Biocybernetics and Biomedical Engineering
2008, Volume 28, Number 4, pp. 43–50

Double Sort Algorithm Resulting in Reference Set of the Desired Size

MARCIN RANISZEWSKI*

Technical University of Łódź, Computer Engineering Department, Łódź, Poland

An algorithm for obtaining the reduced reference set that does not exceed the desired size is presented. It consists in double sorting of the original reference set samples. The first sort key of the sample *x* is the number of such samples from the same class, that sample *x* is their nearest neighbour, while the second one is mutual distance measure proposed by Gowda and Krishna. The five medical datasets are used to compare the proposed procedure with the RMHC-P algorithm introduced by Skalak and the Gowda and Krishna algorithm, which are known as the most effective ones.

Keywords: reference set reduction, mutual distance measure, representative measure, Gowda and Krishna algorithm, Hart's algorithm, Skalak's RMHC-P algorithm, Double Sort Algorithm

1. Introduction

The reference set reduction is a term strongly connected with the 1-NN classification [1, 2]. 1-NN rule is very popular, simple and effective method of classification, but it has also the disadvantages. The main problem with 1-NN classifiers is the amount of samples in reference set. With growth of the reference set size, requirements of computer memory and time of classification grow up. Hence, it is very important to find a satisfying way to weaken the mentioned disadvantages. One of the solutions is the reference set reduction. The reduced reference set, as the new reference set, should fulfil two conditions: providing of the similar fraction

* Correspondence to: Marcin Raniszewski, Technical University of Łódź, Computer Engineering Department, ul. Stefanowskiego 18/22, 90-924 Łódź, e-mail: mranisze@kis.p.lodz.pl

Received 15 February 2008; Accepted 01 August 2008

of correct classifications to those obtained with the use of complete reference set, and containing of possibly small number of samples.

2. Consistent Reduced Reference Set – Hart's and Gowda-Krishna Algorithms

Many of the well known reference set algorithms produce the consistent reference set. Reference set consistency means that the 1-NN rule, operating with the reduced set, correctly classifies all the samples from the complete reference set.

The Hart's algorithm (CNN) (historically the first reference set reduction algorithm) presented in [3], produces the consistent reduced reference set in a random way. Hence, each time the algorithm is applied, it results in a different subset. Moreover, in the first phase of the algorithm, in the most cases, the samples not representative for its classes are selected as elements of the reduced reference set. Thus, despite the CNN algorithm results in the consistent subset, this subset consists too many not representative samples and noisy samples, what decreases the fraction of correct classifications.

This disadvantage partially removes the modification of CNN algorithm, introduced by Gowda and Krishna in [4]. The all samples from complete reference set are sorted by growing values of the mutual distance measure (*mdm*) and then the Hart's procedure is applied. The mutual distance measure is calculated in the following way: for the point x, the nearest point y from opposite class is found. The mutual distance measure is the number of points from the same class as point x, which are closer to point y than to point x. The small values of that measure are characteristic for samples, which lie near the class borders. Such kinds of points should be included to the reduced reference set.

The Gowda and Krishna algorithm results in smaller reduced reference set than the Hart's procedure and provides often highest fraction of correct classifications. Furthermore, the obtained reduced reference set is less depended on the primary arrangement of the reference set.

But the condition of consistency often causes too strong reflection of the complete reference set in the obtained reduced set. If the original reference set contains many noisy samples, the condition of consistency causes inclusion of these noisy samples in the reduced reference set. Therefore, often the consistent reference set is more numerous and has worse fraction of correct classifications than the inconsistent reference set. If the requirement of consistency will be removed then it is necessary to establish another stop condition. In the case of approaches based on the consistency, number of samples in the reduced reference set was minimised. Below, the two algorithms producing the reduced reference sets, not exceeding the desired size, are described. Instead of the size of the resulting set the misclassification rate is minimised.

3. Inconsistent Reduced Reference Set – Skalak's RMHC-P Algorithm and Double Sort Algorithm

The Skalak's RMHC-P (Random Mutation Hill Climbing to select prototype set) algorithm was presented in [5]. It has two parameters: k – the number of samples in the reduced reference set and m – the number of mutations. The algorithm is very simple (let X_{red} denotes reduced reference set, X – complete reference set, $f(X_{red})$ – fraction of correct classifications the X using 1-NN rule operating with the X_{red}):

1. Choose randomly k samples from X to X_{red};
2. Let $f_{max} = f(X_{red})$;
3. Mutate a random sample from X_{red} with a random sample chosen from the $X \setminus X_{red}$;
4. If $f_{max} >= f(X_{red})$ discard the mutation, otherwise, let $f_{max} = f(X_{red})$;
5. If the number of mutations exceeds m return X_{red}, otherwise, go to step 3.

The simplicity and easiness in implementation are not the only advantages of the Skalak's RMHC-P algorithm. In the most cases the level of reduction of the reference set can be very high, without loss in the fraction of correct classification. The disadvantages are the lack of unequivocal solution (the random samples are chosen each time the algorithm is started) and the parameter m (additional tests are required to discover the best number of mutations for the current reference set, what is more difficult due to randomness of the algorithm).

Two versions of the double sort algorithm were introduced by author of this paper in [6]. Both versions result in consistent reduced reference set based on the Hart's procedure, applied after double sorting of the reference set samples. In comparison with the Gowda and Krishna modification, both versions of double sort algorithm result in better classification quality and smaller size of reduced reference set.

The double sort algorithm presented in this paper is an algorithm that uses double sorting in following order: the samples are sorted by a decreasing representative measure (rm) and then in the groups of the samples with the same rm, by the mutual distance measure. The representative measure of the sample x is the number of such samples from the same class, that sample x is their nearest neighbour (Fig. 1).

This kind of sorting promotes the samples that represent its own class in the best way, and in the groups of equal representative measures, promotes the samples, which lies near the class borders.

The difference between the current approach and that described in [6] is the additional parameter: the desired maximum number of samples in the reduced reference set (let denote it by k_{max}). The Hart's procedure is interrupted after the k_{max}-th sample is added to the current reduced reference set. Its consistency is, of course, not guarantied. If the k_{max} is greater than the number of samples in the Hart's procedure result, the double sort algorithm results in the consistent reduced reference set (the Hart's procedure ends as in its original version).

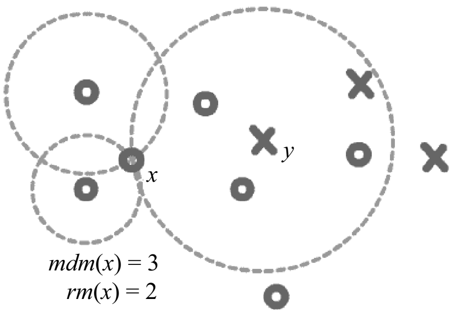

Fig. 1. The representative and mutual distance measures for a point *x*

The Hart's procedure applied in double sorting algorithms has little, but important modification: the samples from the beginning of the sorted reference set will be more likely added to the reduced set than in the original Hart's procedure. The Hart's procedure is stopped, when a new sample is added and the sample presentation starts from the beginning.

4. Experimental Results

The tests were made on five medical datasets:

- BUPA liver disorders [7] (BUPA Medical Research Ltd.) (number of classes: 2, number of attributes: 7 (6 features + class indicator – selector), number of instances: 345) – the first 5 attributes are all blood tests (mean corpuscular volume, alkaline phosphotase, alamine aminotransferase, aspartate aminotransferase, gamma-glutamyl transpeptidase), which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. 6th attribute is the number of half-pint equivalents of alcoholic beverages drunk per day. Each sample constitutes the record of a single male individual.
- Liver [8] (number of classes: 2, number of attributes: 14 (13 features + pixel class indicator), number of instances: 81968) – the dataset comes from ultrasound images that are sections of certain 3D objects found in a human body. Two class of pixels were taken into account: class 1 representing the objects (metastasis) of interest, class 2 denoting the background (liver areas without metastasis).
- Pima Indians Diabetes Database [7] (National Institute of Diabetes and Digestive and Kidney Diseases) (number of classes: 2, number of attributes: 9 (8 features + class indicator – test result), number of instances: 768) – all patients in this database are Pima-Indian women at least 21 years old and living near Phoenix, AZ, USA. Class 1 means a positive test for diabetes and class 0 is a negative test for diabetes. All 8 features are clinical findings: number of times pregnant, plasma glucose concentration a 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm

Hg), triceps skin fold thickness (mm), 2-hour serum insulin (mu U/ml), body mass index, diabetes pedigree function and age (years).

- Wisconsin Diagnostic Breast Cancer (WDBC) (Diagnostic) [7] (number of classes: 2, number of attributes: 32 (ID + class indicator – diagnosis + 30 features), number of instances: 569) – features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Class “M” means “malignant” and class “B” – “benign”.
- Protein Localization Sites (YEAST) [7] (number of classes: 10, number of attributes: 9 (8 features + class indicator), number of instances: 1484) - the paper [9] describes a predecessor to this dataset and its development.

Each dataset was divided 30 times into the training and testing sets of approximately equal size (BUPA, PIMA, WDBC, YEAST) and in proportion 1:3 (LIVER). The training sets were reduced by three algorithms:

- 1 time by the Gowda and Krishna algorithm;
- 10 times by DSA (presented in this paper Double Sort Algorithm) with the different constrains of the desired size of the reduced reference set: the number of samples in the reduced set was set to 1.0, 0.9, 0.8, ..., 0.1 part of the number of samples in the Gowda and Krishna reduced set;
- 10 times by the Skalak’s RMHC-P algorithm with the desired number of samples in the resulting set as in the case of DSA and following numbers of mutations: 300 for BUPA, 5000 for LIVER, 500 for PIMA, 400 for WDBC and 1500 for YEAST. As it was mentioned above, the number of mutations was established experimentally.

The Euclidean metric and classic standardization (based on mean values and standard deviations of features) were used. To estimate the quality of classification, the samples from the testing sets were classified by the use the 1-NN rule with the reduced reference sets.

In Table 1 the results of reductions are presented. Each fraction of correct classifications is an average value, counted from 30 pairs of training and testing sets.

Below are the descriptions of each row in Table 1:

- “Compl._points” – the number of samples in the reference set before reduction.
- “GK_points” – the average number of samples in the Gowda and Krishna reduced set.
- “Compl.” – the average fraction of correct classifications of samples from the testing sets by use the complete training set as the reference set.
- “GK” – the average fraction of correct classification of samples from the testing sets by the Gowda and Krishna reduced set.
- “1.0 DSA”, “0.9 DSA”, ..., “0.1 DSA” – the average fraction of correct classification of samples from the testing sets by the DSA reduced reference set with the number of samples not exceeding correspondently to 1.0, 0.9, ..., 0.1 part of the number offered by the Gowda and Krishna procedure.

- "1.0 Skalak", "0.9 Skalak", ..., "0.1 Skalak" – the average fraction of correct classification of samples from the testing sets by the Skalak's RMHC-P with the number of samples not exceeding correspondently to 1.0, 0.9,..., 0.1 part of the number offered by the Gowda and Krishna procedure.

The last column "Avg" in Table 1 presents average values of rows for all 5 datasets.

All fractions are presented in percentages.

Table 1. Results of reduction

	BUPA	LIVER	PIMA	WDBC	YEAST	Avg
Compl_points	173	27323	384	285	742	–
GK_points	99	1191	182	38	488	–
Compl.	60.8	98.0	69.6	95.0	50.7	74.8
GK	58.2	97.2	65.8	93.1	47.8	72.4
1.0 DSA	58.7	97.3	66.3	92.3	47.0	72.3
1.0 Skalak	59.3	96.6	68.3	94.3	49.0	73.5
0.9 DSA	58.5	97.1	66.2	92.9	46.5	72.2
0.9 Skalak	57.7	96.7	68.0	93.7	48.7	73.0
0.8 DSA	58.4	96.9	66.7	93.1	46.4	72.3
0.8 Skalak	58.6	96.8	68.2	94.0	48.8	73.3
0.7 DSA	58.8	96.6	67.9	93.4	47.0	72.7
0.7 Skalak	58.2	96.5	69.2	94.0	48.8	73.3
0.6 DSA	58.7	96.2	68.6	93.2	47.6	72.9
0.6 Skalak	58.5	96.5	69.3	94.2	49.1	73.5
0.5 DSA	59.3	95.7	69.3	93.5	48.7	73.3
0.5 Skalak	58.6	96.4	69.0	94.4	49.5	73.6
0.4 DSA	58.7	95.0	70.3	93.3	50.6	73.6
0.4 Skalak	59.4	96.1	70.1	94.0	50.5	74.0
0.3 DSA	58.2	94.3	70.8	93.1	52.5	73.8
0.3 Skalak	58.3	95.7	71.0	94.5	51.5	74.2
0.2 DSA	58.8	92.9	71.4	92.4	52.6	73.6
0.2 Skalak	58.9	94.8	71.5	93.8	53.1	74.4
0.1 DSA	59.2	90.9	71.2	88.4	50.0	71.9
0.1 Skalak	60.9	92.7	72.0	85.1	52.4	72.6

5. Discussion

As we can see in Table 1, for the BUPA set, the average fraction of correct classifications for our DSA as well as for Skalak's RMHC-P is, in most cases, higher than that offered by the Gowda and Krishna algorithm, except one case with 57.7% of correct classifications for Skalak's procedure. Differences between fractions of correct classification offered by DSA and RMHC-P do not exceed 2%. Slightly better results can be observed for RMHC-P than DSA.

In the case of the next considered LIVER data set, both algorithms, DSA and RMHC-P, offer slightly worse result than Gowda and Krishna procedure. But this time DSA outperforms the Skalak's algorithm if the desired size of the reduced reference set equals at least 0.7 of the reduced set obtained by Gowda-Krishna algorithm. For stronger reductions, the results of Skalak's algorithm are again slightly better that for DSA and differences, as in the case of BUPA data, do not exceed 2%.

For PIMA data both compared algorithms, i.e. DSA and RMHC-P, give better results than Gowda-Krishna approach and differences between them never exceed 2%. It is interesting that fractions of correct classifications are nearly monotonically higher as reduction is getting stronger. Similar behavior of DSA and RMHC-P can be observed for YEAST data. However, in this case the differences between fractions of correct classifications for DSA and RMHC-P reached the value of 2.4%.

The results for WDBC data, obtained by DSA and RMHC-P were always worse than that received by the use of Gowda-Krishna procedure. The reduction to 0.1 of the part of Gowda-Krishna reference set size causes significant decreasing of the fractions of correct classifications. Furthermore, the difference between mean values of fractions of correct classification, in DSA and RMHC-P increased to 3.3%. For to remaining degrees of reduction it does not exceed the value of 2%.

Looking at the last column of the Table 1, we can notice that the average fractions received for all analyzed datasets suggest that consistency is too strong condition and results in too numerous reduced reference sets (the Gowda and Krishna algorithm).

Renunciation of consistence gives us two, very important, advantages: stronger reduction and highest fractions of correct classifications.

6. Conclusions

The presented double sort algorithm (DSA) is based on double sorting of the samples from reference set, before applying the Hart's algorithm. The samples are sorted by decreasing representative measure (*rm*) and then, in the groups of samples with the same *rm*, by the mutual distance measure. The Hart's algorithm which originally builds the consistent reduced reference set in its use to DSA has been modified twice:

- the Hart's procedure cycle is broken when a new sample is added to the reduced set and the sample presentation starts from the beginning;

- the Hart's procedure is interrupted after the k_{max} sample will be added to the reduced set. k_{max} denotes the maximal allowed number of the samples in the reduced set.

The described DSA results in the inconsistent reduced reference set (only in the case when the desired constrain to the size of the reduced set is set to a sufficiently low value, otherwise, it results in the consistent reduced set). The results of the experiments taken on five medical datasets (BUPA, PIMA, LIVER, WDBC and YEAST) suggest two important advantages of the inconsistent reduced sets obtained from the double sort algorithm and the Skalak's RMHC-P algorithm: the improvement of fractions of correct classifications and the stronger reduction (often about 10 times stronger than by the Gowda and Krishna algorithm).

The DSA fractions are slightly lower than Skalak's ones. However, the Skalak's RMHC-P algorithm has one more parameter than DSA: the number of mutations, which should be established experimentally by use the validation sets. Moreover, DSA results in unequivocal solution, while the Skalak's algorithm constructs the reduced set in randomly way.

References

1. Theodoridis S., Koutroumbas K.: Pattern Recognition – Third Edition. Academic Press – Elsevier, USA, 2006.
2. Duda R.O., Hart P.E., Stork D.G.: Pattern Classification – Second Edition. John Wiley & Sons, Inc, 2001.
3. Hart P.E.: The condensed nearest neighbor rule. IEEE Transactions on Information Theory, 1968, vol. IT-14, 3, 515–516.
4. Gowda K. C., Krishna G.: The condensed nearest neighbor rule using the concept of mutual nearest neighborhood. IEEE Transaction on Information Theory, 1979, v. IT-25, 4, 488–490.
5. Skalak D.B.: Prototype and feature selection by sampling and random mutation hill climbing algorithms. 11th International Conference on Machine Learning, New Brunswick, NJ, USA, 1994, 293–301.
6. Raniszewski M.: Reference set reduction algorithms based on double sorting. Computer Recognition Systems 2, Advances in Soft Computing, Springer Berlin/Heidelberg, 2007, 45, 258–265.
7. Asuncion A., Newman, D.J.: UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science, 2007.
8. Jóźwik A., Kieś P.: Reference set size reduction for 1-NN rule based on finding mutually nearest and mutually furthest pairs of points. Computer Recognition Systems, Advances in Soft Computing, Springer Berlin/Heidelberg, 2005, Vol. 30, 195–202.
9. Nakai K., Kanehisa M.: Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria. PROTEINS: Structure, Function, and Genetics 1991, 11, 95–110.