Using Representative-Based Clustering for Nearest Neighbor Dataset Editing

Christoph F. Eick , Nidal Zeidat, and Ricardo Vilalta
Dept. of Computer Science, University of Houston
{ceick, nzeidat, vilalta}@cs.uh.edu

Abstract

The goal of dataset editing in instance-based learning is to remove objects from a training set in order to increase the accuracy of a classifier. For example, Wilson editing removes training examples that are misclassified by a nearest neighbor classifier so as to smooth the shape of the resulting decision boundaries. This paper revolves around the use of representative-based clustering algorithms for nearest neighbor dataset editing. We term this approach supervised clustering editing. The main idea is to replace a dataset by a set of cluster prototypes. A novel clustering approach called supervised clustering is introduced for this purpose. Our empirical evaluation using eight UCI datasets shows that both Wilson and supervised clustering editing improve accuracy on more than 50% of the datasets tested. However, supervised clustering editing achieves four times higher compression rates than Wilson editing.

1. Introduction

Nearest neighbor classifiers have received considerable attention by the research community (for a survey see Toussaint [3]). Most of the research aims at producing time-efficient versions of the algorithm. For example, several condensing techniques have been proposed that replace the set of training examples O by a smaller set $O_C \subset O$ such that all examples in O are still classified correctly by a NN-classifier that uses O_C.

Data set editing techniques, on the other hand, aim at replacing a dataset O with a, usually smaller, dataset O_E with the goal of improving the accuracy of a NN-classifier. A popular technique in this category is *Wilson editing* [5]; it removes all examples that have been misclassified by the 1-NN rule from a dataset. Wilson editing cleans interclass overlap regions, thereby leading to smoother boundaries between classes. It has been shown by Penrod and Wagner [2] that the accuracy of a Wilson edited nearest neighbor classifier converges to Bayes' error as the number of examples approaches infinity. Figure 1a shows a hypothetical dataset where examples that are misclassified using the 1-NN-rule are marked with circles around them. Figure 1.b shows the reduced dataset after applying Wilson editing.

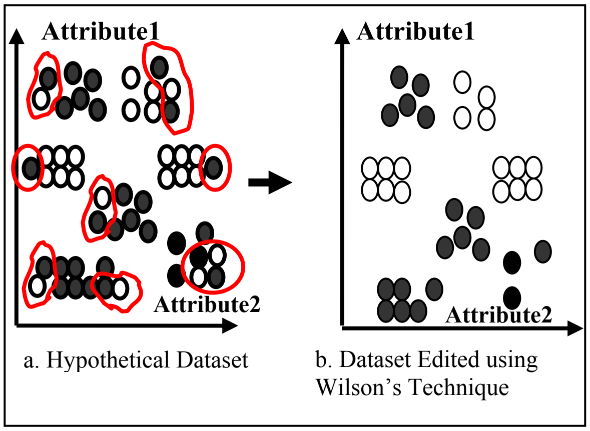


Figure 1: Wilson editing for a 1-NN classifier.

In addition to analyzing the benefits of Wilson editing, this paper proposes a new approach to nearest neighbor editing that replaces a dataset by a set of cluster prototypes using a *supervised clustering* algorithm [6]. We will refer to this editing technique as *supervised clustering editing* (SCE) and to the corresponding nearest neighbor classifier as *nearest representative (NR) classifier*. Section 2 introduces supervised clustering editing. Section 3 discusses experimental results that compare Wilson editing, supervised clustering editing, and traditional, ”unedited” nearest-neighbor classifiers. Section 4 summarizes the results of this paper.

2. Using Supervised Clustering for Dataset Editing

Due to its novelty, the goals and objectives of supervised clustering will be discussed in the first subsection. The subsequent subsections will explain how supervised clustering can be used for dataset editing.

2.1. Supervised Clustering

Supervised clustering deviates from traditional clustering in that it is applied on classified examples with the objective of identifying clusters having not only strong cohesion but also class purity. Moreover, in supervised clustering, we try to keep the number of clusters small, and objects are assigned to clusters using a notion of closeness with respect to a given distance function. Consequently, supervised clustering evaluates a clustering based on the following two criteria:

- *Class impurity, Impurity(X).* Measured by the percentage of minority examples in the different clusters of a clustering X. A minority example is an example that belongs to a class different from the most frequent class in its cluster.
- *Number of clusters, k.* In general, we favor a low number of clusters.

In particular, we use the following fitness function in our experimental work (lower values for q(X) indicate 'better' quality of clustering X**).**

$$q(X) = \text{Impurity}(X) + \beta * \text{Penalty}(k) \qquad (1)$$

Where

$$\text{Impurity}(X) = \frac{\#\text{ of Minority Examples}}{n},$$

$$\text{Penalty}(k) = \begin{cases} \sqrt{\dfrac{k-c}{n}} & k \geq c \\ 0 & k < c \end{cases}$$

With n being the total number of examples and c being the number of classes in a dataset. Parameter β (0< β ≤3.0) determines the penalty that is associated with the number of clusters, k; i.e., higher values for β imply larger penalties as the number of clusters increases.

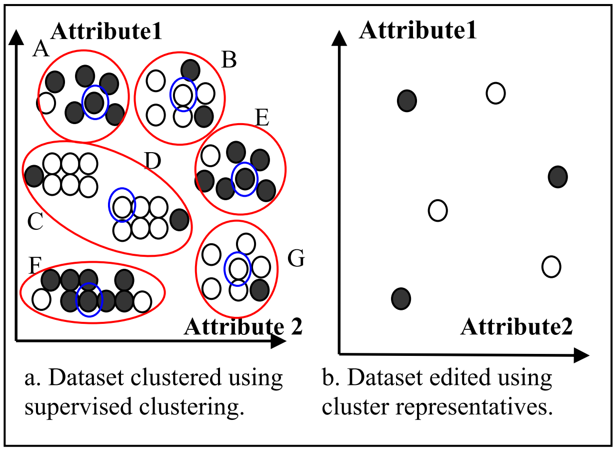
2.2 Representative-Based Supervised Clustering Algorithms

Representative-based clustering aims at finding a set of k representatives that best characterizes a dataset. Clusters are created by assigning each object to the closest representative. Representative-based supervised clustering algorithms seek to accomplish the following goal: *Find a subset O_R of O such that the clustering X, obtained by using the objects in O_R as representatives, minimizes q(X).*

As part of our research, we have designed and evaluated several supervised clustering algorithms [1,6,7]. Among the algorithms investigated, one named Single Representative Insertion/Deletion Steepest Decent Hill Climbing with Randomized Restart (SRIDHCR for short) performed quite well. This greedy algorithm starts by randomly selecting a number of examples from the dataset as the initial set of representatives. Clusters are then created by assigning examples to their closest representative. The algorithm tries to improve the quality of the clustering by adding a single non-representative example to the set of representatives as well as by removing a single representative from the set of representatives. The algorithm terminates if the solution quality (measured by q(X)) of the current solution does not improve. Moreover, SRIDHCR is run r times, reporting the best solution found as its result.

2.3 Using Cluster Prototypes for Dataset Editing

Figure 2 gives an example illustrating how supervised clustering is used for dataset editing. Figure 2.a shows a dataset that was partitioned into 6 clusters using a supervised clustering algorithm. Cluster representatives are marked with small circles around them. Figure 2.b shows the result of supervised clustering editing.



a. Dataset clustered using supervised clustering.

b. Dataset edited using cluster representatives.

Figure 2: Editing a dataset using supervised clustering.

3. Experimental Results

To evaluate the benefits of Wilson editing and supervised clustering editing, we applied these techniques to a benchmark consisting of eight UCI datasets [4] (see. Table 1).

Table1: Datasets used in the experiments.

Dataset name	# of objects	# of attributes	# of classes
Glass	214	9	6
Heart-Statlog	270	13	2
Heart-H	294	13	2
Iris Plants	150	4	3
Pima Indians Diabetes	768	8	2
Image Segmentation	2100	19	7
Vehicle Silhouettes	846	18	4
Waveform	5000	21	3

All datasets were normalized using a linear interpolation function that assigns 1 to the maximum value and 0 to the minimum value. Manhattan distance was used to compute the distance between two objects.

Table 2: Prediction accuracy for the four classifiers.

β	NR	Wilson	1-NN	C4.5
Glass* (214*)				
0.1	0.636	0.607	0.692	0.677
0.4	0.589	0.607	0.692	0.677
1.0	0.575	0.607	0.692	0.677
***Heart-Stat Log* (270)**				
0.1	0.796	0.804	0.767	0.782
0.4	0.833	0.804	0.767	0.782
1.0	0.838	0.804	0.767	0.782
***Diabetes* (768)**				
0.1	0.736	0.734	0.690	0.745
0.4	0.736	0.734	0.690	0.745
1.0	0.745	0.734	0.690	0.745
***Vehicle* (846)**				
0.1	0.667	0.716	0.700	0.723
0.4	0.667	0.716	0.700	0.723
1.0	0.665	0.716	0.700	0.723
***Heart-H* (294)**				
0.1	0.755	0.809	0.783	0.802
0.4	0.793	0.809	0.783	0.802
1.0	0.809	0.809	0.783	0.802
***Waveform* (5000)**				
0.1	0.834	0.796	0.768	0.781
0.4	0.841	0.796	0.768	0.781
1.0	0.837	0.796	0.768	0.781
***Iris-Plants* (150)**				
0.1	0.947	0.936	0.947	0.947
0.4	0.973	0.936	0.947	0.947
1.0	0.953	0.936	0.947	0.947
Segmentation (2100)				
0.1	0.938	0.966	0.956	0.968
0.4	0.919	0.966	0.956	0.968
1.0	0.890	0.966	0.956	0.968

The parameter β of q(X) has a strong influence on the number *k* of representatives chosen by the supervised clustering algorithm. In general, an editing technique reduces the size *n* of a dataset to a smaller size *k*. We define the dataset compression rate of an editing technique as:

$$\text{Compression Rate } = 1 - \frac{k}{n} \tag{2}$$

In order to explore different compression rates for supervised clustering editing, three different values for parameter β were used in the experiments: 1.0, 0.4, and 0.1. Prediction accuracies were measured using 10-fold cross-validation throughout the experiments. Representatives for the nearest representative (NR) classifier were computed using the SRIDHCR supervised clustering algorithm that was restarted 50 times. Accuracies and compression rates were obtained for a 1-NN-classifier that operates on subsets of the 8 datasets obtained using Wilson editing. We also computed prediction accuracy for a traditional 1-NN classifier that uses all training examples when classifying a new example. Finally, we report prediction accuracy for the decision-tree learning algorithm C4.5 that was run using its default parameter settings. Table 2 reports the accuracies obtained by the four classifiers evaluated in our experiments. Table 3 reports the average dataset compression rates for supervised clustering editing and Wilson editing. It also reports the average, minimum, and maximum number of representatives found on the 10 runs for SCE.

If we inspect the results displayed in Table 2, we can see that Wilson editing is a quite useful technique for improving traditional 1-NN-classfiers. Using Wilson editing leads to higher accuracies for 6 of the 8 datasets tested and only shows a significant loss in accuracy for the Glass dataset. The SCE approach, on the other hand, accomplished significant improvement in accuracy for the Heart-Stat Log, Waveform, and Iris-Plants datasets, outperforming Wilson editing by at least 2% in accuracy for those datasets. It should also be mentioned that the achieved accuracies are significantly higher than those obtained by C4.5 for those datasets. However, our results also indicate a significant loss in accuracy for the Glass and Segmentation datasets.

More importantly, looking at Table 3, we notice that SCE accomplishes compression rates of more than 95% without a significant loss in prediction accuracy for 6 of the 8 datasets. For example, for the Waveform dataset, a 1-NN classifier that only uses at an average 28 representatives outperforms the traditional 1-NN classifier that uses all 4500 training examples[1] by 7.3% (76.8% to 84.1%).

As mentioned earlier, Wilson editing reduces the size of a dataset by removing examples that have been misclassified by a *k*-NN classifier, which explains the low compression rates for the Iris and Segmentation datasets. Condensing approaches, on the other hand, reduce the size of a dataset by removing examples that have been classified correctly by a nearest neighbor classifier. Finally, supervised clustering editing reduces the size of a

[1] Due to the fact that we use 10-fold cross-validation, training sets contain 0.9*5000=4500 examples.

dataset by removing examples that have been classified correctly as well as examples that have not been classified correctly. A representative-based supervised clustering algorithm is used that aims at finding clusters that are dominated by instances of a single class, and tends to pick as cluster representatives[2] objects that are in the center of the region associated with the cluster.

Table 3: Dataset compression rates for SCE and Wilson editing.

B	Avg. *k* [Min-Max] for SCE	SCE Compression Rate (%)	Wilson Compression Rate (%)
Glass* (214*)			
0.1	34 [28-39]	84.3	27
0.4	25 [19-29]	88.4	27
1.0	6 [6 – 6]	97.2	27
***Heart-Stat Log* (270)**			
0.1	15 [12-18]	94.4	22.4
0.4	2 [2 – 2]	99.3	22.4
1.0	2 [2 – 2]	99.3	22.4
***Diabetes* (768)**			
0.1	27 [22-33]	96.5	30.0
0.4	9 [2-18]	98.8	30.0
1.0	2 [2 – 2]	99.7	30.0
***Vehicle* (846)**			
0.1	57 [51-65]	97.3	30.5
0.4	38 [26-61]	95.5	30.5
1.0	14 [9-22]	98.3	30.5
***Heart-H* (294)**			
0.1	14 [11-18]	95.2	21.9
0.4	2	99.3	21.9
1.0	2	99.3	21.9
***Waveform* (5000)**			
0.1	104 [79-117]	97.9	23.4
0.4	28 [20-39]	99.4	23.4
1.0	4 [3-6]	99.9	23.4
***Iris-Plants* (150)**			
0.1	4 [3-8]	97.3	6.0
0.4	3 [3 – 3]	98.0	6.0
1.0	3 [3 – 3]	98.0	6.0
Segmentation (2100)			
0.1	57 [48-65]	97.3	2.8
0.4	30 [24-37]	98.6	2.8
1.0	14	99.3	2.8

4. Conclusion

The goal of dataset editing in instance-based learning is to remove objects from a training set in order to increase the accuracy of the learnt classifier. This paper evaluates the benefits of Wilson editing using a benchmark consisting of eight UCI datasets. Our results show that Wilson editing enhanced the accuracy of a traditional nearest neighbor classifier on six of the eight datasets tested, achieving an average compression rate of about 20%. It is also important to note that Wilson editing, although initially proposed for nearest neighbor classification, can easily be used for other classification tasks. For example, a dataset can easily be "*Wilson edited*" by removing all training examples that have been misclassified by a decision tree classification algorithm.

In this paper, we introduced a new technique for dataset editing called supervised clustering editing (SCE). The idea of this approach is to replace a dataset by a subset of cluster prototypes. Experiments were conducted that compare the accuracy and compression rates of SCE, with Wilson editing, and with a traditional, unedited, 1-NN classifier. Results show SCE accomplished significant improvements in prediction accuracy for 3 out of the 8 datasets used in the experiments, outperforming the Wilson editing based 1-NN classifier by more than 2%. Moreover, experimental results show that for 6 out the 8 datasets tested, SCE achieves compression rates of more than 95% without significant loss in accuracy. In summary, surprisingly, high accuracy gains were achieved using only a very small number of representatives for several datasets. In general, our empirical results stress the importance of centering more research on dataset editing techniques.

5. References

[1] Eick, C., Zeidat, N., and Zhao, Z., "*Supervised Clustering – Algorithms and Applications.* submitted for publication.

[2] Penrod, C. and Wagner, T., "*Another look at the edited nearest neighbor rule*", IEEE Trans. Syst., Man, Cyber., SMC-7:92–94, 1977.

[3] Toussaint, G., "*Proximity Graphs for Nearest Neighbor Decision Rules: Recent Progress*", Proceedings of the 34th Symposium on the INTERFACE, Montreal, Canada, April 17-20, 2002.

[4] University of California at Irving, Machine Learning Repository. http://www.ics.uci.edu/~mlearn/MLRepository.html

[5] Wilson, D.L., "*Asymptotic Properties of Nearest Neighbor Rules Using Edited Data*", IEEE Transactions on Systems, Man, and Cybernetics, 2:408-420, 1972.

[6] Zeidat, N., Eick, C., "*Using k-medoid Style Algorithms for Supervised Summary Generation*", Proceedings of MLMTA, Las Vegas, June 2004.

[7] Zhao, Z., "*Evolutionary Computing and Splitting Algorithms for Supervised Clustering*", Master's Thesis, Dept. of Computer Science, University of Houston, May 2004.

[2] Representatives are rarely picked at the boundaries of a region dominated by a single class, because boundary points have the tendency to attract points of neighboring regions that are dominated by other classes, therefore increasing cluster impurity.