Computer Science
School of Informatics, Computing, and Engineering
Indiana University, Bloomington, IN, USA
Datamining B565 Fall 2017

Kaggle Competitions: Spooky Author Identification Statoil/C-CORE Iceberg Classifier Challenge

John Stein[1]*

Executive Summary Briefly explain the kaggle competition and datamining as a solution. Briefly explain each problem, the solution, and results.

Keywords
Keyword1 — Keyword2 — Keyword3

[1] *Computer Science, School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*
***Corresponding author**: jodstein@iu.edu

Contents

1. Introduction

- Briefly, but more completely than in the Executive Summary, explain the kaggle competition – how to get to it, *etc.*.
- Discuss datamining abstractly and how it fits as a solution to the kaggle competition.
- Briefly, but more completely than in the Executive Summary, explain the two problems in two different subsections. [1]

Author Identification

What is the problem to be solved? What is the data? How is goodness quantified? This should not be too technical, but we can say, for example, given three authors $A = \{a_1, a_2, a_3\}$ and selections of their individiually corresponding works $S_{a_1}, S_{a_2}, S_{a_3}$, we are constructing a probability mass function f over A that is applied to a text t written by one of the authors, but unlabled:

$$f(A|t,S) = \{p_{a_1}, p_{a_2}, p_{a_3}\} \tag{1}$$

A text is simply a passage from one of the author's works:

> Once upon a midnight dreary, while I pondered, weak and weary, Over many a quaint and curious volume of forgotten lore—While I nodded, nearly napping, suddenly there came a tapping, As of some one gently rapping, rapping at my chamber door."'Tis some visitor," I muttered, "tapping at my chamber door— Only this and nothing more."

from Edgar Allen Poe's, *The Raven*. You must find out whether it's only prose – and how much of the original structure is maintained. Likely you will not take in text directly – so this must be described.

Statoi/C-CORE Iceberg Classifier Challenge

Use footnotes sparingly, but you must use at least two in this document[1].

[1] A footnote provides some ancillary information that's somewhat misplaced in the text, *i.e.*, the reader can skip it without missing any critical information. On the other hand, a dedicated reader appreciates the extra information. A footnote might clarify information in an informal way. Although possible, putting mathematics in a footnote is a bit odd.

2. Datamining

- What is datamining? Datamining seeks to inform decision making by answering questions about, exploring or discovering relationships in, or providing annotation of data.
- What does it yield? Datamining yields information that is not possible or practical to directly observe, based on data that is practical to observe. This information may be predicted trends in human behavior, perceived relationships or clusters within high-dimension data, or a prediction of future events. In all cases, the yielded information is deemed valuable to someone who is then enabled to make decisions based on that information.
- What are the general steps?
  - The first (and most difficult) step in datamining is developing the problem statement: What decision or action is the datamining effort required to inform or enable, and what must the datamining effort yield in order to satisfy that requirement?
  - The second step is data acquisition. Acquisition of data is non-trivial. It can be difficult, expensive, and in some cases even prohibited or restricted by law. Therefore, the miner must be intentional about pursuing the acquisition of data that is both needed (from the problem statement) and available, whether by manual observation, survey, or automated sensing/recording.
  - The third (and most time consuming) step in datamining is data pre-processing. The data must be analyzed, including identifying unknown, missing, or outlier records, examining attributes and their domains, and visualizing the data for first-order behaviors, trends, and distributions. The data must then be integrated and/or cleaned, including dealing with unknown, missing, or bad data, enriching the data from other sources, and/or transforming the data to work better in the algorithm.
  - The fourth step is to actually mine the data (i.e. answer questions, explore or discover relationships, or annotate). The methodology, assumptions, models, and parameters of how the mining is performed can depend on the objective, the data itself, and the miner themselves.
  - The fifth and final step in datamining is to interpret the output of the mining step and either validate that it meets the stated objective/requirement of the problem statement, or determine that changes to the pre-processing or mining steps are warranted and start again with a modified approach.
- What is clustering *vs.* classification? Classification refers to the task of predicting the label of some unobserved target attribute given a set of observed attributes for some record. The label is generally a single member of a finite list of possible labels. Clustering refers to the task of identifying or discovering relationships among data. In other words, are there properties or measures for which subsets of data can be interpreted as similar or dissimilar in a meaningful way? Clustering aims to discover these 'clumps' or similar data as well as the properties or measurements from which the similarity can be determined. It is useful to think of Classification as a supervised learning task because it is notionally possible to produce a training set with labeled records, and Clustering as an unsupervised learning task because the number and meaning of the discovered clusters is not known apriori (if they were, they could be labeled and it would be called Classification).
- What is a loss function? A loss function is a function which attempts to assign a real value to some undesirable property of an intermediate outcome during the mining step. This value (known as cost or loss) may represent error, variance, or bias of a model, clustering, or classifier. The actual value of loss is usually not meaningful. Instead, the goal is generally to minimize the loss over some parameter or variable during an optimization step.

Data Preprossing

- What are the steps, and what challenges does each present? The first step in data preprocessing is data analysis. Data analysis can include examining histograms, bar/pair-plots, frequency tables, etc. The goal is to get familiar with the data attributes and their domains, frequency of values, trends and distributions, and first-order behaviors. From the analysis, the two primary remaining tasks are selection and modification. Selection (i.e. sampling and attribute selection) refers to deciding which subset or records and/or which subset of attributes for each record to use for the mining task. Modification (i.e. aggregation, dimensionality reduction, discretization, and transformation) refers to changing existing, creating new, or transforming, attributes and their values.
  - Sampling refers to choosing a subset of the total data from which to develop the analysis technique - often because working on the entire dataset would result in more resource investment than it would improve the performance of the technique. The challenge with sampling is that, especially for high-dimensionality data, the data data can quickly become sparse, resulting in a greater potential for over-fitting due to the small data size and high dimensionality (degrees of freedom).
  - Attribute Selection refers to selecting which attributes to include in the analysis and which to exclude, based on some preliminary analysis. Manual attribute selection (or de-selection) is challeng-

ing because humans can only perceive date in two, three, or perhaps four dimensions. One can only reasonably expect to act on the most obvious of attribute properties without impacting the analysis in unexpected ways.

  - Aggregation refers to manually grouping several attributes or attribute values together, preferably while preserving some higher-level relationship among those being grouped. Aggregation can be relatively straightforward in some cases, but in other cases it can be very difficult to group attributes or values in a way that recognizes the essential common property of its members.
  - Dimensionality Reduction refers to the creation of fewer new attributes which are combinations of the many old attributes. Two common methods include Principal Component Analysis (PCA) and Neural Networks. The primary challenge with dimensionality reduction is that the meaning of the original attributes are all but lost as the data are mapped to the new, reduced feature space.
  - Discretization refers to the assignment of values from a continous-valued attributes into one of finite-many bins.
  - Transformation refers to the mapping of old attribute values to new attribute values - generally reversible.

  Finally, data with unknown, erroneous, or missing values must be handled. Some of these handling techniques may be implemented in the mining algorithm(s) themselves, but some may be performed prior to the mining step. Unknown/missing data handling is more challenging for some algorithms (i.e. regression) than it is for others (i.e. decision trees). Some algorithms need each record to have good data, or they will simply not work. Other algorithms can be made to handle unknown/missing data in a reliable way and achieve good performance. Still other algorithms can operate on records having missing data with no extra effort. The primary challenge, especially for high-dimensionality data, is that the impact of the handling technique can be different for every attribute, and must be understood and chosen carefully.

- What is the general load (time, space, \$) for preprocessing? Computationally, preprocessing may consume $O(n)$ or $O(nlogm)$ time where n is the number of records to process and m may be search steps if matching, inclusion, or sorting operations are required. Preprocessing may consume anywhere from constant space to $O(n)$ space, if the records are processed one at a time and stored back into their original location, or copies are made. Practically speaking, preprocessing can consume upwards of 80% of the labor hours, and therefore 80% of the funding, of the overall datamining effort. That is because it is the most human-intensive step of the process. Automation of the preprocessing step, although possible, is of limited benefit if the datamining problems vary significantly in size, scope, data representation, purpose, etc. The ability to predict the preprocessing impacts of data across problem spaces is indeed an analytics problem unto itself.

Mining, Interpretation, and Action

- Briefly discuss the top 10 algorithms.
- Does datamining tell us what to do?
- What are some new types of problems in datamining?

3. Author Idenfication: Full Problem Description

- Define problem
- Formally describe problem–inputs, outputs, training and testing method

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

$$\cos^3\theta = \frac{1}{4}\cos\theta + \frac{3}{4}\cos 3\theta \tag{2}$$

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Data Analysis

- Describe the data in full detail–from its raw form to the transformation
- Provide summary statistics and relationships

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Methods

- Discuss the algorithm you've chosen, *e.g.*, why you chose it
- Provide some background material on your method that shows you are well-acquianted with it
- Challenges to your method
- What software and hardware did you use, packages, *etc.*
- Final structure of data after preprocessing
- Present training and testing as some combination of text and visualizations

Algorithm 1 shows an extended k-means.

Algorithm 1 This is a caption for the algorithm. k-means* over Δ

```
INPUT data $\Delta$, blocks k, distance $\mathbf{d} : \Delta^2 \to \Re_{\geq 0}$
OUTPUT centroids $C_1, \ldots, C_l$
%% assume that a centroid is a pair $(v, X)$
%% $v \in \Re^m$ and (a possibly empty) $X \subseteq \Delta$
%% heap $H \subseteq \Delta$
randomly construct k centroids $\mathsf{C}^0 = \{C_1^0, C_2^0, \ldots, C_k^0\}$
$i \leftarrow 0$
%% $\Delta_{HE}$ represents HE data, $\Delta = \Delta_{HE} + \Delta_{LE}$
$\Delta_{HE} \leftarrow \Delta$
repeat
    for $\mathbf{x} \in \Delta_{HE}$ do
        for $C_j^i \in \mathsf{C}^\mathsf{i}$ do
            %% assign data to centroid/heap that is nearest
            %% $\sigma \Rightarrow d$
            $C_j^i.H.insert(\mathbf{x}, d)$, where $\min\{\mathbf{d}(\mathbf{x}, C_j^i.v)\}$
        end for
    end for
    $\Delta' \leftarrow \emptyset$
    for $C_j^i \in \mathsf{C}^\mathsf{i}$ do
        %% recalculate centroid as average of over $C.H$
        $C_j^{i+1}.v \leftarrow \sum_{\mathbf{x} \in C_j^i.X} (\mathbf{x}/|C_j^i.X|)$
        $\Delta' \leftarrow C_j^i.H.flush(\sigma)$
        $\mathsf{C}^{\mathsf{i}+1} \overset{\cup}{\leftarrow} \{C_j^{i+1}\}$
    end for
    $i \leftarrow i + 1$
    $\Delta_{HE} \leftarrow \Delta'$
until threshold on $\mathsf{C}^{\mathsf{i}-1}$
```

Results

- Dispassionately describe your results both quantified and qualified
- Do you deem this successful
- What do the results suggest
- What were challenges

Remember, we're interested in the journey, so simply because an approach failed doesn't mean failure if you discuss the failure!

Summary and Future Work

- Briefly summarize project and outcome
- What would you do differently in the future?

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

4. Iceberg: Full Problem Description

Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

We used 10-fold...

Table 1. Training and Test Results

Name		
Data	Description	*Sum*
1	"Hello..."	7.5
2	"Goodbye..."	2

Citations and Subsubsection

Word Definition

Concept Explanation

Idea Text

Acknowledgments

Put people, Grants, *etc.* that helped contribute to the success and completion of the work. Be generous.

References

[1] Mark Twain. *Huckleberry Finn*.