Tandem Fusion of Nearest Neighbor Editing and Condensing Algorithms - Data Dimensionality Effects

Abstract

In this paper, the effect of the dimensionality of data sets on the exploitation of synergy among known nearest neighbor (NN) editing and condensing tools is analyzed using a synthetic data set. The synergy is exploited through a tandem mode of fusion approach that combines the proximity graph (PG) based editing scheme and the minimal consistent set (MCS) condensing technique. These two methods were selected on the basis of prior experience to representatively evaluate the effect of the data dimensionality. The algorithm level fusion of PG editing and MCS condensing is experimentally shown to be a powerful implement across the range of data dimensionality.

1. Introduction

The NN classifier is arguably one of the simplest pattern classification algorithms devised, next only to the closest centroid approach. In its classical manifestation, given a set of n previously labeled prototypes (namely, training set), this classifier assigns any given sample to the class indicated by the label of the closest prototype in the training set. More generally, the k-NN rule maps any sample to the pattern class most frequently represented among the k closest neighbors. The popularity of this classifier arises in part from its extreme simplicity and in part from its near-optimal asymptotic behavior. Cover and Hart [1] showed that the infinite sample error is less than twice the error of the Bayes classifier.

Nevertheless, these classifiers have certain well-known deficiencies. The applicability of the NN rules to real-time problems, with a large set of prototypes of high dimensionality, can become prohibitive because of the immense computational loads required for searching the nearest neighbors of each sample in the data set. Another important drawback comes from the fact that training sets may contain noisy or erroneously labeled prototypes, which generally lead to a decrease in performance. This is why a considerable amount of effort has been devoted to the analysis of the NN classification rules since the 70s, as shown in [2]. For example, editing and condensing methods have been developed to pick out an appropriate subset of prototypes with computational efficiency and accuracy as their objectives. The synergy between these two groups of tools has been recently explored in [3] by application of these NN techniques in various tandem modes of algorithm fusion.

In particular, condensing algorithms aim at selecting the minimal subset of prototypes that lead to approximately the same performance as the NN rule using the whole training set [4–8]. On the other hand, editing approaches eliminate erroneously labeled prototypes from the original training set and "clean" the overlapping among regions from different classes [9–13]. While condensing is primarily focused on reducing the number of prototypes in order to gain a computational advantage, the main goal of editing is to improve recognition accuracy by producing a more sterilized training set.

In the earlier study [3], the benefits of using specific combinations of some NN editing and condensing techniques were empirically illustrated. The results reported therein should be interpreted as a first step toward a more complete understanding of how to exploit the synergy between editing and condensing. Accordingly, the aim of the present paper is to examine the effect of the dimensionality of the patterns on this synergy exploitation using a synthetic data set. More specifically, this experimental analysis focuses on the joint application of the PG based editing with the MCS condensing for the NN rule, since this combination appears to provide an excellent trade-off between classification accuracy and training set size reduction.

2. Proximity graph based editing

PG editing [11] is based on the concepts of Gabriel Graph (GG) and Relative Neighborhood Graph (RNG) [14]. The method consists of applying the general idea of Wilson's editing algorithm [13] but using the graph (Gabriel or relative) neighbors of each sample, instead of the Euclidean distance-based neighborhood, in order to estimate whether a sample is mislabeled or not. In a few words, the simplest PG

based editing can be summarized as follows: after computing the graph neighborhood of every sample in the original training set, discard those samples that are misclassified by their graph neighbors (instead of their k nearest neighbors). In practice, the general PG based editing algorithm can be written as follows:

PG Editing Algorithm

Step 1 Construct the PG corresponding to the original training set, X.

Step 2 For each $x_i \in X$ do:

- Discard x_i if there is a majority of graph neighbors from a different class.

This editing technique provides some advantages as compared to conventional methods. Firstly, it considers the neighborhood size as a characteristic which depends on each one of the prototypes in the training set. Secondly, PG editing provides some kind of information about prototypes close enough but homogeneously distributed around a given sample, which can be specially interesting to detect outliers close to the inter-class or decision boundaries. A more detailed description of PG editing can be found in [11].

3. Minimal consistent set condensing

MCS selection [4] is based on the concept of Nearest Unlike Neighbor (NUN) subset, which can be looked upon as an optimal descriptor of the inter-class boundaries. The NUN of a sample $x \in$ class A, is the sample $y \notin A$ of the shortest distance from x. The properties of NUN sets and other related topics are widely covered in [2, 4].

Based on this concept, it is possible to see that for every given sample, the sufficient condition for its correct classification (i.e., for consistency) is the presence within MCS of a sample from its own class that is closer than its NUN. Obviously, many samples independently satisfy this sufficiency condition for each given sample under consideration. This can be looked upon as a vote of confidence cast by the given sample and received by such closer-than-NUN samples. The sample with the most such votes therefore represents the prime candidate for inclusion in a MCS. Once this is picked, all the samples which were the voters contributing to the selection of the candidate for MCS can be disregarded from further consideration and the vote counts of other candidates are reduced to reflect this. The candidate with the maximum votes after this update becomes the next most effective MCS sample. This process is repeated until full consistency is achieved.

This condensing approach derives a consistent subset that is aimed to be minimal in size. In other words, the method results in a minimal (or, at worst, close to minimal) subset that guarantees the correct classification of all the prototypes in the original training set. Moreover, MCS selection always leads to a unique solution irrespective of the order of presentation of the data. Further, consistency property is assured at every stage of the iterative process.

4. Data sets and experiments description

In this section, the effect of the data dimensionality on the synergy exploitation is studied using a synthetic database with high overlap [15]. This consists of a collection of seven data sets that correspond to the same problem but with dimensionality ranging from 2 to 8. All the features are designed to be equally effective in terms of their discrimination potential thus making the analysis a function of only the size of the subset used and not of its specific features. The samples are divided into two classes representing multivariate normal distributions with zero mean and standard deviation 1 and 2 in all dimensions, respectively. Each class contains a total of 2,500 samples.

Here the classification results are averaged over five different random partitions (2,500 training samples and 2,500 test samples) of each one of the seven initial data sets, to obtain the overall error rate estimates. Thus the experiment consists of applying the NN rule to each of the seven test sets, where the training portion has been preprocessed using PG based editing, MCS condensing, and both applied in a tandem fusion mode.

From each trial, the reduction in the training set size and the classification accuracy are calculated. The percentage reduction of training samples required gives a direct measure of the amount of computational savings due to the number of prototypes resulting from each technique. The recognition accuracy provides a check on the ability of the algorithms to select the most "efficient" prototypes (for example, high classification accuracy may result from retaining more prototypes instead of selecting fewer efficient samples). In order to assess the performance relative to these two competing goals simultaneously, a performance measure in terms of the normalized Euclidean distance between each (reduction, recognition) pair and the origin (0% reduction, 0% recognition) is defined. Using this measure, the "best" technique is deemed as the one that has the largest measure, i.e., farthest from the origin.

5. Experimental results

The first significant result (see Figure 1) of this empirical analysis is that while editing techniques achieve less and less reduction in the training set size as the data dimensionality increases, MCS offers increasing gains in the size reduction rate as the dimensionality increases. The synergistic ap-

proach of PG based editing first followed by MCS condensing will thus have contributions of two opposing trends. As expected, this results in a near cancellation of the two trends with the set size reduction becoming fairly insensitive to the dimensionality of the data set. Of course, this reduction under PG and MCS is greater than under either PG or MCS by itself.

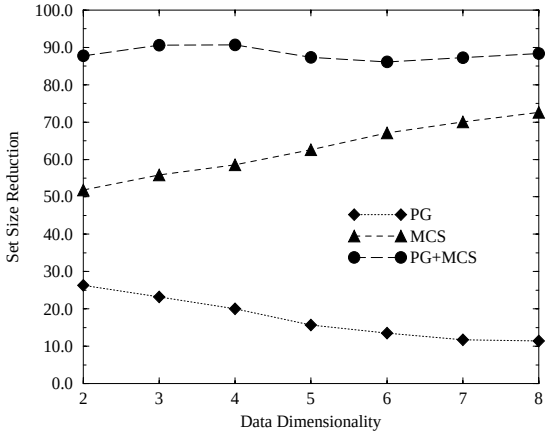

Figure 1. Set size reduction vs. data dimensionality.

Examining the other critical factor of interest (see Figure 2), namely recognition accuracy, the results show that, as is to be expected, the accuracy increases with data dimensionality almost without exception under all the cases, including original training set (OTS), PG, MCS, as well as PG with MCS (PG+MCS). While PG, as is to be expected of any editing technique, offers an improvement over OTS at every stage, the increase is more significant in the mid ranges. MCS, as is true of any condensing algorithm whose main thrust is reduction in computational load even though at some sacrifice in accuracy, has poorer performance compared to the OTS more or less consistently across the range. It is heartening to note that the combination of PG with MCS retrieves all of the losses in accuracy inflicted by MCS and performs more or less close to that of PG alone.

Thus the algorithm level fusion proves to be a powerful tool across the spectrum of data dimensionality. The resulting accuracy is considerable higher than that of OTS and a size reduction rate close to 90% is achieved, which consequently provides the desired decrease in computational loads in the operational phase. The performance measure plotted in Figure 3 visually illustrates the fact that the best option corresponds to the combination of PG editing with MCS condensing.

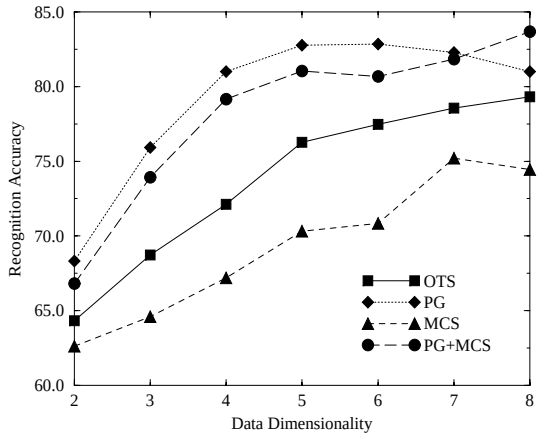

Figure 2. Recognition accuracy vs. data dimensionality.

6. Conclusions and further work

The effect of data dimensionality on the exploitation of synergy between NN editing and condensing techniques has been studied in the present paper. This analysis is to be viewed as an extension of the prior work [3] and is focused on examining the benefits of fusing specific editing and con-

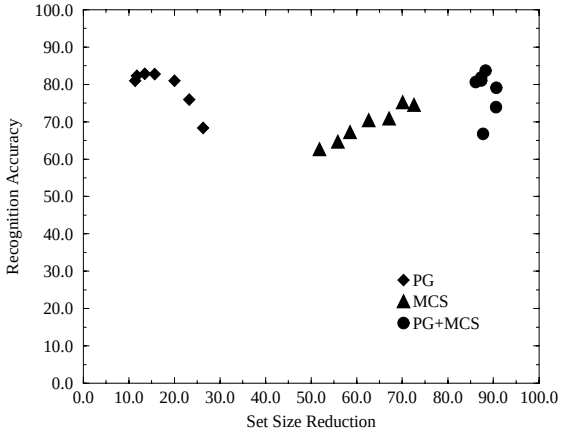

Figure 3. Set size reduction vs. recognition accuracy at different data dimensionality values.

densing tools. In this study, some preliminary conclusions were drawn. Firstly, PG editing generally achieves high enough recognition accuracy, but retains a very large number of prototypes. It is to be admitted that the use of PG based editing without any condensing is for all practical purposes of little value. Secondly, the combination of PG editing with MCS condensing produces the best results in terms of balancing set size reduction for implementation purposes with classification accuracy. On the other hand, the experimental results supported the conjecture that editing should be done before condensing.

From the experiments carried out here, it is apparent that editing schemes provide less and less reduction in the set size as the data dimensionality increases. This is in clear contrast to the behavior of condensing approaches. With respect to recognition accuracy, the results show that the performance improves with data dimensionality under all the implementations. Finally, the most significant point to be noted is the fact that tandem fusion of PG editing and MCS condensing algorithms constitutes a well balanced trade-off between classification accuracy and training set size reduction. This results in maximum set size reduction and close to highest recognition accuracy across the entire data dimensionality range.

Future plans include investigation of the fusion potential of other editing and condensing tools for achieving even better performance (reduction, recognition). This should help to draw more definitive conclusions with regard to the benefits of using distinct prototype selection techniques in a tandem fashion.

References

[1] T.M. Cover and P.E.Hart, "Nearest neighbor pattern classification", *IEEE Trans. on Information Theory*, 13(1):21–27, 1967.

[2] B.V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, IEEE Computer Society Press: Los Alamitos, CA, 1990.

[3] B.V. Dasarathy, J.S. Sánchez and S. Townsend, 'Nearest neighbor editing and condensing tools – synergy exploitation", *Pattern Analysis and Applications*, (in press), 1999.

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

[5] K. Fukunaga and J.M. Mantock, "Nonparametric data reduction", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6(1):115–118, 1984.

[6] G.W. Gates, "The reduced nearest neighbor rule", *IEEE Trans. on Information Theory*, 18(3):431–433, 1972.

[7] K.C. Gowda and G. Krishna, "The condensed nearest neighbor rule using the concept of mutual nearest neighborhood", *IEEE Trans. on Information Theory*, 24(4):488–490, 1979.

[8] P.E. Hart, "The condensed nearest neighbor rule", *IEEE Trans. on Information Theory*, 14(3):515–516, 1968.

[9] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Prentice Hall: Englewood Cliffs, NJ, 1982.

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

[11] J.S. Sánchez, F. Pla and F.J. Ferri, 'Prototype selection for the nearest neighbour rule through proximity graphs", *Pattern Recognition Letters*, 18(6):507-513, 1997.

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

[13] D.L. Wilson, 'Asymptotic properties of nearest neighbor rules using edited data", *IEEE Trans. on Systems, Man, and Cybernetics*, 2(3):408–421, 1972.

[14] J.W. Jaromczyk and G.T. Toussaint, "Relative neighborhood graphs and their relatives", *Proc. IEEE*, 80:1502–1517, 1992.

[15] P.M. Murphy and D.W. Aha, *UCI repository of machine learning databases*, Dept. of Information and Computer Science, University of California, Irvine, CA, 1991.