산업통계에 관한 국제 심포지움

ISBIS 2016 on Statistics in Business and Industry

참가 결과 보고

통계청
Statistics Korea

산업통계에 관한 국제 심포지움 참가결과보고(요약)

□ **공식명칭:** ISBIS 2016(International Symposium on Business and Industrial Statistics)

□ **주제:** 사업체자료의 데이터마이닝, 시계열자료 분석 및 응용

□ **참가인원:** 약 100명

□ **출장자:** 경제통계국 산업동향과 통계주사 안다영

□ **회의기간 및 장소:** 2016. 6. 8. ~ 6. 10.(3일간), 스페인 바르셀로나

○ 카탈루냐 공과대학교*

* Northern campus of the Universidad Politecnica de Catalunya Barcelona TECH

□ **활동:** 키노트 세션 및 트랙별 세션 참가

□ **주요 내용 및 시사점**

▌주요 내용

- 빅데이터의 중요성을 강조한 통계적 공정관리 기법 및 웹서비스 플랫폼에 대한 통계 모델 소개
- 가법모형, 스펙트럼 분석 등을 응용한 예측 및 적용 등

▌시사점

- 통계학의 최신 연구 동향을 이론 탐구에 그친 것이 아닌, 특정 사업체 생산의 예측 정확도 및 수익성 향상을 위해 활용
- 반도체 등 주요 제조업체의 공장을 디지털화하고 스마트 팩토리를 구현하여 산업용 IoT(IIoT)를 통한 산업별 빅데이터 분석 및 의사결정 가치의 중요성을 강조
- 빅데이터 분석 성능을 향상시키기 위한 수학적 알고리즘 적용 등 이론과 실무의 조화가 인상적

산업통계에 관한 국제 심포지움 참가결과보고

Ⅰ 국외출장 및 워크숍 개요

1. 출장 개요

□ 출장 목적

○ 경제통계의 정확성이 주요 이슈로 부각되고, 점점 응답을 얻기 어려워지는 조사환경에서 통계 작성을 위한 최신 연구 기법 및 실무 활용에 대한 사례 분석이 요구됨

○ 사업체 통계에 관한 국제회의 참가를 통해 사업체 빅데이터 분석, 시계열 자료 처리 등 통계방법론에 대한 이해 및 연구동향 파악

○ 경제통계 관련 해외전문가 및 통계실무자와의 토론을 통해 정보 공유 및 인적 네트워크 구축 가능

□ 출장자

○ 경제통계국 산업동향과 통계주사 안다영

□ 출장 기간 및 출장지

○ 2016. 6. 7. ~ 6. 12.(6일간), 스페인 바르셀로나

□ 활동

○ 키노트 세션 및 트랙별 세션 참가
- 통계적 공정관리 기법 및 시계열 분석, 데이터마이닝 세션 등

□ 세부 일정

구분		시간대	주요내용
1일차	6.7.(화)		인천→바르셀로나
2일차	6.8.(수)	08:15~11:30 12:00~13:30 15:00~16:30 17:00~18:00	○ Registration & Keynote Address ○ Big Data Analytics ○ Modeling Electricity Demand ○ Time Series in Industry
3일차	6.9.(목)	08:30~11:30 12:00~13:30 15:00~16:30 16:30~18:00	○ Applied Stochastic Models in Business & Industry ○ Business Statistics ○ y-BIS Special Session ○ Time Series Modeling and Prediction
4일차	6.10.(금)	08:30~10:00 10:00~11:30 12:00~13:30 15:00~16:30	○ Keynote Address ○ Applicable Statistics ○ Statistical Theory ○ Stochastic Modeling
5일차	6.11(토)~12(일)		바르셀로나→인천

2. 워크숍 개요

□ 공식 명칭

○ International Symposium on Business and Industrial Statistics (focus on) Data Mining, Time Series and Applications

- 세계통계기구(ISI: International Statistics Institute) 산하의 산업통계를 중심으로 한 통계적 기법에 관한 국제회의(매년 개최)

□ 회의 기간 및 장소

○ 2016. 6. 8. ~ 6. 10.(3일간), 스페인 바르셀로나 카탈루냐 공과대학교

□ 워크숍 목적

○ 사업체 빅데이터 컴퓨팅의 시간 단축 등 효율적인 데이터마이닝, 예측력 향상을 위한 일반적인 시계열 분석 기법의 응용 등을 소개

II 주요 발표 내용

□ 기업의 생산성 및 품질향상을 위한 통계적 공정관리

○ 통계적 공정관리(SPC) 기법은 산업에서 데이터 모니터링 등을 통해 품질 향상 및 비용 절감 위한 핵심 역할을 함

- 반도체 설비의 SPC 적용이 대표적: 반도체 하나를 만들기 위해 수백 가지의 공정 및 단위 공정별 수천 개의 센서 존재, 실시간으로 발생하는 빅데이터의 통계적 관리가 필수

- 일반적인 SPC 방법으로 1956년 웨스턴 일렉트릭社에서 발간한 SPC Handbook에서 제시한 방법을 활용 중이나 최근의 데이터 환경은 더 정교한 분석과 의사결정이 요구됨

○ 변화점분석(Change-point Analysis)과 다변량분석을 응용하여 확률모형과 우도비검정에 기초한 누적합 관리도(control chart)의 개선된 형태를 새로운 SPC 기법으로 제안

· 품질 변수의 평균이 정규분포에 근사한다는 가정하에 구축되는 슈하르트 관리도(Shewhart, SPC의 가장 간단한 형태),

· 공정 평균의 변동에 민감한 누적합 관리도(CUSUM: Cumulative SUM),

· 지수가중 이동평균 관리도(EWMA: Exponentially Weighted Moving Average) 등이 전통적 SPC 기법으로 알려져 있음

○ 다양한 SPC 기법들은 제조 과정에서 발생하는 품질 특성치를 모니터링 하기 위해 활용되며, 제품의 주요 품질 변수의 변동을 줄이는데 도움이 됨

○ 개선된 CUSUM 관리도는 공급망자료, 보증 및 수명자료, 반도체 공정자료의 모니터링 체계 설계 등 사업체에서의 대규모 모니터링 시스템에 활용

□ 단변량 스펙트럼 분석의 예측력 구간

○ 단변량 스펙트럼 분석(Singular Spectrum Analysis(SSA))는 시계열 분석에서 광범위하게 사용되나 신뢰구간의 구성에 대한 이론적 접근이 부족함

- SSA는 주성분분석에서 파생된 기법으로 시계열의 주축을 변화시켜 자료의 경향성을 파악하는 방법

○ 시계열분석에서는 예측의 정확성을 평가하는 예측구간의 역할이 중요하므로, 잔차의 분위수와 체비셰프 부등식을 활용한 신뢰구간의 두 가지 타입을 실제 및 시뮬레이터 데이터에 적용하여 비교함

□ 링크드인 광고 플랫폼에 대한 통계모델

○ 링크드인(Linked in)은 구인구직 및 기업정보 서비스에 SNS 기능을 합친 비즈니스 전문 업체로, 이용자들의 비즈니스 네트워크를 기업 간 거래(B2B) 사업에 접목이 가능

- 산업분야, 직무, 회사 규모, 위치, 성별 나이 등에 따라 원하는 타겟, 예산, 클릭 및 노출 당 가격 등을 개인이 설정하여 광고 캠페인을 만들 수 있음

- 따라서 플랫폼 제공 업체에서는 광고 타겟팅과 콘텐츠 추천을 제공하기 위한 통계분석이 매우 중요

○ 링크드인 광고 플랫폼에 일반적인 로지스틱 회귀와 같은 기계학습 알고리즘 외, 실시간 예측 환경 안에서 대규모 통계 모형설정이 적용되는 시스템에 대한 오버뷰 제공

- ADMM(alternating direction method of multipliers) 알고리즘을 활용한 빅데이터 시뮬레이션, 고차원 공변량 분석을 소개(분석 속도가 빠름)

○ 특히, 톰슨 샘플링 알고리즘은 장기 광고 고객의 유지 및 신규 광고 캠페인의 균형을 더 효율적으로 유지하는데 도움을 줌

- 톰슨 샘플링은 베이즈 연산을 통한 무작위 확률 매칭을 사용, 구글 애널리틱스(웹 로그 분석 서비스)에서 활용하는 것이 대표적

○ 광범위한 오프라인 실험과 온라인 A/B 테스트를 통해 이 시스템이 상당한 이익, 예측 정확도 향상, 수익성, 클릭률(또는 전환율)*에 유의미한 효과가 있음을 보여줌

* Click-through rate, CTR: 클릭수를 광고가 보여진 횟수(노출, Impression)로 나눈 것을 백분율로 나타내는 것, 즉 어떤 사용자가 온라인광고 노출 대비 클릭을 한 횟수

- A/B 테스트는 웹 사이트 방문자를 임의로 두 집단으로 나누고, 한 집단에게는 기존 사이트를 보여주고 다른 집단에게는 새로운 사이트를 보여준 다음 새 사이트가 기존 사이트에 비해 좋은지를 정량적으로 평가하는 방식, 주로 웹서비스에서 사용

□ 가법모형을 활용한 버스 소요시간 예측

○ 일반적으로 버스 소요시간은 교통체증, 날씨, 해당 지역 행사가 영향을 미치는데 다른 측면(요일 또는 시간대)으로는 변수들 간의 관계에 따라 직·간접적으로 영향을 미침

- 소요시간 예측변수는 대부분 상황에서 정확도를 높이기가 어려움

○ 이 논문에서는 가법모형을 사용해 버스소요시간 모형을 자유롭게 설계할 수 있는 프레임워크를 개발

- 특히, 예측함수의 평활함수로 설정된 선형·비선형항의 합으로 소요시간 모형을 설정하는데, 이 모형은 설계면에서 매우 유연

○ 리우데자네이루시의 버스운행자료를 GPS 자료에 적용하여 비교한 결과 안정적으로 예측시간의 우수성을 입증

□ 시간공변량을 활용한 주거용 전력 발전의 세미 마르코프 모델

○ 겨울철 가정용 소형 열병합발전으로 생산된 전기에너지 자료를 확률과정으로 분석하여, 그리드(전력망)에 공급되는 전력량을 평가

- 한 개의 열병합발전기로 생산된 전기에너지 시계열*을 베이지안 추론을 통해 설명

* 시간에 따라 의존하는 공변인(시간공변량)을 적용한 가속화 고장 시간 모형(Accelerated Failure Time model, 모수적 생존분석 기법)으로 얻어진 자료

○ 시간공변량을 활용한 교대재생과정(alternating renewal process)의 stationary trajectories를 마르코프 체인을 사용해 시뮬레이션

- 시뮬레이션 결과는 이탈리아 정부에 의해 수집된 자료에 적용하여 예측

□ 베이지안 비모수 마르코프 스위칭 확률변동성 모형에 대한 파티클 러닝

○ 금융 자료의 베이지안 비모수 확률변동성(SV) 모델의 추정에 대한 SMC* 알고리즘 설계에 관한 내용을 소개

* Sequential Monte Carlo, 순차적 몬테 카를로

○ 특히, 파티클 러닝(Particle Learning(PL))이라고도 불리우는 최신의 파티클 필터를 사용함

- 파티클 필터는 베이지안 접근법에 기반하여 시간 의존적 시스템 모델을 추정하고 순차적으로 갱신하는 방법, 마르코프 체인 몬테 카를로 일괄처리법을 순차적으로 유사하게 만든 것

- 온라인 유형의 추론에서도 활용 가능

○ 파티클 방법을 비모수 확률변동성 모형에 대한 일반적 마르코프 체인 몬테 카를로(MCMC) 방법과 비교함

○ 사후분포는 관측된 새로운 데이터로 업데이트 되는데, MCMC 방법은 비용이 많이 드는 단점

○ 본 논문에서는 비모수 확률변동성 모형으로 마르코프 스위칭 모형을 제안하고 파티클 러닝을 사용하여 시뮬레이션

○ 두 가지 비모수 확률변동성 모형(스위칭 모형과 스위칭 없는 모형)을 실제 금융 시계열자료에 적용하여 비교

○ 시뮬레이션 결과 마르코프 스위칭 모형이 더 높은 예측력(분포의 꼬리 부분에서)을 보여줌

□ 다차원 계층적 베이지안* 프레임워크를 통한 희귀사건의 추정률

---

* hierarchical Bayesian: 만약 모수들이 서로 연관되어 있다면 모수들의 결합확률 모형에서 이들 사이의 의존 관계를 반영해야 하므로 모형화를 계층적으로 하는 것

○ 야후에서는 다차원 추론을 수행하기 위해 기존 계층 및 몇 가지 선택 기능을 사용하여 희소 데이터에 대한 희귀사건의 발생률을 추정하는 문제에 대한 연구가 진행 중임

○ 특히 여러 세분화된 계층에서 상황정보를 파악하는 (광고주, 게시자, 사용자) 튜플에 대한 클릭률 추정 문제에 포커스

- 일반적으로 클릭률은 낮고, 계층·차원의 적용이 일치하기 어려움

○ 이러한 어려움을 극복하기 위해 텐서곱 분해를 이용한 3차원 클릭률(또는 전환율)*의 결합 사전인자를 분해하고 다차원 계층적 베이지안 프레임워크를 제안

* Click-through rate, CTR: 클릭수를 광고가 보여진 횟수(노출, Impression)로 나눈 것을 백분율로 나타내는 것, 즉 어떤 사용자가 온라인광고 노출 대비 클릭을 한 횟수

- 차원별 특성을 모형화하기 위해 각 차원별 특정 프레임워크를 설정

- 광고주 및 게시자 차원에 대한 새로운 계층적 사전인자를 고려하고, 사용자 차원에 대한 기능 의존적 혼합 모형을 고려함

○ 그 외, 추정에 대한 맵리듀스*와 스파크**를 통한 알고리즘을 제안했는데 이는 확장성 및 대규모 데이터마이닝 적용 부분에서 탁월

* MapReduce: 구글에서 대용량 데이터 처리를 분산 병렬 컴퓨팅에서 처리하기 위한 목적으로 제작하여 2004년 발표한 소프트웨어 프레임워크

** Spark: SQL과 스트리밍 데이터, 머신 러닝을 한 곳으로 통합·운영할 수 있는 오픈소스 소프트웨어(OSS)로, 맵리듀스를 벗어나 다양한 작업을 빠르게 처리

○ 제안한 다차원 계층적 베이지안 프레임워크를 실제 야후 광고 캠페인 플랫폼에 적용한 결과, 클릭 성향 분석에서 매우 드문 성향을 구별해낼 수 있었음을 증명

Ⅲ 시사점

□ 통계학의 최신 연구 동향을 이론 탐구에 그친 것이 아닌, 특정 사업체 생산의 예측 정확도 및 수익성 향상을 위해 활용

○ 반도체 등 주요 제조업체의 공장을 디지털화하고 스마트 팩토리를 구현하여 산업용 IoT(IIoT)를 통한 산업별 빅데이터 분석 및 의사결정 가치의 중요성을 강조

- 일반적인 SPC 방법을 개선하여 더 정교한 분석과 의사결정이 가능한 개선된 형태의 CUSUM 관리도 제안

○ SNS 등 온라인서비스의 주요 수입원인 광고캠페인 플랫폼 분석 모델을 개선하여 매출 상승에 기여

- 링크드인, 야후 등에서는 웹서비스 이용자의 성향, 특히 소수 성향까지 분석 가능한 모형 구축(베이지안 접근)으로 광고 타겟팅 및 콘텐츠 추천의 정확도 향상

□ 빅데이터 분석 성능 및 시계열예측력 향상을 위한 알고리즘 적용 등 이론과 실무의 조화

○ 전통적 시계열분석 방법의 이론적 탐구를 통한 예측구간 설정 및 모형개발로 정부 및 사업체 통계에 활용

- 단변량 스펙트럼 분석(SSA)의 예측구간 설정, 가법모형을 활용한 대중교통 소요시간 예측, 마르코프 모형을 활용한 전력생산량 예측 등

○ 빅데이터 시뮬레이션의 분석속도 향상을 위한 ADMM, 톰슨 샘플링 알고리즘 등을 적용

□ 경제 및 산업통계에 관한 심도 깊은 이론 및 실무적 접근으로 업무 담당자의 전문성 향상에 기여

○ 각국의 주요 산업(반도체 등)에 관한 이슈, 산업통계 분야별 연구 중인 방법론 및 적용, 각 기업 실무자와의 교류 등을 통해 자기계발 및 담당 업무의 전문성 제고에 기여

- 연구논문의 경쟁이 있는 분위기로 자료 공유가 쉽지 않으나 연구내용이 심도 깊고, 국제적인 대기업의 빅데이터 및 예측력 등에 관한 기법 소개 등이 인상적임