비즈니스에 있어서 데이터마이닝의 현재와 미래

CRM 시대에 승리하기 위한 전략과 비즈니스 모델

글 / 와시오 다카시

I 데이터마이닝 실용화 위한 컨테스트

데이터마이닝 기술은 1990년 전반에 걸쳐 기초연구가 본격화돼 최근 2~3년간 부쩍 산업계에서 실용화되기 시작했다. 필자가 조사한 범위에서는 금융이나 유통 등의 분야에서 마케팅 조사를 하는데 적용하는 것이 주류였으며, 기타 제조업이나 통신 서비스 업종에서도 품질관리 및 고객관리 등에 적용시키고 있다. 이처럼 데이터마이닝 기술이 급속한 보급을 보이고 있는 배경에는 데이터가 점차 디지털화 되어 가고 데이터 베이스 축적에 의해 그 대상이 되는 데이터를 풍부하게 얻을 수 있게 되었다는 점이다. 또한 기존의 통계기법과 새로운 데이터마이닝 기법이 다양하게 이용할 수 있게 되었으며, 이와 같은 구조를 통합한 마이닝툴이 시판된 것이라 할 수 있다.

한편으로 학술 연구적인 면에서 다수의 연구자가 공통의 데이터에 여러 가지 기존 혹은 새로운 데이터마이닝 기법을 적용해서 각 기법의 특징을 비교 평가하려는 데이터마이닝 컨테스트가 1990년대 중반부터 널리 성행되었다. 학술 연구 분야에서는 새로운 데이터마이닝 원리나 기법을 제안하는 것만으로도 일정한 성과로 인정되는데, 연구자는 스스로 제안하는 기법에 유익한 데이터만을 평가에 사용하려는 경향이 있다. 여기에서 각 제안 기법의 장점이나 단점, 습관이라 할 수 있는 특징을 공통 데이터 제공자인 동시에 이용자의 입장에서도 어떤 특정분야의 전문가로부터 공정하게 비교하고 평가받는 장으로서 데이터마이닝 컨테스트가 행해지고 있다. 이와 같은 공통된 벤치마크 데이터에 의한 각 기법의 비교평가는 기계학습 분야에서 이전부터 행해지고 있다. 기계학습의 벤치마크 데이터에서는 교차검증법(Cross-validation)

혹은 부트스트랩법을 이용해서 트래이닝 데이터의 학습에 기초해 테스트 데이터에 대한 어느 정도 정확한 추정이 가능한가를 정답률이라는 명확한 평가지표에 따라 비교한다. 이에 비해 데이터마이닝 컨테스트에서는 공통 데이터를 제공한 전문가에게 얼마나 유용한 지식을 줄 수 있는가를 전문가의 경험적 주관에 의해 판단할 수 있다. 이 때문에 기법간의 비교 평가의 객관성의 확보보다 이용자의 필요에 입각한 평가를 목표로 하는 경우가 많다.

데이터마이닝 컨테스트는 학술 연구 범위 내에서도 이용자의 니즈를 중심으로 한 기법 평가를 표방하고 있다는 점에서 실용화의 접점을 산업계에서 찾을 수 있다. 그러나 현 상황에서는 산업계의 실용목적과 명확하게 연계된 컨테스트가 실시되는 예는 아주 적다. 이에 따라 본 원고는 데이터마이닝 기술의 산업계 적용 상황과 그 문제점을 기술하고, 그에 대한 데이터마이닝 컨테스트가 어떠한 기여를 하는지, 그 가능성과 과제를 고찰해보기로 한다.

II 데이터마이닝 기술의 산업계 적용과 문제점

[표-1] 각 분야 데이터마이닝의 대표적 적용사례

금융분야
· 마케팅 분야
잠재적인 주택론 신청 고객의 추정
고객에 따른 은행상품의 적절한 편성(크로스세일즈)의 설계 및 제시 지원
생명보험의 잠재적 해약후보고객의 발굴, 효과적인 다이렉트 메일 수신인 후보 고객의 발굴
· 업무특화분야
소비자론 여신 심사의 반 무인화 룰의 발굴
고객에 따른 위험세분형의 자동차 보험의 설계 및 제시 지원, 증권 고객과 영업사원과의 갈등 예측
사채 차이 추측, 신용카드의 부정이용 성향 추정
유통 소매 분야
약국 체인 판매 데이터에 의한 우량고객 발굴
투입시 판매현황에 기초한 신제품 판매 예측
신제품의 히트 요인분석, 품질의 판매현황 요인 분석, 우유 판매량의 예측
소비자 구매행동 패턴 분석, 여러 판매조건 하에 따른 판매 패턴 분석
제조 분야
홈페이지에서의 고객 의견 수집에 의한 차세대 신제품 개발 (커스터머리렉션 마케팅).
고객의 제품 크레임 정보와 제조정보의 대조에 의한 설계. 제조현장에서의 품질관리 요구 발굴
제조현장의 제조조건과 제품 검사결과의 대조에 의한 제조공정의 개선.
통신 분야
홈페이지 열람 정보로부터 개별고객의 프로프파일링과 고객 경향 분석
전화회선망 관리를 위한 부하 상황 파악 및 장해 진단
전화망 사용 수요 마케팅을 위한 통신 트래픽 데이터 분석.
고객의 통화 성향에 의한 전화회선 부정사용 검출
계산기 시스템에 의한 엑서스 로그에 기초해 부정 엑세스 검출

적용의 현상

데이터마이닝의 산업계 적용의 현상을 개관해 보면 공개된 자료를 기초로 가장 많은 적용 사례를 보이는 곳이 금융 쪽이다. 적용 사례는 마케팅 분야와 각종 금융업무에 특화된 분야로 크게 나뉜다. 미국에서는 1994년부터 유통업이나 금융업에서 데이터마이닝 사례가 보고됐다. 뉴럴네트워크, 코호넨 네트, 크러스터링 분류 결정목, 러프 집합, 중회귀 분석 등에 다양한 데이터마이닝 기술이 이용되고 있으며, 금융업에서 데이터마이닝이 적용된 대표적 사례를 [표-1]에 나타내 보았다. 데이터마이닝 적용의 효과는 각 사례마다 차이가 있겠지만, 유용한 결과를 얻은 사례도 많다. 마케팅 분야에서는 막대한 고객 리스트로부터 특정 후보리스트를 찾아낼 필요가 있다. 생명보험을 해약할만한 잠재적인 고객명단이나 효과적인 다이렉트 메일을 수신할만한 후보 고객들을 찾는 마이닝에서 업계의 효율과 질의 개선효과를 얻을 수 있다. 또 업무특화 분야에서는 여신 심사 반 무인화 룰을 적용하여 소비자론 무인 신청기를 개발하거나, 막대한 신용카드 사용기록을 통해 부정이용 성향을 발굴하는 등 실적을 올리고 있다.

통신분야에서는 소매부문의 마케팅을 위한 데이터마이닝 적용이 주류이고, POS 데이터를 이용한 유통전반의 업무지식의 도출, 점포 내에서의 판매촉진용 지식의 도출, 유망고객의 발굴 등이 행해지고 있다. 데이터마이닝 기술로서는 분야 결정목, 바스켓 분석, 중회귀 분석, 상관분석 등이 이용되고 있다. 이제까지 공공연하게 보고되는 대표적인 적용 사례를 [표-1]에 나타냈다. 이처럼 유통 분야에서도 여러 가지 시도를 하고 있는데, 판매현황 예측 등에 대한 적용 사례가 충분히 성공했다고 말할 수는 없다. 우량 고객을 발굴하거나 각종 성향을 파악하고 분석하는 면에서는 효과를 올리고 있는 사례가 많은데, 새로운 데이터마이닝 기술보다도 전체적인 경향을 파악하는 종래의 통계적인 기법에 의거하는 경우가 많다. 이것은 금융분야에 비해 상품이나 소매 조건, 고객행동 유형이 훨씬 다양하기 때문이며 고객이나 구매사례를 파악하기 쉬운 형태로 구별해내기가 어렵기 때문이다.

제조분야에서 데이터마이닝을 적용시키는 것 역시 다른 분야들처럼 진전을 보이고 있다. 사내 문서나 매뉴얼 검색, 마케팅이 대부분이고 다른 업종과 서로 비슷한 목적으로 기술 적용을 한다. 그 결과 현 상황에서는 제조업 고유

의 데이터마이닝 적용사례가 그다지 많아 보이지 않는다. [표-1]에는 제조업 고유라고 할만한 적용사례를 정리했다. 첫 번째는 커스터머 릴레이션(고객관련 마케팅)에의 적용이고, 주요 전자제품 회사에서 시도하고 있는데 아직 시행 단계의 벽을 넘지 못하고 있다. 아래 두 가지 품질이나 공정관리의 적용은 앞으로 광범위한 적용 가능성을 보유하고 있으며 지금도 실용화가 추진되고 있는 사례이다. 이와 같은 적용사례에서는 사례 베이스 검색이나 텍스트 마이닝, 바스켓 분석, 분류 결정목 등의 최신 데이터마이닝 기술이 이용되어 효과를 거두고 있다.

통신분야에서는 [표-1]에 나타난 것처럼 주로 인터넷의 고객 마케팅이나 전화망 관리 분야에 데이터마이닝 기술이 이용되고 있으며 사용 기술은 분류결정목, 바스켓 분석, 베이지안 네트, 뉴럴 네트, 텍스트 마이닝, 각종 통계적 기법 등 여러 범위에 걸쳐 있다. 통신분야에서는 풍부한 전자화 데이터 축적이 있기 때문에 데이터마이닝 적용 범위가 넓다. 특히 부정사용이나 부정 엑세스 검출 등 막대한 통신 로그로부터 특징적인 성향을 발굴하는 적용은 성공을 거두고 있다.

사용기술과 실시체제의 실상

데이터마이닝을 산업계에 적용시킨 실상에 대해서는, 2001년 3월에 개최된 인공지능학회 제14회 AI 심포지움 [데이터마이닝은 이렇게 활용하라!]에서 각각 몇 가지가 산업을 조사하며 검토했다. 여기서는 필자가 조사하고 위 심포지움을 통해 명확해진 내용에 대해서 정리를 했다.

위에서 기술한 적용사례를 포함해 산업계에서 이용되고 있는 데이터마이닝 기술은 다양하다. 그러나 반드시 1990년대에 개최된 최첨단 기술만이 이용되는 것은 아니며, 전통적인 방법의 통계 기법들도 활용되고 있다. 어떤 기술이 이용되는가는 각 사례의 목적이나 니즈, 데이터 사양뿐만이 아니라 데이터마이닝 툴의 개발자나 유저에 의해서도 좌우된다. 특히 일본에서는 제조, 통신 분야의 개발자나 유저들이 기술적인 축적을 서로 공유하기 때문에 시판 툴을 이용하는 것과 함께 각종 기술을 테스트하고 그 중에서 가장 좋은 것을 선택해, 대상적용에 커스터머이즈, 튜닝한 툴이나 시스템을 스스로 구축하는 경우가 많다. 이것은 효과적인 데이터마이닝을 실현하는데 단독적인 기술뿐만 아니라 데이터의 전후의 처리도 포함하여, 각 사례에 적용한 기존의 여러 가지 기술의 편성조건 설정을 필요로 하는 것이 많기 때문이다. 이에 비해 금융이나 유통 분야에서는 사례에 적용되는 툴을 유저 스스로가 개발하는 경우가 적고 제조나 통신 분야에 비교하면 기성의 시판 툴을 이용해서 데이터마이닝을 실시하는 경우가 많다. 시판 툴 중에는 충실한 기능을 가진 것도 많아 그로 인해 다양한 기술의 적용이 행해지고 있는 상황이다.

데이터마이닝 기술을 개발하고 사용하는 체제도 사례에 따라 많이 다르다. 앞서 말한 것처럼 제조나 통신분야에서는 대상 데이터 수집, 기술개발이나 시스템 개발, 사용까지를 모두 스스로 행하는 경우가 많다. 그러나 금융이나 유통 분야까지 포함하면 컨설턴트나 개발 기업과 함께 시스템을 개발하거나 기획 편성하는 경우도 있으며, 단독으로 시판 툴을 구입해 앤드유저로서 그대로 이용하는 경우도 많다. 또한 대상 데이터를 수집하는 데에도 스스로 수집한 데이터에 마이닝을 적용하는 경우가 많고, 목적에 따라서는 필요한 데이터를 구입해서 분석을 행하는 경우도 있다.

기술적 및 실무적 문제점

산업계 적용에 대한 실상에는 몇 가지 문제점이 떠오르고 있다. 일부는 앞서 언급한 심포지움에서 검토되고 있다. 그 하나는 [데이터 수집 장해]이다. 현장에 축적된 데이터를 이용할 때에는 데이터가 특정의 마이닝을 목적으로 해서 축적된 것이 아니기 때문에 목적 달성에 필요한 정보를 포함되지 않은 경우가 자주 있다. 이와 같은 목적과 데이터 내용의 부적합은 실제로 마이닝 분석을 실행하지 않으면 명확하게 할 수 없는 경우도 많다. 또 경우에 따라서는 데이터의 수정도 필요하나 데이터마이닝 때문에 데이터를 다시 수집하는 것은 비용적으로 맞지 않아 저 비용의 보조 데이터 수집 수단의 확보나 기존 데이터에서 필요한 정보를 추정하는 기법 등의 기술이 중요한데, 현 상황에서는 아직 미개척 단계이다. 데이터마이닝 연구나 그 응용에서는 마이닝의 본체 기술만이 주목되고 있어, 데이터 수집법의 충분한 검토 및 개량, 보조 처리 등 데이터 수집기술이 성공을 여는 열쇠라 해도 과언이 아니다.

또 하나의 문제점은 데이터마이닝을 실시할 때 시판되고 있는 툴을 구입해서 앤드유저가 사용하면 뭐든지 된다고 하는 [시판 툴 만능주의]가 만연해 있다는 점이다. 앞서 말한 대로 단순한 경우를 제외하면 효과적인 데이터마이닝에는 풍부한 지식과 경험을 기초로 각 사례에 적용된 여러 가지 기술의 편성이나 설정조건을 알아볼 필요가 있다. 마이닝의 목적이나 대상 데이터의 내용은 사례마다 천차만별이다. 현 상황의 시판 툴은 개별기술

이나 그것에 맞는 환경을 제공하는 것으로, 사례에 맞게 즉시 적절한 기술의 결합에 의한 처리 기획이나 각종 성능 평가지표, 튜닝파라미터 설정까지는 가르쳐 주지 않는다. 데이터마이닝을 실무에 적용할 때는 앤드유저가 충분한 지식이나 경험을 축적하기 위한 시간이나 자본을 투입하고, 컨설턴트나 개발 회사와 밀접한 연계체제를 구축하는 등의 투자가 필요하다.

이 문제에 관련해서 데이터마이닝 기술을 제공하는 연구개발자측에서도 [연구 분야의 분석 문제]가 가로 놓여 있다. 데이터마이닝 기술은 인공 지능이나 데이터베이스 통계 등 복수 분야의 기초 기술에 기인하는데, 이들 개별 분야의 연구 개발자의 연계가 반드시 잘 된다고는 말할 수 없고, 기술이 따로 따로 제공되고 있는 경향이 있다. 또 데이터마이닝을 실시하는 기획 전체를 통한 각종 기술의 편성에 있어서, 적합성이나 각종 목적이나 데이터 내용에 기인한 처리 기획의 체계화 등에 관계하는 연구는 아직 손도 안 된 상태이다. 앤드유저의 입장에서 보면 필요한 보조 기술, 각종 판단 지표, 체계적 마이닝 기획의 축적 등에 관련된 연구가 필요하다고 느껴질 것이다.

III 데이터마이닝 컨테스트의 가능성과 과제

데이터마이닝 컨테스트의 현 상황

데이터마이닝 컨테스트는 주로 데이터마이닝 기법의 학술 연구와 기술 개발의 견지에서 여러 연구팀이 표준 벤치마크 데이터에 여러 가지 기술을 적용해 유용지식을 발굴하는 기획이다. 또한 기법이나 분석 기획의 이점, 문제점 등의 특징을 명확하게 잡아서 이론이나 기술을 개선하고 신주제를 발굴 하는 것으로 세계 각지에서 실시되고 있다. [표-2]는 이에 관련된 대표적인 기획을 모은 것이다. 첫 번째는 컨테스트에서는 없지만 미국 UC Irvine 교에 있어서 데이터마이닝을 테스트한 다수의 산업용 내지는 그에 준하는 데이터와 그것을 제공한 분야의 전문가들의 연락처를 함께 공개하고 있다. 연구자나 기술자는 이것을 이용해 기술을 테스트하거나 개량할 수 있다.

이와 같은 상설 아카이브 외에도 [표-2]에 나타난 많은 컨테스트가 국제회의 등에서 정기적으로 행해지고 있다. 특히 미국의 KDD 국제회의 등에서 매년 실시하는 컨테스트는 산업분야나 과학 기술 분야의 실 데이터에 관해 참가자가 일정 기간 내에 가지고 있는 기술력을 총동원해 데이터를 분석하는 것으로 지극히 실천적 내용을 갖는 컨테스트로 알려져 있다. 또 이들 데이터는 앞서 말한 UCI KDD 아카이브에도 수록되어 있다. 일본을 포함해 아시아나 유럽에서도 산업 분야에 관련된 컨테스트가 이뤄지고 있다. 아시아 태평양을 중심으로 하는 PAKDD 국제회의 및 일본내의 인공지능학회 지식 베이스 연구회에서는 해마다 수차례에 걸쳐 산업 및 과학기술 분야의 벤치 마케팅 공개 및 컨테스트가 행해져 왔다. 또 유럽의 PKDD 국제 회의에서도 년 1회 비율로 산업 및 과학 기술 분야에 관한 컨테스트가 행해지고 있다(주9). 이들 모두 발굴 지식의 정도뿐만 아니라 데이터를 제공한 전문가가 실리적인 면이나 흥미차원에서 참가팀이 발굴한 지식의 유용성을 비평하는 특징이 있다. 예를 들어 이것이 주관적인 면이 많다고 해도 전문가에게 유용하게 느껴지는 지식을 제공하는 것이 데이터마이닝의 주요 목적이라고 생각되기 때문이다.

[표-2] 대표적 데이터 아카이브 및 데이터마이닝 컨테스트

데이터마이닝
미국 UC LRVINE교 UCI KDD 아카이브
마케팅이나 인터넷 관련, 엑세스 제어, 인공위성 화상 등 많은 산업용 내지는 그것에 준하는 데이터, 각각 제공한 분야의 전문가 연락처 제공
데이터마이닝 컨테스트
미국 KDD 국제회의 KDD CUP
이익 최대화를 목표로 한 다이렉트 메일 반송 패턴 발굴이나 네트워크에서의 부정접근을 가장 효율적으로 검출하는 지식 발굴 등, 산업 분야 및 과학 기술 분야의 실 데이터
아시아 태평양 PAKDD 국제회의 Discovery Challenge
의료 질환 동 정 지식의 발굴, 화학 물질 독성과 분자구조의 관계지식의 발굴 등 산업 분야 및 과학기술 분야의 실 데이터
유럽 PKDD 국제회의 Discovery Challenge
금융 분야에 있어서 고객 제공 서비스 결정 지식의 발굴, 의료 질환 동 정 지식의 발굴 등 산업 분야 및 과학 기술 분야의 실 데이터
일본 국내 인공 지능 학회 지식 베이스 연구회
의료질환 동 정 지식의 발굴, 화학 물질 독성과 분자 구조의 관계 지식의 발굴, 항체 단백질 특성에 관한 지식의 발굴 등 산업 분야 및 과학 기술 분야의 실 데이터

산업계에 가져다주는 기여의 가능성과 한계

앞서 말한 컨테스트 대부분이 산업계의 실 데이터를 대상으로 실제 문제에 입각하는 분석과 기술상의 여러 가지 문제점 색출할 기회를 제공한다. 이와 같은 컨테스트는 실제로 검증을 통하여 기존 기술이 개량될 뿐만 아니라 새로운 기술을 개발하는 실마리를 얻을 수도 있다. 또 현장 전문가의 이해 양식에 적용된 지식 표현이나 지식 내용 발굴의 필요성을 추구하는 실천적 관점에서 기술개발을 추진하는 것도 가능하다.

따라서 이와 같은 컨테스트는 연구자와 산업계 기술

자, 앤드유저와의 접점을 확대한다. 이것은 데이터마이닝 기술을 산업계에 이전시키는 과정을 촉진시키는 동시에 산업계측의 니즈나 문제제기 등 연구자의 일상적인 활동에서는 얻을 수 없는 지식을 제공한다. 또한 연계를 통해 새로운 연구에 대한 동기부여와 소재가 발굴되어 데이터마이닝 기술의 연구 개발이 폭넓게 다져지는 계기가 되리라 기대한다.

한편, 현 상황의 컨테스트 내용에는 불충분한 점도 많이 존재한다. 앞서 말한 대부분의 컨테스트가 발굴 지식에 대한 전문가의 주관적 유용성이나 흥미 위주로 기술을 평가하기 때문이다. 과학 분야에서는 그같은 기준 아래 데이터마이닝이 적용되는 것도 있지만, 산업계의 실제 적용보다 구체적인 목적이나 평가 기준에 의한 마이닝이 요구되는 것이 대부분이다. 목적에 따른 지식을 얻을 수 있는지 없는지는 불필요한 지식의 혼입이 어느 정도이며 그것이 이익으로 연결되는가 등 사례 내용에 따른 보다 구체적이고 직접적인 기준으로 컨테스트 평가가 병행되어야 할 것이다.

또 컨테스트는 각종 기술들의 특성을 서로 비교하는 취지에서 다수의 참가팀을 구성하며 여러가지 기술을 적용할 수 있는 일반적인 데이터나 목적을 설정해야 할 것이다. 그러나 실제 산업계에서는 적용 가능한 기술이 한정되어 있기 때문에 특정한 데이터나 목적을 위주로 한 데이터마이닝이 주류를 이룬다. 따라서 컨테스트 내용이 산업계의 사례에서 보여지는 특성과는 반드시 일치한다고 볼 수 없을 뿐더러 산업계 현장에서 직접적으로 유용한 지식을 얻는다고는 할 수 없다.

더욱이 산업계 현장의 데이터나 개인 정보 데이터의 공개, 그와 같은 데이터에 정통한 전문적인 노하우는 공개되기 어려운 경우가 많다. 그 때문에 데이터마이닝 컨테스트에서도 데이터 수집 장해가 발생하는 경우가 있다. 앞서 말한 UCI KDD 아카이브와 같은 수집과 공개적인 노력도 되어 있지만 산업계의 다양한 현장 상황을 포함한 사례의 집적과는 아직 먼 상황이다.

마지막으로 많은 컨테스트가 직면하고 있는 데이터마이닝 순행상의 자원제약을 지적해 보면 대부분의 컨테스트에서는 인터넷을 통해서 사례 데이터나 그 사양 정보의 배포가 행해져, 몇 개월 간에 각 참가팀들이 분석 결과를 논문으로 모아 투고해 마지막 컨테스트 미팅에서 전문가로부터 평가를 받는다. 따라서 참가 팀과 전문가들 사이에 밀접한 연계기간은 아주 짧다고 할 수 있다. 그 이유는 전문가가 일상 업무에 바빠 봉사적인 측면에서 컨테스트 평가나 의론에 충분한 시간을 가질 수 없다는 점과 참가팀들도 전문가의 의견을 반영한 기술개량이나 분석에 충분한 시간을 확보할 수 없다는 점을 들 수 있다.

도전해야 할 과제

산업계에 데이터마이닝을 적용하고 컨테스트가 기여하는데 대두되는 문제점은 3가지 과제로 집약된다.

하나는 산업계의 실사례를 폭넓게 정리하고 유형화해서 각 유형이 갖는 사례의 특성을 효율적으로 포괄하는 컨테스트 예제를 만들고 그에 관한 전문가 집합을 준비하는 것이다. 이것에 의해 현장사례의 목적이나 다양한 데이터 내용에 입각한 데이터마이닝 기술을 평가할 수 있고 이에 대한 개선책이나 새로운 연구자료를 발굴할 수 있게 될 것이다.

두 번째는 산업계의 유형사례에 입각해서 개별 마이닝 기술뿐만 아니라 그것들의 효과적인 편성과 기획 차원에서 연구개발을 중시하는 것이다.

이를 위해서는 컨테스트 평가자로서 데이터 분야의 전문가 외에도 각종 처리기술에 정통한 다수의 연구자를 참여하게 하는 것이 효율적이다.

이들 연구자에게는 마이닝 기획에 쓰여지는 각종 기술의 타당성이나 상호 정합성을 기술적 견지에서 음미하는 것이 요구된다.

마지막으로 산업계에 기여하는 관점에서 앞으로의 데이터마이닝 컨테스트에 요구되는 실시형태에 대해서 말하고 싶다.

지금의 컨테스트에는 데이터 공개에서 컨테스트 미팅까지 몇 개월의 시간이 소요되는데, 참가팀과 평가 전문가가 밀접하게 논의해 분석을 진행하는데는 시간적 여유가 너무 적다고 생각된다. 따라서 컨테스트 데이터를 상설 아카이브로 하든지 일년 전에 공개해 충분한 시간적 여유를 확보해야할 것이다. 더 나아가

각 참가팀과 전문가들 사이의 논의사항을 다른 팀에게도 공개해서 논의가 중복되는 일이 없이 서로를 자극할 수 있도록 하는 연구가 필요하다.

산업계에 있어서 데이터마이닝의 전망

산업계에서 데이터마이닝 기술을 이용한 성공사례가 많이 보고되고는 있으나, 아직 내용면이나 관련 인재들이 한정되어 있어서 발전해 나가는 진행단계로 볼 수 있다.

앞으로는 경험 축적에 의해 데이터마이닝 순행에 필요한 여러 가지 지식이나 경험의 중요성이 인식될 것이다. 그렇게 되면 우수한 컨설턴트나 개발 기업의 성장, 앤드유저나 개발기술자에의 지식이나 경험의 침투 및 시판 툴 기능확장이나 개선 등 산업계에 적용될 수 있는 분야가 더욱 확대될 것이다. 전반적으로 인식이 확장된다면 시판 툴이 모든 것을 해결해주리라 하는 만능주의는 소멸될 것이다.

적절한 마이닝 기획을 설계 지원하는 기술개발이나 성공적인 사례가 축적되어 시판 툴에 짜 넣는 한편, 앤드유저나 컨설턴트, 개발기술자가 효과적인 마이닝 기획을 편성하는 능력을 배양함으로써 툴과 인재 육성이라는 두 가지 측면에서 질 높은 마이닝이 가능하게 된다.

더 나아가 데이터마이닝 분야의 연구가 심화 확립되어 데이터마이닝 기술 전체의 바람직한 모습을 염두에 두고 개발하는 연구자나 기술자도 증가해 가리라 예상된다.

긴 안목에서 보면 연구분야의 분단 문제도 해소될 것이다. 그러나 데이터마이닝이 산업계 현장에서 실로 유용한 기술이 되는데 저해요소가 되는 것이 데이터 수집 장해다. 이것은 데이터 결여의 문제이고 이 문제를 정공법으로 메우기 위해서는 데이터 수집 비용이라는 장벽에 직면하게 된다. 더욱 심각한 것은 데이터마이닝 기술 연구자들조차 이 문제의 중요성을 명확하게 인식하지 못한다는 점이다.

문제를 해결하는데는 보충 데이터를 수집하는 방법을 정비하고 결여정보를 다른 데이터로부터 추정하는 기술 등 문제를 감소시킬 가능성도 충분히 있으니 추후의 연구개발이 기대된다.

어쨌든 현 상황에서는 시판 툴과 기존 데이터만 있으면 누구라도 간단하게 효과적인 데이터마이닝이 가능하다는 것은 아니다. 따라서 붐 조성에 편승한 과대 선전은 삼가야 할 것이고 현 상황의 기술 가능성과 한계에 관한 올바른 인식을 가질 필요가 있다.

한편, 학술연구와 산업계 현장의 접점에 대해서는 앞으로 다양화가 진행되리라 예상된다. 데이터마이닝 연구는 어떤 목적을 실현시키기 위해 다양한 기술을 편성하는 종합 공학적인 연구분야이다. 학술연구와 산업계 실천 거리가 가까운 분야라 할 수 있다.

현 상황에서는 양자의 구체적인 접점이 데이터마이닝 컨테스트이다. 주로 연구자나 산업계 기술자간의 접점이고, 또 기술 정보 발신의 장으로 한정되어 있다. 그러나 한편으론 산업계의 앤드유저의 목소리를 연구에 반영시킬 기회 제공도 필요하다.

앞서 언급한 AI 심포지움은 그와 같은 장의 하나로서 앞으로의 전개가 기대된다.

산업계에서 제시한 실천 사례보고 및 거기에서 보여지는 니즈나 과제를 발신해 앤드유저와 연구자, 기술자가 하나가 되어서 검토할 수 있는 장의 중요성이 증가될 것이다.

참고문헌

1) Bay. S. D : UCI KDD 아카이브. 데이터마이닝 연구와 실험을 위한 대규모 데이터 집 합의 아카이브, 정보처리, Vol. 42, No 5, PP, 462-466(May 2001)
2) 오노 키요시 : 금융업에 있어서 데이터마이닝의 응용. 제 18회 일본 . SAS 유저회 연구 발표논문집, PP, 159-171 (1999)
3) (주)니케이(日經)리서치 : POS 데이터에 대한 데이터마이닝 사례집 (June 2000)
4) 호리 소우타: 전기제품의 시장품질 감시 시스템 데이터마이닝 기술의 응용, 인공지능 학회지, Vol. 15, No. 5, PP 813-820 (2000)
5) Hashirnto, K, et al. : probabilistic Modeling of Alarm Observation Delay in Network Diagnosis, Proc. of PRICAL 2000, PP, 734-744 (2000)
6) 인공 지능 학회 제 14회 AI 심포지움 자료 : 데이터마이닝은 이렇게 활용하라! , SIG - J - A004 (http://www.soc.nacsis.nacsis.ac.jp/jsa:/AI-sympo.html) (2001)
7) Kohavi, R. et al. : KDD컵 2000주최자 보고서 : 본질을 파악해 정보처리, Vol. 42, No. 5, PP, 445-453 (May 2001)
8) 스즈키 에이노신 : 일본 아시아에 있어서 데이터마이닝 컨테스트, 정보처리, Vol. 42, No. 5, PP. 457-461 (May 2001)

(2001년 3월 31일자)