Biocybernetics and Biomedical Engineering
2010, Volume 30, Number 1, pp. 31–40

The Edited Nearest Neighbor Rule Based on the Reduced Reference Set and the Consistency Criterion

MARCIN RANISZEWSKI*

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

In this paper a new editing procedure for the Nearest Neighbor Rule (NN) is presented. The representativeness measure is introduced and used to choose the most representative samples of the classes. These samples constitute a reduced reference set. An edited reference set is created from all the training set samples (including samples from the reduced set), which are correctly classified by the NN rule operating with the reduced set. The performance of the presented method is evaluated and compared with five other well-known editing techniques, on five medical datasets.

Keywords: editing techniques, nearest neighbor rule, consistency criterion, reference set reduction, representativeness measure

1. Introduction

The Nearest Neighbor (NN) rule is a very popular and effective method of Pattern Classification [1, 2]. The NN rule is a particular case of the *k* Nearest Neighbor rule [1, 2] (*k*-NN) operating with $k = 1$. Despite of the fact that the fundamentals of this rule were presented in the early 50s [3,4] (further developed in [5]), its simplicity and robustness cause that the NN rule is still widely used in many serious applications.

The rule does not require a training phase (the whole training set is treated as a reference set in further classification phase) and the classification error is never beyond the double value of Bayesian classification error for a sufficiently large training set [1].

The main disadvantages of the NN rule are the computational load and memory requirements for large training sets and worse classification results for datasets with noisy and atypical samples.

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

Received 09 July 2009; accepted 04 November 2009

One of the solutions of the former problem is reduction of the training set: only certain samples are chosen from an original training set and these samples form the reduced reference set. First of all, the reduced reference set should be much smaller than the original one and (if possible) it should provide a similar or higher fraction of correct classifications than that obtained with the complete training set.

The possible solution to the problems posed by such disadvantages as noisy and atypical samples in the training set is the editing of the training set: noisy, mislabelled and atypical samples are removed from the reference set. The remaining samples constitute the edited reference set. This set should achieve more accurate and reliable classification results than the original one. This is the main aim of editing strategies in contrary to the reduction techniques.

In this paper a new editing procedure is presented. It is based on the results obtained by an application of the reference set reduction algorithm proposed in the present study and uses the consistency criterion (defined in section 3).

The second section of this paper contains brief descriptions of well-known editing methods. In the third section the proposed editing procedure is presented. In the next two sections results of the tests are described. All the algorithms from the second section were implemented and compared with the proposed editing method on five well-known medical datasets. The last, sixth section contains the conclusions.

2. Well-known Editing Algorithms

In 1972 Wilson proposed a modified version of the *k*-NN – the Edited Nearest Neighbor (ENN) rule [6]. The ENN rule is simply the *k*-NN rule operating with an edited reference set. The edition procedure proposed by Wilson is based on classification of each sample using the *k*-NN rule with the training set (for a given odd value of *k*). Misclassified samples are remembered and removed from the reference set after classification of all samples. The remaining samples constitute the edited reference set.

Tomek in 1976 described two new editing procedures based on Wilson's ENN rule: the Repeated ENN (RENN) rule and All *k*-NN rule [7]. In the RENN rule the ENN edition is repeated until no more samples are removed from the edited reference set. In All *k*-NN rule each sample from a reference set is classified using the *k*-NN rule operating with the training set (for each value of *k* from 1 to a given maximum, e.g. in All 3-NN rule to the value of 3). Each misclassified sample is marked. After the classification of all the samples, the marked samples are removed. Like in Wilson's ENN rule the remaining samples constitute the edited reference set. It is worth noticing that All 1-NN rule and the ENN rule are the same editing methods.

Four years later, in 1980, Devijver and Kittler proposed the MULTIEDIT algorithm [8]. It can be described in the following few steps (initially the training set is denoted as *S*):

1. Diffusion: make a random partition of samples from S into N separable subsets $S_1, S_2, ..., S_N$.
2. Classification: Classify samples from S_i using NN rule with $S_{(i+1) \bmod N}$, (where $x \bmod y$ denotes the remainder of division of x by y), $i = 1, 2, ..., N$.
3. Editing: Remove all the samples, that were misclassified at step 2.
4. Confusion: Pool all the remaining samples into a new set S.
5. Termination: If the last I iterations produced no editing, exit with the final solution S, otherwise go to step 1.

The experiments with MULTIEDIT, made by authors, have demonstrated that this editing procedure produced the homogeneous clusters: each cluster consisted of the samples from only one class.

Kuncheva in 1995 described a heuristic method based on Genetic Algorithms (GA). Every subset Y of the training set is represented as a binary string called "chromosome" (i-th bit is set to one if the i-th sample is in the reduced set and otherwise to zero). Each bit in a chromosome is initially set to one with the predefined probability, called the reduction rate. In the reproduction phase, two offspring chromosomes are produced by every couple of parent chromosomes (the crossover point is randomly selected and the right parts of parent chromosomes are exchanged) and then each bit of each offspring chromosome alternates (mutates) with a given mutation rate. Kuncheva proposed a fitness function J based on the number of neighbors leading to the correct classification:

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

where m denotes the number of correctly classified samples from the training set using only the edited reference set Y to find the k nearest neighbors (the sample s is classified using k-NN with Y-$\{s\}$), k_i, $i = 1,2,\ldots,m$ – the number of neighbors leading to the correct classification of the i-th sample and n – the number of all samples in the training set.

3. The Editing Procedure Based on the Reduced Reference Set and the Consistency Criterion

The editing procedure called EAC as an abbreviation of Editing Algorithm based on Consistency, proposed in this paper, consists in finding the samples which are correctly classified by a certain subset of the training set initially chosen as a representative. The term "representative sample" means the sample x being a nearer neighbor than the nearest neighbor from the opposite class to many samples from the same class as x in the training set. The measure of sample representativeness was introduced in [10]: the representativeness measure of the sample x is the number of samples (voters), which lie nearer *x* than their nearest neighbours from the opposite class. In Fig. 1 the representativeness measure of sample *x* equals 3.

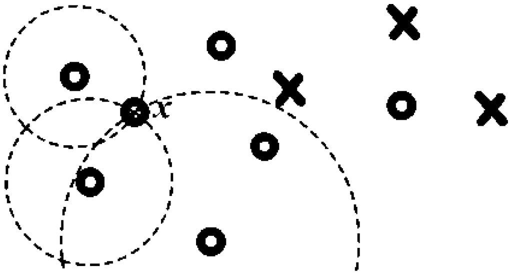

Fig. 1. The representativeness measure for the sample *x* equals 3

The sample with the higher representativeness measure (with more voters) is more representative for its class.

The reduction algorithm which creates a subset containing the representative samples has one parameter *R*. It is a minimal measure of representativeness. Only the sample with the greater representativeness measure than *R* can be added to the reduced reference set. The greater the value of *R* is, the smaller the number of samples in reduced reference set is. The value of *R* should be established experimentally.

The algorithm can be described in the following steps (two flags are assigned to each sample from the reference set: "not in/in the reduced reference set" and "available/unavailable"):

1. Mark all samples in training set as "not in the reduced reference set" and "available".
2. Count the representativeness measure for all samples marked as "available" and "not in the reduced reference set". Samples marked as "unavailable" can not be the voters.
3. If the greatest representativeness measure is less or equal to *R*, go to step 6.
4. Choose the sample with the greatest representativeness measure (the ties break randomly) and mark it as "in the reduced reference set". Mark all its voters as "unavailable".
5. If there is at least one sample marked as "not in the reduced reference set" and "available", go to step 2.
6. The samples marked as "in the reduced reference set" constitute the reduced reference set.

It is obvious that the value of *R* should be fitted to achieve the best classification accuracy of the resultant reduced set. There can be different "best" value of *R* for different data sets. For this reason, the above algorithm ought to be implemented in the version, which returns a few reduced reference sets, each set corresponding to the different value of R, e.g. ten reduced sets for R decreasing from 9 to 0 by a step of 1. The proposed improvement can be realized by a simple modification of the third step of the algorithm:

3. If the greatest representativeness measure is less or equal R, return the reduced reference set (all samples marked as "in the reduced reference set" constitute the reduced reference set) and if $R = 0$, end the algorithm, otherwise, let $R = R - 1$ and once more execute step 3.

Hence, e.g. if the initial value of R (let's denote it as R_{init}) is set to 9, the proposed algorithm returns the reduced reference sets for $R = 9$, $R = 8$, ..., $R = 0$ (if the step 6 will not be executed). If for some value of $R \neq 0$, the step 6 will be executed (there will be no more samples marked as "not in the reduced reference set" and "available"), the reduced reference sets for smaller values of R will be the same as that returned in step 6.

Now, from all of the reduced subsets generated for the given dataset, the "best" reduced reference set (let's denote it *bestRM*) can be chosen using one of the well-known techniques of classification quality estimations (e.g. ten-fold cross--validation [11]).

In practice, the value of R_{init} may be permanently set to 9 for any datasets. The above presented, improved version of the algorithm is called RM as an abbreviation of Representativeness Measure.

In our further considerations, a consistency criterion defined in [12] will be used. The set X is consistent with the set Y, when all samples from Y are correctly classified using the NN rule operating with the samples from X.

The edited reference set is built in the following way: each sample from the training dataset, which is correctly classified using the NN rule operating with *bestRM* is added to the edited reference set. Hence, the *bestRM* is consistent with the created edited reference set.

It is intuitively understood that, if the *bestRM* contains representative samples and offers the best classification quality, the edited reference set is created from samples, which are correctly classified by *bestRM* and for this reason they are not noisy and not atypical.

4. Experimental Results

The tests were made on five medical datasets:

- BUPA liver disorders (BUPA) [7] (BUPA Medical Research Ltd.) (number of classes: 2, number of attributes: 6, number of instances: 345) – the first 5 attributes are all blood tests (mean corpuscular volume, alkaline phosphatase, 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.

- Parkinson's Disease Data Set (PARKINSONS) [7] (number of classes: 2, number of attributes: 22, number of instances: 195) – biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each of 22 features (average vocal fundamental frequency, maximum vocal fundamental frequency, minimum vocal fundamental frequency, 5 measures of variation in fundamental frequency, 6 measures of variation in amplitude, 2 measures of ratio of noise to tonal components in the voice, 2 nonlinear dynamical complexity measures, signal fractal scaling exponent, 3 nonlinear measures of fundamental frequency variation) is a particular voice measure, and each sample corresponds to one of 195 voice recordings from these individuals. The main aim of the data is to discriminate healthy people from those with PD.
- Pima Indians Diabetes Database (PIMA) [7] (National Institute of Diabetes and Digestive and Kidney Diseases) (number of classes: 2, number of attributes: 8, number of instances: 768) – all patients in this database are Pima-Indian women at least 21 years old and living near Phoenix, AZ, USA. The classes correspond to positive and negative test for diabetes. All 8 features are clinical findings: number of times pregnant, plasma glucose concentration at 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), triceps skin fold thickness (mm), 2-hour serum insulin (μU/ml), body mass index, diabetes pedigree function and age (years).
- Wisconsin Diagnostic Breast Cancer (WDBC) (Diagnostic) [7] (number of classes: 2, number of attributes: 30, 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. The two classes are: "malignant" and "benign" type of the diseases.
- Protein Localization Sites (YEAST) [7] (number of classes: 10, number of attributes: 8, number of instances: 1484. The paper [9] describes a predecessor to this dataset and its development.

Stratified ten-fold cross-validation was used for each experiment [11]. To estimate the classification quality the NN rule was used. All tests were made on Intel Core 2 Duo, T7250 2.00 Ghz processor with 2 GB RAM.

The training datasets were edited with six algorithms: ENN, RENN, All *k*-NN, MULTIEDIT, GA and the proposed EAC. Before EAC editing, the reduction with RM was made. The ENN, RENN and GA were tested for $k = 1$ (ENN1, RENN1, GA1) and for $k = 3$ (ENN3, RENN3, GA3), All *k*-NN for $k = 3$ (All3NN) and RM for $R_{init} = 9$. In MULTIEDIT the parameter N was set to 3 and the parameter I to 5 due to small size of datasets. In GA (in accordance with [9]) the number of iterations was set to 200, the reduction rate to 0.8, number of chromosomes to 50 and the mutation rate to 0.05. After RM reduction, for each dataset *bestRM* is chosen using stratified ten-fold cross-validation (for both BUPA and PIMA the "best" value of *R* is equal 3, for YEAST and WDBC: $R = 2$ and for PARKINSONS: $R = 1$). Subsequently, as it was described in the Section 3, *bestRM* is used to build the edited reference set in EAC.

All the algorithms were implemented in Java.

Table 1. The test results: classification qualities. All fractions are presented in percentages. The number under mean value for specific dataset is a standard deviation. Column called "compl." means the classification quality using the NN rule operating on complete training set, the last row "avg" presents average values in columns from all five datasets

	compl.	ENN1	ENN3	RENN1	RENN3	All3NN	MULTI-EDIT	GA1	GA3	bestRM	EAC
BUPA	62.61	63.73	63.74	65.18	64.03	64.32	58.87	63.49	67.50	68.07	71.56
	7.30	5.79	7.02	5.89	8.01	6.86	8.98	8.63	7.98	9.28	7.06
PIMA	67.20	68.49	71.37	68.10	72.67	71.36	70.71	69.02	71.63	74.23	74.23
	3.99	3.83	4.72	4.17	4.04	4.01	5.19	3.91	4.05	4.52	4.13
PARKINSONS	84.49	84.04	84.04	84.07	84.57	83.54	81.07	84.51	84.49	85.60	86.09
	6.49	4.09	4.09	3.94	4.43	4.26	5.24	5.76	5.94	7.37	5.18
WDBC	91.19	92.25	92.61	92.77	93.49	92.61	90.50	91.55	92.78	93.49	93.31
	4.16	2.14	1.68	2.35	1.72	1.68	3.78	2.40	3.43	2.07	1.67
YEAST	53.04	56.06	56.11	56.80	56.38	56.59	52.91	53.22	54.65	57.08	57.82
	4.70	3.96	5.44	4.34	4.78	4.79	4.20	4.04	5.18	5.10	4.24
avg	71.70	72.92	73.57	73.39	74.23	73.68	70.81	72.36	74.21	75.69	76.60
	5.33	3.96	4.59	4.14	4.60	4.32	5.48	4.95	5.32	5.67	4.45

Table 2. The test results: reduction levels (fractions of discarded samples). All fractions are presented in percentages. The number under mean value for specific dataset is a standard deviation. The last row "avg" presents average values in columns from all five datasets

	compl.	ENN1	ENN3	RENN1	RENN3	All3NN	MULTI-EDIT	GA1	GA3	bestRM	EAC
BUPA	0.00	38.07	38.62	41.35	43.74	41.84	75.17	48.80	51.47	96.30	30.82
	0.00	1.47	2.00	1.58	2.38	1.96	5.37	3.72	1.88	0.47	1.89
PIMA	0.00	32.16	30.02	35.26	33.75	36.37	56.12	49.06	51.48	95.53	24.16
	0.00	1.05	0.99	1.04	1.25	1.11	1.46	1.75	2.45	0.27	0.86
PARKINSONS	0.00	15.62	16.41	16.64	17.55	17.27	32.14	46.72	44.10	88.89	11.05
	0.00	1.29	0.80	1.16	1.17	0.95	4.24	2.99	5.50	0.92	1.01
WDBC	0.00	8.36	8.05	8.77	8.20	9.26	13.85	48.76	49.72	96.52	5.92
	0.00	0.52	0.48	0.58	0.41	0.50	0.47	1.65	2.86	0.18	0.26
YEAST	0.00	47.87	46.53	51.49	52.31	50.58	76.69	48.68	49.78	95.26	40.66
	0.00	0.84	0.88	0.84	0.93	1.07	2.13	1.95	1.26	0.25	1.43
avg	0.00	28.42	27.92	30.70	31.11	31.06	50.79	48.40	49.31	94.50	22.52
	0.00	1.03	1.03	1.04	1.23	1.12	2.73	2.41	2.79	0.42	1.09

5. Discussion

The experimental results are presented in Fig. 2. The dashed line indicates the average fraction of correct classifications using the complete training set as the reference set. The bestRM results are not presented in Fig. 3, because RM is the reference set reduction algorithm and we deal with reference set editing.

The results of ENN, RENN and All *k*-NN results are very similar. Their average fractions of correct classifications are between 72.9% (for ENN with $k = 1$) and 74.2% (for RENN with $k = 3$) and average reduction levels between 27.9% (for ENN with $k = 3$) and 31.1% (for RENN with $k = 3$ and All *k*-NN with $k = 3$).

The Kuncheva's GA and MULTIEDIT results are characterized by high reduction levels (from 48.4% to 50.8%). However, this is not so important in editing methods. MULTIEDIT offer the worst average fraction of correct classifications, even worse than the average fraction obtained for the complete reference set. It is possible that with the other values of parameters MULTIEDIT results would be better. GA with $k = 3$ results in better fractions of correct classification (approx. 74.2%) than with $k = 1$ (and the same for RENN with $k = 3$).

The highest average fraction of correct classification is offered by the proposed EAC: 76.6% (the results are approx. over 3% better than the best results of the other methods: RENN and GA with $k = 3$). However, EAC offers the worst reduction level: approx. 22.5%.

The reference set reduction algorithm: bestRM offers reduction level of approx. 94.5% and the fraction of correct classification of approx. 75.9%, higher than almost all the tested editing procedures (except the proposed EAC).

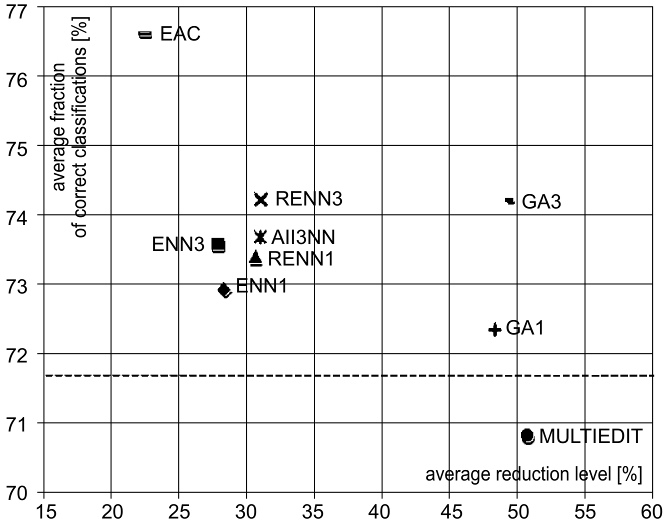

Fig. 2. The experimental results of the NN editing strategies

Almost all the editing methods (except MULTIEDIT) improve the fraction of correct classification in comparison with the NN rule operating with the complete reference set.

The training phase of all algorithms was very short (a few milliseconds or seconds), except Kuncheva's GA (a few minutes or even 2 hours for YEAST dataset).

6. Conclusions

From the experimental data presented above the following conclusions can be drawn, in relation to the proposed EAC editing technique:

- the highest average fraction of correct classifications (in comparison with ENN, RENN, All *k*-NN, MULTIEDIT and Kuncheva's GA);
- short time of the training phase in comparison with Kuncheva's GA;
- only one parameter R_{init} (that can be permanently set to 9 for any dataset) in opposition to MULTIEDIT and Kuncheva's GA;
- the unique solution in opposition to MULTIEDIT and Kuncheva's GA;
- simultaneous receiving of the reduced reference set as well as the edited reference set, so one can select which of these two solutions is more satisfactory.

References

1. Duda R.O., Hart P.E., Stork D.G.: Pattern Classification – Second Edition. John Wiley & Sons, Inc, 2001.
2. Theodoridis S., Koutroumbas K.: Pattern Recognition – Third Edition. Academic Press – Elsevier, USA, 2006.
3. Fix E., Hodges J.L.: Discriminatory analysis – nonparametric discrimination: Consistency properties. Project 21-49-004, Report No. 4, USAF School of Aviation Medicine, Randolph Field, TX, USA, 1951, 261–279.
4. Fix E., Hodges J.L.: Discriminatory analysis — nonparametric discrimination: Small sample performance. Project 21-49-004, Report No. 11, USAF School of Aviation Medicine, Randolph Field, TX, USA, 1952, 280–322.
5. Cover T.M., Hart P.E.: Nearest neighbor pattern classification. IEEE Trans. Inform. Theory, 1967, IT-13, 21–27.
6. Wilson D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. on Systems, Man and Cybern., 1972, 2, 408–421.
7. Tomek I.: An Experiment with the edited nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, 1976, SMC-6, 6, 448–452.
8. Devijver P.A., Kittler J.: On the edited nearest neighbor rule. Proc. 5th Internat. Conf. Pattern Recognition, 1980, 72–80.
9. Kuncheva L.I.: Editing for the k-nearest neighbors rule by a genetic algorithm. Pattern Recognition Letters, 1995, 16, 809–814.
10. Raniszewski M.: Reference set reduction algorithms based on double sorting. Computer Recognition Systems 2: 5th International Conference on Computer Recognition Systems CORES'07, Springer-Verlag, Berlin-Heidelberg, 2007, 258–265.

11. Kohavi R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc. 14th Int. Joint Conf. Artificial Intelligence, 1995, 338–345.
12. Hart P.E.: The condensed nearest neighbor rule. IEEE Transactions on Information Theory, 1968, IT-14, 3, 515–516.
13. 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.
14. Nakai K., Kanehisa M.: Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria. PROTEINS: Structure, Function, and Genetics 1991, 11, 95–110.