Sensitivity of the Elucidative Fusion System to the Choice of the Underlying Similarity Metric

Belur V. Dasarathy
Dynetics, Inc., P. O. Box 5500
Huntsville, AL. 35814-5500, USA
Belur.d@dynetics.com

Abstract - In a recent study, a new concept, namely elucidative fusion systems, was proposed and demonstrated using case based reasoning as the fusion tool. Elucidative fusion systems are designed to offer not only optimally fused decisions but also elucidate the relative contributions made by the different data sources (sensors) to the fused decisions. In the earlier study, the concept was illustrated using the classical Euclidean metric for measuring the dissimilarity of cases under the case based reasoning model chosen for the fusion process. In this study, multiple metrics are tested and evaluated to determine the sensitivity of the elucidation process to the choice of the underlying metric. Euclidean and two other metrics, all of which are special cases of the general Minkowski distance metric, are chosen for this purpose. An audio-visual system for the recognition of spoken French vowels is again used as the example for illustration of the associated concepts and analysis.

Keywords: Elucidative fusion systems, metric sensitivity

1. Motivation

Case based reasoning (CBR) driven decision systems are by their very nature sensitive to the metric employed in the assessment of the similarity between cases. Fusion systems that employ such case-based reasoning are accordingly sensitive to the underlying similarity or distance metrics. Sensitivity of the fusion system performance, in terms of the decision quality, to the distance metric chosen was studied just recently [1] in fair detail. The concept of elucidative fusion system (EFS), put forth during the past year [2], provides a means of assessing the relative influence of the different information sources on the fused decision. A joint focus on these two studies together brings to the fore the question "how does the similarity metric affect not only the fused decision but also the relative influence of the different information sources on the fused decision process?" This paper, in essence, represents an initial attempt at answering this additional aspect of the question: once again using the audio-visual fusion system employed in the earlier studies as the illustrative vehicle for such investigations. Section 2 offers a brief overview of the elucidative fusion system and the underlying concepts. Section 3 presents a short review of similarity metrics. Section 4 describes the fusion system used here for illustrative purposes. Section 5 details the results of the experimental investigation. Discussion and conclusions are included in the last section.

2. Elucidative Fusion System

The concept of elucidative fusion systems was introduced recently [2] as a means of understanding the relative contributions of the individual data sources (sensors) in the context of fusion of the information provided by these sources. As discussed in the earlier paper, the specifics of the fusion system model define the elucidation process. While parametric, non-parametric, and similarity-driven models were listed as conceivable alternatives, the latter one was chosen for strictly illustrative purposes and discussed in more detail. The similarity-driven model, such as the case-based reasoning (CBR) whose origins can be traced back to nearest-neighbor pattern classification techniques [3], requires a definition of the measure of closeness between the cases in the database and the incoming to-be-identified case. The specifics of how the relative influence of the different sources is defined and measured within this CBR framework have already been presented in detail in the earlier study [2] and as such are not repeated here. The next section briefly reviews the array of similarity metrics that has been conceived to measure the similarity or more appropriately measure of separation.
3. Similarity Metrics

The measure of closeness or separation is closely tied to a definition of the distance metric underlying such a measure. As is well known numerous metrics have been proposed in the literature. The metric most often employed is the Euclidean metric. However, this metric is mostly suited to data scenarios with continuous valued attributes only. The next most common metric, also used mostly in the case of continuous-valued attributes is the city block or Manhattan distance metric. The two are similar in that they sum up the distances along the different attributes, but differ in the specifics of how the distances are summed. The third metric, which is less frequent in its usage, is the chessboard or Chebychev metric [4]. This is significantly different in that it only considers a single attribute that has the maximum separation. All the three are special cases of the Minkowski distance with the Minkowski parameter \( r \) being 2 for Euclidean, 1 for City-block, and \( \infty \) corresponding to the Chebychev metric. Other variations to the Euclidean such as the Box-Cox metric [5] have also been considered in the context of case based reasoning in its original form, nearest neighbor techniques. Various other distance metrics [6] have indeed been proposed in the literature to address cases wherein the attributes may have discrete or symbolic values, or a mix thereof. Examples of these include the Value Difference Metric (VDM), Modified Value Difference Metric (MVDM), Heterogeneous Euclidean-Overlap Metric (HEOM), Heterogeneous Value Difference Metric (HVDM), Interpolated Value Difference Metric (IVDM), Discretized Value Difference Metric (DVDM), and Windowed Value Difference Metric (WVDM). Since the data sets employed here were continuous valued, this analysis was limited to the three special cases of the Minkowski metric listed above.

4. Fusion System Description

For illustrative purposes, we employ here the same fusion system investigated in the prior studies. The system employs a combination of audio and video sensors for the recognition of orally presented French vowels. The data set consists of 100 samples of each of the 10 French vowels. These vowels - [i, e, æ, y, œ, ɜ, u, o, ø, a] are orally pronounced by a single speaker separately one at a time. The audio data consists of twenty features while the visual has only three features. White Gaussian noise is added to the audio signal sampled at 16 kHz. The noise is generated at eight signal-to-noise ratios (SNR): no noise (infinite dB), 24 dB, 12 dB, 6 dB, 0 dB, -6 dB, -12 dB, and -24 dB. The 20 audio features were obtained by an FFT on the first 64 ms of the audio signal and correspond to the dB values in 20 1-bark wide channels between 0 and 5 kHz. The three visual parameters correspond to the three aspects of the inner contour of the lips: horizontal width, vertical height, and area. The entire data set was divided into two subsets one each for training and testing with fifty samples from each vowel class. While only a single partitioning was used in these experiments, experience suggests that it is advisable to repeat these experiments over several partitions to derive an estimate of the variance in these results. The training was limited to only four of the SNR cases: no noise (infinite dB), 24 dB, 12 dB, and 0 dB. The training data set, from an audio perspective, effectively has 2000 samples (50 samples x 4 noise levels x 10 classes). From a video perspective, the number of samples is only 500 (50 samples x 10 classes) although for implementation purposes the same video data gets included multiple times along with the audio data at each noise level in defining the 2000 training feature vectors. The 6 dB SNR case represents an interpolation scenario since it is not specifically included in the training but lies within the range used for training. The other three SNR cases: -6 dB, -12 dB, and -24 dB represent extrapolation or off-nominal scenarios. The reader can look up the original reference [7] for additional details regarding the data acquisition process and other related matters including details of the extensive application specific contextual methods of implementing extrapolation paradigms.

5. Experimental Results

In view of the limitations of the publication space, we shall first present the results averaged over all the ten vowel classes and follow it up with results for only two of the ten individual classes. Admittedly, the results when examined over individual classes are not similar since there is a greater variability at that level and consequently less informative as well. Figure 1 shows the average (over all ten classes) relative influence of the two sensors under the 3 different metrics at different noise levels. As can be observed therein, at very high noise levels wherein
no training information was available, there is little difference among the metrics. This is to be expected since the noise, which affects only the audio input, forces the fused decision to mainly rely on the video information only, and consequently there is little wiggle room in the influence of the individual sensor to distinguish among the 3 metrics. However, as the noise level decreases, some distinctions among the metrics begin to appear. The video information has relatively less influence under the Euclidean metric than under the other 2 metrics. This is perhaps caused by the imbalance in the number of features provided by the video (3) and audio data (20). With fewer features, the contribution of a significant dissimilarity in any one feature (emphasized by the squaring under the Euclidean) is more dominant for the video than is the case for audio. On the other hand, the video influence is maximum, under the chessboard metric, wherein the decision is based on a single maximum difference. It is interesting to note, that at SNR of 6 dB, which was another instance wherein no explicit training was made available (interpolation regime), the relative differences among the metrics are once again less significant. At SNR levels greater than 6 dB, i.e., at much lower noise levels, the influence of the audio is observed to be more than that of the video under the Euclidean. Under the city-block metric, the audio and video just balance out only at zero noise level. Under the chessboard metric, the influence of the video remains higher than that of the audio even at zero noise. Figure 2 portrays the variance in the average influence (over the 10 classes). It is clear that the variance is higher for the Euclidean by a convincing margin at all the noise levels. The chessboard and city-block metric have nearly equal variance at high noise levels while at low noise levels, the chessboard metric has the least variance.

Figures 3 and 4 portray the corresponding results observed for class 1 decisions. Similarly, Figures 5 and 6 illustrate the results that correspond to class 2 decisions. Comparing Figure 3 with Figure 1, it can be seen for Class 1 decisions, the average influence of audio is clearly greater than that of the video at all noise levels even under the Euclidean norm. In all other respects, the behavior for class 1 is similar to the average over all classes. However, comparing Figure 5 with Figure 1, the relative influence of audio as compared to video across the lower end of the noise spectrum is less clearly defined and quite mixed. Correspondingly, the performance for class 2 is indeed poor. Comparing the variance measures for the average across classes (Figure 2) with that for Class 1 (Figure 4) and Class 2 (Figure 6), unlike for the average and Class 2 alone, the variance for Class 1 is by no means so clearly higher under the Euclidean norm. Thus the average and variance of the relative influences of audio and video data across classes falls in form somewhere in between that for Class 1 and Class 2, but much closer to that observed for Class 2 than for Class 1. Publication space considerations do not permit more detailed discussion on a class-wise basis, which may or may not be significantly informative from a statistical perspective.

6. Discussions and Conclusions

The study investigates the sensitivity of the elucidation process to the choice of the metric employed by the case based reasoning model deployed by the fusion system. A real-world example, an audio-visual system for the recognition of orally pronounced French vowels, is used here for illustrating the analysis. Comparatively speaking, behavior patterns under the city-block and Chebychev metrics are closer to each other than that under the Euclidean. This is particularly noticeable in the variance of the average influence of the two sensors on the fused decision. The system is mostly insensitive to the choice of the metric at high noise levels wherein no training information was provided, i.e., under the extrapolation regime. Further experimentation with more diverse data sets is necessary to be able to draw more definitive conclusions. Additional studies that encompass a wider array of metrics that cover discrete and symbolic valued attribute data scenarios are to be recommended.

7. Acknowledgments

The author wishes to place on record his gratitude to Dr. Pascal Teissier and his associates at ICP and LTIRF, Grenoble, France, who provided the audio-visual data sets used in these investigations.
Figure 1. Average (over all the 10 vowel classes) relative influence of audio and video information on the fused decision at the various noise-levels under the different metrics.

Figure 2. Variance in the average (over all the 10 vowel classes) relative influence of audio and video information on the fused decision at the various noise-levels under the different metrics.
Figure 3. Average relative influence of audio and video information on the fused class 1 decisions at various noise-levels under the different metrics.

Figure 4. Variance in relative influence of audio and video information on the fused class 1 decisions at various noise-levels under the different metrics.
Figure 5. Average relative influence of audio and video information on the fused class 2 decisions at various noise-levels under the different metrics

Figure 6. Variance in relative influence of audio and video information on the fused class 2 decisions at various noise-levels under the different metrics
7. References


