Case Selection and Interpolation in CBR Retrieval

Brian Knight, Miltos Petridis, Fei Ling Woon,
University of Greenwich, CMS, London, UK

Abstract. In this paper, several existing interpolation methods of use in CBR are discussed and compared. Interpolation in CBR is normally applied to a retrieval set of cases which are 'near to' a given target set in the problem domain. The interpolation method is then used to select an appropriate solution value from a solution domain. The main factors examined here, governing the accuracy and power of the interpolation, are the selection of cases for interpolation and the method of interpolation. Two selection criteria are examined: selection by nearest neighbours and selection by divergence algorithms. Three interpolation methods examined are examined: nearest neighbour, distance weighted nearest neighbour, linear regression and a generalised regression method, suitable to nominal values. Experimental results on three case-bases are presented for comparison. These are a: a real valued 2-dimensional sinusoidal random valued function, the classical iris case base, and the travel case base. The results show that linear regression is best for dense case bases, but is limited to real continuous problems. For general CBR usage, divergence selection can improve accuracy by a factor of 2, and that generalised regression can additionally improve accuracy also by a factor of 2.

Keywords: Case-Based Reasoning, Interpolation, Regression

1 Introduction

The objective of the experiments reported here has been to estimate the impact of various interpolation methods in estimating a solution for a given target from a variety of case bases. We used four different case bases: a standard smoothly varying function proposed by Ramos and Enright [Ramos and Enright, 2001], the classic iris case base, the travel case base [http://www.ai-cbr.org/cases.html, and a commercial sales database.

The accuracy of any interpolation depends upon how we select cases for the interpolation. Initially, we might just take the nearest neighbours and use these.

The experiments have been aimed at comparing 2 factors:

1. The selection method for cases with which to interpolate. This can be using k nearest neighbours, or kNN in conjunction with a diversity algorithm
2. Interpolation method. This can be distance weighted nearest neighbour, or generalised regression. In fact, we also examined 1NN, but this in all experiments is worse than DWNN.

2 Test of Interpolation methods in real domain

This case base is a random selection of points over the unit square as problem space. The solution space is generated using equation (1). For this case base, we expect that increasing the size of the case base will reduce the error, eventually towards zero, whatever the interpolation method.

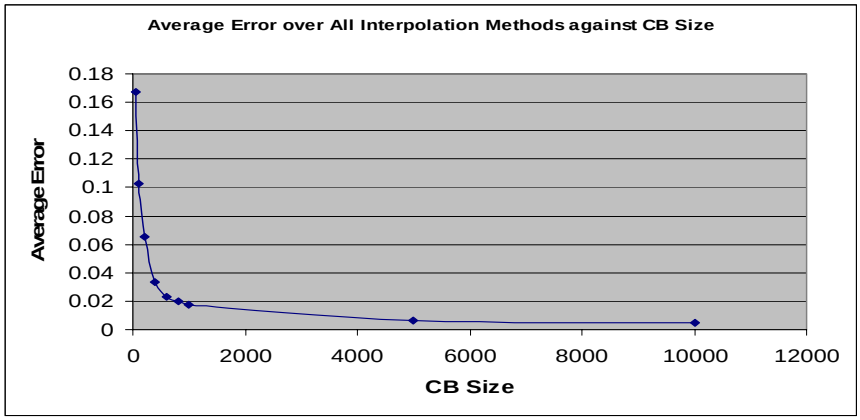

Figure 1

Figure 1 shows the average behaviour of error against CB size. Error here is calculated using leave-one-out. We would expect that a good interpolation method will allow a much smaller CB for a given error tolerance.

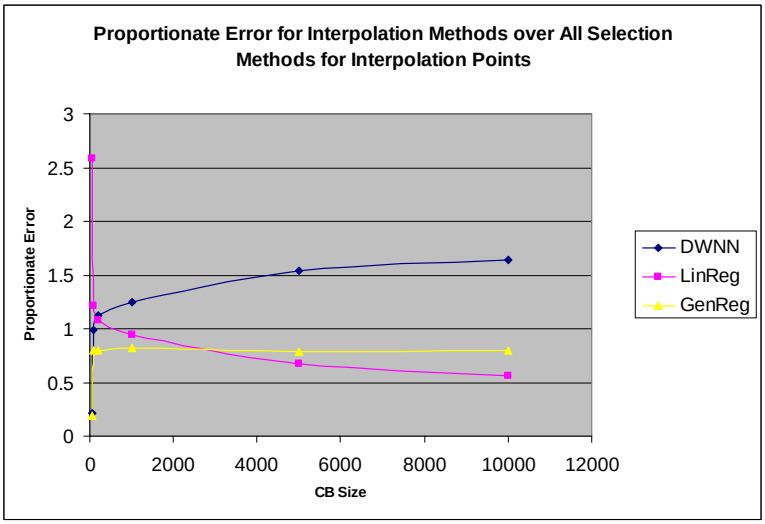

Figure 2

Figure 2 show the effectiveness of the interpolations regardless of selection method. These results show that for large case bases, as we would expect, linear

regression does best. Generalised regression is not quite so accurate , and DWNN is over twice the error. For very small case bases however, DWNN and GR outperform linear regression. The reason for this is that the linear assumption is not a good one for this function over large areas of the unit square. We could conclude that GR and DWNN are better methods in rapidly changing areas of a case base.

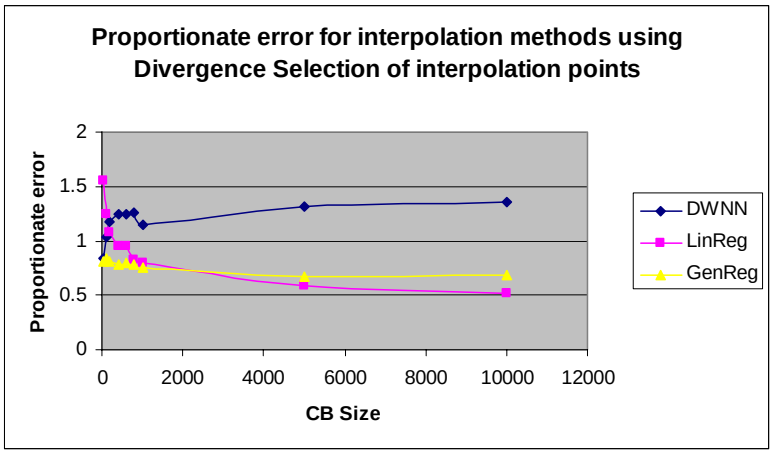

Figure 3

However GR does very much better than DWNN. In fact an error reduction to 60% of DWNN Error.

Figure 3 shows the error for the three interpolation methods using a divergence selection method. Surprisingly, GR seems to do even better for divergent selection sets than for average selection sets. This is a surprise because the way GR works seems to be rather independent of selection set.

Notice also, that GR does very much better than DWNN **and** Linear Regression for small case bases.

In the case of divergent selection, GR does even better, approaching Linear Regression in performance.

We see that using GR interpolation gives 50% of the error of DWNN. To put this in perspective, this is equivalent to more than doubling the size of the case base.

Figure 4 shows a summary of the effect of diversity selection over all the interpolation methods and case base sizes. This graph shows how much better selecting on a divergent set is to the usual kNN set. Divergence selection is nearly twice as good for large case bases.

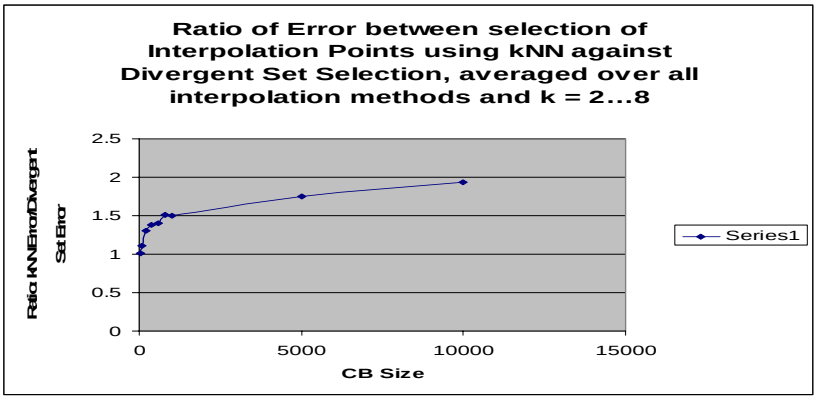

Figure 4: Impact on interpolation error, by using kNN selection compared to Divergent set selection

Test of Interpolation methods for the Iris Case Base

. In this section we illustrate the interpolation procedures with reference to the well-known Irises classification problem [Fisher, 1936]. In this database, there are 50 cases of each of 3 types of iris : setosa, virginica and versicolor. Each case is characterized by four real attributes representing petal and sepal width and length.

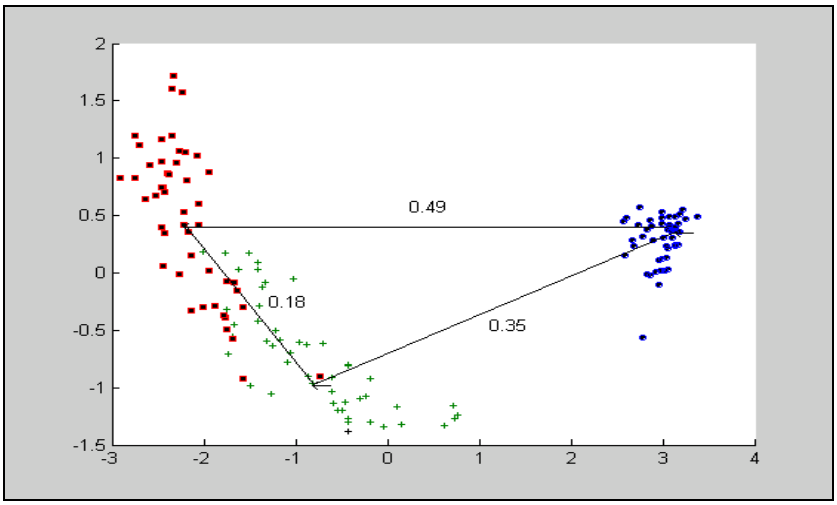

Figure 5: a principal components plot of the case base.

For the iris case base, practically any method gives good results for large case bases. However for small case bases, GR seems to outperform DWNN quite considerably. Random 10 fold cross validation tests were performed. The results are shown in figure 6. They show that selecting diverse interpolation points and using generalised regression give better accuracy than using nearest neighbour selection and DWNN. Diverse selection does not seem to help DWNN. This is probably because diversity will introduce more distant cases into the interpolation set. Because of the inverse distance weighting, DWNN will tend to ignore these extra cases. On the other hand, GR does not use distance weighting, and will use the extra cases on an equal basis.

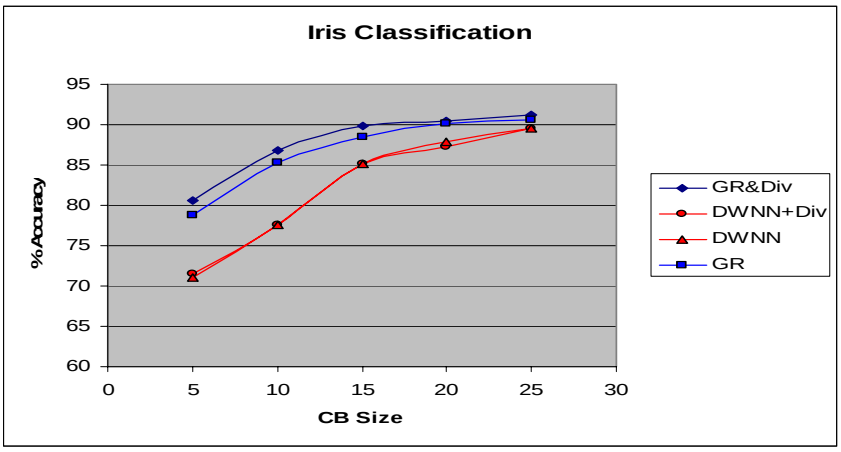

Figure 6 Interpolation on the Iris database

Test of Interpolation Methods using the Travel Case-base

In this section, the interpolative method is tested on a benchmark case base from the travel domain [http://www.ai-cbr.org/cases.html]. The problem investigated here is that of predicting a hotel for a given package holiday. We divide the domain attributes into the problem domain: X = {holiday type, destination region, duration, accommodation type, …}, and the solution domain Y = {Hotel}. The case base consists of 1024 package holidays. For the problem space we define distance according to a weighted sum of attributes with equal weight. For the Y space, we derive a metric on Y defined by its region and class of accommodation.

We use bounded-greedy diversity technique proposed by Smyth and McGinty [Smyth and McGinty 2003] to generate a diverse set of candidate cases. We then used the diverse set for interpolation, using DWNN and GR.

From the original case base we have chosen 1000 cases for experiments. 300 cases are chosen randomly as target problems. These cases are unseen target problems, not in the case base. The remaining 700 cases are used to form experimental case bases. We divide 700 cases into 7 independent case bases ranging from 100, 200, 300 to 700. This enables us to examine the predictive power of each retrieval method using various case base sizes, on the 300 unseen target problems. Following Smyth & McKenna [Smyth and McKenna 1998], we use a similarity threshold as criterion for correct prediction. If the predicted value is within the similarity threshold, that counts as correct prediction. In the experiments below, we take the threshold as 100%.

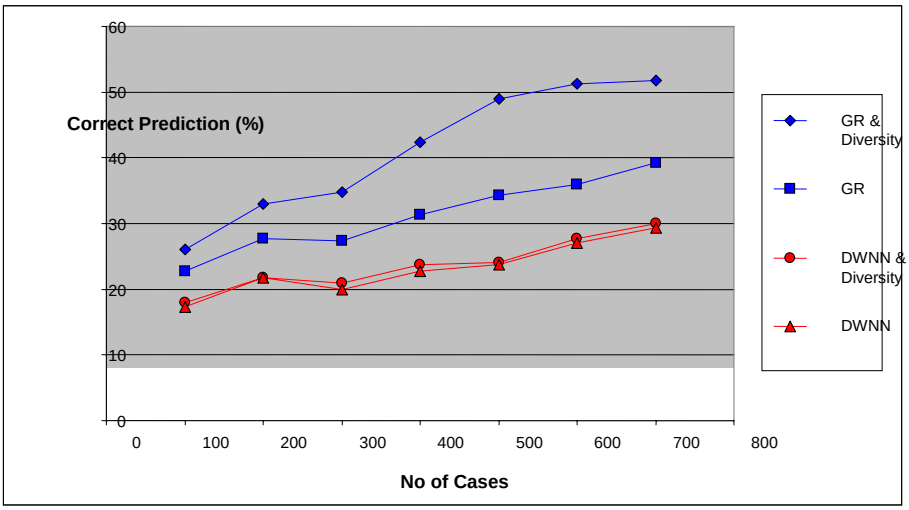

Figure 7. Comparing the correct prediction accuracy (%) of retrieval methods using both diverse retrieval sets and nearest neighbour retrieval sets, on 300 unseen target problems

Test of Interpolation Method on a prediction problem

Another test of the method has been conducted during a business intelligence project, which formed part of a collaboration between Greenwich University and Paperflow Ltd. This research has been conducted under the UK Department of Trade and Industry Knowledge Transfer Partnership scheme 10100. This test involved an operational customer database, and involved the classification of customer based on web activity over a period of time. The problem was to predict monthly order quantities of fast moving items, from previous monthly order frequencies. The solution domain here is real, but the problem domain is a mixture of integer order quantities, and discrete variables such as month and product. Figure 8. Shows the Accuracy of the three methods of interpolation.

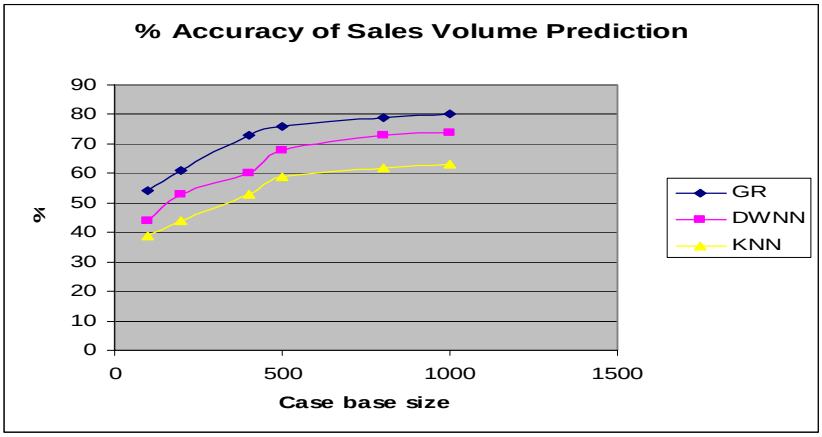

Figure 8: Accuracy of interpolation methods on Product Sales database

Conclusions

In this paper, we have described experiments relating to the accuracy and efficiency of methods of interpolation. The experiments have examined both the selection method for cases to form a basis for interpolation and the interpolation scheme itself. The conclusions are that a diversity scheme for selection and the method of generalised regression can improve the can improve accuracy independently of each other, and thus represents an optimum choice overall.

References

[Knight and Woon 2003] Knight B and Woon F L Case Base Adaptation Using Solution-Space Metrics. IJCAI 2003: 1347-1348

[Knight & Woon 2004] Knight B & Woon F L *"Case Based Adaptation Using Interpolation Over Nominal Values"*, Proceedings of **AI-2004,** The 24th Specialist Group on Artificial Intelligence (**SGAI**) International Conference on Innovative Techniques and Applications of Artificial Intelligence, Research and Development in Intelligent Systems XXI, Cambridge, UK, December 2004, pp.73-86, Springer-Verlag, London, UK.

[Ramos and Enright, 2001] Ramos, G. A. & Enright, W. (2001) Interpolation of Surfaces over Scattered Data. *Visualization, Imaging and Image Processing*

VIIP2001, *Proceedings of IASTED*, Marbella, Spain, 3-5 September. ACTA Press, pp.219-224

[Smyth and McKenna 1998]. Smyth, B. & McKenna, E Modelling the Competence of Case-Bases. Proc of 4th European Workshop on Case-Based Reasoning, EWCBR-98, Dublin, Ireland, September. Springer-Verlag, Berlin, 1998, pp 208-220 (Lecture Notes in Artificial Intelligence no. 1488)

[Smyth and McGinty 2003]. Smyth B. and McGinty L The Power of Suggestion. Proceedings of 18th International Joint Conference on Artificial Intelligence IJCAI-03, Acapulco, Mexico, 9-15 August. Morgan Kaufmann, San Francisco, CA, 2003, pp 127-132

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

[Smyth and McClave 2001]. Smyth B. and McClave P Similarity vs. Diversity. Proc of 4th International Conference on Case-Based Reasoning, ICCBR-01, Vancouver, BC, Canada, July/August. Springer-Verlag, Berlin, 2001, pp 347-361 (Lecture Notes in Artificial Intelligence no. 2080)