Effect of Attribute Relevance on Feature Weighting and Selection in Nearest Neighbour Classification Rules*

J.M. Sotoca, J.S. Sánchez, and F. Pla

Abstract. This paper explores the benefits of using feature weighting and feature selection in situations with different degree of relevance in the attributes. Therefore, some experiments are performed over both synthetic and real data sets, studying the behaviour of these techniques when irrelevant attributes and/or features with varying relevance are present.
Keywords: Feature weighting; Feature selection; Nearest neighbour classifier.

1 Introduction

The Nearest Neighbour (NN) classification rule constitutes one of the most appealing examples of this kind of learning algorithms. This technique and its extension to k neighbours (or k-NN rule, in which a sample is classified by taking the majority vote from its k closest instances) combine their simplicity in implementation with an appropriate behaviour in their expected performance.

In practice, an important drawback of any distance-based classifier comes from the possible inclusion of irrelevant, redundant, interacting, or noise attributes, which can drastically degrade the resulting classification performance. It is well-known that NN classifiers are especially sensitive to irrelevance since the corresponding distance function calculates an average similarity measure across all the attributes [1].

In this paper, we are interested in comparing the feature selection algorithms [5] with the feature weighting approaches [2] when analysing the degree of attribute relevance. This is an important topic especially in the case of distance-based classification schemes due to the influence of each particular attribute on the distance computation. We concentrate our study on the k-NN rule because of its well-known good practical behaviour in real-world tasks.

2 Feature Weighting vs. Feature Selection

In the machine learning literature, two degrees of feature relevance have been suggested. Thus, a feature is said to be relevant if it is either weakly relevant or strongly relevant; otherwise, it can be considered as irrelevant [3].

In general, the application of the feature weighting methods give a performance superior to that of the feature selection methods in problems where some features are

* This work has partially been supported by grants TIC2000-1703-C03-03 and DPI2001-2956-C02-02 (Spanish CICYT), P1-1B2002-07 (Fundació Caixa Castelló - Bancaixa), and IST-2001-37306 (IST Project European Union).

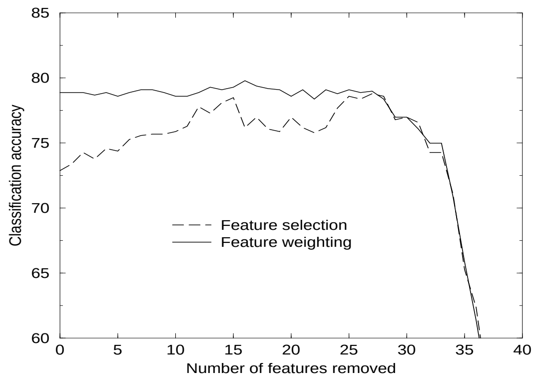

Fig. 1. Performance of ReliefF when an increasing number of features are removed. The dotted line represents the accuracy when assigning a weight value of 1 to the features selected and 0 to those removed (feature selection). The solid line represents the accuracy when assigning the weight value to the features picked up and 0 to those discarded (feature weighting).

more relevant than others [7]. In Fig. 1, the well-known ReliefF algorithm [4] is employed to obtain a relevance ordering on the Waveform+40 dataset with 21 relevant features and 19 irrelevant attributes. The weight vector is used to progressively discard the feature with the lowest weight value following a Sequencial Backward Selection (SBS) scheme. The feature weighting approach obtains a better NN accuracy than the feature selection scheme when discarding irrelevant attributes and also weakly relevant features. It is also important to remark that the removal of strongly relevant features has a similar behaviour in both feature weighting and feature selection.

3 Empirical Results and Discussion

In this section, we basically investigate differences between feature selection and feature weighting under different degrees of relevance. From the UCI Database Repository [6], twelve artificial and real data sets have been selected to analyse the classification performance of feature weighting and selection. The main characteristics of the data sets are summarized in Table 1, in which the number of irrelevant features are given in brackets. To increase statistical significance of the results in domains with a limited number of instances, 5-fold cross-validation has been employed.

Five procedures have been employed to carry out the experiments just described. First, *ReliefF* in Table 2 corresponds to using the computed weights for the features that have not been discarded at each stage. Second, *ReliefF(0,1)* is the binary version of ReliefF, that is, the weights for the attributes that have not been eliminated are now set to 1. These two approaches compute the weights only once at the initial stage (i.e., when no attribute has yet been discarded) and these define the unique relevance ordering along the whole process. On the contrary, *ReliefF-W* and *ReliefF-W(0,1)* differ from the previous approaches in the fact that, after eliminating one attribute, we have to recompute the feature weights and therefore, redefine the new ordering of feature

Table 1. The data sets used in the experiments.

	Features	Classes	Instances
Led+17	24 (17)	10	2,000
Waveform , Waveform+40	21 , 40 (19)	3	5,000
Monk1 , Monk2 , Monk3	6 (3) , 6 , 6 (3)	2	556 , 494 , 768
Diabetes	8	2	768
Glass	9	6	214
Heart	13	2	270
Vowel	10	10	528
Vehicle	18	4	848
Wine	13	3	178

relevance. Finally, *leave(0,1)* corresponds to the results obtained by simply estimating the accuracy of each single feature and then defining the order of attribute relevance according to such an accuracy rate.

Table 2. Accuracy of the weighted NN classifier and dimensionality of the best feature subset.

	ReliefF		ReliefF-W		ReliefF(0,1)		ReliefF-W(0,1)		leave(0,1)	
	% class.	Dim.	% class.	Dim.	% class.	Dim.	% class.	Dim.	% class.	Dim.
Led+17	**100**	6.0	**100**	6.0	**100**	6.0	**100**	6.0	**100**	5.0
Waveform	**79.5**	16.4	79.4	16.2	79.3	15.4	**79.6**	16.8	79.6	18.4
Waveform+40	80.4	19.6	**80.6**	26.4	79.7	14.4	79.7	16.0	**79.9**	16.0
Monk1	**100**	3.0	**100**	3.0	**100**	3.0	**100**	3.0	**100**	3.0
Monk2	**83.5**	5.0	**83.5**	5.0	**81.6**	6.0	**81.6**	6.0	**81.6**	6.0
Monk3	93.5	2.2	**97.6**	3.4	97.8	4.4	**97.8**	4.2	97.2	3.4
Diabetes	71.2	6.0	**73.1**	5.2	72.7	5.2	73.2	4.8	**74.3**	4.0
Glass	70.0	4.8	**71.8**	4.0	71.9	5.4	73.4	5.2	**80.2**	4.4
Heart	80.6	5.2	**84.3**	7.8	80.9	5.6	**82.5**	5.4	81.7	7.4
Vowel	98.1	8.4	**98.5**	9.4	98.0	9.2	**98.0**	8.9	97.8	8.8
Vehicle	71.2.	12.0	**71.3**	13.8	71.3	11.2	71.1	13.4	**72.8**	13.4
Wine	97.1	6.8	**97.1**	6.6	98.2	6.8	98.2	6.0	**99.4**	8.4

From the results in Table 2, some preliminary conclusions can be drawn. Firstly, comparing *ReliefF* with *ReliefF(0,1)*, one can observe that the former obtains better accuracy results over the databases with a clear difference in the degree of feature relevance. This is the case of the Waveform+40 and Monk2 data sets, in which there exists a number of weakly and strongly relevant features. On the contrary, when all the attributes have similar relevance, the feature selection approach provides a higher performance than the feature weighting scheme. With respect to the dimensionality of the best feature subset, the results here reported suggest that there are not significant differences between these approaches.

On the other hand, focusing on the differences between the methods that compute only one weight vector at the initial stage and those that recompute it after discarding one feature, it is possible to see that the highest improvement is obtained over the feature selection approaches. For example, in the Glass database, *ReliefF(0,1)* provides a 71.9% of accuracy while *ReliefF-W(0,1)* gives a 73.4%. Analogously, in the Heart database, the classifi cation accuracy of *ReliefF(0,1)* is 80.9% and *ReliefF-W(0,1)* reaches a percentage of 82.5%. Finally, when all features are relevant (this is the case of Glass, Vehicle, and Wine databases), the employment of the *leave(0,1)* approach gives the best results in terms of classifi cation accuracy. Nevertheless, it is also obvious that this option has a very high computational cost.

4 Conclusions and Future Work

This paper compares the behaviour of feature weighting and selection when analysing the degree of attribute relevance. This is an important topic especially in the case of distance-based classifi cation shemes, as it is the case of the weighted k-NN rule. From the experiments here carried out, it can be observed that the use of weights in the subset of features obtains a higher accuracy when there is a number of weakly and strongly relevant features in the data sets. On the contrary, when all the attributes have equal or similar relevance, the feature selection approaches provide a higher performance than the feature weighting algorithms.

On the other hand, analysing the differences between *ReliefF* and *ReliefF-W*, one can see that the latter obtains higher accuracy in the case of feature selection. This supposes to give a better ordering in the subset of relevant features. When all features have a similar relevance, the employment of the *leave(0,1)* approach provides the best results in terms of classifi cation accuracy.

References

1. Aha D.W.: Tolerating noise, irrelevant and novel attributes in instanced-based learning algorithms. International Journal of Man-Machine Studies, **36** 2 (1992) 267–287.
2. Cost S. and Salzberg S.L: A weighted nearest neighbor algorithm for learning with symbolic features. Machine Learning **10** 1 (1993) 57–78.
3. John G.H., Kohavi R. and Pfleger K.: Irrelevant features and the subset selecton problem. In: Proceedings of the 11th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco (1994) 121–129.
4. Kononenko I.: Estimating attributes: analysis and extension of relief. In: Proceedings of 7th European Conference on Machine Learning, Springer-Verlag, Catania, Italy (1994) 171–182.
5. Kudo M. and Sklansky J.: Comparison of algorithms that select features for pattern classifiers. Pattern Recognition **33** (2000) 25–41.
6. Murphy P.M.: UCI repository of machine learning. Department of Information and Computer Science, University of California, Irvine, CA.
7. Wettschereck D., Aha D.W., Mohri T.: A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review **11** (1997) 273–314.