Data Reduction Using Classifier Ensembles

J.S. Sánchez[1*] and L.I. Kuncheva[2]

1- Dept. Llenguatges i Sistemes Informàtics, Universitat Jaume I
Av. Sos Baynat s/n, 12071 Castelló de la Plana, Spain
E-mail: sanchez@uji.es

2- School of Electronics and Computer Science, University of Wales
Bangor, Gwynedd, LL57 1UT, United Kingdom
E-mail: l.i.kuncheva@bangor.ac.uk

Abstract. We propose a data reduction approach for finding a reference set for the nearest neighbour classifier. The approach is based on classifier ensembles. Each ensemble member is given a subset of the training data. Using Wilson's editing method, the ensemble member produces a reduced reference set. We explored several routes to make use of these reference sets. The results with 10 real data sets indicated that merging the reference sets and subsequent editing of the merged set provides the best trade-off between the error and the size of the resultant reference set. This approach can also handle large data sets because only small fractions of the data are edited at a time.

1 Introduction

One of the most widely studied non-parametric classification approaches is the k-Nearest Neighbour approach (k-NN). Given a set of N previously labelled instances (training set, TS) in a d-dimensional feature space, the k-NN classifier assigns an input sample to the class most frequently represented among the k closest instances in the TS, according to a certain similarity measure. A special case of this rule is when $k = 1$ where an input sample is assigned to the class of its closest neighbour. The set of instances used for k-NN is also termed *the reference set*.

Various works have been devoted to reducing the computational burden of k-NN. The challenges posed by modern practices of automatic data collection demand new efficient approaches to data reduction. One of the problems is that the data cannot be loaded in full into the computer memory. Inspired by this challenge, in this paper we propose a new data reduction approach based on classifier ensembles.

2 Data reduction techniques

Editing (or filtering) approaches [4, 7, 8, 11–13] eliminate mislabelled instances from the original TS and "clean" possible overlapping between regions from

*This work has been supported in part by the Spanish Ministry of Education and Science under grant PR2006–0217.

different classes. Wilson [12] introduced the first baseline editing algorithm. It consists of two steps. First a leave-one-out k-NN is applied to TS to label each instance. Second, the mislabelled instances and removed from TS and the remaining instances are the new reference set. The result from this algorithm amounts to smoothing the decision boundaries between the classes.

Condensing techniques [1, 3, 6, 9, 10] aim at selecting a the smallest possible subset of training instances without a significant degradation of classification accuracy. Other terms, such as *pruning* or *thinning*, have also been used for this group. Hart's algorithm [6] is the earliest attempt at minimizing the number of stored instances by retaining only a *consistent* subset of the original TS. A subset $S \subset T$ is said to be consistent iff all of T is correctly classified using 1-NN with S as the reference set. Although there are many consistent subsets, we are interested in the one with the minimum cardinality, called the *minimal* consistent subset. Hart's algorithm does not guarantee finding a minimal consistent subset of TS.

As the two approaches target different disadvantages of 1-NN, Wilson's and Hart's methods are perceived as complementary and often applied in succession. First Wilson's method is applied to "clean" the data, and the Hart's method eliminates the objects which are not important for preserving the (smoothed) boundary.

3 Data reduction using classifier ensembles

We introduce a general methodology for data reduction which avoids the need that all data be stored in the computer memory at the same time. The basic scheme can be summarized as follows. First, the original training data are split randomly into h disjoint subsets (or bags) $T_1, T_2, \ldots, T_h$. The size of each subset can be chosen to be a percentage of the original TS size. The chosen data reduction algorithm is applied to each bag T_j $(j = 1, 2, \ldots, h)$, thus obtaining h reduced subsets, $R_1, R_2, \ldots, R_h$.

From this stage onwards, different strategies for reassembling a final reference set could be applied.

- *Edited Ensemble (Voting)* Once the reference sets are available, an ensemble of h classifiers can be constructed. The classifier corresponding to reference set R_j $(j = 1, \ldots, h)$ gets as input a feature vector $\mathbf{x} \in \Re^d$, and assigns it to one of the problem classes using k-NN. The final decision is made by simple majority voting between the h assigned labels. The reduction in size can be measured by the fraction of selected instances, which we will call *storage rate*. The smaller the value, the better. Thus the storage rate of *Voting* is $\frac{\sum_{j=1}^{h} |R_j|}{|TS|}$, where $|\cdot|$ denotes cardinality.

- *Merged Reference Set (Merging)* We can merge the h subsets in order to obtain a final reduced set, say $M = \bigcup R_j, (j = 1, 2, \ldots, h)$. As different bags of instances are selected from TS, R_j are disjoint subsets and the storage rate is

equivalent to that of the Voting method.

• *Edited Merged Set (Merging+Editing)* The merged set M may contain similar instances selected by several ensemble members. Instead of keeping them all we propose to run a further round of editing on M using the classical editing methods. This will improve on the storage rate without much degradation of accuracy.

It has been argued that ensembles of k-NN are not efficient unless different features are used by each ensemble member. Our pilot experiments also showed that *Voting* was inferior to the two merging methods. Therefore we included in the experiments only *Merging*, and *Merging+Editing*.

We propose that the ensemble-based strategies provide a reasonable compromise between accuracy and storage rate compared to classical data reduction methods. The major benefit is that, within the ensemble approach, only a small fraction of the data set is being edited at a time.

4 Experiments and discussion

Nine methods were examined in the experiment

Standard	Ensemble
• Standard 1-NN (no data reduction)	• Merging (M)
• Random search (RND)*	• Merging+Wilson (MW)
• Wilson (W)	• Merging+Wilson+Hart (MWH)
• Hart (H)	• Merging+Hart (MH)
• Wilson followed by Hart (WH)	

*2% of the data was randomly chosen and evaluated as the candidate-reference set. The best candidate out of 100 trials was returned as the selected reference set.

All the experiments were carried out on 10 data sets (Table 1) taken from the UCI Machine Learning Database Repository (`http://www.ics.uci.edu/~mlearn`) and the ELENA European Project (`http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA/`). A 10-fold cross-validation was applied for all data sets with stratified sampling for the folds. The folds were kept the same for all methods. None of the datasets was normalised which accounts for some discrepancies with the error rates reported elsewhere. Each ensemble member uses Wilson's method to select the reference set R_i from the presented training set T_i.

Table 2 shows the error rates achieved with the selected reference sets and Table 3 shows the respective storage rates.

To visualize the performance of a data reduction algorithm we can use a scatterplot of the storage rate $\mathcal{S}$ versus the error rate $\mathcal{E}$ of k-NN with the selected reference set. Points close to the origin (0% storage, 0% error) signify good reduction methods compared to points further up the diagonal line from (0,0) to (1,1). A Pareto-optimal subset of methods can be constructed. This

Table 1: A brief summary of the experimental datasets

	c	d	N		c	d	N
Cancer	2	9	699	Heart	2	13	270
Diabetes	2	8	768	Phoneme	2	5	5404
Gauss	2	2	5000	Sonar	2	60	208
German	2	24	1000	Vehicle	4	18	846
Glass	6	9	214	Wine	3	13	178

c: number of classes; d: number of features; N: number of data points

Table 2: Error rates of the data reduction methods

	Standard					Ensemble			
Datset	1-NN	RND	W	WH	H	M	MW	MWH	MH
Cancer	4.98	4.08	3.49	3.93	6.87	3.93	3.49	5.09	4.80
Diabetes	31.88	28.79	28.39	28.52	35.40	29.31	27.35	27.75	28.92
Gauss	35.18	34.04	30.62	31.26	36.08	29.54	27.76	28.28	30.64
German	33.50	33.30	30.60	30.90	37.50	29.20	29.80	30.50	30.70
Glass	26.72	51.65	36.90	34.63	27.58	38.78	37.85	37.87	38.44
Heart	44.07	36.67	34.44	35.93	45.56	34.07	32.59	32.59	34.07
Phoneme	9.25	22.45	11.31	12.49	11.51	12.34	13.84	14.82	13.03
Sonar	16.75	45.76	20.70	22.60	17.60	23.05	37.25	38.79	24.59
Vehicle	34.77	50.12	38.86	40.23	36.45	39.42	42.23	43.34	39.67
Wine	23.85	35.67	27.89	29.32	27.45	28.77	27.27	27.76	28.91

set contains non-dominated data reduction methods. Method i is called non-dominated iff there is no other method j such that $\mathcal{S}_j \leq \mathcal{S}_i$, $\mathcal{E}_j \leq \mathcal{E}_i$, and one of these two inequalities is a strict inequality. Scaling the axes of the scatterplot may account for different importance of the size reduction and error components of the performance. However, the Pareto optimal set will remain unchanged for any such scaling.

Since the error rates are very different for the different datasets, using the average error across the data sets will be inadequate. Instead we calculate ranks for the methods. For each data set the best method receives rank 1, and the worst receives rank 9. As there are two criteria - error rate and storage rate - each method will receive two ranks. Let $r_e(i,j)$ and $r_s(i,j)$ be the error and the storage ranks for method i, respectively evaluated on dataset j. Figure 1 displays the 9 methods in the space spanned by the average rank on storage rate $\left(\bar{r}_s(i) = \frac{1}{10}\sum_j r_s(i,j)\right)$ and the average rank on error rate $\left(\bar{r}_e(i) = \frac{1}{10}\sum_j r_e(i,j)\right)$.

In order to evaluate the relative quality of the reduction methods with respect to one another we calculated the "distance from the origin" in terms of the ranks

$$D(i) = \frac{1}{10}\sum_{j=1}^{10}\sqrt{r_e(i,j)^2 + r_s(i,j)^2}$$

Table 3: STORAGE RATES OF THE DATA REDUCTION METHODS

	Standard					Ensemble			
Datset	1-NN	RND	W	WH	H	M	MW	MWH	MH
Cancer	100.00	1.95	96.63	3.18	10.54	96.16	95.33	2.04	4.06
Diabetes	100.00	1.88	69.41	10.78	52.62	64.99	57.58	4.11	15.75
Gauss	100.00	2.00	67.58	9.87	53.38	66.26	58.66	2.94	16.95
German	100.00	2.00	68.49	13.32	54.83	66.30	59.26	5.20	17.42
Glass	100.00	3.11	65.34	12.95	45.91	39.64	38.08	4.87	7.82
Heart	100.00	1.65	64.53	15.02	59.42	57.78	46.75	5.76	20.78
Phoneme	100.00	1.99	89.15	9.34	23.52	81.23	77.21	4.83	10.32
Sonar	100.00	1.60	81.44	16.90	34.28	56.79	42.35	7.75	21.23
Vehicle	100.00	1.97	62.94	15.26	52.95	47.01	38.32	4.74	14.97
Wine	100.00	1.88	70.13	9.38	40.94	60.19	53.94	3.19	10.56

The results were as follows:

D(1-NN)	=	10.4845	D(RND)	=	7.6688	D(W)	=	8.7718
D(WH)	=	6.1320	D(H)	=	8.4192	D(M)	=	8.2454
D(MW)	=	7.0332	D(MWH)	=	$\boxed{5.5544}$	D(MH)	=	6.8475

The method closest to the origin was found to be MWH followed by MH. This demonstrates the advantages of ensemble-editing to standard data reduction methods. The Pareto optimal set of methods is indicated in Figure 1 by a dashed line. The set consists of RND, MWH, MH and W. While RND and W are at the two ends of the set where one of the two criteria is quite predominant, MWH and MH are in the middle offering a valuable trade-off.

5 Concluding remarks

We propose to use a classifier ensemble approach to select a reference set for the 1-NN classifier. The experiments indicated that the best trade-off between size and error is achieved using merging of the individual reference sets constructed by the ensemble members followed by editing of the merged set. There are two hyper parameters of the ensemble approach - the ensemble size and the proportion of the data used for each ensemble member. The ensemble size was fixed in our study to 25 [???]. The whole data set was used to form the training sets T_i, which means that each ensemble member was presented with 4% of the data set. It is interesting to investigate how the hyper parameters affect the quality of the ensemble approach for data reduction. The ensemble approach also addresses the challenge of handling large data sets whose processing in a batch mode may be a problem.

References

[1] J.R. Cano, F. Herrera, M. Lozano, Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study. IEEE Trans. on Evolutionary Computation **7** (2003) 561–575.

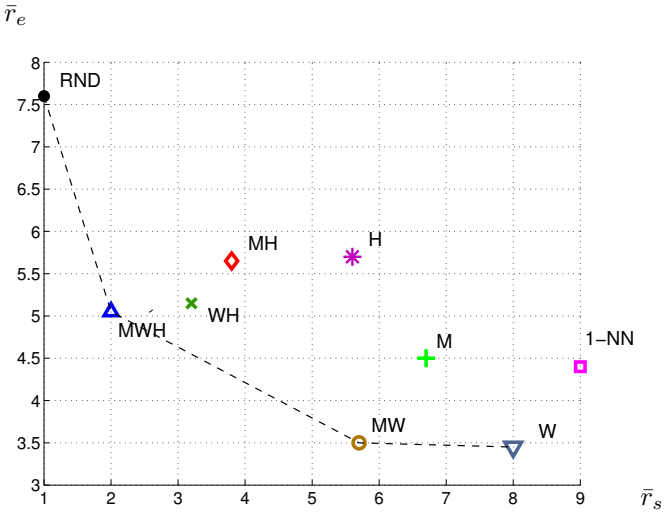

Fig. 1: Error-versus-size plot for the 9 data reduction methods using average ranks.

[2] C.L. Chang, Finding prototypes for nearest neighbor classifiers. IEEE Trans. on Computers **23** (1974) 1179–1184.

[3] B.V. Dasarathy, Minimal consistent subset (MCS) identification for optimal nearest neighbour decision systems design. IEEE Trans. on Systems, Man, and Cybernetics **24** (1994) 511–517.

[4] P.A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs, NJ (1982).

[5] S. Geva, J. Sitte, Adaptive nearest neighbor pattern classification. IEEE Trans. on Neural Networks **2** (1991) 318–322.

[6] P.E. Hart, The condensed nearest neighbor rule. IEEE Trans. on Information Theory **14** (1968) 515–516.

[7] J. Koplowitz, T.A. Brown, On the relation of performance to editing in nearest neighbor rules. Pattern Recognition **13** (1981) 251–255.

[8] L.I. Kuncheva, Editing for the k-nearest neighbors rule by a genetic algorithm. Pattern Recognition Letters **16** (1995) 809–814.

[9] R.A. Mollineda, F.J. Ferri, E. Vidal, An efficient prototype merging strategy for the condensed 1-NN rule through class-conditional hierarchical clustering. Pattern Recognition **35** (2002) 2771–2782.

[10] G.L. Ritter, H.B. Woodruff, S.R. Lowry, T.L. Isenhour, An algorithm for a selective nearest neighbour decision rule. IEEE Trans. on Information Theory **21** (1975) 665–669.

[11] I. Tomek, An experiment with the edited nearest neighbor rule. IEEE Trans. on Systems, Man and Cybernetics **6** (1976) 448–452.

[12] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data sets. IEEE Trans. on Systems, Man and Cybernetics **2** (1972) 408–421.

[13] W.R. Wilson, T.R. Martinez, Reduction techniques for instance-based learning algorithms. Machine Learning **38** (2000) 257–286.