SHYSTER and the Calculation of Distance between Cases

James Popple

Department of Computer Science
Australian National University

Abstract

The SHYSTER project is concerned with the development of a hybrid legal expert system. The case-based part of SHYSTER chooses the cases upon which to base its opinion by measuring the similarity between previously-decided cases and the case in question.

Statisticians make use of many different similarity measures, however—for SHYSTER—each of the commonly used similarity measures reduces to one of three measures. These measures can be weighted so that some facts in the cases can be given more importance than others.

SHYSTER uses one of these weighted measures (a distance measure) to construct it opinion. Early testing shows that SHYSTER's approach to finding the distance between cases allows it to produce useful advice. Further testing will attempt to validate this approach empirically.

CR Categories and Subject Descriptors: **I.2.1** [Artificial Intelligence]: Applications and Expert Systems—*law*; **I.2.5** [Artificial Intelligence]: Programming Languages and Software—*expert system tools and techniques*; **J.1** [Computer Applications]: Administrative Data Processing—*law*.

Key Words and Phrases: case-based reasoning, expert systems, law.

This research is supported by an Australian National University PhD Scholarship, funded by the Centre for Information Science Research.

1 Introduction

The SHYSTER project is concerned with the development of a hybrid legal expert system: one which uses rule-based techniques to represent statute law and case-based techniques to represent case law. A prototype has been developed, implementing the case-based part of SHYSTER (Popple 1990, 1991, 1992).

SHYSTER reads programs written by a legal expert in a case law specification language. Each program specifies an area of case law in terms of the leading cases in that area, and the facts—the *attributes*—in each of those leading cases. In each case, each of the attributes can have a value of YES, NO or UNKNOWN.

SHYSTER interrogates the user as to the facts in the case in question—the *instant case*—and constructs a legal opinion as to the likely result in the instant case, based on the leading cases.

In order to construct its opinion, SHYSTER needs to determine which of the leading cases are most similar to the instant case. To do this, it assigns a weight to each of the attributes, and calculates the distances between the instant case and each of the leading cases. The smaller the distance between two cases, the more similar those two cases are.

SHYSTER calculates two different types of distance: the *known distance* is defined as the sum of the weights of every attribute for which those two cases have different known values; the *unknown distance* is defined as the sum of the weights of every attribute for which either of the two cases has an unknown value.

A known distance of zero indicates that the two cases are identical—at least as far as the known attributes are concerned. A large unknown distance indicates that the values of some important (i.e. heavily weighted) attributes were unknown, casting some doubt on the accuracy of the known distance calculation for that case. The unknown distance can be thought of as a measurement of possible error: it is the maximum distance that could be added to the known distance if all of the unknown attribute values were known.

SHYSTER's method is not the only method of measuring similarity; statisticians make use of many different similarity measures. However, because

- SHYSTER's known values are *binary* (each YES is treated as a "1", each NO as a "0"), and
- SHYSTER deals with the *relative* distance/similarity between cases (it quantifies the distance between cases, and has regard to the fact that a case j is further from a case k than a case ℓ, but not how much further),

each of the commonly used similarity measures reduces to one of three measures.

This technical report discusses those commonly used similarity measures and how they are reduced for SHYSTER. It also discusses how these measures can be weighted to give some attributes more importance than others.

2 Similarity measures

Cluster analysis is the separation of data into groups on the basis of "similarity". Entities are grouped so that two entities in the same group are more similar than two entities in different groups. Similarity measures are used to quantify the similarity between every entity and every other entity. SHYSTER does not perform cluster analysis, but it does quantify the similarity between the instant case and each of the leading cases. This process takes $O(n)$ time, where n is the number of cases, whereas cluster analysis takes $O(n^2)$ time.

Many different types of similarity measure have been developed. The most widely used measures are called *metrics*.

> The quantitative estimation of similarity has been dominated by the concept of metrics; this approach to similarity represents cases as points in a coordinate space such that the observed similarities and dissimilarities of the points correspond to metric distances between them ... The dimensionality of the space is determined by the number of [attributes] used to describe the cases.[1]

For a similarity measure to be a metric it must satisfy the following four criteria:[2]

- Given two cases j and k, $d_{jk} = d_{kj} \geq 0$, where d_{jk} is the distance between case j and case k.
- Given three cases j, k and ℓ, $d_{jk} \leq d_{j\ell} + d_{k\ell}$ (this is called the *triangle inequality* or the *metric inequality*).
- Given two cases j and k, if $d_{jk} \neq 0$, then j is not identical to k.
- Given two identical cases j and j', $d_{jj'} = 0$.

Many researchers have argued against the use of similarity measures which do not meet these criteria. Aldenderfer and Blashfield (1984) point out that

> [Measures] that are not metrics may not be jointly monotonic; that is, the values of different [measures] used with the same data will not necessarily vary conjointly, raising the disturbing issue that these [measures] could suggest quite different relationships among the entities.[3]

[1]Aldenderfer and Blashfield 1984, p. 18, emphasis omitted.
[2]Aldenderfer and Blashfield 1984, p. 18. Everitt 1974, p. 56 lists all but the third criterion.
[3]Aldenderfer and Blashfield 1984, pp. 18–19.

However, it is not essential that a similarity measure be a metric. For example, the Pearson product-moment correlation coefficient ("a popular similarity measure",[4] discussed in §2.3) is not a metric.

There are four different kinds of similarity measures: distance measures, association coefficients, correlation coefficients, and probabilistic similarity coefficients.

Strictly speaking, probabilistic similarity coefficients do not actually calculate the similarity between two cases. They take into account the distribution of the frequencies of the attribute values over all the cases, and are calculated during the formation of clusters.[5] Probabilistic similarity coefficients are inappropriate for use with SHYSTER because it does not perform cluster analysis.

The three kinds of similarity measure appropriate for SHYSTER are distance measures, association coefficients, and correlation coefficients.

2.1 Distance measures

Two popular distance measures are Euclidean distance and Manhattan distance. The Euclidean distance d_{jk} between two cases j and k is defined as

$$d_{jk} = \sqrt{\sum_{i=1}^{n} \left(A_{ij} - A_{ik}\right)^2}$$

where A_{ij} is the value of the ith attribute for the jth case,[6] and n is the number of attributes.[7] The Manhattan distance is defined as

$$d_{jk} = \sum_{i=1}^{n} \left|A_{ij} - A_{ik}\right|.$$

Both Euclidean distance and Manhattan distance are specific examples of the class of metric distance functions known as Minkowski metrics. The Minkowski metric is defined as

$$d_{jk} = \left(\sum_{i=1}^{n} \left|A_{ij} - A_{ik}\right|^K\right)^{1/K}$$

where K is some constant: for Euclidean distance, $K = 2$; for Manhattan distance, $K = 1$. In the binary case (e.g. SHYSTER),

$$\left|A_{ij} - A_{ik}\right|^K = \begin{cases} 0, & \text{if } A_{ij} = A_{ik}; \\ 1, & \text{if } A_{ij} \neq A_{ik}. \end{cases}$$

[4] Aldenderfer and Blashfield 1984, pp. 19.

[5] Sneath and Sokal 1973, pp. 140–5; Aldenderfer and Blashfield 1984, p. 33.

[6] Attribute values are often standardized before Euclidean distance is calculated. Standardization is not necessary in SHYSTER as all of the attribute values are of the same type.

[7] In SHYSTER, n is the number of known pairs of attribute values. Attribute pairs with one or two UNKNOWNs are ignored for these purposes; they are taken into account in the calculation of unknown distance.

Hence, for binary attribute values, the Minkowski metric reduces to

$$d_{jk} = {\Delta_{jk}}^{1/K}$$

where Δ_{jk} is the number of differences in the corresponding attribute values of case j and case k.

For increasing values of Δ_{jk}, d_{jk} always increases regardless of the value of K. As SHYSTER is concerned only with the relative distance between cases, it does not matter which value of K is used. Choosing $K = 1$ gives

$$d_{jk} = \Delta_{jk}.$$

This distance measure is a metric as it satisfies all four metric criteria. It ranges from 0 to n: the smaller the value of d_{jk}, the nearer case j is to case k.

2.2 Association coefficients

Association coefficients are used to describe similarity between cases with binary attributes. These coefficients are usually expressed in terms of a, b, c and d; the two-way association table in figure 1 defines these variables.

Many different association coefficients have been proposed—Aldenderfer and Blashfield say that there are more than thirty.[8] According to Everitt (1974) the reason for this proliferation of coefficients is:

> ... uncertainty over how to incorporate negative matches [d] into the coefficients, and also whether or not matched pairs of [attributes] are equally weighted, or carry twice the weight of unmatched pairs, or unmatched pairs carry twice the weight of matched pairs.[9]

Everitt gives several examples of association coefficients. Jaccard's coefficient,

$$S_{jk} = \frac{a}{a+b+c},$$

is one measure of the association between two cases j and k. It ignores d—the number of NO/NO pairs. This coefficient was developed for use in applications where it would be inappropriate to treat two cases as being similar for lacking the same features as well as sharing the same features.[10] Similarly,

$$S_{jk} = \frac{2a}{2a+b+c} \qquad \text{and} \qquad S_{jk} = \frac{a}{a+2(b+c)}$$

[8] Aldenderfer and Blashfield 1984, p. 28.

[9] Everitt 1974, p. 51.

[10] Gower's coefficient (another measure of similarity) is identical to Jaccard's coefficient in the binary case (Aldenderfer and Blashfield 1984, p. 31).

		case k YES	case k NO
case j	YES	a	b
	NO	c	d

a is the number of YES/YES pairs;
b is the number of YES/NO pairs; etc.

Figure 1: Two-way association table defining a, b, c and d.

ignore negative matches—and give double weight to matched pairs and unmatched pairs, respectively. And

$$S_{jk} = \frac{a}{a+b+c+d}$$

gives no positive significance to negative matches.

Coefficients such as these are inappropriate for use by SHYSTER because the occurrence of NO/NO pairs is as important as the occurrence of YES/YES pairs. An attribute value of NO does not indicate the lack of a feature; it means that the answer to the attribute's question is "no". Rephrasing attribute questions could turn all YES/YES pairs into NO/NO pairs.

The other two example coefficients that Everitt gives are

$$S_{jk} = \frac{a+d}{a+b+c+d} \qquad \text{and} \qquad S_{jk} = \frac{2(a+d)}{2(a+d)+b+c}.$$

The first, called the simple matching coefficient, is the ratio of matching pairs to non-matching pairs. The second gives matched pairs twice the weight of unmatched pairs.

Coefficients such as these are appropriate for use by SHYSTER because they do not distinguish between YES/YES pairs and NO/NO pairs.

As $\Delta_{jk} = a+d$ and $n = a+b+c+d$, coefficients of this kind can be generalized to

$$S_{jk} = \frac{K\Delta_{jk}}{(K-1)\Delta_{jk}+n}$$

where K is some constant. For increasing values of Δ_{jk}, S_{jk} always increases regardless of the value of K. Because SHYSTER is concerned only with relative measures of similarity, it does not matter which value of K is used. The simplest is $K = 1$, which gives

$$S_{jk} = \frac{\Delta_{jk}}{n}.$$

Note that n may vary within the same area of law, because n is the number of pairs of attributes with known values: i.e. YES or NO, but not UNKNOWN. Hence, S_{jk} is not proportional to d_{jk} as derived in §2.1. This association coefficient satisfies all four metric criteria. It ranges from 0 to 1: the smaller the value of S_{jk}, the nearer case j is to case k.

2.3 Correlation coefficients

According to Aldenderfer and Blashfield, the most popular correlation coefficient is Pearson's product-moment correlation coefficient.[11] This coefficient is defined as

$$r_{jk} = \frac{\sum_{i=1}^{n} \left(A_{ij} - \bar{A}_j\right)\left(A_{ik} - \bar{A}_k\right)}{\sqrt{\sum_{i=1}^{n} \left(A_{ij} - \bar{A}_j\right)^2 \sum_{i=1}^{n} \left(A_{ik} - \bar{A}_k\right)^2}}$$

where n is the number of attributes, A_{ij} is the value of the ith attribute for the jth case, and $\bar{A}_j$ is the mean of all attribute values for the jth case.

In general, this correlation coefficient has several limitations.[12] It is not a true metric.[13] Furthermore,

> ... the use of [this] method to calculate the correlation of cases does not make statistical sense, because one must obtain the mean value across different [attribute] types rather than across cases, as in the standard use of the method. The meaning of the "mean" across these [attributes] is far from clear.[14]

Although this is true in general, in SHYSTER all attributes are of the same type. Hence it *does* make sense to calculate the mean across the attributes.

Despite its drawbacks this coefficient has been widely used. It ranges from -1.0 to 1.0: the larger the value of r_{jk}, the nearer case j is to case k.

3 Weighting similarity measures

The question of whether to weight attributes in the calculation of similarity measures is controversial. Aldenderfer and Blashfield warn that,

> While the concept of weighting is simple, its practice is difficult, and very few guidelines exist. Williams [1971] describes five types of weighting, the most common being the *a priori* manipulation of [attributes]. Sneath and Sokal (1973) argue strongly against *a priori* weighting, and suggest that the appropriate way to measure similarity is to give all [attributes] equal weight.[15]

[11] Aldenderfer and Blashfield 1984, p. 22.

[12] Aldenderfer and Blashfield 1984, pp. 23–4; Everitt 1974, pp. 53–4.

[13] Identical cases do not have a coefficient of zero, and the coefficient often fails to satisfy the triangle inequality.

[14] Aldenderfer and Blashfield 1984, p. 23.

[15] Aldenderfer and Blashfield 1984, p. 21, emphases added.

However, Everitt points out that

> ... the consequences of choosing one or other of the plethora of similarity indices are in many cases equivalent to the adoption of different schemes of [attribute] weighting, and so the concept of "equal weighting" is not as simple as it seems at first sight. It should also be remembered that [attributes] not included in the analysis are effectively given a zero weighting compared with those included.[16]

It is beyond doubt that some attributes in a given area of law are more important than others. SHYSTER uses *a priori* weighting of attributes.[17] The known and unknown distances between cases are both sums of the appropriate attribute weights.

Each of the three similarity measures discussed above can be weighted to take account of the importance of each attribute.

The weighted distance measure d'_{jk} is

$$d'_{jk} = \sum_{i=1}^{n} \left| A_{ij} - A_{ik} \right| \times w_i$$

where w_i is the weight of the ith attribute. It ranges from 0 to $\sum_{i=1}^{n} w_i$: the smaller the value of d'_{jk}, the nearer case j is to case k.

The weighted association coefficient S'_{jk} is

$$S'_{jk} = \frac{\sum_{i=1}^{n} \left| A_{ij} - A_{ik} \right| \times w_i}{\sum_{i=1}^{n} w_i}.$$

It ranges from 0 to 1: the smaller the value of S'_{jk}, the nearer case j is to case k.

The weighted correlation coefficient r'_{jk} is

$$r'_{jk} = \frac{\sum_{i=1}^{n} \left(A_{ij} \times w_i - \bar{A}'_j \right) \left(A_{ik} \times w_i - \bar{A}'_k \right)}{\sqrt{\sum_{i=1}^{n} \left(A_{ij} \times w_i - \bar{A}'_j \right)^2 \sum_{i=1}^{n} \left(A_{ik} \times w_i - \bar{A}'_k \right)^2}}$$

where $\bar{A}'_j$ is the weighted mean of all attribute values for the jth case. It ranges from -1.0 to 1.0: the larger the value of r'_{jk}, the nearer case j is to case k.

[16] Everitt 1974, p. 50.

[17] SHYSTER assigns a weight to each attribute, equal to the inverse of the variance of the values of that attribute across all the cases. This approach is justified by Popple 1990, 1992.

4 Conclusion

Statisticians make use of many different similarity measures, but—for SHYSTER—most of the commonly used similarity measures reduce to one of these three formulae for the similarity/distance between two cases j and k:

- distance measure: $$d_{jk} = \Delta_{jk}$$
- association coefficient: $$S_{jk} = \frac{\Delta_{jk}}{n}$$
- correlation coefficient: $$r_{jk} = \frac{\sum_{i=1}^{n}\left(A_{ij} - \bar{A}_j\right)\left(A_{ik} - \bar{A}_k\right)}{\sqrt{\sum_{i=1}^{n}\left(A_{ij} - \bar{A}_j\right)^2 \sum_{i=1}^{n}\left(A_{ik} - \bar{A}_k\right)^2}}$$

where Δ_{jk} is the number of differences in the corresponding attribute values of case j and case k, n is the number of attributes, A_{ij} is the value of the ith attribute for the jth case, and $\bar{A}_j$ is the mean of all attribute values for the jth case.

These measures can be weighted to take into account the importance of each attribute:

- weighted distance measure: $$d'_{jk} = \sum_{i=1}^{n}\left|A_{ij} - A_{ik}\right| \times w_i$$
- weighted association coefficient: $$S'_{jk} = \frac{\sum_{i=1}^{n}\left|A_{ij} - A_{ik}\right| \times w_i}{\sum_{i=1}^{n} w_i}$$
- weighted correlation coefficient: $$r'_{jk} = \frac{\sum_{i=1}^{n}\left(A_{ij} \times w_i - \bar{A}'_j\right)\left(A_{ik} \times w_i - \bar{A}'_k\right)}{\sqrt{\sum_{i=1}^{n}\left(A_{ij} \times w_i - \bar{A}'_j\right)^2 \sum_{i=1}^{n}\left(A_{ik} \times w_i - \bar{A}'_k\right)^2}}$$

where w_i is the weight of the ith attribute, and $\bar{A}'_j$ is the weighted mean of all attribute values for the jth case.

When choosing an appropriate similarity measure for a particular set of data, it is important to have regard to the sort of data, and to what is to be measured. However,

> ...when all is said and done, the validation of a similarity measure ... in a given field has so far been primarily empirical ...[18]

[18]Sneath and Sokal 1973, p. 146.

SHYSTER bases its opinions on the known and unknown distances. Both of these distances are variations on the weighted distance measure d'_{jk}. SHYSTER also calculates values of d_{jk}, S_{jk}, S'_{jk}, r_{jk} and r'_{jk}. These five similarity measures are used only when SHYSTER has difficulty determining which is the most similar case using the known and unknown distances.

As detailed by Popple (1992), early testing shows that SHYSTER's approach to finding the distance between cases allows it to produce useful advice in two different areas of case law. Future work will attempt to validate empirically SHYSTER's use of known and unknown distance to choose the cases upon which to base its opinion.

Acknowledgements

The author would like to thank David Cullen, Ross Cunningham and Brendan McKay for their comments and suggestions.

References

ALDENDERFER, Mark S. and BLASHFIELD, Roger K. 1984, *Cluster Analysis*, Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 07-044, Sage, Beverly Hills. ISBN 0-8039-2376-7.

EVERITT, Brian 1974, *Cluster Analysis*, Social Science Research Council Reviews of Current Research, no. 11, Heinemann Educational Books, London. ISBN 0-435-82297-7.

POPPLE, James 1990, *SHYSTER: A Hybrid Legal Expert System*, Technical Report TR-CS-90-06, Department of Computer Science, Australian National University, Canberra.

POPPLE, James 1991, *Detecting Attribute Dependence in the SHYSTER Legal Expert System*, Technical Report TR-CS-91-04, Department of Computer Science, Australian National University, Canberra.

POPPLE, James 1992, *SHYSTER and the Authorization of Copyright Infringement*, Technical Report TR-CS-92-08, Department of Computer Science, Australian National University, Canberra, May.

SNEATH, Peter H. A. and SOKAL, Robert R. 1973, *Numerical Taxonomy*, W. H. Freeman, San Francisco. ISBN 0-7167-0697-0.

WILLIAMS, W. T. 1971, "Principles of clustering", *Annual Review of Ecology and Systematics*, vol. 2, pp. 303–26. ISSN 0066-4162.