1 : 설명변수 6개		
	분류정확도	분류기 개수
Decision Tree	72.30%	1
신경망	70.86%	1
뎀스터-쉐퍼융합	72.79%	2
베이지안융합	71.23%	2
로지스틱융합	72.30%	2
bagging (신경망)	72.70%	5
	72.41%	10
bagging (Decision-Tree)	74.78%	5
	73.80%	10
클러스터링방법 (신경망)	73.94%	3
클러스터링방법 (Decision Tree)	76.10%	3

표 5. 4625개의 Training 데이터에 근거한 조건부 확률

신경망		Decision-Tree	
$P(\text{물피}_N\|\text{물피})$	$P(\text{신체}_N\|\text{신체})$	$P(\text{물피}_D\|\text{물피})$	$P(\text{신체}_D\|\text{신체})$
0.67	0.85	0.66	0.82
$P(\text{신체}_N\|\text{물피})$	$P(\text{물피}_N\|\text{신체})$	$P(\text{신체}_D\|\text{물피})$	$P(\text{물피}_D\|\text{신체})$
0.33	0.15	0.34	0.18

전체 수로 나눔으로써 구할 수 있었다. <표 5>를 이용하여 나머지 3469개의 Test자료를 가지고 식 (3)에 적용시켜 베이지안 방법에 의한 사후확률값을 구하고 이를 바탕으로 분석된 분류정확도는 <표 4>에 정리된 바와 같다.

로지스틱방법에서는 4625개의 Training자료를 이용하여 파라미터를 추정하였고, 파라미터들은 $\alpha=0.01$에서 모두 유의하였다. 추정된 파라미터를 적용한 모델은 식 (12)와 같다.

$$P(\text{신체}|\text{물피}_N,\text{신체}_D) = \frac{1}{1+\exp(4.1343-0.2796P(\text{물피}_N)-2.1767P(\text{신체}_D))} \quad (12)$$

이 모델을 적용하여 Test자료로 분류한 결과 분류정확도는 72.30%이었다. 그 밖에 개선된 융합기법을 찾기 위해 $\max(0, P(I_N)+P(I_D)-1)$, $P(I_N)P(I_D)$, $\min(P(I_N), P(I_D))$, $2P(I_N)P(I_D)/(P(I_N)+P(I_D))$, $\text{root}(P(I_N)P(I_D))$, $(P(I_N)+P(I_D))/2$, $\max(P(I_N), P(I_D))$, $P(I_N)+P(I_D)-P(I_N)P(I_D)$ (Madanl *et al.* (1998))과 같은 일반적으로 사용할 수 있는 융합기법을 적용하여 본 결과 위의 방법 모두가 뎀스터-쉐퍼방법과 유사한 결과인 72.79%의 분류정확도를 가지고 있었다.

데이터의 오분류 특성을 알아보기 위해 <표 6, 7, 8>과 같이 오분류 데이터의 형태를 분석하였다. 베이지안방법에 의한 오분류수는 총 3469개의 Test 자료 중 998개이었다. <표 6>의 형태에서 볼 수 있듯이 신경망과 Decision-Tree가 실제사고를 같은 형태의 사고로 분류하였을 때, 베이지안방법은 오분류가 전혀 없음을 볼 수 있다. 전체적인 오분류 형태를 보았을 때 신경망과 Decision-Tree가 사건을 모두 같은 범주로 오분류하였을 때 베이지안방법도 오분류할 비율이 전체 오분류의 약 77%를 차지하는 것을 볼 수 있다. 뎀스터-쉐퍼방법에 의한 오분류수는 총 3469개의 Test 자료 중 944개이었다. <표 7>의 형태에서 볼 수 있듯이 베이지안의 오분류형태와 거의 유사하다.

로지스틱방법에 의한 오분류수는 총 3469개의 Test자료 중 961개 이었다. <표 8>의 형태에서 보면 위의 두 기법과는 다르게 물피\신체\물피\신체, 신체\물피\신체\물피의 오분류의 비율이 0%임을 볼 수 있다. 전체적인 오분류 형태를 보았을 때 위의 두 기법과 유사하게 신경망과 Decision-Tree가 사건을 모두 같은 범주로 오분류하였을 때 로지스틱방법도 오분류할 비율이 전체 오분류의 약 80%를 차지하는 것을 볼 수 있다.

본 논문에서는 위의 오분류형태를 살펴본 결과 각 분류기가

표 6. 베이지안방법이 오분류를 했을 경우 데이터 형태의 비율

실제분류\신경망분류\Decision-Tree분류\베이지안분류			
물피\신체\신체\신체	23.35%	물피\신체\물피\신체	8.12%
신체\물피\물피\물피	53.91%	신체\신체\물피\물피	2.20%
신체\신체\신체\물피	0 %	물피\물피\신체\신체	10.52%
물피\물피\물피\신체	0 %	신체\물피\신체\물피	1.90%

표 7. 뎀스터 쉐퍼방법이 오분류를 했을 경우 데이터 형태의 비율

실제분류\신경망분류\Decision-Tree분류\뎀스터-쉐퍼분류			
물피\신체\신체\신체	24.68%	물피\신체\물피\신체	6.57%
신체\물피\물피\물피	56.99%	신체\신체\물피\물피	2.33%
신체\신체\신체\물피	0%	물피\물피\신체\신체	7.42%
물피\물피\물피\신체	0%	신체\물피\신체\물피	2.01%

표 8. 로지스틱방법이 오분류를 했을 경우 데이터 형태의 비율

실제분류\신경망분류\Decision-Tree분류\로지스틱분류			
물피\신체\신체\신체	24.25%	물피\신체\물피\신체	0%
신체\물피\물피\물피	55.98%	신체\신체\물피\물피	10.52%
신체\신체\신체\물피	0%	물피\물피\신체\신체	9.56%
물피\물피\물피\신체	0%	신체\물피\신체\물피	0%

오분류하였을 때 융합된 결과도 오분류한다는 분류특성을 고려하여 분류정확도를 향상시키고자 bagging, arcing을 사용하였고 클러스터링방법을 제시하였다. 이를 위하여 총 11564개의 자료 중 Training자료로 8095개를 할당하였고 각 분류기에 수정된 Bootstrap resample을 이용하여 각 분류기당 Training자료의 40%를 할당하였다. arcing을 위하여 Training자료에 포함되지 않은 나머지 60%인 4857개를 validation 자료로서 할당하고 분류에러율을 구하였다. 그리고 전체 Training 자료 외에 Test자료로서 3469개의 자료를 사용하였고, 분류기는 신경망과 Decision Tree를 사용하였으며, 분류기의 개수는 5개, 10개로서 bagging, arcing에 적용하였다. 분류정확도 결과는 Decision Tree 5개를 사용하여 bagging을 적용하였을 때 74.78%이었고, 10개를 사용하였을 때는 74.50%이었다. 신경망 5개를 사용하여 bagging을 적용하였을 때는 72.70%이었고, 10개를 사용하였을 때는 72.41%이었다. arcing을 적용하기 위하여 각 분류기의 validation자료를 이용하여 분류에러율을 구해본 결과 분류기의 개수에 관계없이 거의 유사한 결과를 산출하였다. arcing의 분류정확도 결과는 bagging의 분류정확도 결과와 같은 결과를 산출하였다.

5개의 신경망을 사용하여 bagging을 실시한 분류분석결과를 예로 들면 각각 neural 1(73.08%), neural 2 (71.92%), neural 3 (73.36%), neural 4 (73.19%), neural 5 (72.01%), bagging (72.70%)으로 나타났다. 앞의 결과에서 볼 수 있듯이 bagging한 결과는 각 분류기의 분류정확도의 평균치로서 나타남을 알 수 있다. 따라서, 데이터가 입력될 때 가장 잘 분류하는 분류기를 선택하여 입력되는 것, 즉 그 데이터에 적합한 분류기를 선택하여 분류하는 필요성이 대두되었다. 이러한 필요성으로 클러스터링 방법을 제시하였고 이 방법을 위하여 8095개의 전체 Training 데이터에 대하여 k-평균클러스터링 기법을 이용하여 입력데이터를 3개로 클러스터링을 하였다. 각 클러스터데이터에 대하여 신경망과 Decision Tree를 이용하여 Training을 시켰으며, Test자료에 근거한 분류결과 <표 4>에서와 같이 분류정확도가 향상됨을 볼 수 있다.

5. 결 론

위의 분석결과에서 볼 수 있듯이 분별력에서는 뎀스터-쉐퍼방법과 베이지안방법, 로지스틱모형에 근거하여 융합한 경우가 신경망, Decision-Tree보다 더 우수한 분석결과를 보였다. Test자료에서 베이지안방법이 뎀스터-쉐퍼방법보다 평균 13.81%, 로지스틱방법보다 22.52%, 신경망보다 평균 17.67%, Decison-Tree보다 평균 22.56%의 분별력이 개선된 결과를 볼 때 베이지안 방법이 분별력 면에서는 가장 우수한 결과를 보였다. 신경망, Decision-Tree, 뎀스터-쉐퍼방법, 베이지안, 로지스틱방법의 5가지 기법별 분류정확도의 비교에서는 뎀스터 쉐퍼방법의 도입이 약간의 개선은 가져왔지만 거의 유의한 차이가 없는 것으로 나타났다. bagging, arcing을 실시한 결과 분류정확도가 향상되는 것을 볼 수 있다. 또, 본 연구에서 제시된 클러스터방법을 제시하여 본 결과 다른 기법에 비하여 분류정확도가 향상되었다. 차후 다변량분석기법을 이용한 데이터융합 등 더 우수한 분류정확도를 산출할 수 있는 기법 및 앙상블구조를 개발하는 것이 필요하다.

참고문헌

기아정보시스템 (1997), Data Fusion 알고리듬 개발, 아주대학교 교통연구센터.

김미영 (1993), 교통사고 자료분석 기법 비교검토 및 우리나라 사례분석, *서울대 석사논문*.

손소영, 신형원 (1998), 데이터 마이닝을 이용한 교통사고 심각도 분류분석, *대한교통학회지*, 16(4), 187-194.

오윤성, 고양선 (1992), 대형사고 영향요인의 판별모델 구축에 관한 연구, *대한교통학회지*, 10(3), 173-180.

이일병, 임병현 (1990), 한국의 교통사고 예측 모형 개발에 관한 연구, *대한교통학회지*, 8(1), 73-88.

최대우, 구자용, 박헌진, 박재석 (1999), On the Improvement of classification accuracy using combining learners, *데이터마이닝 연구회 세미나 자료*.

Blanco, Y., Zhu, H. and Peter, A. B. (1999), A Study in the Combination of Two Consumer Credit Scores, *Decision Sciences Institute 5th International Conference*, 1, 558-561.

Breiman, L. (1996), Bagging, Boosting, and C4.5, ftp://ftp.stat. berkeley.edu/pub/users/breman.

Buede, D. M. and Girardi, P. (1997), A Target Identification Comparison

of Bayesian and Dempster-Shafer Multisensor Fusion, *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 27(5), 569-577.

Choi, D. B., Ko, H. S. and Ahn, B. H. (1998), On Multisensor Data Fusion using Attribute Association for Intelligent Traffic Congestion Information Inference, *5th World Congress on Intelligent Transport Systems*.

Dar, I. M. and Vachtsevanos, G. (1989), Feature Level Sensor Fusion for Pattern Recognition using an Active Perception Approach, *Proceedings of International Conference on Telecommunications*, 4, 60-165.

Gose, E., Johnsonbaugh R. and Jost, S. (1996), Pattern Recognition and Image Analysis, Prentice Hall.

Hall, D. L. and Llinnas, J. (1997), An Introduction to Multisensor Data Fusion, *Proceedings of IEEE*, 85(1), 6-23.

Hall, D. L. (1992), Mathematical Techniques in Multisensor Data Fusion, Artech House, Boston·London.

Lee, S. W. (1995), Multilayer Cluster Neural Network for Totally Unconstrained Handwritten Numeral Recognition, *Neural Networks*, 8(5), 783-792.

Kam, M., Zhu, X. and Kalata, P. (1997), Sensor Fusion for Mobile Robot Navigation, *Proceedings of IEEE*, 85(1), 108-119.

Madanl, K., Cheblra, A., Bouchefra, K. and Maurln, T. (1998), Reynaudmn, R., Hybrid neural-based decision level fusion architecture: application to road traffic collision avoidance, *Society of Photo-Optical Instrumentation Engineers*, 37(2).

Merz, C. J. and Pazzani, M. J. (1999), A Principal Components Approach to Combining Regression Estimates, *Machine Learning*, 36, 9-32.

Opitz, D. W. and Maclin, R. F. (1997), An Empirical Evaluation of Bagging and Boosting for Artificial Neural Networks, *International Conference on Neural Networks*, 3, 1401-1405.

Quinlan, J. R. (1996), Bagging, Boosting, and C4.5, *http://www.cse.unsw.edu.au/~quinlan/*.

Schapire, R. E. (1990), The Strength of Weak Learnability. *Machine Learning*, 5, 197-227.

Shaw, S. and Garvey, T. (1992), Evidential Signal Processing For Low-level Sensor Fusion, *Proceedings of the 1992 International Conference on Acoustics, Speech, and Signal Processing*, 4, 61-64.

Sohn, S. Y. and Shin, H. W. (2000), Data mining for road traffic accident type classification. *to appear in Ergonomics*.

Xufeng, P., Lien, G. and Xiaolei, L. (1997), The Research of Automobile Transmission System Fault Diagnosis Based on Vibration Signal, *The 9th International Pacific Conference on Automotive Engineering*, 117-123.

Xu, J. and Croft, W. B. (1999), Cluster-based Language Models For Distributed Retrieval, *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99)*.