Experiments on Probability based Similarity Measures Applied to Image Similarity

C. Fang[1], S. Visa[2], M. Ionescu[3], and A. Ralescu[1]

[1]*Department of Computer Science, ML 0030*
University of Cincinnati, Cincinnati, OH, 45237-0030, USA
fangcg@email.uc.edu, Anca.Ralescu@uc.edu

[2]*Computer Science Department*
Wooster College, Wooster, OH, USA
svisa@wooster.edu

[3] *Systems Pathology Company, LLC, Seattle, WA, USA*
mirceaionescu@systempath.com

Abstract

We consider a probability based approach according to which the similarity of two values (in the same domain) is the probability of value pairs whose components are rather apart than the two values under consideration. Similarities across the attributes of the heterogeneous data are combined using Fisher transformation. Results of applying this approach to an image retrieval problem are also presented.

1 Introduction

Heterogeneous data are multidimensional data whose components lie in different domains. More precisely, if $x = (x_1, \ldots, x_n)$ denotes a multidimensional data point, where $x_i \in D_i$. We say that x is heterogeneous if $\forall i \neq j$, $i, j, = 1, \ldots, n$, D_i and D_j are different. The key word in the preceding statement is the word *different*and the treatment of heterogeneity will depend on what is meant by it. The components x_i may lie in different spaces (e.g. real, categorical, ordinal, sets), or their domains may all be subsets of the same space. In the former case, the meaning of *different* is rather obvious, while in the latter is not necessarily so. For example, suppose that all the domains are subsets of the real numbers, i.e. $D_i \subset \Re$. Then one way to define $D_i$different from D_j is to say that $D_i \cap D_j = \emptyset$, or, by relaxing this requirement, that they overlap very little.

This study considers a particular case of heterogeneity, when data may come from the same domain, even same range of values, but from different underlying distributions.

2 Probability based approach to similarity evaluation versus Euclidean distance

For ease of notation X denotes an attribute taking values in a domain D_X, and $a, b \in D_X$ denote two of its values. We are interested in evaluating the similarity between a and b. In the probability based approach X is assumed to be a random variable with values in a space D_X endowed with a distance measure d, and with distribution function F. That is, for $x \in D_X$, $F(x) = P(X \leq x)$. Then, following [2], [5], and [4], for two values a and b their probability based similarity is defined as

$$Sim_F(a, b) = P(d(X, Y) > d(a, b)) \qquad (1)$$

where X, Y are independent identically distributed (iid) according to F. Note that in order to use (1) to calculate $Sim_F(a, b)$ one must first find the probability distribution of $d(X, Y)$. For example, if $D_X = \Re$, and $d(X, Y) = |X - Y|$ the distribution of $|X - Y|$ must be computed. The complexity of this computation depends on the distribution function F. Alternatively, using a well-known result from probability theory, according to

which if X has distribution F, $F(X)$ has distribution $U[0,1]$, then (1) can be replaced by

$$Sim_U(a,b) = 1 - (F(a) - F(b))^2 \tag{2}$$

While equations (1) and (2) define a similarity measure, the extent to which they may agree depends on how close the distributions F and U are.

3 Heterogeneity due to different underlying distributions

Consider that a multidimensional data set of points $\mathbf{x} = (x_1, \ldots, x_n)$, where $x_i \in D_i = D$, endowed with a distance d, and x_i distributed according to distribution function F_i, where $F_i \neq F_j$ when $i \neq j$. The most common way of evaluating proximity is directly from d. For example, using the Euclidean distance, d_E and upon normalization the similarity between a and b can be defined as

$$Sim_E(a,b) = 1 - \frac{d_E(a,b)}{M} \tag{3}$$

where $M = max\{d_E(x,y) \,|\, x,y \in D\}$ is the maximum distance between values of X. To take into consideration the distribution underlying the data, the similarity must be defined directly from this distribution as described below. For illustrative purposes we consider a small example data set of with $N = 5$ data points described by $k = 3$ attributes. Moreover the data points are identical as shown below

$$data = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{array} \begin{array}{c} \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ \begin{pmatrix} 10 & 10 & 10 \\ 14 & 14 & 14 \\ 18 & 18 & 18 \\ 25 & 25 & 20 \\ 35 & 35 & 35 \end{pmatrix} \end{array} \tag{4}$$

The matrix of distances between these data points when the Euclidean distance is used is

$$d_E = \begin{pmatrix} 0 & 6.9282 & 13.8564 & 25.9808 & 43.3013 \\ 6.9282 & 0 & 6.9282 & 19.0526 & 36.3731 \\ 13.8564 & 6.9282 & 0 & 12.1244 & 29.4449 \\ 25.9808 & 19.0526 & 12.1244 & 0 & 17.3205 \\ 43.3013 & 36.3731 & 29.4449 & 17.3205 & 0 \end{pmatrix} \tag{5}$$

and hence that of similarities computed according to (6) is

$$Sim_E = \begin{pmatrix} 1.00 & 0.84 & 0.68 & 0.40 & 0 \\ 0.84 & 1.00 & 0.84 & 0.56 & 0.16 \\ 0.68 & 0.84 & 1.00 & 0.72 & 0.32 \\ 0.40 & 0.56 & 0.72 & 1.00 & 0.6 \\ 0 & 0.16 & 0.32 & 0.60 & 1.00 \end{pmatrix} \tag{6}$$

Assume now that each attribute has a different underlying distribution: a_i has a Normal distribution with mean μ_i and variance σ_i^2. Table 3 shows the probability densities values for the three attributes for specific choices of μ_i, σ_i. Let $G_{\mu_i,\sigma_i} = Prob(X \leq x)$ denote

Attribute	μ	σ	p_1 10	p_2 14	p_3 18	p_4 25	p_5 35
a_1	5	3	0.033	0.002	0	0	0
a_2	10	6	0.067	0.053	0.0273	0.0029	0
a_3	20	8	0.023	0.038	0.048	0.041	0.009

the cumulative distribution function for X distributed according to Normal distribution mean μ_i, and variance σ_i^2. Then using $d(x,y) = |x-y|$, equation (1) becomes:

$$\begin{aligned} &Sim_{N(\mu_i,\sigma_i^2)}(x,y) = \\ &= P(|X-Y| > |x-y|) = 1 - P(|X-Y| \leq |x-y|) \\ &= 1 - P(-|x-y| \leq X - Y \leq |x-y|) \\ &= 1 - [G_{\mu_i,\sigma_i}(|x-y|) - G_{\mu_i,\sigma_i}(-|x-y|)] \\ &= 1 - G_{0,\sigma_i^2}(|x-y|) + G_{0,\sigma_i^2}(-|x-y|) \end{aligned} \tag{7}$$

The similarities S_1, S_2, and S_3, between the values of attributes a_1, a_2, and a_3, when the underlying distribution corresponding to each attribute is used in equation (7) are shown in Figure 1.

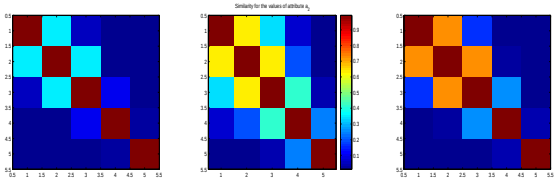

Figure 1. Similarity computed for all values of attributes a_1, a_2, and $a3$.

3.1 Aggregation of similarities across attributes

Recall that according to the probability based approach, the similarity between two values of an attribute is, in fact, a probability. To aggregate similarity values across attributes means, therefore, to aggregate these probabilities. In this study, Fisher transformation χ^2_{2k} shown in equation (8) is used:

$$\chi^2_{2k} = \sum_{i=1}^{k} \log \frac{1}{S_i^2} \tag{8}$$

The values for χ^2_6 obtained for the five data points

and three attributes of this example are then:

$$c = \begin{pmatrix} 0 & 3.672 & 9.243 & 24.115 & 57.4689 \\ 3.672 & 0 & 3.672 & 14.791 & 42.376 \\ 9.243 & 3.672 & 0 & 7.659 & 29.625 \\ 24.115 & 14.791 & 7.659 & 0 & 12.807 \\ 57.469 & 42.376 & 29.625 & 12.807 & 0 \end{pmatrix} \quad (9)$$

Finally, the overall similarity matrix is obtained from the corresponding p-values using equation 10:

$$Sim_C(i,j) = 1 - \chi^2_{2k}(c(i,j)) \quad (10)$$

where χ^2_{2k} denotes the cumulative distribution function for the χ^2 distribution with $2k$ degrees of freedom. Equation (11). For the example data set the overall similarity matrix is

$$Sim_C = \begin{pmatrix} 1.0000 & 0.721 & 0.1604 & 0.0005 & 0.0000 \\ 0.7210 & 1.0000 & 0.7210 & 0.0219 & 0.0000 \\ 0.1604 & 0.7210 & 1.0000 & 0.2642 & 0.0000 \\ 0.0005 & 0.0219 & 0.2642 & 1.0000 & 0.0462 \\ 0.0000 & 0.0000 & 0.0000 & 0.0462 & 1.0000 \end{pmatrix} \quad (11)$$

Figure (2) shows the similarity matrix Sim_C side by side with Sim_E.

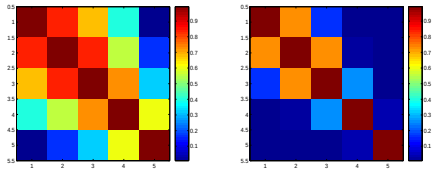

Figure 2. Similarity matrices for the example data set, based on the Euclidean distance (left) and probability based (right).

4 Image Retrieval based on Probability-based Similarity Measures

This section presents results of using the probability-based similarity in content based image retrieval [6], based on both color and texture image features: normalized color histogram, and normalized rotation-invariant local binary pattern (LBP) histogram [3] . Therefore, an image such encoded is a heterogeneous data point. The 1,000 image dataset is selected from PSU 10,000 low resolution web-crawled misc database [1]. The experiments below show that the probability-based similarity plus the aggregation using the p-value of the χ^2-distribution produce better retrieval results than the retrieval based on single image feature (color or texture).

The experiment consists of two stages, offline training and online query.

4.1 Offline Training

In this stage, the probability-based similarity is estimated from image feature vectors, for each of its heterogeneous features. Each attribute takes as values histograms, which upon normalization can be considered discrete probability distributions. Therefore, the distance between two histograms is evaluated using a distance between probability distributions, in particular, here the Jeffrey's distance (12)(a symmetric version of the Kullback-Leibler distance) is used.

$$\begin{aligned} d_{KL}(p,q) &= \textstyle\sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \\ d_J(p,q) &= d_{KL}(p,q) + d_{KL}(q,p) \end{aligned} \quad (12)$$

where $p = \{p_1, \ldots, p_n\}, q = \{q_1, \ldots, q_n\}$ are two histograms. Therefore, the first step is to compute d_{KL} separately, for all possible pairs of color features, and texture features. This generates two 1000×1000 distance matrices. By sorting all elements in similarity matrix, the empirical distribution function of d_{KL} is estimated. Figure 4.1 shows these distributions. Using equation (1) the similarity along color and texture attributes can be easily computed. Figure 4.1 shows the similarity matrices thus computed. A modified version of binary search can boost the search time to $O(\log N)$.

Finally, the χ^2_4 defined in (8) is used to combine these

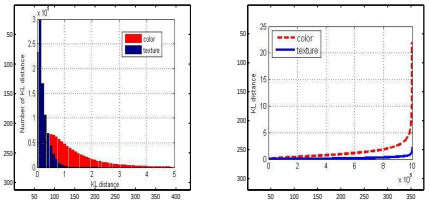

Figure 3. Sorted values of d_{KL} for the color and texture attributes (left) and distribution histogram for d_{KL} distance for these two features (right).

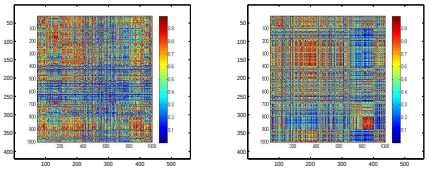

Figure 4. Similarity matrix for color feature (left), and texture feature (right).

two heterogeneous similarities, according to formula. This step is not necessary in building up an efficient CBIR system, but for the current study it can help to visualize how the images in database are clustered. The resulting similarity matrix is shown in Figure 5. Note that this stage requires $O(N^2)$ complexity.

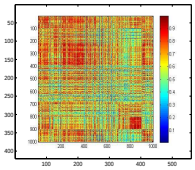

Figure 5. The overall similarity matrix based on color and texture.

4.2 Online Query Retrieval

For the query image, first the system extracts the color and texture features. The probability-based similarities between the query image and each image in dataset, along each of its heterogeneous features are then computed. Aggregation via the χ^2 p-value is then computed. The top k most similar images from the image based are returned. A significant improvement in retrieval quality can be illustrated by the results for three queries, "baby" query (image id 462), "moon, space" query (image id 620), and "sun" query (image id 777). The retrieval results are shown in Figures 6, 7, and 8 respectively.

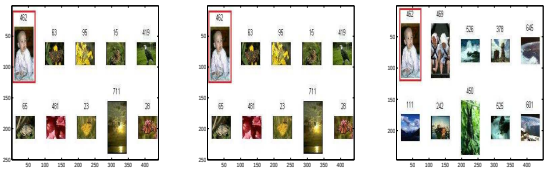

Figure 6. Retrieval results for the query "baby image" (image id = 462): color only (left), texture only (middle) and overall (right).

5 Conclusions

We presented a probability-based approach to the similarity between heterogeneous Similarity across attributes is obtained using the p-value of the Fisher transformation. Experimental results for image retrieval suggest that the approach is very effective. Further work is necessary to fully explore the benefits and limitations of this approach.

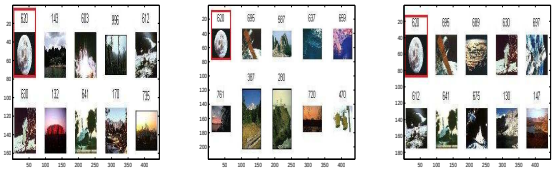

Figure 7. Retrieval results for the query "moon, space" image (image id = 620): color only (left), texture only (middle) and overall (right).

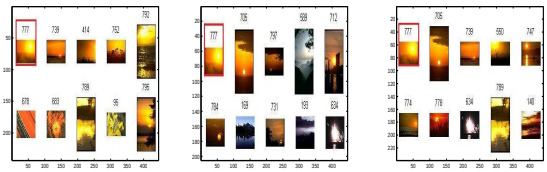

Figure 8. Retrieval results for the query "sun image" (image id = 777): color only (left), texture only (middle) and overall (right).

6 Acknowledgment

This work was partially supported by the Department of the Navy, Grant ONR N000140710438 and OBR Graduate Fellowship.

References

[1] http://wang.ist.psu.edu/docs/related.shtml.

[2] S. Le and T. Ho. Measuring the similarity for heterogeneous data: An ordered probability-based approach. *LNAI*, 3245:129–141, 2004.

[3] T. Ojala, M. Pietikinen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[4] S. A. Popovici. On evaluating similarity between heterogeneous data. Master Thesis, University of Cincinnati, 2008.

[5] A. Ralescu, S. Popovici, and D. Ralescu. On evaluating the proximity between heterogeneous data. In *Proceedings of the Nineteenth Midwestern Artificial Intelligence and Cognitive Science Conference, MAICS-2008, Cincinnati, Oh, USA, April 12-13, 2008*.

[6] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and J. R. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(PART 12):1349–1380, 2000.