Comparing Apples and Oranges: Methods for Comparing the Incomparable

David J. Marchette*    Carey E. Priebe†    Younger Park‡

Abstract

In pattern recognition one is often presented with the difficult problem of comparing observations taken under very different conditions, and sometimes with very different sensors. In this paper we consider a methodology for mapping observations taken from different sensors into a space in which they can be compared. This mapping makes use of a pairing of observations, or sets of observations, from the different sensors, and computing distances between the data and these sets, or bridge points. The distances then form a nonlinear projection of the observations into a space in which all can be compared. In situations in which the relationship between the observations and their respective bridge points are similar under the two sensors, this projection provides a useful space in which to perform pattern recognition or other data analysis. We illustrate the approach with two examples, one using hyperspectral data and the other a text processing application.

1 Introduction

In pattern recognition one is often presented with the difficult problem of comparing observations taken under very different conditions, and sometimes with very different sensors. A classic example of this is biometric identification of non-cooperative subjects using, say, faces or voices. It is often the case that the data used to build such a detection and classification system (for example passport photos) will be very different from those used in the production system. For another example, consider text document processing. In many text processing applications, one often cannot directly compare documents. This is because the main features used for pattern recognition are word statistics, which can only be (easily) compared for those words that are contained in both documents. Thus, methods are needed to bridge this gap, and this is the topic

*Naval Surface Warfare Center, Code B10, Dahlgren, VA 22448. marchettej@snc.navy.mil
†Johns Hopkins University, Department of Applied Mathematics and Statistics, Baltimore, MD 21218. cep@jhu.edu
‡Johns Hopkins University, Center for Imaging Science, Baltimore, MD 21218. youngser@jhu.edu
of this paper. We seek mappings from sensor space(s) to \( \mathbb{R}^d \) which allow the comparison of the observations from different sensors.

The main idea comes from initial work by Cowen and Pribe [1997] and Marchette and Pribe [2003]. In these papers the authors consider distances to observations as features. We will use these ideas in two ways. In the Hyper-spectral example of Section 2 we will identify observations from two sensors, in effect treating the observations from the different sensors as being from the same object. By utilizing the distances to these selected observations we map the data from two distinct sensor spaces to a space in which the observations can be compared. In Section 3 we look at a text processing problem, in which some of the documents cannot be directly compared. By considering distances to groups of documents, we produce a mapping to a space in which all the documents are comparable, and the classes of interest separate relatively well, as we will see.

2 Incomparable Sensors

Hyperspectral imagers are capable of simultaneously measuring light in a large number of wavelength bands. The HyMap sensor considered in this section collects 126 bands.

We will consider data extracted from a HyMap image corresponding to pixls from seven landcover classes: runway (asphalt), water, swamp, oak, pine, scrub (brush and miscellaneous small trees) and grass. A total of 700 observations were used in this experiment, 100 observations from each class.

The data are split into two subsets, \( \mathcal{A} \) and \( \mathcal{B} \). For \( \mathcal{A} \) only the first half of the spectrum is retained; this corresponds to a sensor that only collects bands 1–63. For subset \( \mathcal{B} \) only the last half of the spectrum is retained; this corresponds to a sensor that only collects bands 64–126.

The two subsets are now non-comparable, without further processing. We produce a two dimensional projection as follows: from each set we take one pine pixel and one oak pixel. Denoting the pine pixel from sensor \( \mathcal{A} \) as \( P_A \) and the oak pixel as \( O_A \), and similarly for sensor \( \mathcal{B} \), the new features are then the distances to the two pixels: for \( x \in \mathcal{A} \), the coordinates of \( x \) are then

\[
(d(x, P_A), d(x, O_A)),
\]

and similarly for \( x \in \mathcal{B} \). Putting these together, Figure 1 shows the plot of the projection. Note that we have recaptured the classes to a large extent: water is close to water, grass is close grass, etc., even when considering observations from different sensors.

The pixls selected in this example were arbitrary. These points are referred to as “bridge points”, and we are computing the distance between each observation and it’s associated bridge point.

We can take this idea to an extreme by computing the distances to the subsets containing all the pine and all the oak pixels; setting the distance between
Figure 1: Projection of the two sensors (indicated by symbol) onto the axes corresponding to the distances to the single pine and oak pixels.

For a point $x$ and a set $S$ as $d(x, S) = \min_{s \in S} d(x, s)$, we have:

$$\left( d(x, \bigcup_{A} P_A), d(x, \bigcup_{A} O_A) \right),$$

and similarly for $x \in B$. This is depicted in Figure 2. Note that the vegetation stay together (for the most part), while the non-vegetation landcover classes are separated. This indicates that the vegetation observations have a similar relationship to oak and pine in both sensors, while runway and water have a different relationship under the two different sensors.

This is more clearly shown in Figures 3 and 4. In Figure 3 we see the result of projecting onto the distances to the runway and water classes. Note that the for the other classes the relationships are perpendicular: the relationship of swamp to water in sensor 1 is opposite to that in sensor 2. Figure 4 shows the data
Figure 2: Projection of the two sensors (indicated by symbol) onto the axes corresponding to the distances to the union of the pine and oak pixels, respectively.

Figure 3: Projection of the two sensors (indicated by symbol) onto the axes corresponding to the distances to the union of the runway and water pixels, respectively.
from the two sensors as projected onto their first two principal components. The effect is as if the water pixels were held fixed and the other pixels were rotated by 90 degrees. This explains the “X” pattern in Figure 3. The information which sensor 1 collects about water is in a sense orthogonal to that collected by sensor 2, relative to the other classes.

Consider the simulated data depicted in Figure 5. Here we have two fictional sensors that each make ten observations. These are depicted in the top plots, with a single observation chosen as the distinguished point. The distance from each point to the distinguished point is computed and plotted in the middle plot. Note that because the relationship between the observations and distinguished point in one sensor is approximately the same as those in the other sensor, the mapping to one dimension is approximately the same.

In effect, this procedure is identifying the two distinguished points. This results in a folding of the two sensors together; in the bottom of Figure 5 is a notional depiction of this. The two points are identified via the dotted line, resulting of the closing of the “fan” (or “butterflies wings” if you prefer), resulting in observations which have the same relationship (distance) to the distinguished point being identified.

We can take this idea to an extreme as follows. Construct a seven dimensional data set where for each observation the seven features are the distances to the seven classes. Thus for a given observation $x$ we have $(d(x, C_1), d(x, C_2), \ldots, d(x, C_7))$, where $C_i$ denotes the observations from class $i$ for the appropriate sensor. To visualize this, we have projected these observations to two dimensions, via principal components, and this is shown in Figure 6.

There are several interesting things in this picture. The classes associated with vegetation tend to group together, regardless of sensor, and they are also somewhat (although not perfectly) separated from each other (with the exception of Oak and Pine, which are always difficult to separate). The water and runway pixels associated with the different sensors do not overlap, but they are disjoint from the vegetation. Most importantly, it would be easy enough to construct a classifier on these data that would be sensor independent.

3 Text Documents

3.1 Science News Corpus

Our Science News (SN) corpus consists of 1160 documents from 1994-2002. The corpus was obtained from the SN website on December 19, 2002. Two parsers were used for processing the html-based SN articles; one for the years 1994-1999 and one for the years 2000-2002. The use of two parsers was necessitated by the use of two different style sheets on the website. The translation from the html-based articles to the straight text versions of the articles was not without difficulty. A validation effort was undertaken which indicated that the corpus as parsed is satisfactory for our purposes.

On the SN website http://www.sciencenews.org/ a set of 33 “Editor's
Figure 4: Principal components projections for the two data sets in the hyperspectral example. Figure (a) shows the data in $\mathcal{A}$ and (b) depicts the data in the set $\mathcal{B}$. 
Figure 5: The top two figures show ten observations each taken from two different fictional sensors. The solid dots correspond to the distinguished points, and the lines connecting the other points to the distinguished points illustrate the distances calculated and depicted in the middle figure. In the bottom figure the observations from the two sensors are combined to depict the bridging effect of identifying the two distinguished points.
Choice” articles were categorized into eight categories. We used the same eight categories in our categorization of the remaining articles. A breakdown of the SN corpus by category is contained in Table 1. (Some documents have been assigned to more than one category; 1047 documents are assigned to one and only one category, but this decision was fairly arbitrary.) A single individual performed the categorization, and this was done ad hoc, without specific guidelines as to how to categorize the documents.

With 1,303,243 total words in the corpus — 41,328 distinct words — the documents are approximately 1000 words long on average.

3.2 Text Features

<table>
<thead>
<tr>
<th>Topic Area</th>
<th>Number of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anthropology and Archeology</td>
<td>61</td>
</tr>
<tr>
<td>Astronomy and Space Sciences</td>
<td>131</td>
</tr>
<tr>
<td>Behavior</td>
<td>93</td>
</tr>
<tr>
<td>Earth and Environmental Sciences</td>
<td>170</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>254</td>
</tr>
<tr>
<td>Mathematics and Computers</td>
<td>90</td>
</tr>
<tr>
<td>Medical Sciences</td>
<td>316</td>
</tr>
<tr>
<td>Physical Sciences and Technology</td>
<td>162</td>
</tr>
</tbody>
</table>

Table 1: Subcorpora.

extraction. Given a corpus $\mathcal{D}$ of documents, a collection of “admissible words” are extracted. This lexicon $\mathcal{L}$ corresponds to those words not on a stopper list of common words. All the words have been stemmed to eliminate prefixes and such. For each document $d \in \mathcal{D}$, each word $w \in d \cap \mathcal{L}$ is given three features: the number of times $w$ appears in the document, and two entropy-based scalers. Let $c_{dw}$ be the number of times word $w$ appears in document $d$, and $N = \sum_{i,j} c_{ij}$, the total number of words from the lexicon $\mathcal{L}$ appearing in the corpus, counting repetitions. Set $f_{ij} = c_{ij}/N$. The mutual information between document $d$ and word $w$ is computed as

$$m_{dw} = \log \frac{f_{dw}}{\sum_j f_{jw} \sum_i f_{di}}. \quad (1)$$

This is discounted to eliminate infrequent words, resulting in the feature:

$$F_{dw} = \frac{c_{dw} \min(\sum_j c_{jw}, \sum_i c_{di})}{(c_{dw} + 1)(\min(\sum_j c_{jw}, \sum_i c_{di}) + 1)} m_{dw}. \quad (2)$$

Words whose features fall below a threshold ($\tau$) are dropped (in document $d$, only words $w$ for which $F_{dw} > \tau$ are kept). Only those documents which retain at least one word after this reduction are kept. In our experiments we use a threshold value of $\tau = 0$.

Note that this feature vector is corpus dependent. Since the components $\sum_j f_{jw}$ and $\sum_i c_{jw}$ are summed over the documents within the corpus, these provide a corpus dependence. Also $N$, the total number of words in the lexicon which appear in the corpus is also clearly corpus dependent. This dependence means that reducing the corpus will change the inter-document relationships.

Let $T(\mathcal{D})$ denote the total number of distinct words in the corpus $\mathcal{D}$ and $s_i(a, b) = I\{c_{ai} \cdot c_{bi} > 0\} = I\{c_{ai} > 0\}I\{c_{bi} > 0\}$ indicate whether the $i$th word is shared between documents $a$ and $b$, so that $s(a, b) = \sum_{i=1}^{T(\mathcal{D})} s_i(a, b)$ is the number of shared words between documents $a$ and $b$. 

9
For $a, b \in \mathcal{D}$ we consider the cosine-dissimilarity

$$
\rho(a, b) = 1 - \frac{a \cdot b}{|a||b|} \\
= 1 - \frac{\sum_{i=1}^{T(\mathcal{D})} s_i(a, b) F_a F_b}{\sqrt{(\sum F_a^2)(\sum F_b^2)}}.
$$

(3)

Since the denominator is always nonzero, this is always finite. However, there are two ways the numerator of Equation (3) can be zero: if the vectors are perpendicular or the documents do not share words. We wish to distinguish these cases, and so we set $\rho(a, b)$ to an arbitrary number larger than 2 in this latter case. With this modification, $\rho$ is a dissimilarity and not a distance. Note that $\rho$ is fundamentally corpus dependent, through both the $s_i$ (the words retained) and the features which themselves depend on the corpus. When necessary, we will denote this dependence via a subscript: $\rho_D$.

We could eliminate the “shared word” function $s_i(a, b)$ from the notation through the convention that any word that does not appear in the document obtains a feature value of 0. We prefer not to. Making this explicit allows us to differentiate between a case where the word simply has a mutual information of 0 and where the word does not appear in the document. For document pairs $d_1, d_2$ that do not share words, we say the pair are not comparable, and set $\rho_D(d_1, d_2) = \infty$.

There are two problems with performing analysis on the text corpus. First, the dissimilarity is a true dissimilarity, and so metric methods, such as principal components, cannot be directly applied. Second, the dissimilarity $\rho_D$ can (and does, on the Science News corpus) produce infinite distances. In order to process these data we need to somehow transform them into a metric space. We do this in two steps. First, we set any infinite distances to twice the maximum of the finite distances: $2 \cdot \max_{\rho(d_i, d_j) < \infty} \{\rho(d_i, d_j)\}$. Then we perform multidimensional scaling to produce observations in $\mathbb{R}^d$ whose interpoint distance matrix most closely matches the dissimilarity matrix. With these observations, we can then perform dimensionality reduction (via principal components, or any other metric technique), and any other processing we wish.

The principal components plot for the Science News data is shown in Figure 7. Note that there are three main groups to these data: a (mostly) Medicine wing, a (mostly) Astronomy wing, and a group consisting of the other classes. Note further that the Physical Sciences are closely related to the Astronomy, and the Life Sciences and Behavior are closely related to the Medicine class. These relationships are as we would expect.

Now let us project the observations in much the same way we did for the Hyperspectral data in Figure 6. We construct eight dimensional data, where each column corresponds to the distance to the associated class. Since in these data all observations can be compared with at least one member of each class, we can do this even though some pairs of documents are incomparable. The
Figure 7: Principal components projections for the Science News corpus.

The result of applying principal components to these eight dimensional data is shown in Figure 8.

In Figure 8, we see that the different classes separate very nicely (with the exception of Earth Sciences and Physical Sciences). Compare this with Figure 7. Thus, using the relationships between the classes, we have found a (nonlinear) projection to two dimensions that does a good job of separating most of the classes. Further, the documents that were previously incomparable (due to having no words in common) become comparable in this projection.

A final note on this example. A pairs plot of the eight dimensional data is not terribly illuminating. Any of the single pairs of projections has all classes highly overlapped, in fact quite a bit more overlapped than in Figure 7. Thus, considering all the projections at once is necessary to discover the clustering.
Figure 8: Principal components projection applied to the distances for the Science news corpus. The bottom figure depicts the observations in the upper right corner of the top figure.
4 Discussion

Comparing output from (apparently) in comparable sensors requires some method for mapping the sensor output into a space in which they can be compared. This mapping cannot be arbitrary, but must retain some of the relationships between objects in the original sensor spaces. We have discussed one approach to this, involving the projection to interpoint or set distances. The assumption being that the objects in the sensor space are in roughly the same relationship to each other, and all that is required is a mapping that respects these relationships.

If it is known that each observation in one sensor space corresponds to an observation in the other, the approach of “bridging” by identifying individual points seems natural. Even when the observations are simply from the same class, as was the case in the Hyperspectral example with Oak and Pine, we observed that the resultant mapping is of value.

The actual features in the sensor spaces have been ignored, beyond the fact of their use to compute the distances. Since we are looking at only distances between subsets of points, this means that internal relationships between points may be lost. Methods for allowing the incorporation of this internal information is an area of further research.

The observations used in the hyperspectral example were chosen arbitrarily. It would be preferable to choose these to be “optimal” in some sense. To this end, one might consider the following approach. Let $T$ be any statistical test of goodness-of-fit. So, given sets of observations $Z_1, Z_2$, $T(Z_1, Z_2)$ is large if the observations are from the same distribution (as measured by $T$), small otherwise. Given observations $X$ from sensor 1 and $Y$ from sensor 2, set $d(X, x)$ to be the vector of distances from $x$ to each observation in $X$. Similarly with $Y$. We choose the points $x \in X$ and $y \in Y$ to identify via:

$$ (x, y) = \arg \max_{(x, y)} \{ T(d(X, x), d(Y, y)) \}. $$

Thus, we choose the observations defining the mapping by whether they map the data equally (in distribution, as measured by $T$). In a classification problem, $T$ might be a classifier instead of a goodness-of-fit statistic. This is an area of further research as well. It may be relatively simple to extend this approach to multivariate projections (selecting multiple pairs of projections), however extending it to sets (in which each pair becomes a pair of sets) is difficult, and combinatorially intractable. The work of Marchette and Priebe [2003], while in an entirely different context, may hint at some ideas for approaching this more difficult problem.

References


Acknowledgement: Sponsored by the Defense Advanced Research Projects Agency (DARPA) under “Novel Mathematical and Computational Approaches to Exploitation of Massive, Non-physical Data”, ARPA Order No. P246, Program Code 3E20. Issued by DARPA/CMO under Contract No. MDA972-03-C-0014 to AlgoTek, Inc. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either explicitly or implicitly, of DARPA or the U.S. Government. Approved for Public Release, Distribution Unlimited.