ZESZYTY NAUKOWE POLITECHNIKI ŁÓDZKIEJ
Nr 1080 ELEKTRYKA, z. 122 2010

MARCIN RANISZEWSKI
Faculty of Electrical, Electronic, Computer and Control Engineering
Technical University of Lodz

METHODS OF STRONG REDUCTION AND EDITION OF A REFERENCE SET FOR THE NEAREST NEIGHBOUR RULE

Reviewer: **Professor Adam Jóźwik**

Manuscript received: 18.06.2009

The article summarises a doctoral dissertation proposing new methods of a reference set reduction and edition for the Nearest Neighbour Rule (NN).The presented methods are designed to accelerate NN and to improve its classification quality. The algorithms use the concept of the object representativeness. The obtained results were compared with the results provided by well-known and popular reduction and editing procedures.

1. INTRODUCTION AND THESES

The Nearest Neighbour Rule (NN Rule) [3,6,22] is one of the most popular classifiers offered by Pattern Recognition. It is very simple and intuitive. It classifies a sample to a class of its nearest neighbour sample from a reference set, i.e. a set of samples stored in the computer memory during a classification phase (for the standard NN Rule it is a complete training set). NN Rule has a very important theoretical property; it has been proven that NN classification error is never beyond the double classification error of the Bayesian classifier for sufficiently large training sets [6].

On the other hand, NN Rule has a serious disadvantage; for a large number of samples in the reference set the classification time can be too long to be accepted. Additionally, a classification error can be unnecessarily increased if a training set consists of many noisy samples, i.e. samples from wrong measurements, incorrectly classified or atypical. A well-known solution of these problems is the reduction and edition of a reference set [26]. The

reduction algorithms include methods which reduce a size of a reference set by choosing only the most informative samples. A reduced reference set should preserve the classification quality of the reference set. The editing procedures include methods which improve the classification quality mainly by removing noisy samples.

The aim of the research was to develop new methods of strong reduction and effective edition of the reference set for NN Rule (better than the well-known procedures). The main idea of almost all of the constructed algorithms, proposed in the theses, is a representativeness of a sample (described in the second section).

The theses of a doctoral dissertation were formulated as follows:

Thesis 1: Application of the sample representativeness enables construction of effective reference set reduction algorithms for the Nearest Neighbour Rule.

Thesis 2: Determination of the reference set subset such that the reduced set, created with the use of the representative measure, is consistent with, constitutes an effective method of the reference set edition for the Nearest Neighbour Rule.

2. THE REPRESENTATIVE MEASURE

The main idea, which almost all of the developed algorithms are based on, is the representativeness of a sample. It is expressed by the representative measure [15].The sample $\underline{x}$ has the representative measure rm($\underline{x}$) equal to p, when p samples (called voters) from the same class as $\underline{x}$ lie nearer $\underline{x}$ than their nearest neighbours from the opposite class (Fig. 1).

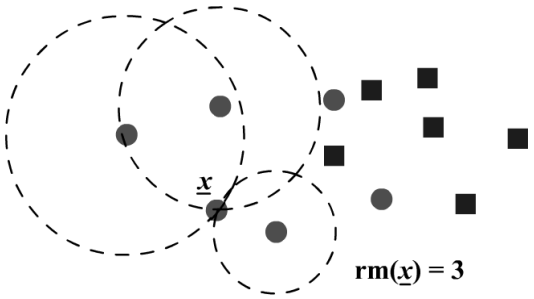

Fig. 1. The representative measure of the sample $\underline{x}$

The higher the value of representative measure is, the more representative is a sample for its class. It is assumed that a reduced reference set should consist only of samples from the centers of classes or samples near borders between classes. The samples from the centers of classes have high values of the representative measure. It can be said that they are highly representative. In the created algorithms such samples will be recognized and used to build the reduced or edited reference sets.

3. DEVELOPED ALGORITHMS

Within thesis 1 the following reduction algorithms were constructed and implemented:

1. The double sort algorithm with the desired size (DSAds) [16] – the algorithm initially sorts all the samples from a reference set with the use of two keys: the representative measure (rm) and Gowda and Krishna's Mutual Neighbourhood Value (mnv) [8]. The samples are sorted with the decreasing values of their rm, and in groups of the same value of rm, with the increasing values of mnv. Next, the modified Hart's Condensed Nearest Neighbour Rule [9] (mCNN) is applied. A little but important modification of CNN consists in breaking the iterations when a misclassified sample is added to a reduced reference set and starting the presentation of samples from the beginning. It causes that samples with the highest value of rm are more likely to be chosen. Additionally the mCNN is stopped, when the desired size of the resultant set is achieved (DSAds stop condition). The algorithm can be started with the different values of a desired reference set size *ds*. In this manner several reduced sets can be obtained. Then, a reduced set with the highest classification quality can be chosen as a resultant reduced set. It is convenient to perform DSAds for the increasing values of *ds* (up to a number of samples – let denote it as ds_{max} – in consistent reduced set generated by the mCNN without DSAds stop condition) because then the algorithm may be implemented to continue building of a reduced set for each value of *ds*.
2. The sequential algorithms:
   a. Sequential Reduction Algorithm (SeqRA) [19] – it is a typically random algorithm and it constitutes the basis to the Sequential Double Sort Algorithm (discussed below). SeqRA starts with a reduced reference set initially consisting of randomly chosen samples, one from each class. Next, sequentially from each class the randomly chosen sample is added or removed from a reduced reference set only if the addition or the removal increases the classification quality of a reduced set (estimated on a complete training set). The procedure is stopped if any adding or removing a sample does not increase the classification quality, i.e. a reduced set with the locally highest classification quality has been built.
   b. Sequential Double Sort Algorithm (SeqDSA) [11,17] – it is based on SeqRA. The main difference is initial sorting of a reference set according to the same order as in DSAds. Then, the first sample from a reference set is added to the initially empty reduced set. Next, according to the double sort order each sample is temporarily added to a reduced set and if the addition causes the growth of the classification quality, it is confirmed and the phase of sample removal is started. In this phase, according to the reverse double sort order, each sample is temporarily removed from a reduced set and again if the removal causes the growth of the classification quality, it is confirmed and the above described phase of sample addition is performed. The stop condition in this algorithm is the same as in SeqRA.

3. The algorithm based on choosing the most representative samples (Representative Measure Algorithm – RM) – for each sample from the reference set the value of representative measure is counted. Then, a sample with the greatest representative measure rm_{max} is chosen (ties are broken randomly). If $rm_{max} > R$ (R is a non-negative parameter, which defines the minimal representative measure), the sample is added to the initially empty reduced set and the value of representative measure is no longer counted for this sample. Moreover, all voters of that sample are no longer considered in the algorithm. If there is at least one sample in a reference set which the representative measure can be counted for, the algorithm proceeds and again values of the representative measure are recounted for appropriate samples. In the other case, the algorithm ends. More often the condition $rm_{max} > R$ is not fulfilled and then also the algorithm ends. For higher value of parameter R reduced sets are smaller. Similarly to DSAds the RM can be started with different values of R. The set with the highest classification quality can be chosen as a resultant reduced set, called bestRM. It is convenient to perform RM for the decreasing values of R (from some initial value R_{init} to 0) because then the algorithm may be continued to build the reduced set for each value of R.

Within thesis 2 the following editing algorithm was constructed and implemented:

4. The algorithm based on consistency criterion (Editing Algorithm based on Consistency – EAC) [18] – we say that a subset Y of a set X is consistent with X if each sample from X is correctly classified by NN Rule with Y as the reference set. EAC is based on extracting a subset from a reference set which bestRM is consistent with. Hence, EAC depends on a resultant reduced set of RM. The idea of such construction of an edited set is very simple: bestRM consists of the most representative samples from a reference set. The definition of a noisy sample negates the possibility that such a sample could be representative. Hence, it is almost impossible that bestRM consists of noisy samples. If we extract only these samples from the reference set which are correctly classified by NN Rule operating with bestRM (including all the samples from bestRM), we will filtrate the noisy samples, leaving only the most informative ones in an editing set.

4. TESTS

Nine real and one synthetic training sets (well-known in literature) were used in tests: Liver Disorders (BUPA) [1], GLASS Identification [1], IRIS [1], PARKINSONS Disease Data Set [1], PHONEME [21], PIMA Indians Diabetes [1], Statlog (Landsat Satellite) Data Set (SAT) [1], Wisconsin Diagnostic Breast Cancer (WDBC) (Diagnostic) [1], YEAST [1] and WAVEFORM (version1) [1].

Stratified ten-fold cross-validation (NN classification) was used for each experiment [10]. All tests were performed on Intel Core 2 Duo, T7250 2.00 Ghz processor with 2 GB RAM. ,Both the algorithms developed by the author and the algorithms well-known in the literature , were equally implemented in Java 5.0 environment.

The Euclidean metric was used. Samples from training sets were not initially standardized.

All the results are compared with the classification quality of NN Rule operating on a complete training set (marked with the dashed line in Fig. 2 and Fig. 3).

The classification quality and the reduction level (the fraction of removed samples) are presented for each algorithm in percentages.

5. RESULTS OF REDUCTION PROCEDURES

The training sets were reduced by the following algorithms: Hart's CNN [9], Gates' RNN [7], RRCNN [2], Gowda and Krishna's (GK) [8], Tomek's [24], Dasarathy's MCS [4], Skalak's RMHC-P [20], Kuncheva's GA [13,14], Cerveron and Ferri's TS [2], DSAds, SeqRA, SeqDSA and RM (designated as bestRM). In RRCNN the number of permutations was set to 100. In GA the number of nearest neighbours k was set to 1, the number of iterations to 200, the reduction rate to 0.1, number of chromosomes to 20, the crossover rate to 0.5, the mutation rate to 0.025 and a fitness function

$$J(Y) = \frac{n(Y)}{n} - 0.01 \cdot card(Y), \tag{1}$$

where: $n(Y)$ denotes the number of correctly classified samples from a training set using only the reduced reference set Y to find the k of nearest neighbors (a sample s is classified using k-NN with Y-$\{s\}$), n – the number of all samples in a training set and $card(Y)$ – the cardinality of Y. The description of all GA parameters is available in [14]. The implementation of GA is equivalent to that proposed in [14] as well. The values of TS parameters are set as they were proposed in [2] with one difference: the condensed method of creating the initial subset was used for datasets like SAT, PHONEME and WAVEFORM due to a very long phase of constructive initialization required for these training sets. In RMHC-P [20]: $k = 1$, m is equal to average reduced set size obtained by DSAds, SeqRA, SeqDSA and bestRM (for better comparison of the algorithms) and n was set to 200 for IRIS, 300 for GLASS and PARKINSONS, 400 for BUPA, 600 for WDBC, 700 for PIMA, 1000 for YEAST, 2000 for PHONEME and 3000 for SAT and WAVEFORM (the value of n was established experimentally). As it was mentioned in the third section, in DSAds and RM, for each training set the best resultant reduced set was chosen (the set with the highest classification quality) from the group of the generated sets. In DSAds the sets were generated for ds set to the multiple of $0.05 \cdot ds_{max}$. In RM R_{init} was set to 9. The results are presented in Fig. 2.

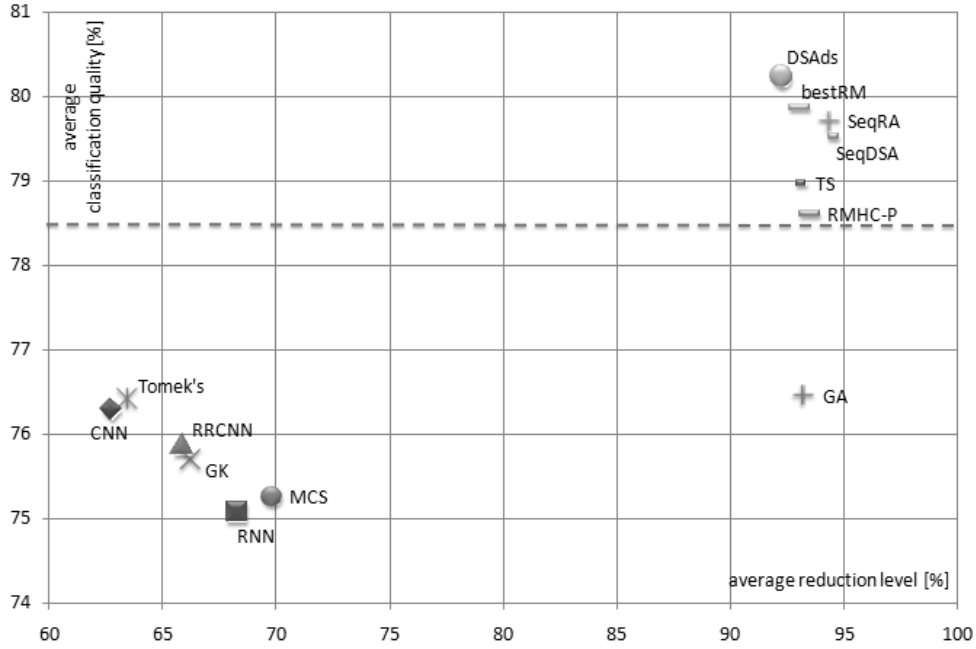

Fig. 2. Experimental results of the reduction methods

As we can see, almost all results of the developed algorithms are better than those of competitive algorithms, in the sense of the average reduction level and classification quality. The sequential algorithms received the highest average reduction level, almost 94.5%, simultaneously with the improvement of the classification quality (in comparison with the quality obtained on complete training sets). DSAds has the highest average fraction of correct classifications equal to 80.2% and also a very high reduction level. bestRM results are very similar.

The presented results are average for all tested training sets. For a particular training set the algorithm with the most satisfying results should be chosen.

The proposed algorithms differ from each other in their properties. The SeqRA has no parameters but it is random and its results are not repeatable. The SeqDSA has also no parameters, its results are repeatable, but the reduction phase for some larger sets can last few hours (SeqDSA reduced SAT for almost an hour). DSAds and bestRM have one parameter (in bestRM it can be permanently set to 9) and need the phase of choosing the best reduced set. On the other hand, when a given size of a reduced set must be achieved DSAds seems to be the right choice.

The results of the TS and RMHC-P are close to the results obtained with the use of the proposed algorithms. But it must be taken into consideration that the former has three parameters and the latter – four parameters. Moreover, TS reduction phase can last very long even for middle datasets like SAT (over 6 hours).

6. RESULTS OF EDITING PROCEDURES

The training sets were edited by six algorithms: Wilson's ENN [25], Tomek's RENN [23], Tomek's All k-NN [23], Devijver and Kittler's MULTIEDIT [5], Kuncheva's GA [12] and proposed EAC. The ENN, RENN and GA were tested with k = 1 (ENN1, RENN1, GA1) and with k = 3 (ENN3, RENN3, GA3), All k-NN with k = 3 (All3NN) and RM (before EAC) with R_{init} = 9. In MULTIEDIT the parameter N was set to 3 and the parameter I to 5 due to small datasets (the meaning of the MULTIEDIT parameters is described in [5]). In GA (in accordance with [12]) the number of iterations was set to 200, the reduction rate to 0.8, number of chromosomes to 50, the mutation rate to 0.05 and the fitness function had the following form:

$$J(Y) = \frac{1}{n}\sum_{i=1}^{n(Y)} k_i, \qquad (2)$$

where $n(Y)$ and n denote the same as in the formula (1), k_i, for i = 1,2,…, $n(Y)$ – the number of neighbors leading to the correct classification of the i-th sample. The results are presented in Fig. 3.

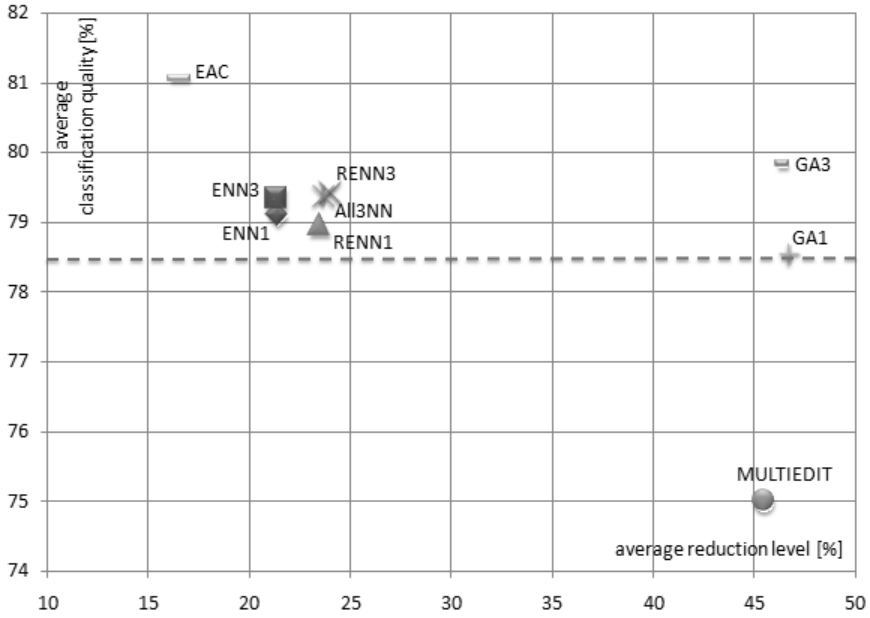

Fig. 3. Experimental results of the editing methods

As we can see, EAC received the highest classification quality, over 81% of correctly classified samples, simultaneously with the lowest reduction level equals to 16.5%. It confirms the assumption that it correctly recognizes and removes noisy samples. Competitive algorithms have improved the classification quality, except the MULTIEDIT, but none of them exceeded the

boundary of 80%. GA and MULTIEDIT reduced the reference set quite strongly which is questionable advantage due to the fact the main objective of these algorithms should be the improvement of the classification quality.

EAC has no parameters and the editing phase is very fast, but it requires initial RM reducing of a reference and a phase of choosing bestRM, what increase the complexity of this method.

7. CONCLUSIONS

To sum up, the proposed algorithms are mainly based on representativeness of samples. The majority of the proposed reduction methods have the following advantages:

- high fraction of correct classifications;
- very high reduction level of a reference set;
- unique solution;
- lack or only one parameter.

The edition algorithm strongly improves the classification quality, simultaneously with the very low reduction level, which confirms the property of the correct removal of noisy samples.

The theses of the research have been confirmed.

REFERENCES

[1] **Frank A., Asuncion A.:** UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[2] **Cerveron V., Ferri F.J.:** Another move towards the minimum consistent subset: A tabu search approach to the condensed nearest neighbor rule. IEEE Trans. on Systems, Man and Cybernetics, Part B: Cybernetics, Vol. 31(3), 2001, pp. 408-413.

[3] **Cover T.M., Hart P.E.:** Nearest neighbor pattern classification. IEEE Trans. Inform. Theory, Vol. IT-13, 1967, pp. 21-27.

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

[5] **Devijver P.A., Kittler J.:** On the edited nearest neighbor rule. Proc. 5th International Conf. on Pattern Recognition, 1980, pp. 72-80.

[6] **Duda R.O., Hart P.E., Stork D.G.:** Pattern Classification – Second Edition. John Wiley & Sons, Inc, 2001.

[7] **Gates G.W.:** The reduced nearest neighbor rule. IEEE Transactions on Information Theory, Vol. IT 18, No. 5, 1972, pp. 431-433.

[8] **Gowda K.C., Krishna G.:** The condensed nearest neighbor rule using the concept of mutual nearest neighborhood. IEEE Transaction on Information Theory, Vol. IT-25, 4, 1979, pp. 488-490.

[9] **Hart P.E.:** The condensed nearest neighbor rule. IEEE Transactions on Information Theory, Vol. IT-14, 3, 1968, pp. 515-516.

[10] **Kohavi R.:** A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc. 14th Int. Joint Conf. Artificial Intelligence, 1995, pp. 338-345.
[11] **Kośla P., Raniszewski M.:** Nowe metody selekcji cech i redukcji zbiorów odniesienia dla klasyfikatora typu 1-NN. Automatyka, Vol. 12(3), 2008, pp. 805-820.
[12] **Kuncheva L.I.:** Editing for the k-nearest neighbors rule by a genetic algorithm. Pattern Recognition Letters, Vol. 16, 1995, pp. 809-814.
[13] **Kuncheva L.I.:** Fitness functions in editing k-NN reference set by genetic algorithms. Pattern Recognition, Vol. 30, No. 6, 1997, pp. 1041-1049.
[14] **Kuncheva L.I., Bezdek J.C.:** Nearest prototype classification: clustering, genetic algorithms, or random search? IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol. 28(1), 1998, pp. 160-164.
[15] **Raniszewski M.:** Reference Set Reduction Algorithms Based on Double Sorting. Computer Recognition Systems 2, Advances in Soft Computing, Vol. 45, Springer Berlin/Heidelberg, 2007, pp. 258-265.
[16] **Raniszewski M.:** Double Sort Algorithm resulting in reference set of the desired size. Biocybernetics and Biomedical Engineering, Vol. 28(4), 2008, pp. 43-50.
[17] **Raniszewski M.:** The Sequential Reduction Algorithm for Nearest Neighbor Rule Based on Double Sorting. Computer Recognition Systems 3, Advances in Intelligent and Soft Computing, Vol. 57, Springer Berlin/Heidelberg, 2009, pp. 221-229.
[18] **Raniszewski M.:** The Edited Nearest Neighbor Rule Based on the Reduced Reference Set and the Consistency Criterion. Biocybernetics and Biomedical Engineering, Vol. 30(1), 2010, pp. 31-40.
[19] **Raniszewski M.:** Sequential Reduction Algorithm for Nearest Neighbor Rule. Computer Vision and Graphics, Second International Conference, ICCVG 2010, Proceedings, Part II, Lecture Notes in Computer Science Vol. 6375, Springer-Verlag Berlin Heidelberg, 2010, pp. 219-226.
[20] **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, pp. 293-301.
[21] The ELENA Project Real Databases [http://www.dice.ucl.ac.be/neural-ets/Research/Projects/ELENA/databases/REAL/].
[22] **Theodoridis S., Koutroumbas K.:** Pattern Recognition – Third Edition. Academic Press – Elsevier, USA, 2006.
[23] **Tomek I.:** An Experiment with the edited nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-6, No. 6, 1976a, pp. 448-452.
[24] **Tomek I.:** Two modifications of CNN. IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-6, No. 11, 1976b, pp. 769-772.
[25] **Wilson D.L.:** Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions On Systems, Man and Cybernetics, Vol. 2, 1972, pp. 408-421.
[26] **Wilson D.R., Martinez T.R.:** Reduction techniques for instance-based learning algorithms. Machine Learning, Vol. 38, No. 3, 2000, pp. 257-286.

METODY SILNEJ REDUKCJI I EDYCJI ZBIORU ODNIESIENIA DLA REGUŁY TYPU NAJBLIŻSZY SĄSIAD

Streszczenie

W artykule zaprezentowano tezy i podstawowe wyniki rozprawy doktorskiej dotyczącej nowych metod redukcji i edycji zbioru odniesienia dla reguły typu najbliższy sąsiad (NN). Przedstawione metody mają na celu przyspieszenie działania reguły NN i poprawę jej jakości klasyfikacji. Zaprezentowane algorytmy w większości wykorzystują pojęcie reprezentatywności obiektu. Wyniki ich działania zostały porównane z wynikami działania innych popularnych algorytmów redukcji i edycji.

Promotor: dr hab. Adam Jóźwik Politechnika łódzka
Recenzenci: dr hab. inż. Leszek Chmielewski,
Instytut Podstawowych Problemów Techniki PAN
Prof. dr hab. inż. Dominik Sankowski, Politechnika Łódzka