Data Representations in Learning

Geetha Srikantan & Sargur N. Srihari

CEDAR, Department of Computer Science, SUNY Buffalo,
Suite 202, UB Commons, 520 Lee Entrance, Amherst,
NY 14228-2567
{geetha,srihari}@{cedar,cs}.buffalo.edu

Abstract

This paper examines the effect of varying the coarse-ness (or fine-ness) in a data representation upon the learning or recognition accuracy achievable. This accuracy is quantified by the least probability of error in recognition also known as the Bayes error rate, assuming that there are finite number of classes into which each data element can be classified. By modeling the granularity variation of the representation as a refinement of the underlying probability structure of the data, we examine how the recognition accuracy varies. Specifically, refining the data representation leads to improved bounds on the probability of error. Indeed, this confirms the intuitive notion that more information can lead to improved decision-making. This analysis may be extended to multiresolution methods where coarse-to-fine and fine-to-coarse variations in representations are possible. Our research was motivated by examining the change in the recognition accuracy of k-nearest neighbor classifiers while the resolution of the data - optical character images - is varied. In this domain, the data resolution is crucial in determining trade-offs in the speed and accuracy of the OCR system.

1 Introduction

1.1 Problem Description

Given a potentially infinite set Ω of distinct objects, the goal in pattern recognition is to find an effective method of classifying these objects into a finite set of classes, Θ. That is, find the mapping $\phi : \Omega \rightarrow \Theta$, such that given any arbitrary member of Ω, ϕ should automatically classify it to a single $\omega \in \Omega$; refer [6].

Formulating this problem in decision theoretic terms, let S be the unknown variable taking values in the finite set Θ, let X be the observation random variable taking values in the finite set Ω and let $A = \alpha$ be a set of decisions that is taken after observing the value x of Ω, via the decision function $\phi \in \Phi$. In the Bayesian approach to pattern recognition the random variable X is used to estimate the value of S, i.e. the observations are used to determine the class identity. We assume *a priori* knowledge of P the probability distribution of Θ and $W = W(x|\theta)$, the conditional probability matrix. In addition there is a loss function which assigns to each (θ, α) pair, either a 0 or a 1; i.e. each decision α made upon observing the value x of X, has a certain cost associated with it; this cost is 0 if the true value of the class identity θ matches α, and is 1 otherwise. The Bayes risk is the minimum risk that can be achieved over all possible decision functions for a given P and W.

In this paper we examine the critical role of the data representation in learning and classification based on data. Various factors determine any data representation such as the size or length of the data vector or matrix, global or local nature of the data, coarse or fine representation. The data representation might capture distinctive features about a function or class of functions (or patterns). Alternatively it could encode the discriminating features between several distinct functions or classes of functions/patterns. We restrict our research to the coarse to fine variation in data representation. We are particularly interested in the coarse to fine representation induced when objects are digitized at different resolutions by the image capturing

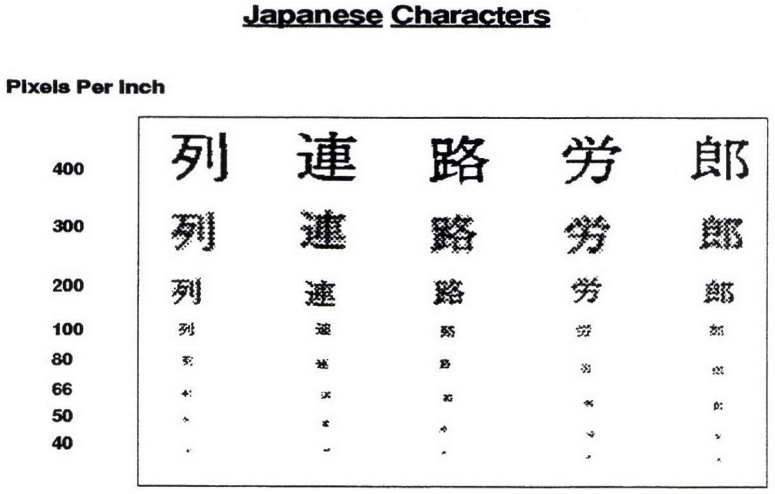

Figure 1: Two-Dimensional Images at Different Resolutions

mechanism. For example, the digitization could result in 200 dots per inch or 300 dots per inch; where the latter is a finer representation than the former. Representations involving multiple levels of data resolution are referred to as multiresolution representations.

1.2 Motivation

Figure 1 illustrates the coarsening phenomenon of reduced resolution for isolated machine-printed Japanese character images. It is observed that at lower resolutions the shape and structure of individual characters is progressively less well-defined. Discriminability of patterns with varying granularity of their representation is harder to establish as a consequence. Our goal is to formalize the relation between discriminability of patterns at varying resolutions.

this study has a particular significance in the domain of optical character recognition (OCR) (an integral component of many document image understanding systems). By using coarse data representation at the OCR level much higher speeds of operation are possible. However, this is at the cost of accuracy in recognition. Trade-offs between speed and accuracy of the OCR component are to be established to aid in system design.

1.3 Previous Work

Much work in sampling/interpolation theory [5, 14] is directed towards determining the minimum sampling rate necessary for signal reconstruction under various criteria. However, in pattern recognition the goal is to discriminate between patterns of distinct classes, rather than exact reconstruction.

Mean recognition accuracy of statistical pattern recognizers, [3], is a related issue. A finite-class pattern recognition problem defined by the number of training elements of each class and the class-conditional probability P for each quantized cell of a measurement space is considered. The *mean recognition accuracy* is then defined as the probability of correct recognition averaged over all sets P and over all sets of training elements of the assumed size. Examining mean recognition accuracy as a function of (a) the pattern measurement complexity (or features) and (b) the data set size, [4] (and references therein) show that peaking in performance for a finite measurement complexity is possible under certain conditions. Under other conditions monotonic improvement in performance with increase in measurement complexity is shown. Several assump-

tions such as independent and binary measurements, recoverability of smaller measurement sets from larger ones [7] have been used in the monotonicity arguments. This body of work examines general measurement spaces without reference to any particular representation such as resolution induced representations.

Theoretical bounds on error probability in pattern recognition are studied firstly to determine analytic bounds in recognition problems and secondly as a guideline for practical pattern recognition system design.

Baird [1] discusses a model-based approach to evaluating the intrinsic error in recognition of patterns digitized at low resolutions. An image defect model is used as a source for generating large databases of machine-printed characters. Wang and Pavlidis [9] have examined the trade-off between quantization and resolution in a character recognition approach. They demonstrate the importance of grayscale in improving recognition performance at low resolutions.

Our approach is to theoretically model the granularity variation in data representation as a probabilistic refinement and extend certain bounds on probability of error based on information-theoretic measures of mutual information and equivocation to this problem. These results are within the Bayesian formalism, without assuming any particular parametric form for the data distribution. While theoretical bounds on error probabilities are of considerable interest, in practice computing these bounds requires knowledge of the underlying probability distribution, which is unknown in many problems.

Theoretical results provide a guideline, however practical system design requires careful experimentation. Rather than assume that the data distribution has a certain parametric form, we use an empirical estimation of the data distribution in our work. Empirical results in character recognition are presented using machine-printed Japanese character images. We demonstrate that the performance of the classifiers improves with resolution, (that is, without peaking in performance for a fixed resolution). This is even when lower resolution signals are not always precisely recoverable from higher resolution signals. Also, information theoretic measures are shown to be useful predictors of error probability or recognition accuracy.

Outline of discussions: Section 2 presents the probabilistic refinement model for the analysis discriminability at varying resolutions. Previous information-theoretic bounds on error probability are extended. Empirical results are described in Section 3. Concluding remarks are discussed in Section 4, followed by a list of references.

2 Probabilistic Refinement Models Varying Data Resolution

We model the variation in granularity of the data as a refinement of the underlying probability structure [2]. Assume that each point x_i in the original space gets subdivided into i_k points, and that the probability measure of x_i is also subdivided among the new points in X'. This is called a refinement [2] P', of P. The refinement of the conditional probabilities W is expressed as W'.

2.1 Refinement vs. Bounds on Probability of Error

We extend previous information-theoretic bounds on error probability under the refinement model. The uncertainty in S given by another random variable X can be assessed by the conditional entropy or equivocation [13]. If X is a random variable (or vector, in our case) over the space V with a well-defined probability distribution $P(x)$, such that for each $x \in V$, $p(.|x)$ is well-defined, then the equivocation is defined as:

$$H(S|X) = -\sum_{x \in V} p(x) \sum_{\theta} p(\theta|x) \log p(\theta|x).$$

Information estimates improve with refinement (for proofs, refer [2, 12]).

Recent bounds on the error probability in terms of the equivocation [8] are extended in this paper. As refinement improves information theoretic estimates [2], an important consequence is the improvement in

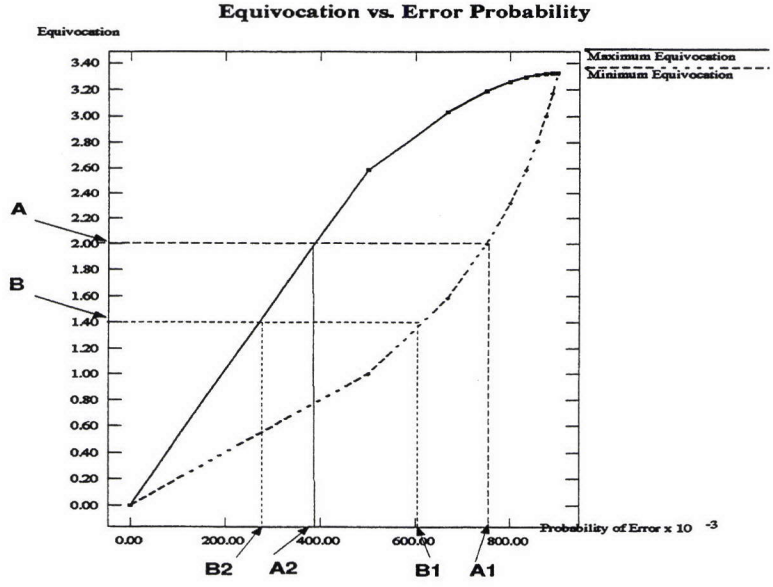

Figure 2: Resolution and Equivocation vs. Probability of Error

the bounds on error probability.

Theorem 2.1 *Bounds on probability of error decrease with increase in sampling rate.*
Proof:
Observe that equivocation decreases with refinement induced by increasing the sampling rate. As $\phi^*()$ and $\Phi()$ are monotonic and continuous functions and defined based on $H(S|X)$, both functions decrease as $H(S|X)$ decreases.

A pictorial proof of this theorem is shown in Figure 2. The relation between the equivocation and probability of error predicts all the achievable pairs of $(H(S|X), \Pi(S|X))$, as shown in Figure 2, for the case of $M = 10$. Equivocation is higher at the point A, which models the low sampling rate data representation, than at B which models a higher sampling rate representation. Correspondingly, the upper and lower bounds on the error probability assuming that equivocation is at A, are shown to be $A1$ and $A2$ respectively. Observe that the upper and lower bounds on the error probability for lower value of the equivocation at B, are reduced to $B1$ and $B2$ respectively. Hence, if $H_2 \leq H_1$, where H_1 is the equivocation at A and H_2 is the equivocation at B, then the possible (h, π) pairs possible can be bounded more tightly for H_2 than for H_1.

2.2 Discussion

The analysis presented above maybe extended to multiresolution and hierarchical goal-directed approaches in artificial intelligence. Essentially low resolution implies higher uncertainty and increasing the resolution reduces the uncertainty. In multiresolution object recognition we see that the bounds on recognition error probability improve with increasing resolution of the data.

While these bounds on error probability are of great interest, they do require a priori knowledge of P the probability distribution on the finite set of pattern classes, and W the conditional probability distribution matrix. Such statistics are seldom known, however we may estimate these probabilities from a training set and use them to predict the error probability. By performing this evaluation at several resolutions of the data, we can study the relation between the recognition accuracy and data resolution. A recent paper [11] examines this issue within the domain of optical character recognition.

Data Description			
# of Classes	Training	Testing	Comments
840	21131	41601	Machine-print Japanese alphabet

Table 1: Japanese Data Set

3 Empirical Results

We now present experimental results in optical character recognition to demonstrate the effect on resolution of discriminability of patterns.

While images of objects may be digitized at several resolutions this is not practical always, especially in applications such as optical character recognition. For this reason, we alter the resolution of the data captures at a fixed resolution using multirate filter theory [10]. These methods do not guarantee exact reconstruction of images at lower resolutions from higher resolution. Our refinement model is appropriate for this study as all we require is a refinement of the space, rather than the stricter assumption of exact reconstructibilty. This differs from previous work relating dimensionality of the data representation to the performance of pattern classifiers, as studied in [3].

The data used in the experiment is described in Table 1. Machine-printed Japanese data set images originally digitized at a resolution of 400 ppi are reduced to resolutions of $\{40, 50, 80, 100, 120, 160, 200, 240, 300, 320, 350, 360\}$ ppi. Directional features [11] are extracted from train and test data. A K-nearest neighbor classifier, with $K = 3$ is used for classification of the test data into one of the 840 classes. That is, each training feature vector is stored with the class identity; each test feature vector is compared with all the stored data to find the K nearest neighbors using a Euclidean distance metric. The most common class among the K neighbors is selected as the class identity for the test image.

The results are shown in Figures 3 and 4. The recognition accuracy (which is the dual of the error probability) is graphed on the y-axis for each resolution (x-axis) in Figure 3. In Figure 4 the mutual information [13] between the feature vectors and class identities is graphed at each resolution. Mutual information, [13], $I(P, W)$ is a measure of dependency between random variables, S and X, defined as follows:

$$I(S;X) = I(P;W) \quad = \quad \sum_{\theta\in\Theta}\sum_{x\in X} P(\theta)W(x|\theta)log\frac{W(x|\theta)}{\sum_{\theta\in\Theta} P(\theta)W(x|\theta)} \tag{1}$$

Where, $P(\theta)$ - probability distribution of θ and $W(x|\theta)$ - is a conditional probability distribution matrix, between random variable $X = x$ and θ. The following relations hold between $I(P;W)$ and $H(S|X)$ [13].

$$I(P;W) = I(S;X) = H(X) - H(X|S) = H(S) - H(S|X)$$

Where, $H(X)$ and $H(S)$ are the self-information of random variables X and S, respectively.

It is observed that the mutual information and recognition accuracy both improve as a function of the resolution. Hence the empirical information estimate is a good predictor of the recognition accuracy or error probability in a practical pattern recognition problem also.

4 Summary

We have demonstrated the effect of varying resolution of data on the discriminability and recognition accuracy achievable. Extension of bounds on error probability under varying resolutions hass been presented. Experiments in optical character recognition demonstrate the effect of resolution on recognition in a practical problem.

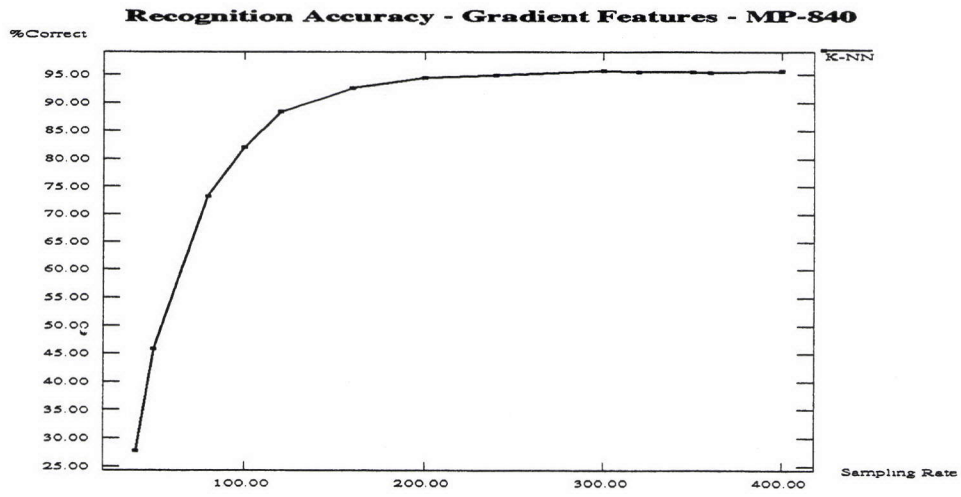

Figure 3: Recognition accuracy vs. Sampling rate - Gradient features: Machine-print Japanese Alphabets (840 classes)

While no exact design equations relating the representation of data to the recognition accuracy, we have presented a quantitative analysis of the relation. The importance of resolution of the data representation in learning or classification has been demonstrated. Our results contribute towards the design of practical pattern recognition systems where choices regarding the data resolution have to be made. While our empirical work is in the domain of optical character recognition, the theory and methods are easily extensible to other problems in object recognition.

References

[1] H. S. Baird. Document image defect models and their uses. In *Proceedings of ICDAR, 1993*, 1993.

[2] R. E. Blahut. *Principles and Practice of Information Theory.* Addison-Wesley, 1990.

[3] A. K. Jain & B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition practice. *Handbook of Statistics - Classification, Pattern Recognition and Reduction of Dimensionality, Ed. P. R. Krishnaiah & L. N. Kanal*, 2:835–855, 1982.

[4] Laveen Kanal & B. Chandrasekaran. On dimensionality and sample size in statistical pattern recognition. *Pattern Recognition*, 3:225–234, 1971.

[5] T. Pavlidis & G. W. Wasilkowski D. Lee. A note on the trade-off between sampling and quantization in signal processing. *Journal of Complexity*, 3:359–371, 1987.

[6] R. O. Duda & P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley, 1973.

[7] W. G. Waller & A. K. Jain. On the monotonicity of the performance of bayesian classifiers. *IEEE Transactions on Information Theory*, 24:392–394, 1978.

[8] M. Feder & N. Merhav. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 1994.

[9] L. Wang & T. Pavlidis. Direct gray-scale extraction of features for character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:1053–1067, 1993.

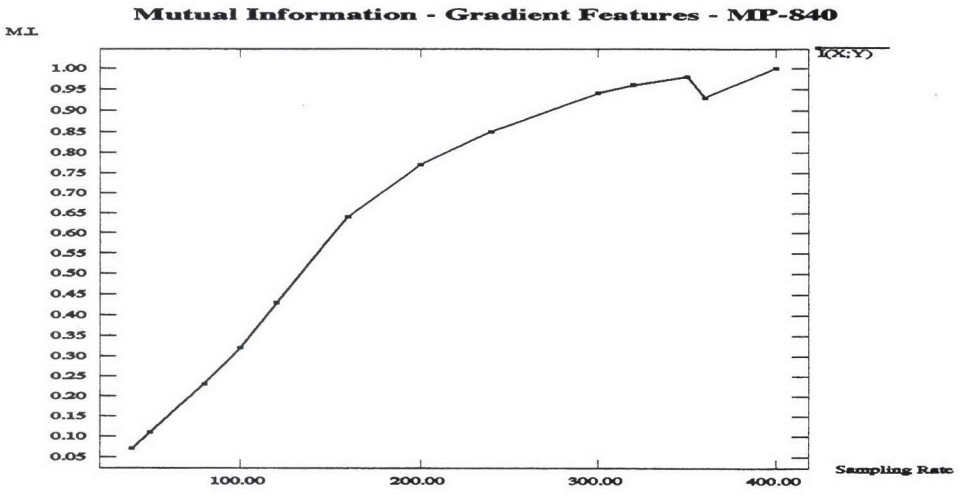

Figure 4: Information measure vs. Sampling rate - Machine-print Japanese Alphabets (840 classes)

[10] R. E. Crochiere & L. R. Rabiner. Interpolation and decimation of digital signals: A tutorial review. *Proceedings of the IEEE*, 69:300–331, 1981.

[11] G. Srikantan & S. N. Srihari. A study relating image sampling rate and image pattern recognition. In *CVPR-94*. IEEE Press, 1994.

[12] G. Srikantan. *Image Sampling Rate and Image Pattern Recognition.* Doctoral Dissertation, Department of Computer Science, SUNY at Buffalo, 1994.

[13] T. M. Cover & J. A. Thomas. *Elements of Information Theory.* John Wiley, 1991.

[14] A. V. Oppenheim & R. W.Schaeffer. *Discrete-time Signal Processing.* Prentice-Hall, 1989.