Compactness and Complexity of Pattern Recognition Problems

Robert P.W. Duin, Pattern Recognition Group
Department of Applied Physics, Delft University of Technology, The Netherlands
duin@tn.tudelft.nl

Abstract

A compactness measure for labeled training sets is defined. Its relation with the complexity of pattern recognition problems is discussed and illustrated by some examples.

***keywords**: compactness, complexity, nearest neighbor error, classifier performance*

1. Introduction

Not all pattern recognition problems are equally difficult, even if they have the same Bayes error. The complexity of the problem depends also on the shape of the classes and thereby on the shape of the decision boundary. We will discuss the possibility of measuring the complexity of a pattern recognition problem by estimating the *compactness* of the problem.

The compactness hypothesis as introduced in the Russian literature in the 1960's, e.g. see Arkedev and Braverman [1] is the fundamental assumption necessary for any generalization from finite training sets. It roughly states that similar real world objects have to be close (having small distances) in their representation space. Wolpert [2] clearly shows by the 'no free lunch theorems' what happens without such assumption: in expectation (over a uniform prior distribution of classification problems) all classifiers are equally good, in particular they are as good the random class selector.

The compactness hypothesis states that distances between the object representations have some meaning. In the next section we will reformulate the original compactness definition in such a way that it can be measured and tested. This definition will be closely related to the Nearest Neighbor (NN) classifier, being the natural distance based classifier. It will be argued and illustrated by some examples that the compactness might be a useful measure for estimating the complexity of a classification problem. More complex problems need a larger training set, more complicated classifiers and an increases training effort.

2. A new definition of compactness

In order to be solvable, a classification problem should not be arbitrary complex, i.e. it should be possible to restrict it to a limited set of object measurements. The original formulation of the compactness hypothesis as given by Arkedev and Braverman [1] attempts to restrict to size of the class boundaries. We think that their definition implicitly assumes a low-dimensional object representation. For that reason a new, more general definition is needed. Moreover, we like to have a testable compactness hypothesis: it should be possible to verify the assumption that a pattern recognition problem is compact or is not compact using the training set.

Any definition for compactness should be based on a given metric $D(x,y)$, which is the distance between objects x and y. Changing the distance measure (e.g. by scaling one of the features) influences the behavior of some classifiers and thereby changes the problem. A natural mechanism directly based on the metric is the nearest neighbor rule. The following definition will be investigated:

The classes in a classification problem are compact if for an arbitrary object it is expected that its distance to an arbitrary object of the same class is smaller than its distance to an arbitrary object of another class:

The *compactness* of a labelled set of objects is defined as:

$$c = \text{Prob}\{D(x_1,x_2) < D(x_1,y) \mid \text{label}(x_1) = \text{label}(x_2), \text{label}(x_1) \neq \text{label}(y)\} \quad (1)$$

If $c > 0.5$, we define the classes to be compact. The following set of properties holds.

Property 1: For two equi-probable classes with identical probability density functions $c = 0.5$.

If the densities of x_2 and y are equal, the probability of $D(x_1,x_2) < D(x_1,y)$ equals the probability of $D(x_1,y) < D(x_1,x_2)$.

◊

Property 2: The the compactness is related to the class overlap ε^* as $c \leq 1 - \varepsilon^*$.

Suppose we want to classify an arbitrary object x using a random training set of one object per class. Let x be classified as the class of its nearest neighbor. According to (1) there is a probability c that this is correct. So $c \leq 1 - \varepsilon^*$ as no classifier can perform better than the class overlap.

◊

Property 3: For two-class problems the compactness is related to $E(\varepsilon_1)$, the expected error of the 1-NN rule with one object per class as $c = 1 - E(\varepsilon_1)$.

This directly follows from the definition (1) if x_1 is an arbitrary test object and x_2 and y are training objects.

◊

Property 4: If the classes are compact ($c > 0.5$) there exists a classifier with $\varepsilon < 0.5$.

This is a direct consequence of property 3.

◊

The compactness for classes with an infinite number of objects cannot be proven using a finite training set. A compactness assumption, however, can be tested using an estimated compactness, in which the probability in (1), estimated by a count over the training set is used as a test statistic.

3. Relation with problem complexity

The compactness (1) measures something different than the classification error. It is, like the Bayes error, a quantity that just depends on the problem and not on the choice for the classifier. The compactness is, however, much easier to estimate. Overlapping classes have a compactness of $c = 0.5$. For non-overlapping classes, with $\varepsilon^* = 0$, we didn't found a lower bound for the compactnes higer than 0. However, only in extreme situations a compactness of somewhat smaller than 0.5 is found. In the next section some examples are given.

As $c \leq 1 - \varepsilon^*$ and for $c = 1 - \varepsilon^*$ the classifier is simple and can be based on just two training objects, we will investigate whether the difference between c and its upper limit

$$q = 1 - \varepsilon^* - c \tag{2}$$

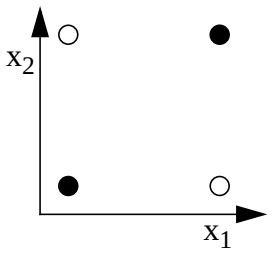
can be interpreted as a measure for the *complexity* of a classification problem. The problem complexity will influence the choice of the classifier complexity (number of degrees of freedom, or VC dimension) and the size of the training set in order to obtain a particular classification performance. More complex problems need more complex classifiers and/or larger training sets.

In this discussion we will restrict ourselves to the size of the training set. If the above defined measure for the problem complexity (2) is useful, then more complex problems need more training objects. As we want to discuss this over problems with different class overlap ε^* we are not interested in the training size itself, but in the speed the error ε limits to ε^*. A simple approach is to assume that learning curves approach ε^* exponentially as

$$\varepsilon = \alpha \exp(-vm) + \varepsilon^* \tag{3}$$

in which α is some constant, v is the so called classification power and m is the sample size. From a given learning curve $\varepsilon(m)$ the quantities of interest, v and ε^* can be estimated. (This assumption is certainly not globally valid. In many problems the classification power v as defined in (3) appears to depend on m).

4. Examples



In the famous XOR problem (see right) half of the objects of one class have a larger distance than all objects of the competing class. So the compactness is $c = 0.5 \times 0 + 0.5 \times 1 = 0.5$. In this extreme example with two non-overlapping classes we find the the same compactness $c = 0.5$ as for two entirely overlapping classes. This is in agreement with the well-known complexity of this problem. The 3-dimensional XOR problem has even a complexity of $c = \frac{7}{16}$, being the smallest compactness we have found so far. This emphasizes that it is not the compactness itself that makes a problem difficult, but its difference with the class overlap.

In the below figures two sets of classification problems are presented. Fig. 1 shows two overlapping normally distributed classes (in the table referred to as ‘normd’), having a Bayes error of 0.025. This doesn’t change by changing the variance of the non-discriminating feature *y*. However, a large *y*-variance makes this classification problem for scale-dependent classifiers like the NN-rule more complex. In fig. 2 it is shown how larger variances correspond with a decreasing compactness. In fig. 3 two non-overlapping classes are shown. Increasing their distance makes their separation easier and corresponds with an increasing compactness as presented in fig. 4.

In table 1 the compactness is given for a number of classification problems, together with three other quantities: the extrapolated value for $\varepsilon(n=\infty)$ for the NN rule , the classification complexity q (2) if $\varepsilon(n=\infty)$ is used as en estimate for ε^* and the classification power v as used in (3) for the NN rule.

We have also performed a set of experiments using 2000 characters from a NIST database. Ten numerical classes were normalized to 16x16 grey value representations as computed by De Ridder [3]. Each class has 200 training objects. The results for the nine classification problems between the character ‘3’ and the other numerals ‘0’ - ‘9’ are given in table 2.

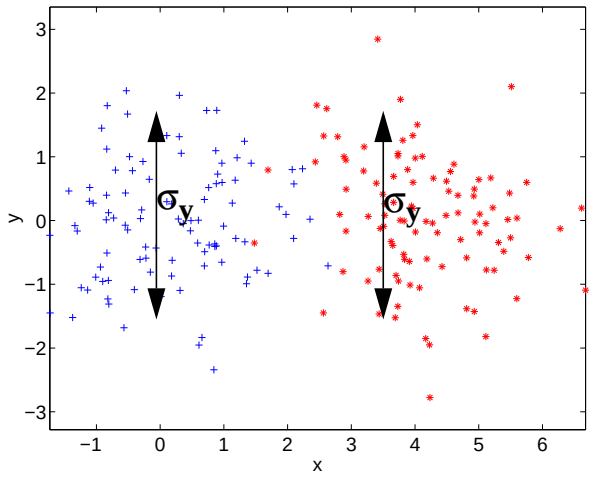


Fig. 1. A set of classification problems of variable complexity ('normd').

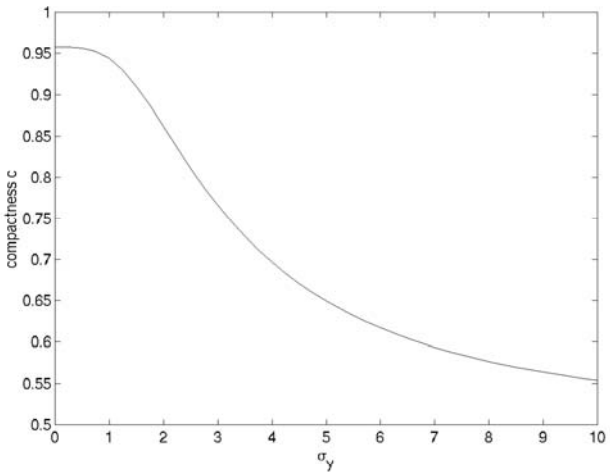


Fig. 2. Compactness as function of σ_y.

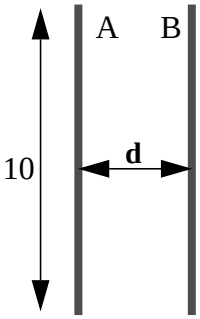


Fig. 3. A set of classification problems of variable complexity.

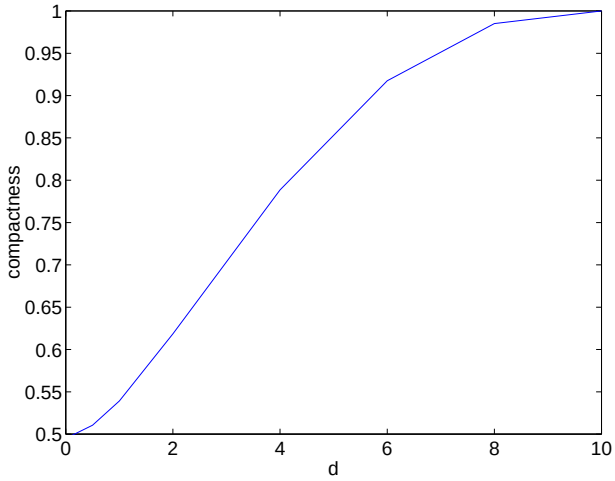


Fig. 4. Compactness as function of d.

From all these experiments it can be concluded that the compactness gives an indication of the problem complexity, but that the global problem based on overall statistics might be different from the problem of removing the last small errors which is studied by estimating the speed of learning for larger sample sizes.

Table 1: Compactness and complexity of some traditional classification problems

problem	samples	features	classes	*c*	ε(*n*=∞)	*q* = 1-*c*-ε*	*v*
normd, σ = 2	200	2	2	0.86	0	0.14	0.01
normd, σ = 6	200	2	2	0.62	0.03	0.35	0.02
Iris	150	4	3	0.93	0.03	0.04	0.03
Imox	192	8	4	0.82	0.05	0.13	0.10
Sonar	208	60	2	0.53	0.16	0.30	0.04
Spirals	194	2	2	0.50	0	0.50	0.007

Table 2: Compactness and complexity of 10 NIST database OCR classification problems

problem	samples	features	classes	c	$\varepsilon(n=\infty)$	$q = 1\text{-}c\text{-}\varepsilon^*$	v
3 <--> 0	400	256	2	0.83	0	0.17	0.001
3 <--> 1	400	256	2	0.86	0	0.14	0.002
3 <--> 2	400	256	2	0.83	0	0.17	0.004
3 <--> 4	400	256	2	0.87	0	0.13	0.004
3 <--> 5	400	256	2	0.72	0.03	0.25	0.018
3 <--> 6	400	256	2	0.88	0	0.12	0.0002
3 <--> 7	400	256	2	0.82	0	0.18	0.001
3 <--> 8	400	256	2	0.73	0.03	0.24	0.020
3 <--> 9	400	256	2	0.84	0.01	0.15	0.023

5. Discussion

There are two reasons for introducing the compactness. One is formally: it gives a fundament for applying generalization methods like classifiers by testing the compactness hypothesis $c > 0.5$. Not for all above examples the classification problems are compact. If we don't have the compactness confidence from the problem itself, the hypothesis $c = 0.5$ (no compactness) would be accepted for the spiral problem, as well as for the sonar problem.

Untill now we didn't study the issues of multi-class problems and non-equally probable classes. They both have their influence on the compactness. For instance, if we have many (n) completely overlapping classes, the compactness as defined in this paper will be equal to $1/n$. See also property 1. Or if we have two classes in a k-dimensional feature space, one in the origin and the other clustered in k points on the k axes all at distance one from the origin, and if the classes have prior probabilities of $\frac{1}{k+1}$ and $\frac{k}{k+1}$ then the compactness equals $c = \frac{2}{k+1}$. So here we have a two-class problem in which c can be arbitrary small. It may be necessary to include the number of classes and the class probabilities in the definitions for the compactness and problem complexity in order to extend the use of these quantities.

Finally it is concluded that still a number of issues has to be solved. We think, however, that the compactness statistic and the problem complexity estimator may give an indication on the difficulty of a classification problem. How this should be used in predicting sufficient sample sizes and selecting classifiers has further investigated.

6. References

[1] A.G. Arkedev and E.M. Braverman, *Computers and Pattern Recognition*, Thompson, Washington, D.C., 1966.

[2] D.H. Wolpert, *The Mathematics of Generalization*, Addison-Wesley, London, 1995.

[3] D. de Ridder, *Shared weights neural networks in image analysis*, Master Thesis, Delft University of Technology, 1996, 1-152.