Canadian Disaster Analysis Using Machine Learning

Junhan Liu
University of Waterloo,
j896liu@uwaterloo.ca
GitHub: https://github.com/jimjimliu/Canadian-Disaster-Datamart.git

Abstract

By using data provided by the Canadian government[1], we are able to view disasters occurred within Canada in the past years from 1900. We use this rather small data set to try to understand a certain inner relationship between a disaster that happened and its recorded features. This model aims to find if a disaster's type is predictable from various feature combinations of the disaster such as money spent on recovering damaged properties, casualties, etc. We employ a set of conventional classifiers such as SVM, random forest, etc. Among these trained models, the best performed is considered as the final model. As a result of evaluating the prediction performance of each model, the best one is GBDT which has a testing accuracy of **87.0%** before re-sampling and **72.3%** after over sampling. The purpose of the task is to find some connections of disasters and the recorded data properties provided by official authorities. Our hypothesis is that different disasters must cause damages at different levels, for example, a mass natural disaster must cost more money and casualties than a fire incident caused by electrical wires. Therefore, the question of "will we be able to know what type a disaster is when we know such as its cost, casualties, etc." is raised. Since at times, some disasters remain unknown even there is a serious investigation. To answer such problems, our model aims to provide some useful insights.

1 Introduction

The occurrence of disasters such as fire, tsunami, earthquake, etc. is increasing since 1900 within Canada as Figure 1 shows. The overall tendency is increasing since the year 1900. Our data set provided by the Canadian government[1] records various kinds of disasters occurred within Canada including natural disasters, technology disasters, etc. The damage caused by disasters is becoming more and more severe in terms of many aspects including the money spent on recovering and human life's loss (Choi et al. 2018). Given the fact that investigations after disasters' occurrences have been carried out, many disasters' types still remain unknown. There are many unknown disasters recorded in our data set.

The task of this short paper is to find out if it is possible and feasible to deduce(predict) a disaster's type when given its recorded features. If it is feasible, it might provide some useful insights for official authorities to better understand a certain kind of disaster and even use that as a baseline to predict future disaster types. The incentive to build this model is based on the hypothesis that different kinds of disasters might have a different cost. For example, natural disasters like tsunami may cause more money spent on recovering properties and more casualties than a normal household fire caused by the ageing of electrical wires. Based on this idea, we shall ask the question: 'is it possible to deduce the type of disaster when given enough feature dimensions such as the money spent on recovering and the number of casualties?'. Due to

[1] https://www.publicsafety.gc.ca/cnt/rsrcs/cndn-dsstr-dtbs/index-en.aspx

the fact that our input data is a rather small set, this model could be treated as a baseline to show that there is a possibility to perform further analysis especially when the input data source is improved and expanded. Our final result reaches an 87% accuracy on predicting an instance into one of the following labels: 'natural', 'technology', 'conflict', and 'unknown'. The accuracy of 72.3% is reached after over sampling since the input training data is badly imbalanced. As above illustrated, our analysis aims to provide useful insights for official authorities to identify a disaster's category faster and more accurate.

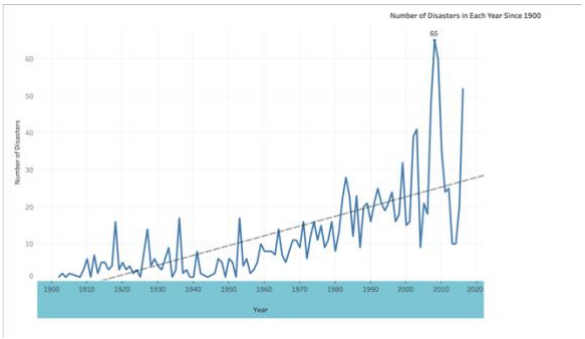

Figure 1: Number of disasters since 1900

The model aims to predict a disaster instance into one of the four labels: 'technology', 'natural', 'conflict', and 'unknown' from a different combination of variables. The possible features are shown in Table 1. From Table 1, we try to pick a feature set that contributes the most to the final testing results. By using all four features as the final feature set, it performs the best and generates the highest accuracy using GBDT.

1.	Population	Number
2.	Fatalities	Number
3.	Evacuated	Number
4.	Cost	Number

Table 1: Potential features and data types

There are 2 sections of the project. The first section performs classification that categorizes an instance into one of the four labels mentioned above. The second part is clustering. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters) (Deshmukh et al., 2016). Before doing anything to the data set, we aim to discover some not-easy-to-see properties of the data. A most frequent case of using unsupervised learning is for explorative analysis to discover if the data are characterized by a small number of representative patterns that can be used to summarize the dataset in a more compact representation. We try to find some patterns or groups using variables: 'population', 'cost', 'fatalities', and 'evacuated' as Figure 2&3 shows.

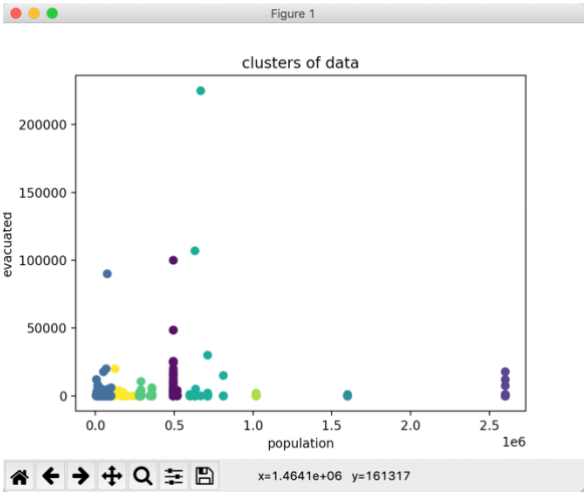

Figure 2: Cluster of people evacuated and population

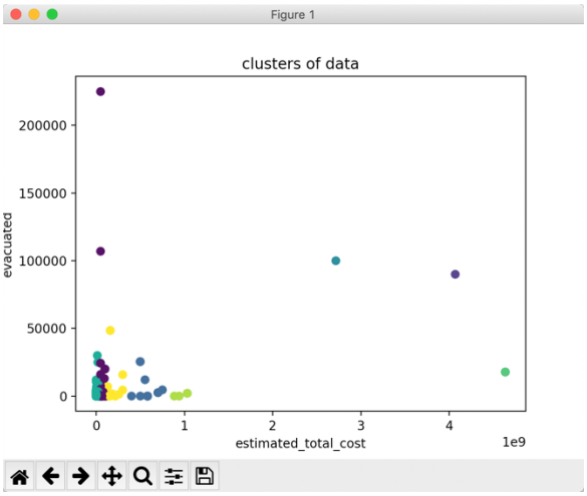

Figure 3: Cluster of people evacuated and the cost

2 Related Work

Similar works regarding disasters analysis have been done by many scholars and students across countries. Project (Choi et al., 2018) describes methods to predict the

estimation of damage caused by heavy rain. Some authors apply state-of-the-art methods to seek better performance. Authors (Mosavi et al., 2018) uses neural networks and MLP in flood prediction.

3 Data

Our input data set is from multiple sources. The first set is provided and recorded by Canadian government [2] which is CanadianDisasterDatabase.csv. The second data set[3] is a .txt file contains most Canadian cities and each city's population record. These two data are combined together as the input data set. As Table 2 shows, the data set is badly imbalanced. The label 'natural' itself takes more than 84% of the entire data set. Re-sampling is needed in order to get a rather useful result. There are 1584 rows of instances.

1.	Natural	0.8459
2.	Technology	0.1332
3.	Conflict	0.017676
4.	Unknown	0.003156

Table 2: Data label distribution

3.1 Missing values

There are a lot of missing values in the original data set. Many values such as 'total estimated cost', 'fatalities', 'population', etc. have empty record. Given the fact that there are only 1584 rows of instances recorded by the Canadian government, the set is very small to do an analysis. Taking the size into consideration, the missing values are filled by the mean value of each category. A simple drop of the row contains missing values would result in an even smaller data set. Therefore, dropping rows is not taken into consideration.

3.2 Data cleaning

The data preparing part follows the ETL process. The data is extracted from multiple resources and joined together as one for further analysis. The date set overall is very organized; however, several columns contains data that are very dirty for analysis. We would like to extract locations from the column 'place' and column 'comments'. This tells us the locations where incidents occurred. All other columns require only simple data cleaning technics such as removing special characters and punctuations.

3.3 Visualization

The data provided by Canadian government records incidents happened within Canada. By using tableau[4] and functionalities provided by the platform, we are able to see how some incidents are grouped together to better understand the data set. Figure 4 shows several summaries of the data set in a more compact representation.

From the figure, we can tell that since 1900, there are more and more incidents happening in Canada till the year 2020. The tendency of disasters occurrence is increasing overall. Ontario has the most disasters since 1900 and the number of incidents is over 200 cases. The province that has the greatest number of fatalities is nova scotia.

From Figure 5, it is very clear to see that most frequently occurred disasters are natural disasters and on top of that most natural disasters happened are in Ontario, Quebec, and British Columbia.

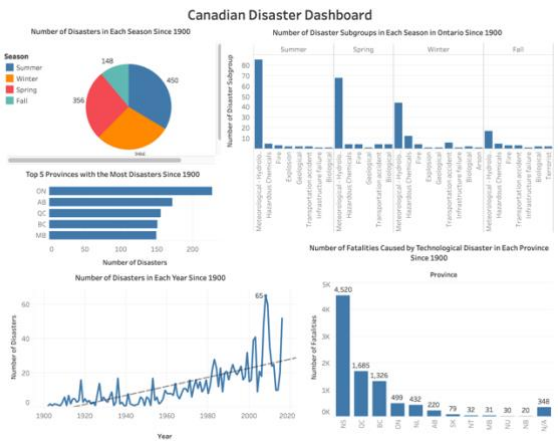

Figure 4: Integrated dashboard

[2] https://www.publicsafety.gc.ca/cnt/rsrcs/cndn-dsstr-dtbs/index-en.aspx

[3] https://worldpopulationreview.com/countries/cities/canada

[4] https://www.tableau.com

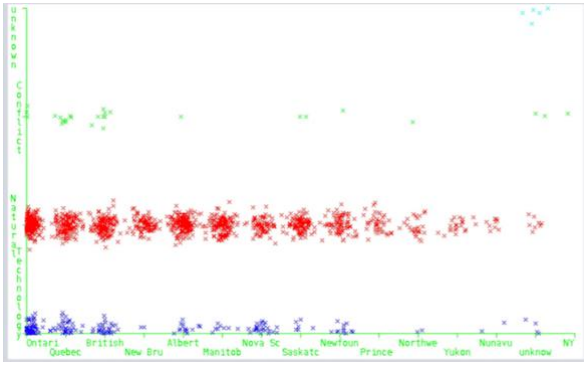

Figure 5: Disaster type clustering

4 Result

4.1 Supervised learning(classification)

The main task of this short paper is to perform a supervised learning to discover the possibility to predict a given instance's disaster type. Table 3 shows our potential features to choose from to form our final feature set. The final feature set is chosen based on how each feature combination contributes to the final testing accuracy. After repeatedly testing the features, the final feature combination is listed in Table 1 above mentioned in the previous section.

Table 3&4 shows the testing accuracy before and after re-sampling. The input data set is small and imbalanced; these are problems affect our training and testing. By using SMOTE[5] technic, we over sample the data set to reach a rather fair result.

Since the original data set is imbalanced, the result before re-sampling seems promising. After over sampling, though it is the only way to generate a rather objective result given the fact that we are unable to expand the data set, the result seems alright. Figure 5 compares several classifiers based on testing accuracies.

City
Population
Province
Season
Fatalities
Evacuated
Estimated total cost

Table 3: potential features

SVM	0.839
GBDT	0.8675
KNN	0.8265
Random forest	0.8422
Ada Boost	0.8328
Decision tree	0.8549
Linear Discriminant	0.8328

Table 4: Classifiers and testing accuracies before re-sampling

SVM	0.4731
GBDT	0.6877
KNN	0.694
Random forest	0.7298
Ada Boost	0.6246
Decision tree	0.5741
Linear Discriminant	0.1199

Table 5: Classifiers and accuracies after over sampling

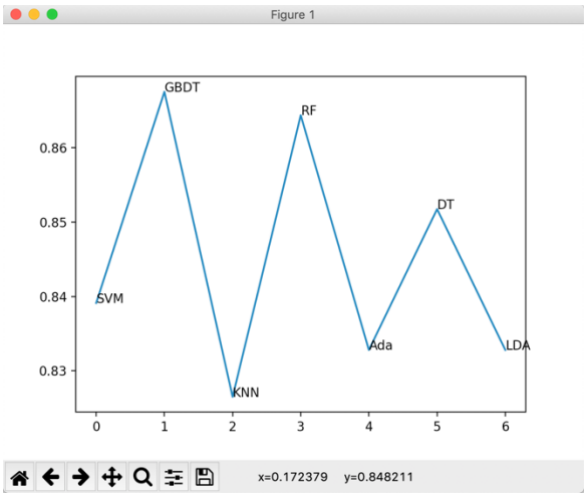

Figure 5: Accuracy line plot

4.2 Unsupervised learning(clustering)

Figure 2&3 shows two possible clusters among many clusters we did which are saved in the data

[5] https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html

folder. The information provided by performing clustering is limited and the two clusters do not provide many valuable insights to our task. We aim to discover groups of similar objects when performing clustering. It is not surprising that clustering does not generate much useful patterns and insights for our task. Our data set is easy to understand and by using 'event group', 'place', 'time' labels, it is quite easy to discover similar groups. Apart from some columns of numeric values, the data set contains a lot of categorical columns for us to discover groups like Figure 5 shows.

5 Conclusion

5.1 Discussion

By using supervised learning and comparing several conventional classifiers, we reach an accuracy of 86.75% using GBDT, and 72.98% using random forest after over sampling. From unsupervised learning, we discover several groups that are composed of similar objects. The result from supervised learning gives us some useful information such that the task is feasible to perform further analysis. Our findings and results are limited to the data set which we collect across the Internet. First, the data set is not large enough to give a very promising and trustworthy outcome. The dimensions recorded by Canadian government does not contribute much useful insight into our overall task. Second, too many missing values are presented in the data set. Dropping rows is not feasible since many rows would be dropped when the data set is small already. These problems limit our ability to discover more useful information using either supervised learning or unsupervised learning.

5.2 Future work

This short paper illustrates that our hypothesis, which properties and recorded disaster details can be used to deduce the type of a disaster, is worth analyzing. The model can be treated as a baseline of the task. We welcome scholars and practitioners to perform further analysis on top of our analysis.

References

Choi, Changhyun, et al. "Development of Heavy Rain. Damage Prediction Model Using Machine Learning Based on Big Data." *Advances in Meteorology*, vol. 2018, 2018, pp. 1–11. *Crossref*, doi:10.1155/2018/5024930.

Deshmukh, M. A., et al. "Importance of Clustering in. Data Mining." *International Journal of Scientific & Engineering Research*, vol. 7, no. 2, 2016, pp. 247–51. *IJSER*, www.ijser.org/researchpaper/Importance-of-Clustering-in-Data-Mining.pdf.

Mosavi, Amir, et al. "Flood Prediction Using Machine. Learning Models: Literature Review." *Water*, vol. 10, no. 11, 2018, p. 1536. *Crossref*, doi:10.3390/w10111536.