CLUSTER IDENTIFICATION FOR SPEAKER-ENVIRONMENT TRACKING

J.T. Wickramaratna, P.C. Woodland

{jtw28, pcw}@eng.cam.ac.uk
Cambridge University Engineering Department, Cambridge CB2 1PZ, U.K.

ABSTRACT
Cluster Identification is introduced as the process of jointly evaluating clustering and labelling schemes for cluster-labelling scheme selection. Normalized Rand and BBN metrics for comparing clustering performances across varied clustering and labelling schemes are presented. The merits of the metrics are evaluated and applied for speaker-environment tracking in Broadcast News.

1. INTRODUCTION
There is a rapid increase in the volumes of recorded speech that require automatic processing. It often becomes necessary to be able to track a speaker through varied noise environments, for applications such as speaker tracking through a telephone archive, and retrieving his speech from a Broadcast News database. It is also useful to successfully group together speakers in varied noise conditions for speaker and environmental adaptation. In such applications where clustering is utilized, it is necessary to be able to evaluate what kind of clustering is carried out by a particular clustering scheme on a particular kind of data. This allows us to select the clustering scheme that provides a suitable clustering, and shows which attributes of the data affect the clusters most. This can be evaluated for a clustering scheme by labelling the clustering units according to varied possible attributes - such as speaker, gender, or speaker and noise condition for a speech database - and evaluating the labelling schemes according to some scoring criteria. This problem of cluster-labelling scheme evaluation is investigated for speaker-environment tracking in US Broadcast News.

In section 2 the problem of evaluating clustering schemes with respect to varied labelling schemes is formulated. In section 3 the metrics of evaluation are introduced and normalized for the evaluation of varied clustering and labelling schemes. In section 4 the merits of the metrics are discussed and experimental results are presented for speaker-environment tracking.

2. THE CLUSTERING PROBLEM
The clustering units are an $S$ number of segments each with one of $N (1 \leq N \leq S)$ labels associated with it. The segments are distributed amongst the $N$ labels according to some distribution and form the segment-label distribution vector $I = (l_1, l_2, \ldots, l_N)$ where $l_i$ is the number of segments with label $L_i$. Given a fixed segment-label distribution vector, the clustering problem is one of assigning the segments to exactly $N$ clusters such that all of the segments labelled $L_i$ are in a single pure cluster labelled $C_i$. This ideal clustering can be represented by a diagonal matrix $M_0 = \text{diag}\{(l_1, l_2, \ldots, l_N)\}$

However, a clustering algorithm may produce results that differ from the ideal in two ways. It may produce a number $K$ of clusters that is not exactly $N$; it may distribute segments with the same label in different clusters and may include segments with different labels in the same cluster. This produces matrices that have more than or less than $N$ columns, and have non-zero off-diagonal elements.

$$
C_1 \quad C_2 \quad \ldots \quad \ldots \quad C_K
$$

$$
L_1 \quad \begin{pmatrix} s_{11} & s_{12} & \ldots & \ldots & s_{1K} \\
L_2 & s_{21} & s_{22} & \ldots & \ldots & s_{2K} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
L_N & s_{N1} & s_{N2} & \ldots & \ldots & s_{NK} 
\end{pmatrix}
$$

$$
i = \sum_{j=1}^{K} s_{ij}, \; i = 1 \ldots N; \quad c_j = \sum_{i=1}^{N} s_{ij}, \; j = 1 \ldots K
$$

A measuring scheme measures to which degree the resulting clustering matrix differs from the ideal clustering.

Secondly, it is important to conclude what a clustering result of a clustering scheme best represents. A result may be evaluated with respect to different labelling schemes, where each scheme considers a different attribute or combination of attributes of the clustering unit. By changing labelling schemes, the segment-label distribution vector $I$ and the number of labels $N$ of the clustering matrix are now allowed to vary.

\textbf{Scheme1} : \quad I = (l_1, l_2, \ldots, l_N) \\
\textbf{Scheme2} : \quad \hat{I} = (\hat{l}_1, \hat{l}_2, \ldots, \hat{l}_N)

$$
\sum_{i=1}^{N} l_i = \sum_{i=1}^{N} \hat{l}_i = S \quad (\text{constant})
$$

A clustering metric is utilized to select, from a number of clustering schemes, the scheme that produces the best set of clusters for the considered fixed labelling scheme. The number of clusters $K$ may vary from one scheme to another, and the measure must be normalized not to introduce an inherent bias with variation in the dynamic range.

Secondly, it is utilized to evaluate the result of a fixed clustering scheme with respect to varied labelling schemes. This shows what kind of clustering is carried out by the selected scheme on the particular kind of data. The measure of evaluation must be normalized not to introduce biases in the dynamic range with variation in the labelling scheme $I$ or the number of labels $N$. This process of joint evaluation allows us to select a scheme that provides a suitable clustering for the data, and shows what attributes of the data affect the clusters most.

Evaluation can be carried out at two levels.

\textbf{Segment based}: The unit of evaluation is the unit weighted segment. An approximation is made that the segments are homogeneous and of approximately equal length.
Frame based: Information on variation in segment lengths must be captured for the successful comparison of clusters with a few large segments and clusters with many shorter segments. This is achieved by considering the frame to be the unit of evaluation by assigning each segment a weight equivalent to the number of frames in it. This problem can be represented by the clustering matrix \( w_{ij} \) (where \( w_{ij} \) is the number of frames labelled \( i \) in cluster \( j \)), the corresponding frame-label distribution vector \( I \) with elements \( l_i = \sum_{j=1}^{K} w_{ij} \), \( c_j = \sum_{i=1}^{N} w_{ij} \), the perfect clustering matrix \( diag(1) \), and the total number of frames \( W \).

3. CLUSTERING METRICS

A number of clustering metrics, such as entropy, variance, Gini and misclassification metrics, exist for determining the purity and efficiency of a clustering result. This paper considers the Rand metric [2] and the BBN efficiency metric [6].

3.1. The Rand Metric

The Rand metric [2] of the clustering matrix \( M \) is defined as:

\[
I_{RAN D} = \frac{1}{2} \sum_{j=1}^{K} c_j^2 + \frac{1}{2} \sum_{i=1}^{N} l_i^2 - \sum_{i=1}^{N} \sum_{j=1}^{K} s_{ij} \quad (1)
\]

The \( I_{RAN D} \) measure has value 0 for ideal clustering and a positive value that reflects the degree to which a \( M \) differs from the ideal \( M_0 \). To adapt the measure for the evaluation of varied labelling schemes, consider the metric result when the segment distribution amongst the clusters is random. Given a segment-label distribution vector \( I \), using the uniform random assignment of the segments of each speaker to \( K \) clusters \( s_{ij} = l_i/K \), it can be shown that

\[
\text{random}(I_{RAN D}) = \frac{S^2}{2K} + \frac{(K-2)}{2K} I' \quad (2)
\]

The graph of Fig.1 shows the \( \text{random}(I_{RAN D}) \) dynamic range variation with \( K \) for the speaker segment-label distribution of 1996 Hub-4 Broadcast News data. For the range of clusters generally produced by a clustering scheme a variation of about 1500 index points is shown. (Table 1) The second term of expression (2) also captures the variation introduced by variation of the labelling scheme. The \( \text{random}(I_{RAN D}) \) readings as the labelling scheme varies are shown for four considered labelling schemes in Table 1.

The expressions for the frame-based metric is given by substituting \( s_{ij} \) with \( w_{ij} \) and \( S \) with the total number of frames \( W \) in (1) & (2), where \( I \) is now the frame-label distribution. In this expression, the term \( W^2 \) is very large and should make \( \text{random}(I_{RAN D}) \) less sensitive to variations in \( K \) and \( I \). The frame-based \( \text{random}(I_{RAN D}) \) vs \( K \) graph is flat; \( \text{random}(I_{RAN D}) \) values for four different labelling schemes considered are shown in Table 2.

Normalised \( I_{RAN D} \) is defined to enable comparison of clusters with varying \( K \) and segment/frame-label distributions:

\[
\text{Norm}(I_{RAN D}) = 1 - \frac{I_{RAN D}}{\text{random}(I_{RAN D})} \quad (3)
\]

3.2. The BBN Metric

The BBN index [6] measures the efficiency of the clustering matrix \( M \) with respect to \( M_0 \). It is defined in terms of \( p_j \), the purity of cluster \( j \) and \( c_j \) the total number of segments in cluster \( j \). The metric is already normalized and is appropriate for comparison of varying clustering and labelling schemes. Normalization is achieved in terms of \( I_{BBN}(M_0) \) for singleton clustering with each segment in a separate cluster, \( I_{BBN}(M_0) \) for perfect clustering and \( I_{BBN}(M_1) \) for one cluster containing all segments, given the total number of segments \( S \) and the segment-label distribution vector \( I \).

\[
p_j = \sum_{i=1}^{N} s_{ij}^2 / c_j \quad (4)
\]

\[
\eta_{BBN} = \frac{I_{BBN}(M) - I_{BBN}(M_0)}{I_{BBN}(M_0) - I_{BBN}(M_1)} \quad (5)
\]

\[
I_{BBN}(M) = \sum_{j=1}^{K} c_j p_j - Q K \quad (6)
\]

\[
I_{BBN}(M_0) = S - Q N \quad (7)
\]

\[
I_{BBN}(M_1) = \frac{I'}{S} - Q \quad (8)
\]

\[
I_{BBN}(M_s) = S(1-Q) \quad (9)
\]

The frame-based BBN measure is defined by substituting \( w_{ij} \) for \( s_{ij} \) and \( W \) for \( S \), and deriving \( I \) and \( c_j \) as appropriate for equations (4) through (8). However, the frame-based \( I_{BBN}(M_s) \) must be experimentally estimated for the case where each cluster has a single segment in it. It is not appropriate to consider the case where each cluster has a single frame as the singleton measure, since this level of clustering is not achievable and causes measure insensitivity to clustering performance improvements.

Q is a user defined parameter representing the trade off between a few large clusters with many mixed labels, and many small clusters where labels may have more than 1 cluster associated with them. For the segment-based evaluation, Q is set to 0.5 to allow \( \eta_{BBN} \) to give a value in \([-1,1]\) where 1 is achieved for perfect clustering, \( I_{BBN}(M_1) \) achieves value -1, and negative values indicate a tendency to have large clusters with mixed labels. When Q achieves a critical value, \( Q_{crit} \), \( I_{BBN}(M_1) \) achieves value 0, and \( \eta_{BBN} \) is normalized to give a value in \([0,1]\). \( Q_{crit} \) varies
Table 1. Segment based evaluation of clustering & labelling schemes for homogeneous data.

<table>
<thead>
<tr>
<th></th>
<th>No. Clust</th>
<th>spkr</th>
<th>spkr -env2</th>
<th>spkr -env3</th>
<th>spkr -music</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No of Labels</strong></td>
<td></td>
<td>77</td>
<td>103</td>
<td>117</td>
<td>116</td>
</tr>
<tr>
<td><strong>Adapt_c:</strong></td>
<td></td>
<td>81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rand(IRAND)</td>
<td></td>
<td>7580</td>
<td>5885</td>
<td>5385</td>
<td>5243</td>
</tr>
<tr>
<td>IRAND</td>
<td></td>
<td>5376</td>
<td>4138</td>
<td>3759</td>
<td>3599*</td>
</tr>
<tr>
<td>norm(IRAND)</td>
<td></td>
<td>0.291</td>
<td>0.297</td>
<td>0.302</td>
<td>0.314*</td>
</tr>
<tr>
<td><strong>Speaker1_c:</strong></td>
<td></td>
<td>92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rand(IRAND)</td>
<td></td>
<td>7423</td>
<td>5723</td>
<td>5221</td>
<td>5078</td>
</tr>
<tr>
<td>IRAND</td>
<td></td>
<td>4286</td>
<td>3346</td>
<td>3167</td>
<td>2965*</td>
</tr>
<tr>
<td>norm(IRAND)</td>
<td></td>
<td>0.423</td>
<td>0.416</td>
<td>0.393</td>
<td>0.416</td>
</tr>
<tr>
<td><strong>Speaker2_c:</strong></td>
<td></td>
<td>165</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rand(IRAND)</td>
<td></td>
<td>6911</td>
<td>5194</td>
<td>4687</td>
<td>4543</td>
</tr>
<tr>
<td>IRAND</td>
<td></td>
<td>4937</td>
<td>3665</td>
<td>3270</td>
<td>3106*</td>
</tr>
<tr>
<td>norm(IRAND)</td>
<td></td>
<td>0.286</td>
<td>0.294</td>
<td>0.370*</td>
<td>0.316</td>
</tr>
<tr>
<td><strong>Adapt_c:</strong></td>
<td></td>
<td>81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\eta_{BBN})Q = 0.5</td>
<td></td>
<td>0.646</td>
<td>0.537</td>
<td>0.511</td>
<td>0.516</td>
</tr>
<tr>
<td><strong>Speaker1_c:</strong></td>
<td></td>
<td>92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\eta_{BBN})Q = 0.5</td>
<td></td>
<td>0.707</td>
<td>0.595</td>
<td>0.564</td>
<td>0.564</td>
</tr>
<tr>
<td><strong>Speaker2_c:</strong></td>
<td></td>
<td>165</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\eta_{BBN})Q = 0.5</td>
<td></td>
<td>0.616</td>
<td>0.511</td>
<td>0.484</td>
<td>0.496</td>
</tr>
</tbody>
</table>

The ‘adapt_c’ scheme clusters by growing a clustering tree that terminates on a minimum occupancy count. The ‘speaker1_c’ scheme grows a tree until the gain from growth falls below a threshold, then recombines nodes that have a distance between the nodes that is less than the average distance within the node. The ‘speaker2_c’ scheme grows a larger tree that terminates on a larger gain growth threshold, and allows fewer recombination of nodes that have an inter-node distance less than the average within-node distance.

4.2. Labelling Schemes [7]

Four different labelling schemes are considered. The environment labels are derived from 3 variables released with the 1996 Hub-4 Broadcast News transcriptions - background music, background speaker and other noise - each marked as Clear, High or Low.

**Spkr scheme:** This scheme labels each segment with its speaker resulting in 77 labels.

**Spkr-env2 scheme:** This scheme gives a label to each speaker and environment condition combination; the environment has 2 attributes on whether there is any noise or no noise in the background. There are 103 resulting labels.

**Spkr-env3 scheme:** This scheme does the same as above with the exception that the environment has 3 attributes on whether there is no noise, or whether the noise level is High or Low (117 labels). According to the release notes some level of subjective human error in marking loudness levels may be expected, and this scheme may suffer from this low confidence in labelling.

**Spkr-music scheme:** This scheme does an analysis that gives priority to music where the environment is analyzed as clear, whether there is music in the background or, where there is no music, contains any other noise. There are 116 labels. It is considered to investigate whether a clustering scheme is able to capture signature tunes or patterns in the music over other background noise. It is noted that 3 segments with low music and high other noise are labelled as music.

4.3. Evaluation of Clustering Schemes

Tables 1 and 2 present the results of the metrics for clustering and labelling scheme evaluation. In the top section, the normalization metric \(\text{norm}(I_{RAND})\), the current Rand metric and normalized Rand metric readings are shown. The two sections below show the [0,1] scaled and [-1,1] scaled BBN metric readings respectively. In each section the clustering schemes, given in bold in the left column, are evaluated according to the 4 labelling schemes. It is noted that according to the Rand measure, the lower the reading the better the ranking, while according to \(\text{norm}(I_{RAND})\) the higher the reading the better the ranking.

Comparing along the columns for segment based clustering scheme evaluation for each fixed labelling, both Rand metrics and the BBN [0,1] metric agree that the ‘speaker1_c’ scheme is the best clustering scheme. However the BBN [-1,1] metric favors the ‘speaker2_c’ scheme albeit by a very small difference. However, in the frame based evaluation that is more sensitive to the length of the segments the BBN [-1,1] metric agrees with the other metrics that the ‘speaker1_c’ scheme is the best clustering scheme for ‘spkr’, ‘spkr-env2’, and ‘spkr-env3’ labelling schemes. In particular for the ‘spkr’ and ‘spkr-env2’ labelling schemes, close observation of the clusters affirms that ‘speaker1_c’ produces the best clusters. The exception is the ‘spkr-music’ labelling scheme, where the BBN [-1,1] metric reads that the ‘speaker2_c’ scheme is better. Hence it is

4.4. EXPERIMENTAL RESULTS

The clustering metrics are used to evaluate three clustering schemes with respect to four labelling schemes. The labelling schemes are defined according to the speaker and the acoustic environment in order to test the sensitivity of the clustering scheme to the acoustic environment. The data clustered are 488 segments of 1996 Hub-4 Broadcast News Transcription development (BNdev97) data homogeneous with respect to the speaker and the environment. [3]
The best ranking labelling scheme according to the Rand metrics is in somewhat better agreement with the observations and the conclusions, it is still unreliable for labelling scheme evaluation.

5. CONCLUSION

In this paper techniques for the joint evaluation and selection of clustering and cluster labelling schemes are investigated. The Rand evaluation metric is normalized to remove the bias arising from variation in the dynamic range with varying number of clusters and labelling schemes; normalization is achieved with respect to the introduced random (I_{RAND}) measure. The Rand metrics and the already normalized BBN metrics are applied to investigate the task of speaker-environment tracking in Broadcast News.

The evaluation concludes that the Rand metrics are appropriate for ranking clustering schemes for a fixed labelling scheme. It also concludes that, while some improvement can be achieved by this normalisation of the Rand metric, the BBN efficiency metrics are preferred for the evaluation of labelling schemes. The experiments reveal that the clustering scheme which grows a tree of moderate depth with re-combinations, while achieving a speaker level clustering of high efficiency (0.760), also tracks the speakers through a noisy/clear environment with a relatively high efficiency of 0.715. This demonstrates that the background noise exerts a high level of influence on this application of speech data clustering.

Further work would involve investigating other metrics that are appropriate for clustering scheme assessment and improving clustering for speaker clusters by prior speech denoising.

Acknowledgements: The authors would like to thank the European Union sponsored CoreTex project for research funding.

6. REFERENCES