The use of Big Data in food safety

大数据在食品安全中的应用

Hans Marvin, Yamine Bouzembrak, Martijn Staats and Esmée Janssen (RIKILT Wageningen UR)
Outline

- Introduction to Big Data
- Use of media in food safety alerts
- Bayesian Networks (BNs) modelling in Big Data research
- Application of BNs to predict food fraud
- Conclusions
Introduction

- Lots of data is being collected and stored, e.g.:
  - Web data, e-commerce
  - Bank/Credit Card transactions
  - E-mail
  - Data recorded by sensors, scanners, other mobile devices.
  - Chemical monitoring data
  - Omics data
Introduction

The FOUR V's of Big Data

From traffic patterns and music downloads to weather history and medical records, data is collected, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity.

Volume

- 2020: 40 Zettabytes of data will be created, an increase of 300 times from 2005.
- 6 billion people have cell phones.
- Most companies in the U.S. have at least 100 terabytes of data stored.

Velocity

- The New York Stock Exchange processes 1 TB of trade information during each trading session.
- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure.
- By 2015, there will be 18.9 billion network connections—almost 2.5 connections per person on earth.

Variety

- 400 million tweets are sent per day by about 200 million monthly active users.
- 38 billion pieces of content are shared on Facebook every month.
- 4 billion hours of video are watched on YouTube each month.

Veracity

- In one survey, 27% of respondents were unsure of how much of their data was inaccurate.
- 1 in 3 business leaders don’t trust the information they use to make decisions.
- Poor data quality costs the US economy around $3.1 trillion a year.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, WEPTEC, QS

By 2011, the global size of data in healthcare was estimated to be 130 exabytes.

As of 2011, it’s anticipated there will be 420 million wearable, wireless health monitors.

IBM
Introduction

Definition of European Commission:

The term "big data" refers to large amounts of different types of data produced with high velocity from a high number of various types of sources. Handling today's highly variable and real-time datasets requires new tools and methods, such as powerful processors, software and algorithms.\(^7\)

\(^7\) Going beyond traditional "data mining" tools designed to handle mainly low-variety, small scale and static datasets, often manually.
Introduction

- The EC has developed a strategy on Big Data and supports a data-driven economy.
  - Big Data as the centre of the future knowledge economy.
  - Supporting open access of data, global connections and e-infrastructure services (www.openaire.eu)

- The Dutch Government stimulates public-private projects to explore Big Data.
  - Big Data Value Center (www.bdvc.nl)
  - Netherlands eScience center www.esciencecenter.nl
Introduction

Typical Big Data workflow

Data Collection
- Internet
- Mobile phones
- Social media
- Omics profiling
- (online) databases

Data Storage & Transferring
- NoSQL
- Operational infrastructure
- MEDLINE/PubMed
- Gene Expression Omnibus (GEO)
- ChemKap

Data Analysis
- Text-mining
- Neural Networks
- Bayesian Networks
- Other pattern-recognition and machine-learning algorithms

Report/Visualisation
- R
- Ciros
- Gephi
- Hugin

Figure adapted from Huang et al. (2015)
Introduction; example Big Data in food safety

- FOSCOLLAB: Global platform for food safety data and information.
- Support risk managers and decision-makers
- Integrate data and information from animal/agricultural, food, and human health
- Promote better data generation

http://www.who.int/foodsafety/foscollab_dashboards/en/
Outline

- Introduction to Big Data
- Use of media in food safety alerts
- Bayesian Networks (BNs) modelling in Big Data research
- Application of BNs to predict food fraud
- Conclusions
Use of media in food safety alerts

The Europe Media Monitor (EMM) provides advanced analysis systems for monitoring of both traditional and social media.

• EMM presents the latest news and classifies it according to subject.
• It is updated every 10 minutes, 24 hours per day.
• It gathers reports from news portals world-wide in 60 languages,
• EMM classifies the articles, analyses the news texts, aggregates the information and issues alerts.
• EMM applies text mining techniques to screen different types of media on the world wide web: websites, databases, blogs, ..etc.
• EMM contain 3 portals: NewsBrief, NewsExplorer and MedISys

(http://emm.newsbrief.eu/overview.html)
Use of media in food safety alerts

MedISys collects food safety news and RIKILT adapted it also to detect food fraud.
Use of media in food safety alerts

Food fraud reports in MedISys (period September 2014 to June 2015; N > 600)
Use of media in food safety alerts

Food fraud reports in MedISys (period September 2014 to February 2015)

- **UK**
- **USA**
- **France**
- **India**
Outline

- Introduction to Big Data
- Use of media in food safety alerts
- Bayesian Networks (BNs) modelling in Big Data research
- Application of BNs to predict food fraud
- Conclusions
Bayesian Networks (BNs) modelling in Big Data research

BN takes into account a number of quantitative and qualitative criteria under uncertainty (Dogan and Aydin, 2011).

BN approach offers a high potential for representing ambiguous knowledge and for performing reasoning under uncertainty.

BN are very powerful for making inferences and drawing conclusions based on available information (Jensen, 1996).

BN can combine expert and domain knowledge that allows flexible inference even with partial and limited information (Lauritzen, 1995).

BN requires less parameters than the conventional naïve method (Dogan, 2012).

BN approach offers a high potential for representing ambiguous knowledge and for performing reasoning under uncertainty.
Bayesian Networks (BNs) modelling in Big Data research

BN is able to combine data from different nature and location
Outline

- Introduction to Big Data
- Use of media in food safety alerts
- Bayesian Networks (BNs) modelling in Big Data research
- Application of BNs to predict food fraud
- Conclusions
Application of BNs to predict food fraud

- (Bouzembrak and Marvin, 2015) used BNs prediction model to aid the border inspector to detect the type of food fraud.
Application of BNs to predict food fraud

Prediction of food fraud type in RASFF
Application of BNs to predict food fraud

Application of BNs to predict food fraud

Products related to food fraud in RASFF

- Meat and meat product: 20%
- Nuts, nut products and seeds: 15%
- Fish and fish products: 10%
- Prepared dishes and snacks: 5%
- Other food product / mixed: 5%
- Poultry meat and poultry meat products: 5%
- Fruits and vegetables: 5%
- Milk and milk products: 5%
- Confectionery: 5%
- Bivalve molluscs and products thereof: 5%
- Crustaceans and products thereof: 5%
- Cereals and bakery products: 5%
- Fats and oils: 5%
- Eggs and egg products: 5%
## Application of BNs to predict food fraud

### Food fraud types in RASFF

<table>
<thead>
<tr>
<th>Fraud type</th>
<th>Description</th>
<th>Count of notification type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC</td>
<td>Improper or fraudulent or missing or absent of health certificate</td>
<td>370</td>
<td>44%</td>
</tr>
<tr>
<td>Illegal importation</td>
<td>Illegal or unauthorized import or trade or transit</td>
<td>256</td>
<td>31%</td>
</tr>
<tr>
<td>Tampering</td>
<td>Adulteration or fraud or tampering</td>
<td>126</td>
<td>15%</td>
</tr>
<tr>
<td>CED</td>
<td>Improper or expired or fraudulent or missing of common entry document or import declaration</td>
<td>45</td>
<td>5%</td>
</tr>
<tr>
<td>Expired Date</td>
<td>Expired Date</td>
<td>24</td>
<td>3%</td>
</tr>
<tr>
<td>Mislabelling</td>
<td>Mislabelling</td>
<td>13</td>
<td>2%</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td>834</td>
<td>100%</td>
</tr>
</tbody>
</table>
Application of BNs to predict food fraud

Modelling of Fraud in RASFF

Model developed with RASFF data up to 2013
Application of BNs to predict food fraud

USER-INTERFACE

Questions
RASFF
Notification country?
Notification type?
Country of origin?
Product name?
Year?

Food Fraud Type
44.44 % = HC
31.03 % = Illegal importation
14.56 % = Tampering
5.24 % = CED
3.07 % = Expired Date
1.66 % = Labelling
Application of BNs to predict food fraud

CASE: What type of fraud can we expect from meat imported from Belgium

Labelling!
Application of BNs to predict food fraud

Validation of the Model;

Method:
Predict the type of fraud for all reports in RASFF in 2014.
Input values: type of product and country of origin

Results:

● 82% correct
● 51% prediction of new fraud combinations

Higher performance with dynamic model
Bayesian Networks (BNs) modelling in Big Data research

BN is able to combine data from different nature and location
Application of BNs to predict food fraud

Prediction of food fraud using multiple data sources and expert knowledge

- Identify parameters relevant for food fraud (=> nodes in BN model)
- Determine the relationships between these parameters (expert knowledge, literature)
- Identify data sources of these parameters
- Develop the BN model
- Validate the BN model
Application of BNs to predict food fraud

Parameters relevant for food fraud (some examples)

**Economic parameters:**
- Prices of the fraudulent product at the time of detection
- Price spike around the period of detection
- Trade volumes of the product between the country of detection and country of origin
- Complexity of the food chain

**Parameters of the country of origin & detection**
- Indices: corruption index, food safety index, governance index, legal system index, press index, human development index and technology index.

**Detected food fraud cases**
- RASFF, UPS and EMA
Application of BNs to predict food fraud

Preliminary BN model for predicting food fraud; relationships between all parameters
Preliminary BN model for predicting fraud linking 36 data sources (18 databases and 8 expert judgements)
Conclusions

- Big data developments show great potential in food safety research, food safety risk assessment and food safety risk management.
- European Media Monitor MedISys is suitable to collect publications on food safety and food fraud on global level.
- Bayesians Network approach is useful to combine information from a variety of data sources.
- Bayesians network models are useful to predict food fraud.
Thank you
谢谢