Case Base Adaptation Using Solution-Space Metrics*

Brian Knight
University of Greenwich
School of Computing & Mathematical Sciences
London, SE10 9LS, UK
b.knight@gre.ac.uk

Fei Ling Woon
Tunku Abdul Rahman College
School of Arts & Science
Kuala Lumpur, Malaysia
f.woon@gre.ac.uk

Abstract

In this paper we propose a generalisation of the k-nearest neighbour (k-NN) retrieval method based on an error function using distance metrics in the solution and problem space. It is an interpolate method which is proposed to be effective for sparse case bases. The method applies equally to nominal, continuous and mixed domains, and does not depend upon an embedding n-dimensional space. In continuous Euclidean problem domains, the method is shown to be a generalisation of the Shepard's Interpolation method. We term the retrieval algorithm the *Generalised Shepurd Nearest Neighbour* (GSNN) method. A novel aspect of GSNN is that it provides a general method for interpolation over nominal solution domains. The performance of the retrieval method is examined with reference to the Iris classification problem, and to a simulated sparse nominal value test problem. The introduction of a solution-space metric is shown to out-perform conventional nearest neighbours methods on sparse case bases.

1 Introduction

We present in this study a Case-Based Reasoning (CBR) retrieval method that utilises a distance metric imposed on solution space. The motivation for such a method is to extend a powerful interpolative method, already proven in the real domain, so that it applies equally in the domain of nominal values. Interpolative methods are well studied in the real domain, and can give good results from relatively sparse datasets. However, no general interpolative method exists for nominal (discrete) solution domains.

* The support by the University of Greenwich and Tunku Abdul Rahman College (TARC) is acknowledged.

2 An Error Function

The GSNN method will be applied to a general class of problem and solution domains. A distance metric $d_x(x,x_i)$ is here defined on the problem domain X and $d_y(y,y_i)$ on the solution domain Y. For the problem space, the term $\| x - x_i \|$ in the Shepard's method [1968] is generalised to $d_x(x,x_i)$ over X. For the solution space Y, $d_y(y,y_i)$ is used where $y=f(x)$ is the value of y which minimizes the error function:

$$I(y) = \sum_1^n \| y - y_i \|^2 \| x - x_i \|^{-p} / \sum_1^n \| x - x_i \|^{-p}$$

$$\partial I / \partial y = 0 \quad iff \quad y \sum_1^n \| x - x_i \|^{-p} = \sum_1^n \| x - x_i \|^{-p} y_i$$

That this is a minimum follows from the positive definite form: $\partial^2 I / \partial y^2 = \sum_1^n \| x - x_i \|^{-p} / \sum_1^n \| x - x_i \|^{-p}$

The function $I(y)$ depends only upon the Euclidean distance over Y = R and X = R^d. In order to generalise the method completely, we propose the error function:

$$I(y) = \sum_1^k d_y^2(y,y_i) d_x(x,x_i)^{-p} / \sum_1^k d_x(x,x_i)^{-p} \qquad a)$$

Here, the set $\{x_1,x_2,...,x_k\}$ are the k nearest neighbours in the problem space to the point x. $d_x(x,x_i)$ and $d_y(y,y_i)$ are distance on domains $x \in X, y \in Y$. The retrieved value y is the value $y \in Y$ which minimizes the error function /. The GSNN algorithm is given as follows:

$$\hat{f}(x_q) \leftarrow \arg\min_{y \in Y} \sum_{i=1}^{k} w_i d_y^2(y, f(x_i)) \text{ and } w_i = d_x(x_q,x_i)^{-p}.$$

3 Illustrative Example

In this example we illustrate in detail how the method works. We choose the Iris data set [Fisher, 1936]. The problem space is X and $x = (x_1, x_2, x_3, x_4)$ is a point in X. The solution space Y= {setosa, versicolour, virginica}. For the problem space we define distance according to a weighted sum of attributes. For the Y space, we define $d_y(y,y')$ by using the distances between cluster centres to represent the distance between the classes. These distances are shown in the following matrix:

	setosa	versicolour	virginica
setosa	0	.35	.49
versicolour	.35	0	.18
virginica	.49	.18	0

We take two cases, one from *setosa* and one from *virginica:*

x_j= *(44 2.9, 1.4, 0.2),* y_i= *setosa*

x_2 = *(7.2, 3.2, 6, 1.8),* y_2 = *virginica*

We take as target the *versicolour* iris:

x = *(5.5, 2.3, 4, 1.3),*y = *?*

Taking p=l and k = 2, the function *l(y)* is:

$$l(y) = \sum_1^2 d_y^2(y, y_i) d_x(x, x_i)^{-1} / \sum_1^2 d_x(x, x_i)^{-1}$$

$$= (0.36)^{-1} d_y(y, setosa)^2 + (0.35)^{-1} d_y(y, virginica)^2 / (0.36)^{-1} + (0.35)^{-1}$$

Since *I (versicolour)* is minimum, we take y=*versicolour* as the estimated value. This example shows an advantage of the interpolation method in situations where cases are sparse, in that it can correctly predict nominal values not represented in the case base itself.

4 Test on a Simulated Case Base

To examine how the GSNN method might work on real case bases, we simulated case bases of varying density and structure, and used the method to estimate simulated target sets. As a basis for the simulation, we adapted the function used by Ramos and Enright [2001] (i.e. $y = Int(10 \sin 2\pi x_1 * \sin 2\pi x_2)$) to give 21 nominal values, yi,…,y_2i. These 21 nominal values inherited a distance metric from the numeric values: $d(y_i, y_j) = |y_i - y_j|$.

Test 6.1 uses regularly spaced cases at various case densities. This might represent a well organised case base. Test 6.2 uses randomly selected cases, and is intended to represent disorganised sparse case bases. Cases (x_1, x_2, y) are constructed in the domain: $0 \le x_1, x_2 \le 1$, over a regular square lattice, with 10^2,20^2,30^2 points.

[Size	Methods	k=l	k=2	k=3	k=4
100	GSNN	709	501	453	539
	k-NN		769	799	786
	DWNN		710	706	709
400	GSNN	501	308	188	215
	k-NN		616	684	652
	DWNN		499	478	488
900	GSNN	360	251	181	224
	k-NN		450	471	456
	DWNN		360	344	344

Table 1. Errors in estimating a test set of 1000 targets, for regular case bases.

Table 1 shows the result of Test 6.1. These results confirm that GSNN with k > 1 can out-perform both k-NN and DWNN [Mitchell, 1997] for case bases with regular structure. Table 2 shows the results of Test 6.2. The results show that more errors are recorded for random case bases than for regular case bases of equivalent size, whatever the value of k. Once again, the results show that GSNN out-performed the other nearest neighbour methods.

[Size	Methods	k=l	k=2	k=3	k=4
100	GSNN	734	663	653	678
	k-NN		772	843	843
	DWNN		733	737	739
400	GSNN	573	506	511	492
	k-NN		643	695	708
	DWNN		573	583	591
900	GSNN	421	356	359	344
	k-NN		486	547	548
	DWNN		422	435	432

Table 2. Errors in estimating a test set of 1000 targets, for random case bases.

5 Conclusion

In this paper, we have proposed a method for interpolation over nominal values. The method generalises the Shepard's interpolation method by expressing it in terms of the minimization of a function *l(y).* This function relies only on distance metrics defined over problem and solution spaces. The method has an advantage for CBR in that it is applicable to case bases with nominal values in the problem and solution domain where no natural ordering exists. The examples studied indicate that GSNN could be useful in CBR with a sparse set of cases, and particularly where the cases can be organised. Tests show that GSNN is more efficient as a retrieval engine than other nearest neighbour methods. The inclusion of a solution space metric in the GSNN technique could be useful in two areas of CBR: (i) The selection of an optimum case base, (ii) Case based model building, from experimental or numerical modeling exercises. Investigations using numerical models indicate at GSNN would appear to be a promising approach for the construction of efficient case-based models.

References

[Fisher, 1936] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annual Eugenics,* 7, Part II, 179-188(1936).

[Mitchell, 1997] T. Mitchell. Machine Learning. *McGraw-Hill Series in Computer Science,* WCB/McGraw-Hill, 230 - 247, USA 1997.

[Ramos and Enright, 2001] G. A. Ramos and W. Enright. Interpolation of Surfaces over Scattered Data. *Visualization, Imaging and Image Processing Conference, VIIP02.* 2001 IASTED

[Shepard, 1968] D. Shepard. A Two-Dimensional Interpolation Function for Irregularly Spaced Data. *Proceeding of the 23rd National Conference,ACM,517-523,* 1968.