CKME136: Capstone

Literature Review & Data Description

By Geoffrey Clark
geoffrey.clark@ryerson.ca
https://github.com/gffryclrk/toronto.fire.incident.data
June 11th, 2018

Literature Review

Below I read, reviewed and summarized 8 papers pertaining to Fire Incident Data. This was a valuable exercise in familiarizing myself with the dataset. With direction from Dr. Ceni Babaoglu I attempted to answer the following questions (where applicable): Dataset, dataset size, research questions & summary, Machine Learning algorithms, dependent variables and description of exploratory data analysis or subsetting. The papers varied in which of these questions could be answers and so I did my best to provide as much information as possible. Full references are included at the end.

Modeling the risk of structural fire incidents using a self-organizing map

Asgary, A., Naini, AS., & Levy, J. (2012)

In this research the authors use a Self-Organizing Map to classify and asses the risk level of structural fire incidents. Self-organizing maps are a type of Artificial Neural Network and have demonstrated better performance for predicting fire phenomena, compared to other traditional models, according to the authors.

The dataset used contained "more than 12,000" structural fire incidents located in Toronto and occurring in the years 2000 through 2006. There were 16 features in the dataset, all pertaining to structural fire incidents, such as type of property, complex and building as well as fire details, time, date and location. Records with null feature values or a number of repetitive or similar records with a risk level of Very Low were randomly excluded from the dataset to "ensure an unbiased dataset". The training set ultimately included 2808 fire or explosion records while the test set contained 936 randomly selected records.

The class variable was risk levels and each observation was assigned a class; risk varying from very low to very high, based on number of fatalities, injuries and estimated damage. After implementing their model the authors assessed the performance of their classifier based on resolution and topology preservation which are "two numeric evaluation criteria typically used to evaluate the quality of the SOM." The authors also provide a graphical representations of the final map.

Interpreting these results is quite cryptic and confusing without at least an introductory understanding of Artificial Neural Networks and Self-Organizing Maps. The authors report on the different rates of classification error, all of which are under 10%.

Modeling number of firefighters responding to an incident using artificial neural networks

Asgary, A., & Sadeghi-Naini, A. (2013)

In this research the authors' purpose is to model the number of firefighters responding to an incident based on structural fire incident data. The approach was to use Artificial Neural Networks (ANN) to predict number of responders on unseen records after training on seen records (training & test sets). The input data contained 12,500 observations of 16 features pertaining to structural fire incidents in Toronto from 2000 through 2006.

The authors choose to use ANN because the prediction is "complex, non-linear and ill-defined." Further, studies show that "ANN is a suitable model for such type of problems. ANN can handle noisy and incomplete data that is often the case in fire incidents records." The speed of ANN prediction is also attractive to the researchers as they explore the opportunity to implement a data model into the Computer Aided Dispatching systems to provide first responders with increased information at time of call.

The researchers categorized the number of responding personnel into seven categories ranging from less than ten to greater than 25. The categorization error ranges from 10 – 33% on the different categories which can be interpreted as either favourable or unfavourable depending on the requirements of the application. However, considering the range (0-99) and standard deviation (12.40) of the class variable such such accuracy might be an improvement against other models. However I am unaware of similar research using other models for comparison purposes.

Modeling loss and no-loss fire incidents using artificial neural network: Case of Toronto

Asgary, A., Sadeghi-Naini, A., & Kong, A. (2009)

In this research the authors apply an Artificial Neural Network (ANN) to predict loss prediction of fire incidents. The premise is that an effective predictor at call time will improve the ability of emergency services to respond to fire incidents. The dataset used included "13893 records of structural fire or explosion incidents reported to the fire stations over the Toronto area by 811, monitoring agencies, police services, ambulances, civilians etc." The dataset included various features including type of property, building and complex, level on fire, fuel type, possible cause, time and date.

The authors classify the observations into loss and no-loss incidents. The loss incidents are defined as "fire incidents with some form of human or property losses" and no-loss fires are defined as "fires without significant human or property losses." The class variable is therefore categorical (binary).

The results of the research are interpreted in two ways: probabilistic and binary prediction. The ANN predicts portions of input data which could lead to loss and accurately predicts this range. The authors then present a ROC curve; True Positive Rate against False Positive Rate, for the binary predictor. The authors then conclude that the results "indicated a very decent ability of such model [sic] to estimate the probability of an incident being a loss one. The probabilities estimated by the FF-NN were realistic in all cases. Binary prediction results of the FF-NN also presented a promising ability of such system to predict an incident to be a loss or no loss one."

My understanding of the results is that they are in bulk. For example, the ANN probabilistic prediction output is the proportion of input data that would result in a loss incident. I am unsure of the model's accuracy in classifying individual events, such as to provide increased information at response time.

Intelligent security systems engineering for modeling fire critical incidents: Towards sustainable security

Asgary, A., Sadeghi-Naini, A., & Levy, J. (2009)

This research aimed to classify fire incidents into categories of estimated damages using data on Toronto Fire Incidents from the years 2000 to 2006. The categories were defined in the dataset: less than $100, between $100 and $5,000, between $5,000 and $50,000 and greater than $50,000.

The dataset itself contained 12,500 observations of 16 features pertaining to structural fire incidents. The researchers used an Artificial Neural Network as their model for this research and were able to get the predictions of probability of damage within their observed ranges for most values. This led the researchers to conclude that the results were encouraging and that ANN would be an effective classifier for this type of dataset.

Exploratory and inferential methods for spatio-temporal analysis of residential fire clustering in urban areas

Ceyhan, E., Ertŭgay, K., & Düzgün, Ş. (2013)

This is an exploratory study into spatio-temporal analysis of residential fires in an urban setting. The primary aim of the study is "not to evaluate a specific fire clustering pattern in a detailed manner but to provide guidelines to the decision makers for the use of various spatio-temporal data analyses techniques in understanding fire clustering patterns."

The dataset used was obtained with permission from the Ankara Metropolitan Municipality (Turkey) and contains address information of all residential fires reported inside the Çankaya district in 1998 and 2005-2009. The authors were then able to geocode approximately 52% of the addresses using available software.

The exploratory research includes comparisons of residential clustering and fire clustering to determine if higher fire density is merely a representation of residential density. The graphical cluster plots suggest otherwise. The authors also use Diggle's D-function for each year as an objective metric

of cluster density and this metric agrees with the visual representation of clusters that there is a difference between residential and fire clusters.

The authors then go on to find patterns in fires as both a function of time and distance. The results show “significant space-time interaction for the time at year level.” However, further analysis implies a “lack of space-time interaction for the time at the month level” and “mild space-time interaction for the time at the week level.”

Spatial and temporal analyses of structural fire incidents and their causes: A case of Toronto, Canada

Asgary, A., Ghaffari, A., Levy, J. (2009)

This research examines spatial-temporal relationships among Toronto Fire Incident Data. The dataset for the research is fire incident records from 2000 to 2006 obtained by the Ontario Office of the Fire Marshal. The database includes 199,534 fire incidents “out of which 110,261 records had full addresses and were geocoded.” Additionally there were “20,182 geocoded fire records for the city of Toronto out of which 16,172 were identified as structural fires.”

The authors used three different methods in their research: temporal, spatial and spatiotemporal. The three types of analyses were used to find trends among these types of variables in order to provide increased insight into Toronto’s fire dataset. The methodology was quite straightforward: for temporal, begin by subsetting the dataset into times, and then further subset the dataset into the dependent variable of interest. In the case of this research it was cause of fire. For spatial the researchers applied Kernel Density Estimation (KDE) and Average Nearest Neighbors Density (ANND) to calculate fire density statistics based on location. The spatiotemporal techniques involved were a combination of the temporal and spatial data analyses and combined to create such visualizations as map animations of fire densities across time.

The results of this analysis are some trends within the fire incident dataset in Toronto. For example, the temporal results show that “structural fires are more frequent during the weekend and that Wednesdays and Thursdays have the least number of fire incidents.” Further, the “highest number of structural fires occur in the spring and the lowest … in autumn.” The researchers then go on to show density plots of fires across the city, most of which concentrate on downtown. Isolating the effects from increased downtown population density doesn’t appear to be explored as in Ceyhan, et al. The research then turns its focus to spatiotemporal analysis of fires across time and region in which bivariate comaps are used to visualize the data.

Necessity of Fire Statistics and Analysis Using Fire Incident Database - Japanese Case -

Sekizawa, A. (2012)

This is a roughly translated transcript of a Japanese presentation on fire incident data. Although not academic in nature it is still useful for general discussion on fire incidents and pertaining data. The article focuses broadly on different fire incident summary statistics across countries (primarily Japan, USA, UK and Canada) and time (the latter half of the 20th Century until present). The presentation then goes on to discuss a breakdown of victims of fire by age and shows trends across time. This demographic breakdown is important to understand how trends may change with a changing population. Although this wasn't academic research it was quite interesting and useful with a directly relevant topic.

The Implementation and Utility of Fire Incident Reporting Systems: The Delaware Experience

Bergen, G., Frattaroli, S., Ballesteros M.F., Ta, V.M., Beach, C., & Gielen, A.C. (2008)

This research investigates Delaware's implementation of the USA's National Fire Data Center (NFDC). The NFDC is used to collect information on fires through the country so that the data may be used to "create annual reports… identify specific fire problem areas, relocate stations, justify budget requests, and conduct research."

The researchers analyzed data from the Delaware Fire Incident Reporting System (DFIRS) from four modules, Basic, Fire, Structure Fire & Civilian Fire Casualty, between May 2003 and December 2004. They discuss data completeness and the importance of collecting fire data. The researchers display temporal patterns within the data by subsetting fires and injuries by month and day. The researchers notice significantly higher numbers of fire in January and on Sundays. However, "there were no significant differences in injuries and deaths by either month or day of the week."

This paper goes on to explore many subsets within the data including percentages of cause of ignition, area of origin, heat source and type of material first ignited. Furthermore the authors present characteristics of smoke alarms in residential fires and discuss the statistics. There are also breakdowns of the data pertaining to injuries.

Other Sources

CKME136 Capstone Project

Maninder Kohli (2018)

Mani completed his Capstone in the Winter Semester, 2018, for the Ryerson Big Data Analytics & Predictive Analytics Certificate with the research question "Analysis and predictions of fire fighter injuries and Rescues (RESCUES, FF_INJURIES) from a set of variables of interest." The dataset used was the Fire Services Incident Data from the Toronto Open Data Catalog.

Mani and I met in May of 2018 and he showed me his project including some of the steps he had taken to clean and prepare the data as well as challenges he had, which types of machine learning

algorithms he used as well as his results. Although we have different goals with our project there is quite a bit of overlap, including the dataset itself. For this reason I thought it best to include my meeting with Mani in the spirit of full disclosure & academic integrity.

Data Description

Dataset: Fire Services Incident Data
Owner: Toronto Fire Services
Currency: July 2017
Format: XML
Refresh Rate: Annually
Retrieved from: City of Toronto Open Data Catalog on May 12th, 2018

Data Preparation

The first major step in this project was to convert from XML to tabular structure. The tabular structure I chose was .csv because this format is well supported in various database and statistical packages. The conversion was made using XQuery and processed with BaseX, an open-source XQuery processor. The XML structure was as follows:

```
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<FIRE_DATA>
      <INCIDENT>
            <RespondingUnits>
            </RespondingUnits>
      </INCIDENT>
</FIRE_DATA>
```

Within the INCIDENT nodes there were also 100 features pertaining to the fire incident itself. Further, details on the Responding Units were nested within the Incident element itself creating a tree-like XML structure. A flat, table-like, structure was desirable and so the INCIDENT data was exported to .CSV for each year using XQuery and a Bash shell script which automatically changed the input and output files for the XQuery conversion script. Further, the data on Responding Units was also extracted from the XML and placed in flattened .CSV files for later use using the same technique. The unique incident number, from the parent element, was included with each Responding Unit observation so that at a later time it could be used as a KEY to JOIN the Responding Units table with the Incident Table.

Incident Observations by Year

Year	Observations
2011	145365
2012	120545
2013	109576
2014	111794
2015	115664
2016	117426
Total	720370

Table 1

Once available in tabular format the data was loaded into an R data.frame. The entire INCIDENT dataset consists of 720,370 observations of 100 features. The features include various variables that pertain to fire incidents including date and time, a unique incident identifier, the type of fire, fuel of ignition source, firefighter injuries, civilian injuries, other services & ministries contacted, and details on the building and property. Table 1 & 2 show the breakdowns of Incident and Responding units by year, respectively. The first year of Incident records, 2011, has the largest number of observations with the remaining years relatively constant. All years have about a similar quantity of responding unit records.

Type Conversion

The next processing step conducted was converting Date features. The three date features in the dataset are DISPATCH_DATE, ARRIVAL_DATE and INCIDENT_DATE. Both DISPATCH_DATE and ARRIVAL_DATE are identical so ARRIVAL_DATE was removed to reduce redundancy. DISPATCH and INCIDENT dates were then converted to POSIXct object in R.

A great majority of the INCIDENT features are of categorical type and thus were converted to factors. NA values were kept as their own level in the factors for future reference. The rest of the INCIDENT features were converted to the most suitable type; mostly Numeric.

Responding Unit Observations by Year

Year	Observations
2011	275210
2012	261988
2013	270696
2014	276175
2015	281603
2016	277013
Total	1642685

Table 2

NA

This is a sparse dataset with many NA values observed. Some features, such as AGENT_APP_MIN, AGENT_APP_HOUR, AGENT_APP_SEC, CONTROL_HOUR, etc have over 700,000 NA values in a dataset of 720370. This analysis will further explore the meaning of NA values in this dataset. For example, some NA values may exist where the field wouldn't make sense, such as building type for an outdoor fire. Other NA values may be the result of incomplete incident repots at the time of incident. Challenges will undoubtedly occur as a result of having features with so many NA values. A row-wise summary of NA values shows, out of 99 features (after dropping ARRIVAL_DATE), a minimum of 12, max of 70 and median of 58. The distribution of NA values has a left skew with the 1st Quartile starting at 58 NA values. As a result of so many NA values, features and observations may have to be dropped entirely.

Approach

The next steps in this analysis will be to subset the dataset further in order to explore trends and patterns in the data. Multiple papers that were reviewed for the Literature Review section above were able to find breakdowns in the data such as which days, months and hours had the most fire incidents and what the different breakdown of causes of the incidents were. Trends over time will be explored and analyzed.

References

Asgary, A., Naini, AS., & Levy, J. (2012). Modeling the risk of structural fire incidents using a self-organizing map. *Fire Safety Journal, 49,* 1-9.doi:10.1016/j.firesaf.2011.12.007

Asgary, A., & Sadeghi-Naini, A. (2013) Modeling number of firefighters responding to an incident using artificial neural networks. *International Journal of Emergency Services, 2(2),* 104-18.doi:10.1108/IJES-03-2012-0001

Asgary, A., Sadeghi-Naini, A., & Levy, J. (2009). Intelligent Security Systems Engineering for Modeling Fire Critical Incidents: Towards Sustainable Security. *Journal of Systems Science and Systems Engineering, 18(4),* 477-488.doi:10.1007/s11518-009-5121-2

Asgary, A., Sadeghi-Naini, A., & Kong, A. (2009). Modeling loss and no-loss fire incidents using artificial neural network: Case of Toronto. *Science and Technology for Humanity,* 159-163.doi:10.1109/TIC-STH.2009.5444513

Ceyhan, E., Ertŭgay, K., & Düzgün, Ş. (2013). Exploratory and inferential methods for spatio-temporal analysis of residential fire clustering in urban areas. *Fire Safety Journal, 58,* 226-249.doi:10.1016/j.firesaf.2013.01.024

Asgary, A., Ghaffari, A., Levy, J. (2009). Spatial and temporal analyses of structural fire incidents and their causes: A case of Toronto, Canada. *Fire Safety Journal, 45,* 44-57.doi:10.1016/j.firesaf.2009.10.002

Sekizawa, A. (2012). Necessity of Fire Statistics and Analysis Using Fire Incident Database - Japanese Case -. *Fire Science and Technology, 31(3),* 67-75.doi:10.3210/fst.31.67

Bergen, G., Frattaroli, S., Ballesteros M.F., Ta, V.M., Beach, C., & Gielen, A.C. (2008) *J Community Health, 33,* 103-109.doi:10.1007/s10900-007-9070-8

Walmsley, P. (2007) *XQuery.* Sebastopol, CA: O'Reilly Media, Inc.

Ripley, B.D., & Hornik, K. (2001) Date-Time Classes. *R News, 1(2),* 8-11. Retrieved from http://cran.r-project.org/doc/Rnews/Rnews_2001-2.pdf