Clustering Heterogeneous Semi-Structured Social Science Datasets

D.B. Skillicorn\(^1\) and C. Leuprecht\(^2\)

\(^1\) School of Computing, Queen’s University
Kingston, Canada
skill@cs.queensu.ca

\(^2\) Political Science and Economics, Royal Military College of Canada
Kingston, Canada
leuprecht-c@rmc.ca

Abstract

Social scientists have begun to collect large datasets that are heterogeneous and semi-structured, but the ability to analyze such data has lagged behind its collection. We design a process to map such datasets to a numerical form, apply singular value decomposition clustering, and explore the impact of individual attributes or fields by overlaying visualizations of the clusters. This provides a new path for understanding such datasets, which we illustrate with three real-world examples: the Global Terrorism Database, details of every terrorist attack since 1970; a Chicago police dataset, details of every drug-related incident over a period of approximately a month; and a dataset describing members of a Hezbollah crime/terror network within the U.S.

Keywords: clustering, hashing, terrorism, crime, Global Terrorism Database, Chicago policing, Hezbollah

1 Introduction

There are an increasing number of social-science datasets of substantial size (thousands of records and hundreds of attributes). Such datasets typically contain highly heterogeneous information in a semi-structured form (that is, defined records, attributes and fields, but free-form information within them). Excel spreadsheets are often the format of choice.

Social scientists who want to analyze such data face two major difficulties. First, the contents of the fields can be extremely heterogeneous: dates or time intervals, explicitly numerical data (e.g. ages), coded numeric data (e.g. 0 or 1 for presence or absence), categorical data (e.g. country), and short textual fields (e.g. brief descriptions). Standard statistical machinery, such as significance testing, is difficult or impossible to apply to such data. Often entries are missing,
with the corresponding locations left empty. This may mean that the corresponding value does not apply, or that it was not collected, and the dataset size usually precludes manual checking.

Second, the data is large enough that all but the most simple analysis requires tools and techniques in which most social scientists are not trained. As a result there are a number of large datasets, to which knowledge-discovery techniques could be applied, that are actually used only for querying.

The contribution of this paper is the development of techniques that can be used to convert such large, heterogeneous datasets into a usable numerical form, the application of clustering to the resulting data, and forms of analysis by overlaying that make it possible to interpret the meaning of clusters in terms of the domain from which the data comes.

2 Technical Approach

We use the following analysis pipeline: **Hash** each entry to produce a true numeric value, replacing the original values, of whatever form, with numeric values that correspond in some useful way. For consistency, we apply the same algorithm for all input attribute types: we treat each entry as a string, convert each character to a floating-point value and compute the mean of these values. Thus the length of the input representation does not have much impact.

The effect of this hashing function is to capture similarity and difference, rather than magnitude. This does not necessarily preserve the ordering – large magnitudes may be hashed to small numbers, and *vice versa*. Normalize each column of the dataset by converting the values in each column to z-scores, (subtracting the column mean from each entry, and then dividing each entry by the column standard deviation). Since the magnitudes of the hashed entries for each attribute are unpredictable, this maps the values of each attribute to roughly the same range, enabling meaningful comparisons among attributes.

**Cluster** the records using singular value decomposition [3]. If the numeric dataset has \( n \) records and \( m \) attributes, then each record has a natural representation as a point in \( m \)-dimensional space. Singular value decomposition can be regarded as a transformation of this space in a way that reveals its greatest variation. This decomposition can be truncated into some smaller number of dimensions, \( k < m \), with the loss of as little structure as possible. If \( k = 2,3 \) then each of the \( n \) records can be represented as a point in 2- or 3-dimensional space, and directly visualized.

Truncation reduces the effect of any one attribute on the geometric embedding, as long as many attributes are broadly correlated with many other attributes. This property of the SVD means that inconsistencies introduced by the hashing tend not to distort the results as much as might, at first glance, be expected.

**Overlay** the clustered points by colors derived from the values, in each record, of one of the attributes. Repeat for each attribute in turn. If there is an association between point positions and attribute values, this will be visible as a pattern in the overlaid rendering.

3 Datasets

We use three datasets to illustrate the power of the technique. The first is the Global Terrorism Database [4], which contains records describing every terrorist attack between 1970 and 2011 (in the version we use). For each attack, data concerning timing, geographical location, form of attack, form of target, demographics of attackers, motives and claims, weapons used, casualties, and costs are included when appropriate. The dataset contains 104,687 records and 117
attributes. It has been studied by many groups: the paper by Godwin et al. [2] uses a form of visualization; the paper by Enders et al. [1] applies a kind of calibration to adjust for known issues (largely fixed in recent releases) and to examine temporal patterns of frequencies. The paper by Shafiq et al. [5] attempts prediction from the GTD data, but uses only a few (and the easiest to code) attributes.

The second dataset is drawn from a collection of Chicago police reports (https://data.cityofchicago.org/). From this we select only reports of drug incidents. Data concerning time, location, descriptions, and coordinates are included. The dataset contains 35,479 records and 13 attributes.

The third dataset is a collection of records describing 181 members of two connected Hezbollah networks in the U.S.. For each member, 59 attributes were collected. These range from demographics such as place of birth and marital history, to activities associated with the group’s criminal, terrorist, and smuggling activities.

4 Results

The most striking property of Figure 1a, which represents all terrorist attacks over more than 40 years, is that there are clusters (rather than one big blob or points scattered across the entire space). There are 8 main clusters (labelled with the letters A–H), and these clusters are themselves organized into a “double hinge” structure.

In other figures, the variation with respect to some interesting attributes are shown as color variations. Figure 1b shows the variation associated with an attribute about the outcomes for hostages in a kidnapping. Other attributes associated with kidnapping and hostage-taking, associated ransom attributes, and duration of incident attributes, show this same variation. Figure 1c shows the variation associated with whether or not the incident was claimed. This direction is also associated with the number of deaths and wounded among the terrorists, presumably because attribution can be inferred from dead and wounded participants. Figure 1d shows the variation associated with target type and number of targets. These attributes are also strongly correlated with terrorist nationalities, so we can conclude that targeting and terrorist nationality are correlated. This may reflect the geography of opportunity, but may capture something deeper.

We can now assign meanings to the clusters in Figure 1a. Clusters A and B are hostage-taking, while C–H are more conventional attacks. Clusters A, C, E, and G are differentiated from B, D, F, and H by properties of the terrorist groups responsible, primarily whether they were degraded by the attack. Clusters A, B, C, D are differentiated from E and F and also from
Figure 2: Overlays of interesting attributes for the Chicago dataset

G and H by the type of target and number of targets – the top layer are single target events, the second layer (E and F) are two target events, and the third layer are three target events. (In fact, part of the diffuse structure of G and H is caused by the presence of a few four target attacks.)

Figure 2a shows the clustering of Chicago drug incident records. Again, the clustering is strong than might have been expected from such disparate records.

Figure 2b shows the overlay of the beat in which each incident occurred. It is clear that this attribute has little to do with the macroscopic clustering, but that it (and related geography) explain some of the microscopic structure within each cluster. The next four attributes: whether or not the incident was domestic (Figure 2c), whether or not an arrest resulted (Figure 2d), the FBI code associated with the incident (Figure 2e), and the primary description (Figure 2f) show that the macroscopic clustering has a Y shape. Domestic versus non-domestic variation goes from lower-right to upper-left; arrests vary from lower-left to upper-right, and both FBI codes and primary code for the incident vary from left to right, with greater variation in the FBI incident codes. Figure 2g shows that time of day does not correlate with macroscopic clustering but does with the internal structure of every cluster in a way exactly orthogonal to the beat attribute. This hints that more cluster might be detectable in a higher-dimensional clustering.

Figure 3a shows the clustering of individuals in the Hezbollah dataset. Again, there is significant structure. Figures 3b, 3c, 3d, and 3e show common vertical variation shows connections between the kinds of criminal acts carried out by some members, and their demographics related to origin and apparent occupation. Figures 3f and 3g show common horizontal variation related to marriages. One major activity of this group was creating sham marriages as a way to get residency in the U.S. and several members were married multiple times in attempts to make this happen. Unsurprisingly this is correlated with education, since presumably those with better education had other paths to residency status.
5 Discussion

Large, semi-structured, wildly heterogeneous datasets are common in the social sciences. These properties make analytics difficult, and many are treated simply as databases to be queried (“How many cases like this are present?”). However, there is much useful data implicit in such datasets. We have shown that hashing, despite its simplistic assumptions, can convert such data to more readily analyzable form, especially when it is combined with a clustering technique that is insensitive to moderate inconsistencies. The resulting clusterings, as we have shown in the examples, reveal much more implicit structure than might be expected. The use of overlaying makes it possible to infer what aspects of the original data explain the visible variation and clusters. This opens up a large class of data, currently mostly ignored, to sophisticated analysis.

Acknowledgement. The authors would like to thank Hillary Ryde for collecting the Hezbollah network data used here.

References


