대한산업공학회/한국경영과학회 2000 춘계공동학술대회 논문집. 2000년4월21-22일. 경남대학교
Session PA10.1

데이터융합, 앙상블과 클러스터링을 이용한 교통사고 심각도 분류분석

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

손소영, 이성호
(연세대학교 컴퓨터과학 · 산업시스템공학과)

Abstract

계속적인 증가 추세를 보이고 있는 교통량으로 인해 환경 문제뿐 아니라 교통사고로 인한 사상자 및 물적피해가 상당량으로 집계되고 있다. 본 논문에서는 데이터융합 및 앙상블, 클러스터링방법을 이용한 교통사고 심각도 분류분석방법을 제안함으로서 교통사고예방에 기여하고자 한다. 이를 위하여 신경망과 Decision-Tree기법을 이용하여 얻은 물적피해와 신체상해가 발생할 확률을 융합하는 전형적인 데이터 융합기법(뎀스터-쉐퍼, 베이지안 방법, 로지스틱융합방법)을 사용하였다. 또한, 분류정확도를 향상시키고자 Bootstrap 재추출 방법을 이용해 얻어진 여러개의 분류예측 결과중 다수의 분류결과를 선택하는 앙상블(arcing, bagging)기법을 적용하였다. 더불어, 본 연구에서는 클러스터링 방법을 제시하고, 이 방법이 기존의 융합기법, 앙상블기법과 비교한 결과, 분류예측면에서 정확도가 향상됨을 보였다.

1. 서론

산업발달과 국민소득의 증가로 생활수준이 향상됨에 따라 자동차 이용도 급격하게 증가하였으며 운전면허소지자 또한 꾸준한 증가추세를 보이고 있다. 이와같이 계속적인 증가 추세를 보이고 있는 교통량은 교통사고 및 환경 문제를 야기시키고 있다. 특히, 인적피해사고는 부상자와 사망자를 포함하기 때문에 사회문제시 되고 있으며, 이중 교통부상사고는 후유 장애인으로 그 피해가 계속될수 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 1만명이 넘고 있으며 물적피해 역시 상당량으로 집계되고 있다. 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 활용 될 수 있을 것이

도로교통사고 자료처리의 일환으로 손소영,신형원(1998)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 3가지범주(치명적 상해, 경미한 상해, 물적피해)와 2가지범주(신체상해, 물적피해)로 분류하고 기법간의 분류정확도를 비교, 분석하였다. 이와같이 교통사고심각도를 분석함에 있어 기존의 다변량분석기법에서 근래에 부각되고 있는 데이터마이닝 기법등이 사용되는 것을 알 수 있다. 그러나, 손소영, 신형원(1998)의 데이터마이닝 기법을 교통사고 심각도 분류분석에 적용한 결과를 보면 개개 분류기법의 정확도가 높지않음을 알수 있다. 본 연구의 주 목적은 사고심각도 분류기법의 정확도 향상을 도모하고자 데이터융합기법 및 앙상블기법 그리고 클러스터링을 이용하여 사고예방 정책에 기여하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 데이터융합기법, 앙상블기법의 문헌고찰을 하였다. 3장에서는 본 논문에서 다루고자 하는 클러스터링방법, 데이터융합기법, 앙상블기법을 제시하였다. 4장

에서는 실제 사례를 이용하여 본 논문에서 다루고 있는 분류기법들의 성능을 비교하였다. 5장에서는 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 결과값을 산출하는 것이다. 데이터융합의 응용분야로는 자동타겟인식, 전장감시, 전략적 경보방어시스템등의 군사분야 및 복잡한 기계의 모니터, 의학진단, CBM (condition-based maintenance)등의 비군사분야를 포함한다.

현재 여러응용분야에서 데이터융합을 위해서는 다중센서가 주로 쓰이고 센서로부터 감지된 데이터는 다양한 수준에서 조합되고 융합된다. 데이터 퓨전은 크게 연관(association), 추정(estimaiton), 정체규정(identity declaration)으로 나눌수 있다. 특히 목표에 대한 정체를 구별하는 정체규정분야에서는 융합되는 수준에 따라 특징수준융합(feature-level fusion), 결정수준융합(decision-level fusion), 데이터수준융합(data-level fusion)으로 나눌 수 있다. 본 논문에서 사용된 결정수준융합은 각각의 센서가 감지하는 객체의 속성, 위치등에 대한 정보를 융합하는 과정이다. 관련기법으로 가중결정방법, 전형적 추론방법, 베이지안추론(Bayesian), 뎀스터-쉐퍼방법(Dempster-shafer), 로지스틱융합방법(logistic fusion)등이 사용되고 있다. 본 논문의 목적에 따라 도로교통사고 심각도 분류의 정확성을 높이기 위해 결정수준융합의 사용하였다(Moshe(1997), <그림1> 참조).

이 논문에서 사용하고자 하는 결정수준융합의 응용사례를 몇가지 살펴보면 다음과 같다. Buede et al.(1997)은 대공전장에서 사용될 수 있는 ESM(Electronic Support Measure), IFF (Identification Friend or Foe), 레이다의 세가지 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 두가지 방법에 대해 비교한 시뮬레이션 결과로서 베이지안방법이 우수함을 보였다. Blanco et al.(1999)은 조직이나 개개인의 신용도를 평가하는 개별적으로 얻어진 결과들을 종합하기 위해 선형융합, 로지스틱융합, 베이지안 융합모델을 제시하였다.

분류기 앙상블(classifier ensemble)은 서로 다른 분류기(예:인공신경망, DT)들의 결과를 융합한다기보다는 하나의 분류기도 training자료의 성격에 따라 결과가 다르게 나올 수 있다는 점을 감안하여 여러개의 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(1996)). 이들 기법들에 대한 설명에 들어가기에 앞서 분류기 앙상블이 본 논문에서 사용된 예를 도식화하면 <그림 2>과 같다. 이 기법들은 하나의 분류기를 사용하므로서 나타나는 불안정성의 단점을 보완하고자 개발되었다. 분류확률값을 융합하는 데이터융합기법과는 다르게 Bootstrap resample을 이용하여 여러개의 데이터집합을 생성하므로서 각각 Bootstrap집합에 대하여 분류기를 구성한다. 그리고, 임의로 생성된 몇 개의 Bootstrap집합으로 Training된 분류기의 분류결과를 융합하여 분류하는 것을 bagging이라 한다. 그리고, 각각의 분류기에 대한 분류결과만을 적용하는 것이 아니라 분류기의 정확도를 나타내는 가중치를 적용하므로서 융합하는 것을 arcing이라한다. 이러한 앙상블기법들은 여러개의 분류기를 사용함으로서 더 안정되고 정확한 분류를 하고자 시도되었다(Breiman (1996), Opitz et al.(1997)).

앙상블의 응용사례를 살펴보면 다음과 같다. Optiz et al.(1997)은 14개의 서로다른 데이터집합에 한 개의 분류기, 데이터를 모두사용하여 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.

k-평균 클러스터링 알고리즘을 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 새로할당된 점의 평균으로 수정된다. 그후, 위의 과정을 n-k개의 데이터를 k

<표1 > Decision tree 가 선택한 변수들

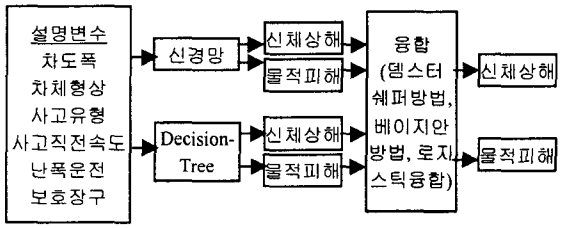
종속변수(2개의범주)	
설명변수명	내용
X34	차도폭(1당)
X43	차체형상(1당)
X49	사고유형
X50	사고직전속도(1당)
X56	난폭운전(1당)
X81	보호장구(2당)

개의 클러스터에 모두할당하고 중심점을 구할 때까지 수행하는 알고리즘이다. k-평균 클러스터링의 응용예를 살펴보면, 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.

3. 개선모형

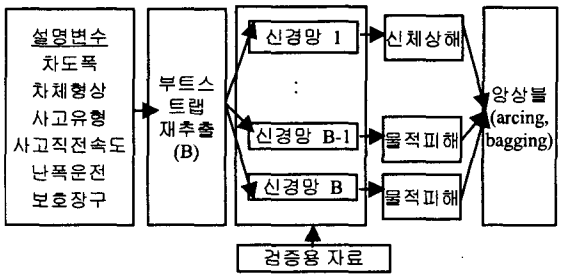
본 장에서는 손소영,신형원(1998)의 교통사고 심각도 분류분석에 사용된 신경망과 Decision-Tree 기법을 각각의 센서로 고려하고 각 경우별 사고심각도(물적피해 또는 신체상해) 예측력을 증가시키기 위해 데이터융합, 앙상블, 클러스터링방법을 이용하고자 한다.

a. 데이터 융합기법으로는 뎀스터-쉐퍼, 베이지안, 로지스틱방법을 이용하였다. 데이터 융합의 적용예를 보면 <그림1>과 같다. 데이터 융합은 분별력과 분류정확도 관점에서 비교분석하였다.



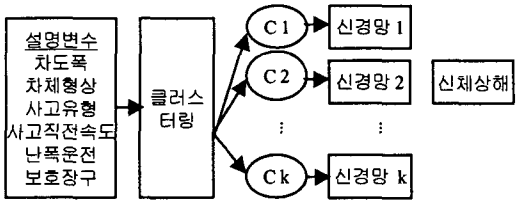
<그림 1> 교통사고 심각도 분류분석을 위한 데이터융합

b. 여러개의 분류기를 융합하여 분류결과를 산출하는 앙상블기법으로는 Bagging(Bootstrap AGGregatING), Arcing(Adaptive Resampling and Combining)을 이용하고자 한다(Breiman(1996), 최대우(1999)). bagging과 arcing은 각 분류기를 통하여 나온 분류결과를 융합하여 입력 데이터의 분산



<그림 2> 교통사고 심각도 분류분석을 위한 앙상블

으로 인한 결과의 편차를 줄여 보고자 하는데 의의가 있다. 입력자료의 분산이 매우 큰 경우에는 분류결과의 ensemble 보다는 입력자료 자체를 군집으로 구분하여 군집별 분류분석을 해주는 것이 나을 수 있다. 본 논문에서는 이런점을 고려하여 클러스터링방법을 제시하였다. 그 과정을 살펴보면, k-평균클러스터링 방법을 사용하여 Traning자료를 적당한 개수로 클러스터링한 후 각 클러스터개수만큼 분류기를 구성한다. 그리고, 클러스터링된 데이터집합을 바탕으로 각 분류기를 Training시킨다. 그 후 Test 데이터가 입력되면 그 데이터가 속하는 클러스터의 분류기를 통하여 분류결과를 산출하게 된다. 따라서, 입력데이터의 패턴을 고려하여 분류결과를 산출하기 때문에 높은 정확도가 예측된다(<그림3>참조).



<그림 3> 교통사고 심각도 분류분석을 위한 클러스터링

각각의 기법으로 실제자료에 이용한 융합된 분류신뢰도를 바탕으로 분류정확도관점에서 각 기법을 비교하였다.

4. 사례 응용

교통사고 자료는 그 양이 매우 방대하며 자료간에 복잡한 상관관계가 있어 분석을 하는데 많은 비용과 시간이 소요된다. 손소영,신형원(1998)의 연구에서는 신경망, Decision Tree, 로지스틱 회귀분석을 이용하여 교통사고 심각도 분류분석을 하였다. 또, 각 기법별 평균 분류정확도의 유의한 성능 차이는 거의 없는 것으로 나타났다. 따라서 본 연구

<표2 > 단일기법과 데이터융합기법의 분별력비교

자료1:설명변수6개		
		분별력(베이지안대비)
단일 분류기	Decision Tree	< 22.56%
	신경망	< 13.81%
융합 기법	뎀스터-쉐퍼	< 17.67%
	베이지안	0
	로지스틱	< 22.52%

에서는 Decision-Tree 및 신경망 분류기법들을 센서로 생각하고 각각의 분류결과 즉.신체상해와 물적피해를 가능한 제안으로 생각하였다. 또, 분류정확성 향상을 도모하고자 앞장에서 제시한 데이터융합기법, 앙상블기법, 클러스터링방법을 이용하였고 분류능력은 분별력과 분류정확도 관점에서 평가하였다.

<표3 > 기법별 분류정확도결과

자료1:설명변수6개		
	분류정확도	분류기 갯수
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**

5. 결론

위의 분석결과에서 볼 수 있듯이 분별력에서는 뎀스터-쉐퍼방법과 베이지안방법,로지스틱으로 융합한 경우가 신경망, Decision-Tree보다 더 우수한 분석결과를 보였다. 신경망, Decision-Tree, 뎀스터-쉐퍼방법, 베이지안, 로지스틱방법 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 제시하여본 결과 다른 기법에 비하여 분류정확도가 향상되었다. 차후 다변량분석기법을 이용한 데이터융합등 더 우수한 분류정확도를 산출할 수 있는 기법 및 앙상블 구조를 개발하는 것이 필요하다.

4. 참고문헌

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

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

Yann Blanco, Hui Zhu, and Peter A. Beling(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.*

Dennis M. Buede and Paul Girardi(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.

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/.*

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).*

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