Decontamination of Training Samples for Supervised Pattern Recognition Methods[1]

Ricardo Barandela and Eduardo Gasca

Lab for Pattern Recognition
Instituto Tecnologico de Toluca, Mexico
rbarandela@hotmail.com
egasca@it-toluca.edu.mx

Abstract. The present work discusses what have been called 'imperfectly supervised situations': pattern recognition applications where the assumption of label correctness does not hold for all the elements of the training sample. A methodology for contending with these practical situations and to avoid their negative impact on the performance of supervised methods is presented. This methodology can be regarded as a cleaning process removing some suspicious instances of the training sample or correcting the class labels of some others while retaining them. It has been conceived for doing classification with the Nearest Neighbor rule, a supervised nonparametric classifier that combines conceptual simplicity and an asymptotic error rate bounded in terms of the optimal Bayes error. However, initial experiments concerning the learning phase of a Multilayer Perceptron (not reported in the present work) seem to indicate a broader applicability. Results with both simulated and real data sets are presented to support the methodology and to clarify the ideas behind it. Related works are briefly reviewed and some issues deserving further research are also exposed.

Keywords: Supervised methods, Nearest neighbor classifier, learning, depuration methodology, generalized edition.

1. Introduction

Traditionally, pattern recognition methods have been sorted into two broad groups: supervised and unsupervised, according to the level of previous knowledge about the training sample identifications in the problem at hand. Much of the research work in the frame of supervised pattern recognition has been almost entirely devoted to the analysis of the characteristics of classification algorithms and to the study of feature selection methods. Recently, however, an increasing emphasis is being given to the evaluation of procedures used to collect and to clean the training sample, a critical aspect for effective automatization of discrimination tasks.

[1] Work partially supported by Project Conacyt 32016-A and by Project Cosnet 744.99P

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 621-630, 2000.

Supervised classifiers´ design is based on the information supplied by the training sample (TS), a set of training patterns or prototypes representing all relevant classes and with correct classes labels. In several practical applications, however, class identification of prototypes is difficult and very costly and, as a consequence, some imperfectly or incorrectly labeled prototypes may be present in the TS, leading to situations lying in between supervised and unsupervised methods, or as they have been called: imperfectly supervised pattern recognition situations. Examples have been reported in medical diagnoses, drawing of pronostic maps of mineral deposits and, particularly, in the interpretation of remotely sensed data. In this later domain, training field selection and the yielding of suitable training statistics have been the concern of several researchers and practitioners, e.g. [1, 8, 10, 19, 29, 40]. Foody [17, 18] discusses to some extent the difficulties introduced into the classification process by those prototypes representing more than one class (e.g., those allocated on the border between classes) or being members of a class not considered when collecting the TS. Mather [32] refers to atypical elements in the TS that may belong to another class or may be hibrid or mixed elements.

The Nearest Neighbor (NN) rule [12] is a supervised nonparametric classifier, whose application does not require any assumption about probabilistic density functions. Some of its main features are described in the next Section. The performance of this classifier, as with any nonparametric method, is extremely sensitive to incorrectness or imperfections of the training sample. The present work introduces a methodology for decontaminating imperfect TSs while employing the NN rule for classification. This methodology can be regarded as a cleaning process removing some suspicious elements from the TS, or correcting the labels of some others and retaining them. Although conceived specifically for the NN rule, initial experiments with a Multilayer Perceptron seem to indicate a broader applicability. Results with both simulated and real data sets are presented to support the methodology and to clarify the ideas behind it. Related works are briefly reviewed and some issues deserving further consideration are also exposed.

2. The NN Rule

This is a classifier that combines conceptual simplicity and an asymptotic error rate bounded in terms of the optimal Bayes error. Let $TS=\{(x_1, \varphi_1), (x_2, \varphi_2), ..., (x_n, \varphi_n)\}$ be the training sample. That is, TS is the set of n pairs of random samples (x_i, φ_i) (i=1,2,...,n), where the label φ may take values in {1,2,...c} and φ_i designates the class of x_i among the c possible classes. For classifying an unknown pattern X with the NN rule, it is necessary to determine first the nearest neighbor x´ of X in the TS. That is, it is necessary to find x´ in TS such that:

$$d(X,x') = \min d(X,x_i) \quad x_i \ \varepsilon \ TS \qquad (1)$$

where d(,) means any suitable metric defined in the feature space. Then, the pattern X is assigned to the class identified by the label associated to x´. Devijver and Kitler [16] have expressed that "the basic idea behind the NN rule is that samples which fall close together in feature space are likely to belong to the same class". A more graphical description is due to [14]: "it is like to judge a person by the company he keeps".

Two other peculiarities of the NN rule have contributed to its popularity: a) easy implementation, and b) known error rate bounds. The computational burden of this classifier, very high with brute-force searching methods, has been considerably cut down by developing suitable data structures and associated algorithms (e.g., [25]) or by reducing the TS size (e.g., [2, 26]).

For improving the NN rule´s performance, Wilson [43] proposed a procedure (Edition technique) to preprocess the TS. The algorithm has the following steps:

1. For every x_i in TS, find the k (k=3 has been recommended) nearest neighbors of x_i among the other prototypes, and the class associated with the larger number of patterns among these k nearest neighbors. Ties would be randomly broken whenever they occur.
2. Edit the TS by deleting those prototypes x_i whose identification label does not agree with the class associated with the largest number of the k nearest neighbors as determined in the foregoing.

The benefits of the Edition technique have been supported by theoretical and empirical results (e.g., [3]). On the other hand, concerned with the possibility of considerable amounts of prototypes removed from the TS, Koplowitz and Brown [31] developed a modification of this technique, the Generalized Edition (GE). Here, for a given value of k, another parameter k´ is defined such that:

$$(k + 1) / 2 \le k' \le k \qquad (2)$$

For each prototype x_i in TS its k nearest neighbors are searched in the remainder of TS. If a particular class has at least k´ representatives among these k nearest neighbors then x_i is labeled according to that class, independently of its original label. Otherwise, x_i is edited (removed). In short, the procedure looks for modifications of the training sample structure through changes of the labels (re-identification) of some training patterns and removal (edition) of some others.

Although none of these two techniques was particularly aimed at facing contaminated training samples, both modify basically the structure of the TS and, therefore, were included in the empirical evaluation to be explained in the Section 4.

3. Incorrections in the Training Sample and Related Works

Traditional approaches to supervised pattern recognition imply the fulfillment of two basic asumptions concerning training samples in order to guarantee accurate identification of new patterns:

1. the set of c classes with representations in the training sample span the entire pattern space
2. the training patterns used to teach the classifier how to discriminate each class are actually members of that class.

Practical experience has shown that in many real applications one or both of these assumptions do not entirely hold and that violation of these requirements strongly degrade classification accuracy. In accordance with this perception, the number of papers and proposals for handling this subject has significantly increased in the last years.

Outlier data is a concept that has been considered in Statistics for some time. It has been defined [41] as: a case that does not follow the same model as the rest of the data. Now the term has come to the fore also in the Pattern Recognition and Data Mining areas. Reports about the effect of these "noisy" patterns when included in the training sample and how to counteract it have been published in [20, 28, 34, 38] among others. Even for unsupervised methods outlier data have been the concern of several researchers, e.g., [22, 30].

However, this term is being employed to cover a broad range of circumstances reflecting some confusion among disimilar situations and a lack of a rigurous and unified concept of outlier data. In general, there are three of these potential situations:

1. noisy or atypical data that can be produced by errors (measuring, capturing, etc), an unfortunate property of many large databases.
2. New unidentified patterns appearing in the classification phase and that do not belong to any of the classes represented in the TS (partially exposed environments: [13, 33]). These cases are usually handled by a reject option [15, 23].
3. Some authors employ the term outlier for denoting mislabed instances in the TS, what constitutes the main focus of the present work. Brodley and Friedl [9] employ a combination of classifiers to filter the training patterns looking for identification and elimination of wrong labeled training cases prior to applying the chosen learning algorithm. Basic differences with the procedure presented in the next section:

    i) they do not consider correcting labes of some contaminated training data
    ii) although they use real data for demonstration purposes, they modified intentionally the labels of some training patterns to simulate a situation that they state as very frequent in these applications (remotely sensed data).
    iii) The filtering method they propose cannot fully overcome the error in the data for noise levels of 30% or greater

4. Depuration of Training Samples

The NN rule, as any other nonparametric pattern recognition method, suffers from an extreme sensitivity to the presence of contaminated training sets [36]. Hence, the importance of building procedures to contend with imperfectly supervised environments. Barandela [2] reports a considerable amount of Monte Carlo experiments to assess some procedures usable for decontaminating training samples. The procedures included were: a) Edition b) Generalized Edition c) Mutual Neighborhood [21] d) Reidentification (after some ideas of Chitinenni [11]) and e) All k-NN, a variant of the Edition technique proposed in [37].

For comparing these procedures, experiments with simulated patterns from two Gaussian populations with different mean vectors and equal covariance matrix were carried out. In every experiment, the TS consisted of 200 prototypes and the independent test sample (used for validation purposes) contained 500 elements, always a half from each class. Five different levels (percentages) of members of the second class were wrong labeled as belonging to class 1. For each of these cases and combinations, 30 replications were done. The averaged results, misclassification percents on the test set, are shown in Table 1. The effect of Generalized Edition on the contaminated training samples was remarkable and its superiority over the rest of the evaluated methods was recorded in more than 98% of the individual replications.

Table 1. Comparison of methods for decontamination (averaged misclassification rates).

	Wrong labels in class 1				
	5%	15%	25%	35%	45%
Original TS	15.0	17.8	20.4	22.9	25.4
Generalized Edition	10.2	10.4	11.5	12.4	14.3

It should be noted that for the decision about the labeling of a prototype (and perhaps the transference to another class or relabeling) or about its edition (elimination from the TS), GE takes into account the labels of the k nearest neighbors of this prototype. That is, for evaluating the correctness of this training pattern label, the procedure is based on the information supplied by the labels of other prototypes which, in turn, can be incorrectly identified. From the results in Table 1 it is clear that the greater the percent of initially mislabeled prototypes the greater the percent of test patterns erroneously classified after the GE application. Nevertheless, the method produces an unquestionable improvement in the classifier´s performance, indicating the achievement of a TS structure with an appreciably reduced amount of wrong labels. This situation seemed to indicate that the reiteration or repetition of the procedure is convenient, because in every further application the environment will be less contaminated [2]. The idea was put into practice with the same simulated data sets, yielding results that are showed in Table 2.

It was found unnecessary to carry out more than three succesive applications. Already after the second application and more clearly after the third one, misclassification rates tend to reach stability, notwithstanding (and that confirms how proper this proposition is) the original amount of wrong labels. In this third application very few remotions and no transferences at all are produced. These results and those of the practical applications below indicate that the procedure will always stop after a finite number ot iterations.

Table 2. Reiteration of Generalized Edition – averaged error rates.

Wrong labels in	Original	Generalized Edition		
Class No. 1	Tr. Sample	1st app.	2nd app	3rd app
5%	15.0	10.2	10.1	10.1
15%	17.8	10.4	10.1	10.1
25%	20.4	11.5	10.2	10.2
35%	22.9	12.4	10.5	10.3
45%	25.4	14.3	10.8	10.6

Practical experiences with several real data sets in the geophysical domain [5] led to a novel methodology for preprocessing and Depurating the TS. The Depuration methodology involves several applications of Generalized Edition until stability has been reached in the structure of the TS and in the estimated error rate (leave-one-out), and then the application of Wilson´s Edition, eventually also reiterated. It has been observed that after the first or second application of Generalized Edition, prototypes elimination stops and transferences (relabelling) number decreases gradually. From the second application onwards these transferences only affect a part of those prototypes whose labels had been changed in the first iteration, producing an oscillatory movement with some prototypes being passed from one class to another and backwards. At each iteration, this "pendulum" effect influences upon an ever decreasing number of prototypes until a steady situation with no more movements is reached. At this point Edition has produced elimination only of those prototypes (or a part of them) that had remained oscillating until the last iterations. These and other characteristics of the Depuration methodology will become more distinct after a brief outline of some of the mentioned practical applications.

Application No. 1 [4] Six features had been measured on a set of 268 unlabeled patterns (strata belonging to a well log data). A clustering method sorted the data set into 4 groups that, since the ultimate purpose was to explore possibilities of gas-oil manifestations, have been regarded as:

- classes 1 and 3, conformed by nonprospective strata
- classes 2 and 4,presenting strong association with the prospective strata of the area

This configuration was taken as the initial TS to be used afterwards for classifying additional strata, building up a semisupervised classification system ([24],[27]). Obviously, this TS should be assumed as imperfectly supervised since, by the very nature of clustering algorithms some prototypes have received unreliable labels. Hence the necessity to employ Depuration process for structure improvement. Generalized Edition was applied twice and the second application produced no remotions and only a few transferences (about 4%). On the other side, the error estimate decreased more than 75%. Edition got also a remarkable improvement in the performance, yet with a small size reduction. As a final task, the processed TS was used to classify, with the NN rule, another 628 strata coming from three different well logs. Resulting assignations were verified using ancillary information and evaluated as more accurate than those got when employing the whole original TS.

Application No. 2 [35] This work was aimed to characterize the stratigraphic sequences present in an oil deposit, on the basis of an evaluation of several petrophysical and geophysical parameters, adding up to 12 features. A TS with 139 prototypes belonging to six classes was available. Peculiarities of the area under study gave place to vagueness and ambiguity in the identification of the prototypes, while skill and intuition of the interpreter played a decisive role in the decisions. Consequently, the TS at hand was adopted as imperfectly supervised, requiring a Depuration process. Most of the resulting transferences affected prototypes located on the borders between classes, shifting them up or down. As already explained, accuracy of these borders definitions had been mistrusted. Geophysicists in charge of the study of the area accepted the depurated TS as more consistent and more convenient for modeling purposes than the original one.

Application No. 3 [6] Data employed in this case came from a previous work [39], aimed to study the processes conditioning the shape of an ophiolitic sequence. The collected sample consisted of 187 prototypes sorted into four classes (according to the lithology). Seven features were recorded for each pattern. Geophysicists well acquainted with the data and with the procedures employed to collect and prepare them, regarded this TS as perfectly supervised and pronounced themselves strongly against any modification. When they accepted to collaborate in the evaluation of the Depuration process (see Table No. 3) merely with exploratory purposes, they demanded severe restrictions about amount and type of transferences to be allowed. Nevertheless, after the process was applied, these same domain experts accepted the depurated TS without hesitations as more accurate and better structured than the original one. Here again, an important part of the transferences involved prototypes located in the borders between different lithologies (classes). Besides, no undesirable transferences were recorded or occured in a very low level.

Table 3. Practical geophysical applications

	Applic. 1		Applic. 2		Applic. 3	
Procedure	No. Patt.	Err est. (%)	No. Patt.	Err est. (%)	No. Patt.	Err est. (%)
Original TS	268	34.0	139	42.4	187	52.9
GE (repeated)	130	1.5	136	2.9	134	6.7
Edition (repeated)	122	0.0	117	0.0	105	0.0

Concerning these applications, some issues may be remarked. Firstly, Depuration has evidenced its benefits in the three possible environments: unsupervised (as a complement to the cluster algorithm), imperfectly supervised, and supervised. In this last case, success of Depuration could be explained by the lack of a precise and well defined distinction (physical and conceptual) among classes, a rather common situation in the praxis, at least in the geoscience fields. As a byproduct, the Depuration process yields a significant reduction of the TS size and, accordingly, of the computational time required for subsequent works involving these data. With the exception of Application No. 2 when the low ratio Dimensionality/TS-size compelled the elimination of superfluous features at the very beginning of the process, feature selection was easier when implemented as an intermediate step. Employment of estimate L for error probability as a guide for conducting the process showed itself as suitable.

5. Discussion

Wilkinson et al. [42] mention adequacy of the training data as one of the factors dictating the performance of any classifier and manifest that it is "very often outside the control of the data analyst". The procedure here exposed evidences that an important contribution can be done for amending some defficiencies of the available TS and to increase its usefulness. The Depuration process has revealed, for both the simulated and the real data applications, significant benefits. Although the above explained real examples have been about classification of geophysical data, it should be clear that the procedure is independent of the origin of the data set in question and can be used in any application involving supervised classification. Actually, an application with remotely sensed data has already been reported [7].

Importance of this decontamination issue deserves further research. Efforts should be concentrated on the feasibility and the convenience of developing pertinent methodologies according to the different possible causes for contamination in the TS. The Depuration procedure manages quite well situations as those already highlighted in the applications above: misidentification of some prototypes due to the difficulty and high cost of collecting the data, or to some characteristics of the application domain that induce vagueness and a lack of clear separation among classes. Inclusion in the TS of prototypes belonging to not considered (untrained) classes and of mixed prototypes (representing more than one class) would require additions to or modifications of the procedure. Implementation of a classifier with a reject option and fuzzy approach (as in [19]) could be useful in this respect. What seems to be evident is that methodologies like the one here exposed lead to computer systems for classification tasks, not only faster than human interpreters. It has also evidenced ability to yield more accurate classification results and, at the same time, to provide a better model for the phenomenon under study. The later advantage is got through the relabeling of some of the training patterns.

References

1. Baker, J.R.,S.A.Briggs,V. Gordon, A.R. Jones, J.J. Settle, J. Townsheed and B.K. Wyatt (1991). Advances in classification for land cover mapping using SPOT HRV imagery, *Int. J. Remote Sensing*, 12 (5), 1071-1085.
2. Barandela, R. (1987). *The NN rule: an empirical study of its methodological aspects*. Unpublished Doctoral Thesis, Berlin.
3. -----------. (1990a). La regla NN con muestras de entrenamiento no balanceadas. *Investigacion Operacional*, X (1), 45-56.
4. ----------. (1990b). Metodos de reconocimiento de patrones en la solucion de tareas geologo-geofisicas. *Ciencias de la Tierra y el Espacio*, 19, 1-7.
5. ----------. (1995). Una metodologia para el reconocimiento de patrones en tareas geologo-geofisicas. *Geofisica Internacional,* 34 (4), 399-405.
6. ----------. (in press*). La practica de la clasificacion con la regla NN*. Editorial Ciencia y Tecnica, La Habana.
7. Barandela, R. and E. Castellanos (1996). La regla NN para la interpretacion de imágenes de percepcion remota. Tercer Taller Iberoamericano Geociencias e Informatica, La Habana.
8. Bolstad, P.V. and T.M. Lillesand (1991). Semi-automated training approaches for spectral class definition. *Int. J. Remote Sensing,* 13 (16), 3157-3168.
9. Brodley, C.E. and M.A. Friedl (1996). Identifying and eliminating mislabed training instances. *AAAI-96 Proc. of the Thirteenth Nat. Conf. On Artificial Intelligence*, AAAI Press.
10. Buchheim, M.P. and T.M. Lillesand (1989). Semi-automated training field extraction and analysis for efficient digital image classification. *Phot. Eng . & Rem. Sensing*, 55 (9), 1347-1355.
11. Chitinenni, C.B. (1979). Learning with imperfectly labeled patterns. Proc. Conf. on *Pattern Recognition and Image Processing*, Chicago.
12. Dasarathy, B.V. (Ed.) (1990). *Nearest Neighbor Norms: NN Pattern classification techniques*. IEEE Computer Soc. Press, Los Alamos, California.
13. -------- (1993). Is your Near Enough Neighbor friendly enough? Recognition in Partially Exposed Fuzzy Learning Environments. *Proc. North American Fuzzy Information Processing Society.*
14. Dasarathy, B.V. and B.V. Sheela (1977). Visiting Nearest Neighbors: a survey of Nearest Neighbors classification techniques. *Proc. Int. Conf. Cybernetics and Society*, Copenhaguen.
15. Denouex, T. (1995). A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25, 5, 804-813.
16. Devijver, P.A. and J. Kittler (1982). *Pattern Recognition - a statistical approach*. Prentice Hall, London.
17. Foody, G.M. (1990). Directed ground survey for improved Maximum Likelihood classification of remotely sensed data. *Int. J. Remote Sensing*, 11 (10), 1935-1940.
18. Foody, G.M. , N.A. Campbell, N.M. Trodd and T.D. Wood (1992). Derivation and application of probabilistic measures of class membership from the maximum likelihood classification. *Phot. Eng. & Rem. Sensing*, 58 (9), 1335-1341.
19. Gopal S. and C. Woodcock (1994). Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Phot. Eng. & Rem. Sensing*, 60 (2), 181-188.
20. Gopalakrishnan, M., V. Sridhar and H. Krishnamurthy (1995). Some applications of clustering in the design of neural networks. *Pattern Recognition Letters*, 16, 59-65.
21. Gowda, K.C. and G. Krishna (1979). Learning with a mutualistic teacher. *Pattern Recognition*, 11, 387-390.

22. Guha, S., R. Rastogi and K. Shim (1998). CURE: An efficient clustering algorithm for large databases. *ACM-SIGMOD Int. Conf. On Management of Data*, Seattle, Washington.

23. Hand, D.J. (1997). *Construction and assessment of classification rules*. John Wiley & Sons, Chichester.

24. Hardin, P.J. (1994). Parametric and Nearest Neighbor methods for hybrid classification: a comparison of pixel assignment accuracy. *Phot. Eng. & Rem. Sensing*, 60 (12), 1439-1448.

25. Hardin, P.J. and C.N. Thomson (1992). Fast nearest neighbor classification methods for multispectral imagery. *The Professional Geographer*, 44 (2), 191-201.

26. Huang, Y.S., K. Liu and C.Y. Suan (1995). A new method of optimizing prototypes for nearest neighbor classifiers using a multi-layer network. *Pattern Recognition Letters*, 16, 77-82.

27. Hung, C.C. (1993). Competitive learning networks for unsupervised training. *Int. J. Remote Sensing*, 14 (12), 2411-2415.

28. John, G.H. (1997). *Enhancements to the Data Mining Process*. PhD Thesis, Stanford University.

29. Kershaw, C.D. and R.M. Fuller (1992). Statistical problems in the discrimination of land cover from satellite images: a case study in Lowland Britain. *Int. J. Remote Sensing*, 13 (16), 3085-3104.

30. Kharim, Y. And E. Zhuk (1998). Filtering of multivariate samples containing 'outliers' for clustering. *Pattern Recognition Letters*, 19, 1077-1085.

31. Koplowitz, J. And T.A. Brown (1978). On the relation of performance to editing in nearest neighbor rules. *Proc. 4th Int. Joint Conf. on Pattern Recognition,* Japan.

32. Mather, P.M. (1999). Computer processing of remotely sensed images - an introduction. Wiley and Sons, Chichester, second edition.

33. Muzzolini, R., Y.H. Yang and R. Pierson (1998). Classifier design with incomplete knowledge. *Pattern Recognition*, 31, 4, 345-369.

34. Ritter, G. And M.T. Gallegos (1997). Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters*, 18, 525-539.

35. Rodriguez, M. and R. Barandela (1989). Aplicacion de algunas tecnicas de reconocimiento de patrones en la caracterizacion estratigrafica del yacimiento Varadero. *Serie Geologica*, 2, 29-38.

36. Sanchez, J.S., F. Pla and F.Ferri (1997). Prototype selection for the Nearest Neighbor rule through Proximity Graphs. *Pattern Recognition Letters*, 18, 6, 507-513.

37. Tomek, I. (1976). An experiment with the Edited Nearest Neighbor rule. *IEEE Trans. Syst., Man and Cyb.*, SMC-6, 448-452.

38. Urahama, K. And Y. Furukawa (1995). Gradient descent learning of nearest neighbor classifiers with outlier rejection. *Pattern Recognition*, 28, 5, 761-768.

39. Valladares, S. (1986). Metodologia para la evaluacion de los colectores y sus propiedades en las rocas pertenecientes al Complejo Aloctono Eugeosinclinal. Doctoral Thesis, La Habana.

40. Warren, S.D., M.D. Johnson, W.D. Goran and V.E. Diersing (1990). An automated objective procedure for selecting representative field sample sites. *Phot. Eng. & Rem. Sensing*, 56 (3), 333-335.

41. Weinsberg, S. (1985). *Applied Linear Regression*. John Wiley & Sons.

42 Wilkinson, G.G., F. Feriens and I. Kenellopoulos (1995). Integration of neural and statistical approaches in spatial data classification. *Geographycal Systems*, 2, 1-20.

43. Wilson, D.L. (1972). Asymptotic properties of Nearest Neighbor rules using edited data sets. *IEEE Trans. Syst., Man and Cyb.,* SMC-2, 408-421