The Knowledge Engineering Review, Vol. 11:3, 1996, 245–252

An integrated approach for different attribute types in nearest neighbour classification

W. Z. LIU

Department of Information Science, University of Portsmouth, Locksway Road, Milton, Hampshire PO4 8JF, UK [Email: liuw@sis.port.ac.uk]

Abstract

The basic nearest neighbour algorithm works by storing the training instances and classifying a new case by predicting that it has the same class as its nearest stored instance. To measure the distance between instances, some distance metric needs to be used. In situations when all attributes have numeric values, the conventional nearest neighbour method treats examples as points in feature spaces and uses Euclidean distance as the distance metric. In tasks with only nominal attributes, the simple "over-lap" metric is usually used. To handle classification tasks that have mixed types of attributes, the two *different* metrics are simply combined. Work by researchers in the machine learning field has shown that this approach performs poorly. This paper attempts to study a more recently developed distance metric and show that this metric is capable of measuring the importance of different attributes. With the use of discretisation for numeric-valued attributes, this method provides an integrated way in dealing with problem domains with mixtures of attribute types. Through detailed analyses, this paper tries to provide further insights into the understanding of nearest neighbour classification techniques and promote further use of this type of classification algorithm.

1 Introduction

The fundamental problem of learning to classify objects has been tackled from many different angles, by reseachers in statistics, artificial intelligence and other fields. One such classification method is the nearest neighbour algorithm. The basic principle of nearest neighbour algorithms is to store a set of training instances and classify a new case by predicting that it has the same class as its nearest stored instance. The set of stored instances is usually called the *training set*. Each instance in the training set is called an *exemplar*. All the cases in the problem domain consist of a class indicator and values on a number of attributes. The machine learning task is to classify new cases (with unknown class membership), according to the information provided in the training set.

The main concepts of nearest neighbour techniques originated with the work by Fix and Hodges (1951, 1952) on non-parametric discriminant analysis. Since then, the study and application of nearest neighbour algorithms have been extensive. The research has led to great advances from the original technique. Some of these advances include edited and condensed nearest neighbour methods (Hart, 1968; Swonger, 1972; Tomek, 1976; Devijver and Kittler, 1980). Most of this work has been reviewed by Dasarathy (1991). Recently, more work has been undertaken by machine learning researchers, in attempting to handle some remaining problems, and enhance the classification performance of nearest neighbour algorithms. The aim of this paper is to review these.

1.1 Problems with conventional nearest neighbour classification algorithms

The basic idea underlying nearest neighbour classification techniques is to treat each case (instance) as a point in the feature space and to classify a new case according to the class of its nearest stored

instances, based on some distance measure. To compute the distance between two instances, the conventional nearest neighbour algorithm usually uses the Euclidean distance measure, which is defined as follows:

$$d = \sqrt{\sum_{i=1}^{N} d_i^2}$$

where N is the number of attributes and d_i is the value difference of the two instances on the ith attribute. For a numeric attribute, the value difference between two values is simply an arithmetic difference. (This is usually calculated on standardised variables rather than the original attributes, in order to put all the attributes on a sort of common scale. If this were not done, then attributes with large variance would dominate the distance computation.) For nominal attributes, the "overlap" metric (or Hamming distance) is usually used. This method simply gives a zero value difference if the two symbolic values match, and gives one otherwise.

There are a few problems working with this conventional approach. The first problem arises from the fact that the Euclidean distance between two instances is calculated from all the attributes indiscriminately, i.e., the technique really involves conditioning on *all* the available variables. Recently, Liu and White (1995) made a comparison between nearest neighbour and tree-based classification techniques, and showed that the nearest neighbour method can suffer from overfitting which, in turn, renders sub-optimal performances in domains where the number of variables is large or the variables are of unequal importance in discriminating between the classes. It is not difficult to see that the worst case would be the situation when a lot of attributes are present, and some of these attributes are irrelevant to the discrimination task. To overcome this problem, attributes really need to be *weighted* (either explicitly or implicitly), according to their discrimination power, in the calculation of the overall distance between two cases.

The second problem is that using the simple "overlap" metric to measure the difference between two values of a nominal attribute can often fail to capture the complexity and subtlety of the problem domain. This is because, even when two nominal values of an attribute are different, it does not mean that they provide different discriminative information. Thus, the approach of simply giving a zero value difference if two symbolic values match and one when they are different, is sub-optimal. As a result, it may yield poor classification performance (Cost and Salzberg, 1993).

To solve the second problem, Stanfill and Waltz (1986) defined a new Value Difference Metric (VDM), to replace the simple "overlap" metric, for measuring the distance between values of symbolic features. This measure is useful because it incorporates the discriminative information of the attributes into a distance metric. This measure was later modified and improved by Cost and Slazberg (1993). This new Modified Value Difference Metric (MVDM) takes into account of exemplar weights, and turns out to be more powerful in measuring the distance between values of symbolic features. Section 2 is devoted to the review of these metrics.

Detailed analyses of these metrics (reported in section 3) reveals that, although the VDM and MVDM do not define weights for attributes explicitly, attribute importance is, however, taken care of by the metric itself. This also provides further insights into the understanding of nearest neighbour classification techniques.

In a domain when numeric-valued attributes are involved, problems still remain. The VDM and MVDM cannot be used, in this case, simply because numeric attributes have too many possible values and some values in *new* cases may not even exist in the training set. One naive scheme of solving this problem is to discretise numeric-valued attributes by dividing them into equal intervals, and then use the MVDM as the distance metric for all the attributes regardless of their type. Recently, researchers such as Ting (1994) studied various methods of discretising numeric-valued attributes. It has been shown that using MVDM, after discretising all numeric-valued variables for domains with mixed type of attributes, can generally lead to better classification performance than using the conventional nearest neighbour techniques. In section 4, some appropriate discretising methods for numeric-valued attributes are described.

Table 1 A general contingency table

	a_1	a_2		a_m	
C_1	n_{11}	n_{12}	...	n_{1m}	n_{1*}
C_2	n_{21}	n_{22}	...	n_{2m}	n_{2*}
⋮	⋮	⋮	...	⋮	⋮
C_v	n_{v1}	n_{v2}	...	n_{vm}	n_{v*}
	n_{*1}	n_{*2}	...	n_{*m}	n_{**}

2 VDM and MVDM

2.1 The Value Difference Metric

The motivation for the development of the Value Difference Metric (VDM) by Stanfill and Waltz (1986) was to define a better metric to replace the Hamming distance, in order to measure the distance between values of symbolic attributes. This measure takes into account the overall similarity of classification of all cases, in the training set, for each possible value of each feature. This means that a matrix defining the distance between all values of an attribute can be derived statistically, based on the information provided by the examples in the training set.

Let us assume a problem domain with v classes $C_1, C_2, ..., C_v$ and N symbolic attributes $A_1, A_2, ..., A_N$. Let's also assume that attribute A_i ($i = 1, 2, ..., N$) has m different values. Based on information provided by examples in the training set, a general contingency table, representing the cross-classification of class and the attribute, can be constructed as shown in Table 1. In this table n_{ij} ($i = 1, 2, ..., v; j = 1, 2, ..., m$) represent the frequency counts of cases with class C_i and attribute value a_j; and:

$$n_{i*} = \sum_{j=1}^{m} n_{ij}$$

$$n_{*j} = \sum_{i=1}^{v} n_{ij}$$

$$n_{**} = \sum_{i=1}^{v} \sum_{j=1}^{m} n_{ij}$$

The distance d between any two values a_i, a_j is defined as:

$$d(a_i, a_j) = \sum_{k=1}^{v} \left| \frac{n_{ki}}{n_{*i}} - \frac{n_{kj}}{n_{*j}} \right| \tag{1}$$

Using this definition, a matrix of value differences for each attribute in the training set can be computed. It is not difficult to see that equation (1) defines a geometric distance (Cost and Salzberg, 1993), i.e., it has the following properties:[1]

- $d(a, b) \geqslant 0, a \neq b$
- $d(a, b) = d(b, a)$
- $d(a, a) = 0$
- $d(a, b) + d(b, c) \geqslant d(a, c)$

The total distance D between two instances is defined as follows:

[1]Perhaps it should be noted that Cost and Salzberg made a mistake in stating that $d(a, b) > 0$ when $a \neq b$. The distance between two different values a and b, as defined in equation (1) can be zero when a and b have exactly the same relative frequencies for all classes.

$$D(X, Y) = \sum_{i=1}^{N} d(x_i, y_i) \tag{2}$$

where X and Y represent two cases. x_i and y_i denote the values of the ith attribute for X and Y, where each instance has N features.

When a new case (with unknown class membership) is tested, distances from this case to each of the cases in the training set are calculated and compared. The class of this case can then be predicted according to the class of the training case with the closest distance from it, as defined in equation (2). In the case of K Nearest Neighbour (KNN), the majority class of these K nearest instances is assigned.

2.2 *The Modified Value Difference Metric*

Cost and Salzberg (1993) modified the value difference metric and proposed the MVDM. The only difference between the VDM and the MVDM is that the latter takes exemplar weight into account. These are based on the idea that some instances in the training set are more reliable classifiers than others. Intuitively, one would like these trustworthy exemplars to have more power than others.

Taking this idea into account, the definition of distance between two instances X and Y now becomes:

$$D(X, Y) = w_X w_Y \sum_{i=1}^{N} d(x_i, y_i) \tag{3}$$

where w_x and w_y are weights assigned to X and Y, respectively. For a new example Y, $w_y = 1$.

In the MVDM, reliable exemplars are given smaller weights, making them appear closer to a new test case with unknown class membership. Basically, the weighting scheme used by Cost and Salzberg (1993) was first adopted in the EACH system (Salzberg, 1989, 1990). The idea is to assign weights to exemplars according to their performance history. The exemplar weight, w_x, is defined as the ratio of the number of uses of that exemplar to its number of *correct* uses. Thus, accurate exemplars will have $w_x \approx 1$. Unreliable exemplars will have w_x bigger than 1, making them appear further away from a new case. The more times an exemplar gives an incorrect classification, the larger its weight grows. If we assume that e is an exemplar, in the training set, with the following two counters:

- *e.used* represents the number of times it is found to be a best match (i.e., the nearest neighbour of a case undergoing classification);
- *e.corr* represents the number of times it is found to produce a correct classification,

then the procedure to produce and modify weights can be described as follows:

1. Find best match e in memory for new instance i. If memory is empty, add i to the memory with initial $e.used = e.corr = 1$. If this case is a best match the next time, its weight is regarded as 1. Continue with next instance.
2. Increase *e.used* by 1.
3. If match is a correct classification, increase *e.corr* by 1.
4. Set $i.used = e.used$ and $i.corr = e.corr$.
5. Add i to the memory and go to the next instance.

The nearest neighbour classification system designed by Cost and Salzberg requires two passes through the training set. The first pass is to construct value difference tables for all the attributes, according to equation (1). In the second pass, the algorithm attempts to classify each case by assigning a classification to it according to its closest neighbour (or neighbours) based on the distance measure defined in equation (3). The system then checks to see if the classification is correct, and uses this information to adjust the weight on the old instance (and initialise the weight for the new case as well), according to the procedure described above. Finally, the new case is stored in

memory. During testing, instances are classified in the same manner, but no modifications are made to memory or to the value difference tables.

3 Importance of attributes

In practice, the method discussed in the previous section, has been shown empirically to be very powerful in dealing with problem domains in which all attributes are symbolic (Cost and Salzberg, 1993; Rachlin et al., 1994). It is also not difficult to see the reason, in theory. By simple mathematical manipulations, the following theorem can be proved.

theorem 1

The distance between any two values of a nominal attribute, as defined in equation (1), *is greater than or equal to zero and less than or equal to* 2, *i.e.*, $0 \leqslant d(a_i, a_j) \leqslant 2$. *The minimum distance is zero only when the two values occur with exactly the same relative frequency for all classifications. The maximum distance occurs when, for all classes, each class co-occurs with only one of the values.*

The proof of this theorem is trivial and is omitted here.

This theorem is useful because it reveals that the distance metric defined in equation (1) can measure the importance of the attributes. On the one hand, when two values of the same attribute occur with the same relative frequency for all classifications, the value difference metric as defined in equation (1) assigns a zero for the distance between these two values. This is just what we expected. If the two values have more or less the same probability of occurrence for all the classes, then it is a sign of lack of association between the attribute and the class, at least for the two values concerned. In such a situation, we would certainly expect that the *superficial* difference between the two values of this particular attribute should contribute little to the computation of the overall distance between these two instances. This provides a useful means of tolerating irrelevant attributes in the distance computation.

On the other hand, when two values of the same attribute occur with substantially different relative frequencies for the different classes, the MVDM regards it as a "signal" that this attribute is important in discriminating between the classes. At the maximum value difference of 2, a difference between the two values of the attribute definitely indicates a difference in the classes to which the two instances belong. In these circumstances, we would expect the value difference on this attribute to contribute heavily to the calculation of the overall distance between the two instances. In this sense, the MVDM also provides a mechanism to measure the importance of attributes. Based on these arguments, it can be claimed that the MVDM is a powerful metric in measuring the importance of attributes.

4 Handling numeric attributes

Using the MVDM, classification tasks can be successfuly dealt with for domains in which all attributes are nominal. To extend the MVDM approach to deal with domains with mixed type of attributes seems to be a simple task. Cost and Salzberg (1993) claimed:

> ...extending it to handle mixed symbolic and numeric data is quite straightforward: the algorithm could use simple differences for numeric features, and value difference tables for symbolic ones.

In fact, combining two different metrics in such a simple way would tend to lead to sub-optimal performance. This is because, again, it involves utilising all the available numeric attributes indiscriminately, some of which may not be important to the discrimination task. In terms of mathematical statistics, this is called "overfitting", which refers to the corresponding process of constructing a predictive model with more parameters than is optimal for best prediction. The adverse effect of overfitting on classification performance is due to the fitting of what is essentially noise. The source of this noise may be measurement error in either the class or the attribute. However, even when an attribute is measured well and contains little or no *intrinsic* noise, it may

discriminate only poorly between the classes. In other words, the association (i.e., correlation) between class and attribute is weak. If so, then for classification performance to be optimal, such an attribute should have only a small influence in the classification procedure.

Another problem with this naive combination of metrics is that, if the numeric attributes were standardised (as is usual in order to put them on the same scale as one another), then they would not be correctly scaled for combination with nominal attributes. This, too, would result in sub-optimal performance of the classification algorithm.

To overcome all these problems, some other approach is necessary. One scheme for dealing with this problem adopted by Rachlin et al. (1994) is to discretise numeric attributes by dividing them into ten equal intervals of equal length. This was reported to give good classification performance. However, this approach is arbitrary and would therefore not be optimal.

Another better approach to discretisation of continuous variables may be found in the literature dealing with the induction of classification trees under uncertainty (e.g. Breiman et al., 1984; Kononenko et al., 1984; Quinlan, 1988; White, 1987; White & Liu, 1990, 1993; Catlett, 1991; Kerber, 1992; Kononenko, 1993; Fayyad & Irani, 1993; Van de Merckt, 1993; Liu & White, 1991, 1994).

One of the most popular approaches is to consider all possible cutting points (between ordered values appearing in the training set) and select the cutting point that gives the greatest value on a particular measure, e.g., information gain (Quinlan, 1986) or, the chi-square test (White & Liu, 1994), etc. The method is recursively re-applied to the subsets of the previous split until a stopping criterion is satisfied. The most commonly used stopping criteria are based on the chi-square significance test (Liu & White, 1994) or the minimum description length principle (Quinlan & Rivest, 1989). Of course, the way that this would be implemented in the context of nearest neighbour classification differs from that used in classification trees. In the latter application, at each node in the tree, the algorithm is searching for the best attribute to branch on, as well as the best point at which to split it. By contrast, in the nearest neighbour approach, discretisation would be conducted to completion on one variable at a time. However, the principle of discretising continuous variables at the points giving maximum association between the derived binary attribute and class remains the same.

Other methods of discretisation include the one described by Van de Merckt (1993), which makes use of a clustering method to convert real-valued attributes into binary splits. Recently, Ting (1994) reviewed these methods and applied them in the context of nearest neighbour classification using MVDM. It has been reported that appropriate discretisation achieves an effect of noise reduction and increases nearest neighbour algorithms' tolerance for irrelevant numeric-valued attributes and, as a result, leads to better classification performance.

5 Conclusions

Based on the discussions and analyses presented earlier, it can be concluded that, when all the attributes are nominal, the MVDM is a good metric to deal with irrelevant attributes and to measure the importance of different attributes.

Careful study of the MVDM also revealed another insight that, in general, any measure which does not utilise the information about class provided by attributes would be sub-optimal. This is one reason why the approach of using simple value difference for numeric attributes and the "overlap" metric for nominal attributes yields poor classification performance, because neither of them takes the concept (i.e., the correlation between class and attributes) into account. Another problem is that using different metrics for different types of attribute means that they are not being handled on a common scale. Based on these arguments, it is now clear why the conventional nearest neighbour approach of simply combining two separate metrics for real-valued and nominal attributes could lead to poor performance. The best approach, so far, is to discretise numeric attributes using some appropriate method and then to use the MVDM for both the discretised numeric attributes and the nominal attributes. This offers an integrated approach to deal with mixtures of attribute types in

nearest neighbour classification. Of course, this does not deny the usefulness of using separate "good" metrics for different type of attributes provided that all the metrics used take correlation with class into account and they all use the same scale.

Perhaps one final point should be made. Now that it has been established that the essence of a good distance metric is the ability to measure the correlation between the class and attributes, there is no reason why some other conventional association measure, such as the information gain (Quinlan, 1986), might not work as well. Such measures might be even better than the value difference metrics studied in this paper. Further investigation is certainly worthwhile.

Acknowledgement

Thanks are due to two anonymous referees for their detailed comments and ideas and also to Allan White and Y Huang for numerous insightful comments and suggestions.

References

Breiman, L, Friedman, JH, Olshen, RA and Stone, CJ, 1984. *Classification and Regression Trees*, Wadsworth.

Catlett, J, 1991. "On changing continuous attributes into ordered discrete attributes". In: Y Kodratoff (ed.), *Proceedings of the European Working Session on Learning*.

Cost, S and Salzberg, S, 1993. "A weighted nearest neighbour algorithm for learning with symbolic features". *Machine Learning* **10** 57–78.

Dasarathy, BV, 1991. *Nearest Neighbour (NN) Norms: NN Pattern Classification Techniques*. IEEE Press.

Devijver, PA and Kittler, J, 1980. "On the edited nearest neighbour rule". In: *Proceedings of the Fifth International Conference on Pattern Recognition*, 72–80.

Fayyad, UM and Irani, KB, 1993. "Multi-interval discretization of continuous valued attributes for classification learning". In: *Proceedings of the 13th International Joint Conference on Artifical Intelligence*, 1022–1027, Morgan Kaufmann.

Fix, E and Hodges, JL, 1951. "Discriminatory analysis—nonparametric discrimination: consistency properties. Project 21-49-004, Report No. 4, USAF School of Aviation Medicine, Randolph Field, TX, 261–279.

Fix, E and Hodges, JL, 1952. "Discriminatory analysis—nonparametric discrimination: small sample performance. Project 21-49-004, Report No. 11, USAF School of Aviation Medicine, Randolph Field, TX, 280–322.

Hart, PE, 1968. "The condensed nearest neighbour rule". *IEEE Transactions of Information Theory* **IT-14** (3).

Kerber, R, 1992. "ChiMerge: discretization of numeric attributes". In: *Proceedings of the Tenth National Conference on Artificial Intelligence*, 123–128, AAAI Press/MIT Press.

Kononenko, I, Bratko, I and Roskar, E, 1984. "Experiments in automatic learning of medical diagnostic rules". Technical Report. Jozef Stefan Institute, Ljubjana, Yugoslavia.

Kononenko, I, 1993. "Inductive and Bayesian learning in medical diagnosis". *Applied Artificial Intelligence* **7** 317–337.

Liu, WZ and White, AP, 1991. "A review of inductive learning". In: IM Graham and RW Milne (eds.), *Research and Development in Expert Systems* VIII, 112–126, Cambridge University Press.

Liu, WZ and White, AP, 1994. "The importance of attribute selection measures in decision tree induction". *Machine Learning* **15** 25–41.

Liu, WZ and White, AP, 1995. "A comparison of nearest neighbour and tree-based discriminant analysis. *Journal of Statistical and Computational Simulation* **53** 41–50.

Quinlan, JR, 1986. "Induction of decision trees". *Machine Learning* **1** 81–106.

Quinlan, JR, 1988. "Decision trees and multi-valued attributes". *Machine Intelligence* **11** 305–318.

Quinlan, JR and Rivest, RL, 1989. "Inferring decision trees using the minimum description length principle. *Information and Computation* **80** 227–248.

Rachlin, J, Kasif, S, Salzberg, S and Aha, D, 1994. "Towards a better understanding of memory-based and Bayesian classifiers". In: *Proceedings of the Eleventh International Conference on Machine Learning*, 242–250, New Brunswick, NJ.

Salzberg, S, 1989. "Nested hyper-rectangles for exemplar-based learning". In: KP Jantke (ed.), *Analogical and Inductive Inference: International Workshop AII'89*, Springer-Verlag.

Salzberg, S, 1990. *Learning with Nested Generalized Exemplars*, Kluwer Academic.

Salzberg, S, 1991. "A nested hyper-rectangle learning method". *Machine Learning* **6** (3) 251–276.

Stanfill, C and Waltz, D, 1986. "Towards memory-based reasoning." *Communications of the ACM* **29** (12) 1213–1228.

Swonger, CW, 1972, "Sample set condensation for a condensed nearest neighbour decision rule for pattern recognition". In: S Watanabe (ed.), *Frontiers of Pattern Recognition*, 511–519.

Ting, KM, 1994. "Discretisation of continuous-valued attributes and instance-based learning". Technical Report, 491, Basser Department of Computer Science, University of Sydney.

Tomek, I, 1976, "An experiment with the edited nearest neighbour rule". *IEEE Transactions on Systems, Man and Cybernetics*, **SMC-6** (6) 448–452.

Van de Merckt, T, 1993. "Decision trees in numerical attributes spaces". In: *Proceedings of the 13th International Joint Conference on Artificial Intellegence*, 1016–1021, Morgan Kaufmann.

White, AP, 1987. "Probabilistic induction by dynamic path generation in virtual trees". In: MA Bramer (ed.), *Research and Development in Expert Systems III*, 35–46, Cambridge University Press.

White, AP and Liu, WZ, 1990. "Probabilistic induction by dynamic path generation for continuous attributes". In: TR Addis and RM Muir (eds.), *Research and Development in Expert Systems VII*, 285–296, Cambridge University Press.

White, AP and Liu, WZ, 1993. "Fairness of attribute selection in probabilistic induction". In: MA Bramer and RW Milne (eds.), *Research and Development in Expert Systems IX*, 209–224, Cambridge University Press.

White, AP and Liu, WZ, 1994. "Bias in information-based measures in decision tree induction." *Machine Learning* **15** 321–329.