Knowledge Intensive Exception Spaces

Sarabjot S. Anand, David W. Patterson, John G. Hughes
School of Information and Software Engineering
University of Ulster at Jordanstown,
Newtownabbey, County Antrim,
Northern Ireland
e-mail: {ss.anand, wd.patterson, jg.hughes}@ulst.ac.uk

Abstract

In this paper we extend the concept of exception spaces as defined by Cost and Salzberg (Cost and Salzberg, 1993), in the context of exemplar-based reasoning. Cost et al. defined exception spaces based on the goodness, in terms of performance, of an exemplar. While this is straightforward when using exemplars for classification problems, such a definition does not exist for regression problems. Thus, firstly we define a measure of goodness of an exemplar. We then use this measure of goodness to compare the effectiveness of exception spaces with a variant that we introduce, called Knowledge Intensive Exception Spaces or KINS. KINS remove the restriction on the geometric shape of exception spaces as defined by Cost et al. We provide a rationale for KINS and use a data set from the domain of colorectal cancer to support our hypothesis that KINS are a useful extension to exception spaces.

Introduction

Instance-based Reasoning (Aha, 1990) (or Exemplar-based (Cost and Salzberg, 1993) or Nearest Neighbour algorithms (Cover, 1967)) employ the principle of lazy learning. These algorithms delay generalisation until a prediction or query instance is presented to them. Thus, they are characterised by large storage requirements, small learning time requirement, large prediction time and robustness in dealing with noise.

Early algorithms using the Nearest Neighbour paradigms suffered from the curse of dimensionality (the presence of a number of irrelevant attributes) which had the dual effect of increased execution time of the algorithm as well as decreased predictive accuracy. These algorithms, known as the 1-Nearest Neighbour (1-NN) retrieved the most similar exemplar (based on the Euclidean distance metric) and allocated the class of the retrieved instance to the target instance, thus, they were not very effective at handling noise.

Since then a number of extensions have been proposed to make these algorithms more robust. Variants of the 1-NN algorithm have been developed along a number of different dimensions. The k-dimension generalises the prediction to be based on more than one of the nearest neighbours, introducing the need for techniques for using a number of possible outcomes (each associated with a different neighbour) to be combined into one. Thus, the concepts of the "most common" and "voting" based prediction techniques were developed. The votes associated with each of the neighbours are based on the distance between the target and retrieved instance, using a kernel function defined on the distance metric. Developments along this dimension lead to better noise tolerance and can be thought of as a generality control on the algorithm. The value of k is normally set using cross-validation (Wetteschereck, Aha and Mohari, 1997).

The attribute weights dimension attempts to deal with the curse of dimensionality. These variants of the k-NN are also known as the *weighted Nearest Neighbour algorithm (wk-NN)*. Wetterchek et al. (Wetteschereck, Aha and Mohari, 1997) provide a survey of various techniques for attribute weight selection for use in conjunction with the k-NN algorithm.

The exemplar weights dimension further enhances the accuracy and noise tolerance of the k-NN algorithm. The idea is to weight each exemplar based on its ability to reliably predict the outcome attribute of an unseen instance (Salzberg, 1990). Reliable exemplars are given small weights (close to 1) while unreliable instances are given large weights (> 1). The rationale is that unreliable exemplars represent either noise or "exceptions" - thus, an exemplar weight greater than 1 is assigned to define a small area within the feature space where generally accepted rules do not apply. Rather than using a continuous exemplar weight (Cost and Salzberg, 1993), Aha et al. (Aha and Kibler, 1989) suggest that exemplars should only be used if they have proven themselves on classifying training examples. The advantage of the approach by Cost et al. however, is that it defines a clustering of the exemplars identifying exceptional exemplars. Such identification of exceptional exemplars may be used to identify optimal exemplar or case bases (Anand et al., 1998).

The distance metric dimension attempts to remove deficiencies within traditional distance metrics especially with respect to handling symbolic attributes. Initial approaches to handling symbolic attributes were based around their conversion into a set of binary attribute values - one attribute for each symbol within the original

symbolic attribute. However, there are a number of problems with this approach. Firstly, there is an unnecessary increase in the number of attributes in the data set ("dimensionality explosion"). Secondly, the intuitive notion of attribute weights as signifying the significance of the attribute within the classification (or regression) problem was lost for the symbolic attribute as a weight was now associated with each value. Thus, it was not possible to compare the significance, for example, of Age versus site of tumour in predicting months of survival in cancer. Anand et al (Anand and Hughes, 1998) have already shown how using symbolic attributes can lead to biased retrieval of neighbours by utilising the overlap distance metric. Enhanced distance metrics that attempt to remove this bias, representing a truer neighbour representation, are enhancements along this dimension (Stanfill and Waltz, 1986, Anand and Hughes 1998, Cost and Salzberg, 1993).

This paper revisits and extends the exemplar weighting ideas presented by Cost et al. (Cost and Salzberg, 1993) and is as a proposed enhancement along the exemplar weights dimension.

Application Domain

The authors have recently been working towards building a prognostic model for colorectal cancer (Anand et al., 1998a). The data set consists of 134 colorectal cancer patients. All patients in this study presented with colorectal cancer between 1973 and 1983 in the Royal Victoria Hospital and the Belfast City Hospital, Belfast, Northern Ireland.

Attribute	Type	Ordered
Sex	Categorical	No
Pathological Type	Categorical	No
Polarity	Categorical	Yes
Tubule Configuration	Categorical	Yes
Tumour Pattern	Categorical	No
Lymphocytic Infiltration	Categorical	Yes
Fibrosis	Categorical	Yes
Venous Invasion	Categorical	Yes
Mitotic Count	Categorical	Yes
Penetration	Categorical	Yes
Differentiation	Categorical	Yes
Dukes Stage	Categorical	Yes
Age	Continuous	Yes
Obstruction	Categorical	No
Site	Categorical	No

Table 1: Description of Attributes in the Data Set

Complete clinical and pathological data were collected on these patients. For each case, details of age, sex, site and obstruction were collected. Histopathological grading according to Jass (Jass, 1987) and Dukes staging (Dukes, 1932) of each tumour was carried out by the same pathologist. Fifteen clinico-pathological features were recorded for each patient. These are described in Table 1. The objective was to use these features to induce a regression model that could predict the number of months the patient is expected to survive after diagnosis of colorectal cancer. Such a model, if accurate, could help in treatment planning and effective management of cancer patients.

Knowledge Intensive Exception Spaces

Cost et al. (Cost and Salzberg, 1993) suggested a scheme for exemplar weighting according to their performance history. In their scheme, the weight assigned to an exemplar was the ratio of the number of uses of the exemplar, to the number of correct uses of the exemplar.

The definition of a correct use of an exemplar is obvious in the case of a classification problem i.e. every time an exemplar is used to classify the target example into a correct class. However, in the case of a regression problem, such a definition is a little more difficult. In this section we will firstly discuss how we defined such a measure of goodness for regression problems. We then describe the concept of knowledge intensive exception spaces or KINS giving the rationale behind them and describing their geometrical interpretation and use.

Defining Goodness of Use of Exemplars

An initial cross-validation run of the k-NN algorithm using the Euclidean distance metric, retrieving five closest neighbours and using a voting scheme for making a prediction resulted in individual neighbour predictive errors shown in Figure 1. The k-NN algorithm used forms part of the MKS data mining toolkit (Anand et al, 1997) under development within the authors' laboratory.

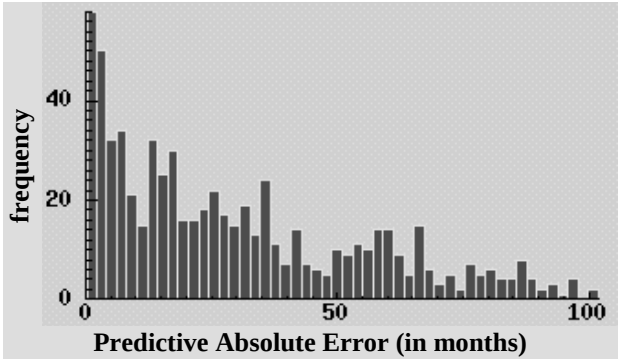

Figure 1: Distribution of Predictive Errors

The distribution of errors may be described using the statistical measure of spread namely quartiles. A *goodness membership function (gmf)* is defined on the quartiles as the degree to which a retrieval of an exemplar producing an error within that quartile may be regarded as a good use of the exemplar. For example, a membership function defined by the values {1, 0.7, 0.3, 0} would imply that a retrieval of an exemplar when it produces an error in the first quartile is certainly good, the second quartile is 0.7 in

magnitude of goodness and the third quartile is 0.3 in magnitude of goodness. Thus when calculating the goodness value or weight of the exemplar we now use the formula:

$$w_x = \frac{\sum_{i=1}^{4} n_x^{\,i}}{\sum_{i=1}^{4} \mu_i * n_x^i}$$

where, μ_i *is the membership value for quartile i*

and n_x^i *is the no. of retrievals of x in quartile i*

We refer to this as the *Weighted Quartile Measure* for allocating exemplar weights.

To illustrate the use of the above measure of goodness as well as to illustrate the effect of the definition of the gmf on performance of the measure we now use three different gmf definitions and see their effect on the exemplar base for colorectal cancer. The gmfs are defined as :

M1: {1, 0, 0, 0}; M2: {1, 1, 0, 0} and M3: {1, 0.7, 0.3, 0}

M1 identifies a usage of an exemplar as good only if it produces an error within the first quartile of the error distribution. M2 identifies as good usage of an exemplar any retrieval that produces an error less than the mean absolute error of the initial k-NN run, while M3 identifies good exemplar usage with varying degrees of magnitude for the first three quartiles.

Exemplar #	Error (in months)	Quartile	Mean Abs. Error
69	11	1	Good
69	0	1	Good
69	35	3	Bad
69	25	2	Good
69	9	1	Good
69	67	4	Bad
69	42	4	Bad

Table 2: Example Usage of an Exemplar

Given the usage of exemplar #69 in Table 2, the weights associated with the exemplar using these gmfs would be:

Using M1: $w_x = 7/3 = 2.33$

Using M2: $w_x = 7/4 = 1.75$

Using M3: $w_x = 7/(3*1+1*0.7+1*0.3+2*0) = 1.75$

Now let us observe the effect that using these different gmfs has on the usage of exemplars in the exemplar-base. Table 3 shows a sample of exemplar weights from the exemplar-base.

As can be seen from Table 3, the stricter weighting regime used in M1 results in higher weights for most exemplars. M2 and M3 produce similar weights, lower than those of M1.

Using M1, exemplar 17, 56 and 59 were not retrieved anymore. Table 4 summarises the exemplar usage for our examples in Table 3 when using weights based on M1. As can be seen, the reduction in usage of exemplars assigned large weights has resulted in an unwarranted increase in the incorrect usage of previously "good" exemplars like exemplar #8. Similar increases can be observed when using M2 and M3-based exemplar weights (Table 5 and 6).

gmf / Exemplar #	M1	M2	M3
8	1	1	1
17	3	2	1.67
40	2	1	1.17
49	1.75	1.16	1.4
50	1.8	1.14	1.26
56	4	4	4
59	4	2	2

Table 3 Exemplar Weights for a subset of Exemplars

Quartile / Exemplar #	1	2	3	4
8	5	5	4	11
17	0	0	0	0
40	1	0	0	0
49	2	1	1	1
50	0	1	0	0
56	0	0	0	0
59	0	0	0	0

Table 4: Exemplar Usage using M1-based Exemplar Weights

Quartile / Exemplar #	1	2	3	4
8	2	5	0	0
17	0	0	0	0
40	14	14	0	0
49	6	6	0	0
50	6	7	0	0
56	0	0	0	0
59	1	1	0	0

Table 5: Exemplar Usage using M2-based Exemplar Weights

Quartile / Exemplar #	1	2	3	4
8	3	5	1	11
17	0	0	0	0
40	9	4	1	2
49	4	1	1	1
50	7	5	0	3
56	0	0	0	0
59	1	1	0	1

Table 6: Exemplar Usage using M3-based Exemplar Weights

An important conclusion of these preliminary results is that the weighted quartile method for measuring goodness of an exemplar is a useful technique to use when defining exception spaces. However, if the membership function is not chosen sensibly the results can be sub-optimal as we saw in the case of the M1 and M3. While M2 produced the best results, in terms of the distribution of errors on the four error quartiles, there is no guarantee that its definition is optimal. Thus a useful addition to the weighted quartile

technique would be its coupling with some kind of optimiser like a genetic algorithm.

Local Exemplar Weights

As described by Cost et al., exemplar weights can be graphically represented as circular boundaries (assuming a 2-dimensional space) defined around exceptional exemplars. The weight assigned to the exemplar defines the radius of the circular boundary using an inverse relationship i.e. the larger the assigned weight, the smaller the radius of the circle. For the exemplar to be retrieved the target example must fall within its boundary region.

The use of a 'global' exemplar weight results in the restriction on the boundary area around the exemplar being circular. We now study the validity of this restriction and present a less rigid boundary to be defined around each exemplar. We refer to this as local exemplar weights as the weights associated with the exemplar vary based on different *retrieval circumstances*. The resulting exception spaces are called *Knowledge INtensive exception Spaces or KINS*, as their definition utilises knowledge about the exemplar performance under various circumstances.

As shown in Table 2, most exemplars are not "globally" bad predictors. As in the case of Exemplar #69, there are three of the seven instances where the exemplar has been used to produce low errors and to discriminate against it by associating a global weight may result in a wholly inappropriate exemplar being retrieved instead in cases where Exemplar #69 was actually appropriate. This can be seen from Tables 3, 4, 5 and 6. Thus, what is required is a method to identify the circumstances where Exemplar #69 is appropriate and when it is not and only assign a large weight to it when its use is inappropriate.

While attribute weights attempt to reconfigure the exemplar space so as to best fit the real-world process being modelled, the reason why exemplars may still be retrieved inappropriately has its roots within the use of the distance metric. Consider the following example where two exemplars are retrieved for the same target example. While one exemplar differs from the target on the attributes "Dukes Stage", "Configuration" and "Site", the other differs on "Tumour Pattern" and "Mitotic Count". It may be the case that due to the global summation used by the Euclidean distance metric, the two exemplars have the same distance from the target associated with them. However, the retrieval of the first exemplar may be wholly inappropriate as in this particular part of the exemplar space, the distance of the target from the exemplar with regards to "Dukes Stage" may invalidate its use. Thus, let us investigate the individual attribute distances for each use of Exemplar #69. Table 7 shows these distances. Columns relate to individual retrievals of Exemplar# 69.

Using these distances as independent variables we may now use a classifier to discover a set of rules that would discriminate between the different Error Quartiles that the usage of the exemplar will result in. The following rules were generated for exemplar #69 using the implementation of C4.5 (Quilan, 1992) within the CLEMENTINE Data Mining Toolkit[2]:

```
if age <= 0.01
then Quartile = 1
if venous = 0 and  age > 0.01
then Quartile = 3
if venous > 0 and  age > 0.01
then Quartile -> 4
Default: Quartile -> 1
```

The rule can now be interpreted as follows: Exemplar #69 is a good exemplar to use if the Age attribute is a near exact match. However, if this in not the case the use of the exemplar produces an error within the third quartile if venous invasion is a perfect match and if it doesn't match either then the error will be in the fourth quartile.

Now an exemplar weight consists of a quadruple representing a weight for each Error Quartile. Note once again that these weights need to be optimised to get the best results. Assume the quadruple to be (1,2,3,4), which by no means is optimal. Even with such a sub-optimal set of weights the results obtained were very encouraging (see next section). Note that a smaller weight is associated with the first quartile and the weights function monotonically increases as the error quartile increases. These rules and the quartile weights together define KINS.

Using such KINS implies that now when a target example is presented to the exemplar-base the following set of steps are followed to retrieve the nearest neighbours. Firstly, the Euclidean distance is calculated between the target example and each exemplar. Next, the attribute distances used in this calculation are used for predicting the Error Quartile or the "appropriateness" of the exemplar to be used in the present context. Finally, the weight associated with the predicted Quartile is multiplied with the Euclidean distance to arrive at a final distance measure for the exemplar.

Now, let us have a closer look at the geometric interpretation of KINS. Consider the case of an exemplar that always produces errors in the same quartile. Without loss of generality we may assume this quartile to be quartile 3. In this case, the only significant weight assigned to the exemplar is 3. As described by Cost et al., such a weight may be represented as a circular region around the exemplar. Thus, exception spaces are a special case of KINS when errors produced by the exemplar always lie in the same quartile of the error distribution. In the more general case, the definition of KINS is based on the inter-attribute distances, and exception spaces of different sizes are drawn around the exemplar. For example, if the age attribute of the target example is only at a small distance from exemplar #69 (<= 0.01), a very large space,

[2] The CLEMENTINE Toolkit is a registered trademark of Integral Solutions Limited, Basingstoke, England.

Attribute	Use#1	Use#2	Use#3	Use#4	Use#5	Use#6	Use#7
Sex	0	0	0	0	0	0	0
Pathological Type	0	0	0	0	0	0	0
Polarity	0	0	0	0	0	0	0
Tubule Configuration	0	0	0.111	0	0	0	0
Tumour Pattern	0	0	0	0	0	0	0
Lymphocytic Infiltration	0	0	0	0.25	0	0	0
Fibrosis	0.25	0	0.25	0	0	0	0
Venous Invasion	0	0	0	0.56	0	0.062	0.062
Mitotic Count	0.027	0.25	0.25	0.11	0	0.027	0.027
Penetration	0	0.062	0	0	0	0.062	0.062
Differentiation	0	0	0.25	0	0	0	0.25
Dukes Stage	0.062	0.25	0.062	0.062	0.062	0.25	0.062
Age	0.0008	0.0008	0.02	0.016	0.058	0.36	0.024
Obstruction	1	0	0	0	0	0	0
Site	0	0	0	0	0	0	0
Error Quartile	1	1	3	1	3	4	4

Table 7 Attribute Distances for each use of Exemplar #69

encompassing the entire exemplar space is defined. However, if age > 0.01 and venous invasion is a perfect match, the region defined as #69's exception space is much smaller. If Venous Invasion isn't an exact match, an even smaller space is defined. Thus, KINS define a much more complex space around exemplars than exception spaces, equivalent to a set of exception spaces.

Experimental Results

Initial tests to observe the effectiveness of KINS used the Euclidean distance metric, no attribute weights, a voting based prediction mechanism and 5 as the value of k. Ten-fold cross validation was used to arrive at the Mean Absolute Errors shown in Table 8. The three gmfs defined in the previous section were used as variants of the Exception Spaces. The aim being, to highlight the effect of the gmf definition on the performance of the algorithm. As can be from Table 8, the definition of the gmf does have an effect on the predictive accuracy of the model. Three variants of KINS were also tested using different exemplar weights for the error quartiles. Once again the weights do seem to have an effect on the predictive accuracy of the model but these differences seem to be much less significant than those in the case of Exception Spaces.

An interesting observation of the results in Table 8 is the large decrease in mean absolute error by using any form of exemplar weighting. This can be partly attributed to the nature of the data set, which is very sparse. Therefore, using no attribute weights was expected to produce large errors in prediction. Using a genetic algorithm to produce an optimal set of attribute weights arrived at the attribute weights shown in Table 9 (Anand and Hughes, 1998). The varied attributes weights explain the high error rate when using an unweighted k-NN.

Exemplar Weight Technique	Variant	Mean Absolute Error
None	-	31.91
Exception Spaces	M1	25.91
	M2	24.64
	M3	25.33
KINS	{1,2,3,4}	21.76
	{1,1,3,4}	21.67
	{1,1.5,3,4}	21.59

Table 8: Mean Absolute Error using different Exemplar Weighting Techniques

Attribute	Weight
Sex	0.74
Pathological Type	0.14
Polarity	0.06
Tubule Configuration	0.22
Tumour Pattern	0.19
Lymphocytic Infiltration	0.53
Fibrosis	0.7
Venous Invasion	0.61
Mitotic Count	0.46
Penetration	0.99
Differentiation	0.03
Dukes Stage	0.72
Age	0.51
Obstruction	0.72
Site	0

Table 9: Attribute Weights generated using a Genetic Algorithm

Table 10 summarises the results produced using exemplar weights along with the attribute weights in Table 9. An interesting result here is that when no exemplar weights are used the mean absolute error arrived at using attribute weights is lower than the mean absolute error produced

using exception spaces in both the unweighted and weighted cases. However, even when using the optimal attribute weights, the mean absolute error achieved by the KINS is lower.

Exemplar Weight Technique	Variant	Mean Absolute Error
None	-	23.21
Exception Spaces	M1	25.72
	M2	26.23
	M3	25.28
KINS	{1,2,3,4}	19.09
	{1,1,3,4}	18.94
	{1,1.5,3,4}	18.73

Table 10: Mean Absolute Error using different Exemplar Weighting Techniques

Conclusions and Future Work

In this paper we revisited exception spaces as defined by Cost and Salzberg. We re-enforced the advantages of assigning weights to exemplars, introducing a method for doing so in the case of regression problems. We then introduced a generalisation of exception spaces called KINS and provided empirical proof of how the definition and use of KINS within lazy learning can prove advantageous in terms of improved accuracy of the developed model.

The results presented in this paper are by no means an end in itself. The authors believe that the results presented in the paper have left a number of open questions that need answered as well as suggesting a number of further developments to the definition of KINS. Extensions to KINS include optimising the weights used for each quartile. Also the optimisation of the goodness membership function needs to be undertaken to enable a true measurement of the advantage of defining KINS rather than simple exception spaces.

The definition of KINS has its disadvantages that need to be further quantified. These include possible Over-fitting of the model produced and increased training costs. Another interesting question that is worth asking is whether the increased training costs still justify such an algorithm being termed as a lazy learning algorithm as well as the loss in terms of the advantages that accrue from lazy learning.

Acknowledgements

The authors would like to thank Dr. Peter Hamilton, Royal Victoria Hospital, for providing the data set used in the study as well as for his continued support and domain expertise during the course of this research.

References

Aha, D.; and Kibler, D. 1989. Noise Tolerant instance-based learning algorithms. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence*, 794-799. Melno Park, Calif..

Aha, D. 1990. A study of instance-based algorithms for supervised learning tasks, Ph.D. diss., University of California, Irvine.

Anand, S. S.; Scotney, B. W.; Tan, M. G.; McClean, S. I.; Bell, D. A.; Hughes, J. G.; and Magill, I. C. 1997. Designing a Kernel for Data Mining. *IEEE Expert* 12(2): 65 - 74.

Anand, S. S.; Patterson, D.; Hughes, J. G.; and Bell, D. A. 1998. Discovering Case Knowledge using Data Mining. In *Proceedings of the Pacific-Asia Conference in Knowledge Discovery and Data Mining*, Springer-Verlag,

Anand, S. S.; Hamilton, P.; Smith, A. E.; and Hughes, J. G. 1998a. Intelligent Systems for the Prognosis of Colorectal Cancer Patients. In *Proceedings of CESA Special Session on Intelligent Prognostic Systems*.

Anand, S. S.; and Hughes, J. G. 1998. Hybrid Data Mining Systems: The Next Generation. In *Proceedings. of the Pacific-Asia Conference in Knowledge Discovery and Data Mining*, Springer-Verlag.

Cost, S.; and Salzberg, S. 1993. A Weighted Nearest Neighbour Algorithm for Learning with Symbolic Features. *Machine Learning* 10: 57-78.

Cover, T.; and Hart, P. 1967. Nearest Neighbour Pattern Classification, *IEEE Transactions on Information Theory*, 13(1): 21-27.

Dukes, C. E. 1932. The classification of cancer of the rectum. *J Pathol Bacteriol*. 35: 323-332.

Jass, J.; Love, S.; and Northover, J. 1987. A new prognostic classification for rectal cancer. *Lancet*. 1333-1335.

Quinlan, J. R. 1992. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers.

Salzberg, S. 1990. *Learning with nested generalised exemplars*, Norwell, MA: Kluwer Academic Publications.

Stanfill, C.; and Waltz, D. 1986. Towards Memory-based Reasoning. *Communications of the ACM*. 29(12): 1213-1228.

Wettschereck, D. 1994. A study of distance-based machine learning algorithms, Ph.D. diss., Oregon State University.

Wettschereck, D.; Aha, D.; and Mohri, T. 1997. A Review of Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms, *Artificial Intelligence Review Journal*.