On the Performance of Edited Nearest Neighbor Rules in High Dimensions

ANDREI Z. BRODER, ALFRED M. BRUCKSTEIN, AND JACK KOPLOWITZ

Abstract—In a Poisson process in n-dimensional Euclidian space the expected distance to the kth nearest neighbor (NN) is

$$E(R_k) = \sqrt{\frac{n}{2\pi e}} + O\left(\frac{\ln n}{\sqrt{n}}\right), \qquad n \to \infty,$$

and the variance is

$$\sigma^2(R_k) = \frac{1}{n}\frac{1}{2\pi e}\left(\Gamma''(k) - \frac{\Gamma'(k)}{\Gamma(k)}\right) + O\left(\frac{\ln n}{n^2}\right), \qquad n \to \infty.$$

This implies that, for sufficiently large n, the k nearest neighbors of any point in the process lie in a thin hyperspherical shell, and that neighbors have disjoint sets of k-nearest neighbors. This makes possible the evaluation of performance for k-NN classification with edited data sets.

I. Introduction

Nearest neighbor rules are well-known discrimination procedures that classify a pattern according to the majority class of the nearest samples in the data set. For the asymptotic case, as the number of samples becomes arbitrarily large, the risk for the k-nearest neighbor (k-NN) rules can be bounded in terms of the Bayes risk [1].

Wilson [9], proposed an editing procedure which tests each sample using the k-NN rule with the remainder of the data. A sample is discarded if it is misclassified by the test. The edited data is then used in a single nearest neighbor rule for classification.

To examine the motivation of editing, consider the Bayes classifier. Define a sample to be of minority class if it is misclassified by the Bayes rule. The proportion $\rho(x)$ of minority samples in a neighborhood of x provides an indication of performance. If the classified data points are locally uniformly distributed then $\rho(x)$ is the probability that for an observation x the single nearest neighbor classifier differs from the Bayes classifier. Editing attempts to remove the minority samples and thus obtain performance closer to the Bayes classifier.

In [5] a generalized kk' edited NN rule is proposed to maximize the remaining data set while reducing $\rho(x)$. The edited set is obtained by taking a sample and its $k-1$ neighbors. If a majority of k' out of k exists the sample is labeled according to the majority class. Otherwise it is deleted.

Wilson [9] evaluates $\rho(x)$, however, it was pointed out by Penrod and Wagner [6] that the editing of each sample is not independent and thus the samples in a small neighborhood are not necessarily uniformly distributed. By considering the one-dimensional case and modifying the editing procedure to use a rule which selects the NN to a sample x from those greater than x, an exact analysis is provided [6]. For this problem the upper bounds on the risk are about six percent worse than if one assumes independence in editing.

Manuscript received January 11, 1984; revised August 29, 1984. This work was supported in part by the Office of Naval Research under contract N00014-81-K0030 and by the National Science Foundation under grants NSF ECS 78-10003 and NSF ENG 78-01557.

A. Z. Broder is with the Department of Computer Science, Stanford University, Stanford, CA 94305.

A. M. Bruckstein was with the Department of Electrical Engineering, Stanford University, Stanford, CA. He is now with the Department of Electrical Engineering, Technicon—Israel Institute of Technology, Haifa 3200, Israel.

J. Koplowitz is with the Department of Electrical and Computer Engineering, Clarkson College of Technology, Potsdam, NY 13676.

In [5], for the purpose of comparing the trade-off between the risk and the size of the edited set, samples are placed into separate groups of k. The kk' editing is then performed independently on each group. This allows for the computation of performance and gives the same result as the assumption of independence in editing each sample.

Another editing method, due to Devijver and Kittler [3], ensures independence in editing by partitioning the data set into two subsets and using the points of the first set to edit the points of the second set, which is subsequently used for classification. The drawback of this procedure is that it needs a data-rich environment.

In this correspondence we show that asymptotically as the dimensionality of the space increases the usual sample editing becomes independent. This makes an accurate calculation of performance in a high-dimensional space straightforward. Thus, with high dimensionality, the grouping in [5] is not necessary for determining the risk and similarly the results by Wilson becomes very close to exact.

Sections II and III determine the mean and variance of the distance to the kth NN. Letting the dimensionality become large we obtain the result that the expected distance to the kth NN does not depend on k and the variance approaches zero. This implies that for sufficiently large n, the k-nearest neighbors lie in a thin hyperspherical shell. In Section IV we show that, with high probability, two points that are neighbors of a sample to be classified are not k-nearest neighbors of each other and, furthermore, have distinct k-nearest neighbors. This implies that with high probability the editing of the neighbors of a sample is an independent process.

II. Distribution of the Distance to the kth Nearest Neighbor

The spatial Poisson point process is defined as a collection of points ("occurrences") randomly distributed according to the following laws [4].

1) The numbers of points in disjoint regions are *independent* random variables.

2) The number of occurrences in a region A, $N(A)$ is a Poisson random variable, that is

$$\Pr(N(A) = k) = \frac{(\lambda v(A))^k e^{-\lambda v(A)}}{k!} \tag{1}$$

where λ is the "intensity" parameter, and $v(A)$ is the volume of the region A.

In the following sections we shall consider the Poisson point process in the n-dimensional Euclidian space.

Consider a fixed point O in space and a region A such that $O \in A$. Initially $A = O$. Let the region A expand in a predetermined fashion (i.e., the expansion is independent of the occurrences of the Poisson process.). At a certain moment the region A will contain exactly k occurrences of the process. Let V_k denote the volume of A at this moment. The distribution of the random variable V_k is easily found from the defining properties 1) and 2) as follows:

$$\Pr(V_k \leq v) = \Pr(k \text{ or more occurences in } v)$$
$$= 1 - \sum_{r=0}^{k-1} \frac{(\lambda v)^r}{r!} e^{-\lambda v}. \tag{2}$$

Now the probability density function (p.d.f.) of V_k is immediately obtained as

$$pV_k(v) = \lambda e^{-\lambda v} \frac{(\lambda v)^{k-1}}{(k-1)!}, \tag{3}$$

0018-9472/85/0100-0136$01.00 ©1985 IEEE

and we can evaluate the moments of this distribution as follows:

$$E(V_k^i) = \int_0^\infty \lambda e^{-\lambda v} \frac{(\lambda v)^{k-1}}{(k-1)!} v^i \, dv = \frac{1}{\lambda^i} \frac{\Gamma(i+k)}{(k-1)!}. \quad (4)$$

In particular $E(V_k) = k/\lambda$, and $E(V_k^2) = k(k+1)/\lambda^2$ so that the variance $\sigma^2(V_k) = k/\lambda^2$.

These results are independent of the number of dimensions and generalize some well-known results for the Poisson process on the line. Note that in the above derivations no assumptions were made about the shape of the region A or about the way of expanding it.

Let P be an occurrence of the process and denote by $S_P(r)$ a "sphere" in the n-dimensional space, centered at P, and with radius r. If R_k is the distance to the kth nearest neighbor of P, then the distribution of this random variable is given by

$$\Pr(R_k < r) = \Pr(N(S_P(r) - \{P\}) \geq k) = 1 - \sum_{j=0}^{k-1} \frac{(-\lambda v(S_P(r)))^j}{j!} e^{-\lambda v(S_P(r))} \quad (5)$$

and the corresponding p.d.f. is

$$pR_k(r) = \lambda \frac{(\lambda v S_P(r))^{k-1}}{(k-1)!} e^{-\lambda v(S_P(r))} \frac{d}{dr} v(S_P(r)). \quad (6)$$

The volume of a n-dimensional sphere of radius r is given by (see [7])

$$v(S_P(r)) = \rho_n r^n \quad (7)$$

where

$$\rho_n = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} \quad (8)$$

which means

$$\rho_n = \begin{cases} \dfrac{\pi^{n/2}}{\left(\frac{n}{2}\right)!}, & \text{if } n \text{ even}, \\[2ex] \dfrac{2^n \pi^{(n-1/2)} \left(\frac{n-1}{2}\right)!}{n!}, & \text{if } n \text{ odd}. \end{cases} \quad (9)$$

Therefore

$$pR_k(r) = \frac{n}{r} \frac{(\lambda \rho_n r^n)^k}{(k-1)!} e^{-\lambda \rho_n r^n} \quad (10)$$

and in particular the p.d.f. of the distance to the nearest neighbor is

$$pR_1(r) = \lambda n \rho_n r^{n-1} e^{-\lambda \rho_n r^n}. \quad (11)$$

III. Expected Value and Variance of the Distance to the kth Nearest Neighbor

Straightforward calculations yield the moments of the p.d.f. of the random variable R_k, the distance to the kth NN.

$$E(R_k^m) = \int_0^\infty r^m \frac{n}{r} \frac{(\lambda \rho_n r^n)^k}{(k-1)!} e^{-\lambda \rho_n r^n} dr = \frac{1}{(\lambda \rho_n)^{m/n}} \frac{1}{(k-1)!} \Gamma(k + m/n). \quad (12)$$

Using the above formula we could obtain again the expected value of the volume of the sphere required to contain k occurrences:

$$E(V_k) = E(\rho_n R_k^n) = \frac{k}{\lambda}. \quad (13)$$

Also, more generally

$$E(V_k^i) = E(\rho_n^i R_k^{ni}) = \frac{1}{\lambda^i} \frac{\Gamma(k+i)}{(k-1)!} \quad (14)$$

which is (4).

From (12) we obtain the expected value and the variance of the distance to the kth NN

$$E(R_k) = \frac{1}{(\lambda \rho_n)^{1/n}} \frac{\Gamma(k+1/n)}{(k-1)!}, \quad (15)$$

and

$$E(R_k^2) = \frac{1}{(\lambda \rho_n)^{2/n}} \frac{\Gamma(k+2/n)}{(k-1)!}, \quad (16)$$

so that

$$\sigma^2(R_k) = \frac{1}{(\lambda \rho_n)^{2/n}} \frac{1}{(k-1)!} \left(\Gamma(k+2/n) - \frac{\Gamma(k+1/n)^2}{(k-1)!} \right). \quad (17)$$

To obtain the asymptotic behavior of the mean and the variance of the distance to the kth nearest neighbor, consider $\rho_n^{1/n}$. Using the Stirling formula for $\ln \Gamma((n/2)+1)$ [2], we obtain

$$\rho_n^{1/n} = \left(\frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} \right)^{1/n} = \sqrt{\pi} \exp\left(-\frac{1}{n} \ln \Gamma\left(\frac{n}{2}+1\right) \right)$$
$$= \sqrt{\pi} \exp\left(-\frac{1}{n} \left(\left(\frac{n}{2} + \frac{1}{2}\right) \ln\left(\frac{n}{2}+1\right) - \frac{n}{2} + O(1) \right) \right) \quad (18)$$
$$= \sqrt{\frac{2\pi e}{n}} \left(1 + O\left(\frac{\ln n}{n}\right) \right), \qquad n \to \infty.$$

Since the gamma function is analytic for positive arguments, using its power series expansion we get

$$\frac{\Gamma(k+1/n)}{(k-1)!} = 1 + O\left(\frac{1}{n}\right), \qquad n \to \infty, \quad (19)$$

and

$$\frac{1}{(k-1)!} \left(\Gamma(k+2/n) - \frac{\Gamma(k+1/n)^2}{(k-1)!} \right) = \frac{1}{n^2} \left(\Gamma''(k) - \frac{\Gamma'(k)^2}{\Gamma(k)} \right) + O\left(\frac{1}{n^3}\right), \qquad n \to \infty. \quad (20)$$

Finally, because λ is fixed, $\lambda^{1/n} = 1 + O(1/n)$.

The above results, used in (15) and (17) generate the following asymptotics:

$$E(R_k) = \sqrt{\frac{n}{2\pi e}} + O\left(\frac{\ln n}{\sqrt{n}}\right), \qquad n \to \infty, \quad (21)$$

and

$$\sigma^2(R_k) = \frac{1}{n} \frac{1}{2\pi e} \left(\Gamma''(k) - \frac{\Gamma'(k)}{\Gamma(k)} \right) + O\left(\frac{\ln n}{n^2}\right), \qquad n \to \infty. \quad (22)$$

Since $E(R_k)$ is asymptotically independent of k, and $\sigma^2(R_k) \to 0$, the k-nearest neighbors to any given point lie, with probability approaching 1, in a thin spherical hypershell for large n.

Some plots of $E(R_k)$ and $\sigma^2(R_k)$ computed according to (15) and (17), for $k = 1$ to 10 and $n = 1$ to 20, clearly exhibit this asymptotic behavior (Figs. 1 and 2). Note that in the plane the variance increases with k, but decreases for all higher dimensionalities (Fig. 3).

IV. The Problem of Common Nearest Neighbors and Independence in Editing

Let Q be the nearest neighbor of the point P. Using the above results we shall argue that, with high probability, P and Q have mutually disjoint k-nearest neighbors, except for P and Q themselves.

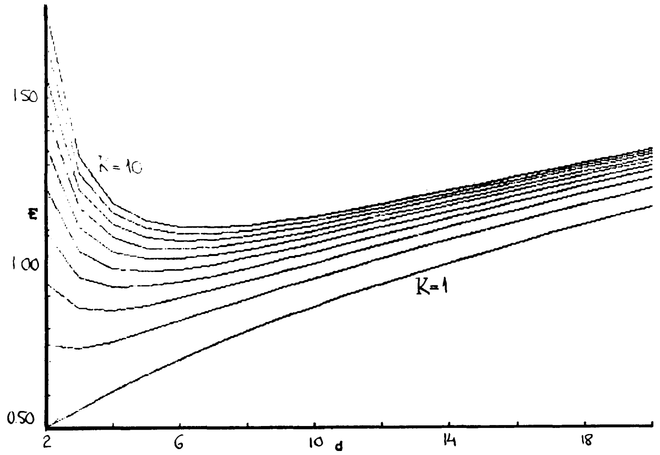

Fig. 1. Expected distance to k-NN as function of dimension.

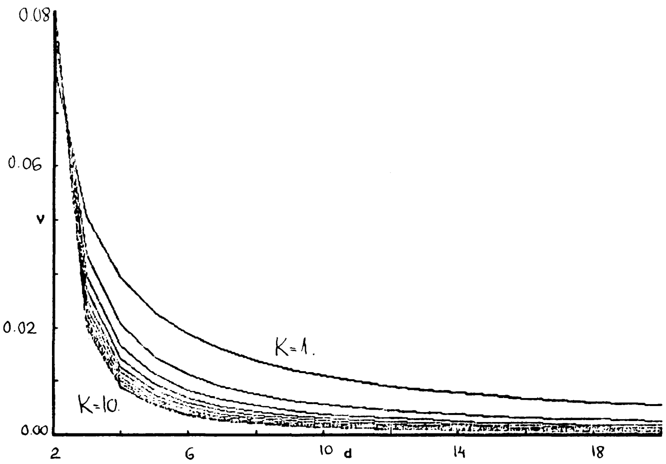

Fig. 2. Variance of distance to k-NN as function of dimension.

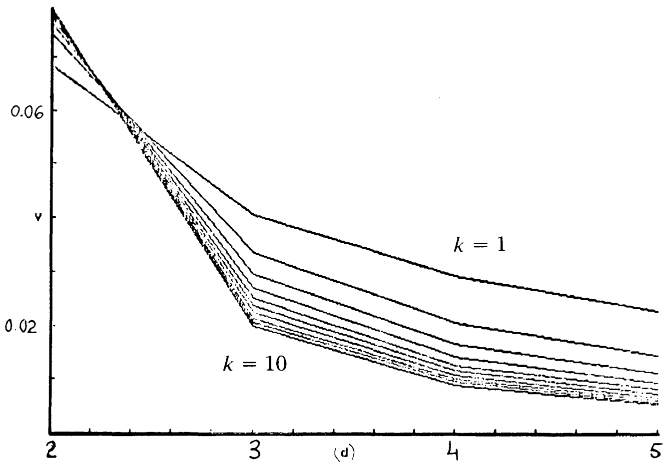

Fig. 3. Variance of distance to k-NN for low dimensions.

Let R_1 be the distance between P and Q. Consider a hemihypersphere H_Q centered at Q with the base tangent to the hypersphere of radius R_1 centered at P. (Fig. 4)

Let's expand H_Q until it contains k occurrences of the point process, and call its corresponding radius R'_k. Since the volume of H_Q is half the volume of a hypersphere of the same radius, by repeating our previous arguments ((5) and (6)) with $\rho'_n = \rho_n/2$ we obtain that

$$E(R'_k) = 2^{1/n} E(R_k) \tag{23}$$

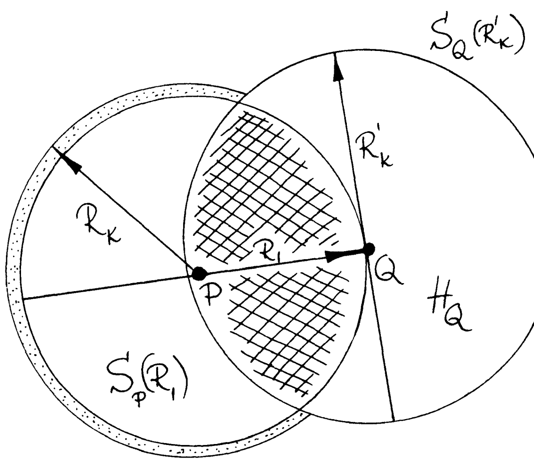

Fig. 4. Hemihypersphere H_Q centered at Q.

and also that the expected volume of the hemihypersphere is again $E(V_k) = k/\lambda$, as predicted by (13) which does not depend on the form of V_k.

We now complete the region H_Q to the hypersphere $S_Q(R'_k)$, which clearly contains all the k-NN of Q. Because Q was the first NN of P we know that $S_Q(R'_k) \cap S_P(R_1)$ does not contain any occurence of the process (the shaded region in Fig. 4). We also know that the k-nearest neighbors of P are in a spherical shell of width $R_k - R_1$.

Given the fact that the k-nearest neighbors of P have a uniform radial distribution in the shell, it follows that the probability of P and Q having a common k-NN is bounded from above by k times the ratio between the area of the sphere $S_P(R_1)$ included in $S_Q(R'_k)$ and the total area of the sphere $S_P(R_1)$. We shall how show that this ratio goes to zero as the dimensionality goes to infinity.

Consider a hypersphere in n-dimensional space, and intersect it with a $(n-1)$ flat (Fig. 5). The area of the resulting hyperdome is computed by dividing it into thin slabs by a series of parallel hyperplanes; the lateral area of each slab is

$$(n-1)\rho_{n-1}(R\sin\theta)^{n-2} R\, d\theta \tag{24}$$

and therefore the dome's total area is

$$(n-1)\rho_{n-1} R^{n-1} \int_0^{\theta_0} (\sin\theta)^{n-2}\, d\theta. \tag{25}$$

Hence the ratio between the area of the dome to the total area of the hypersphere is

$$\eta(\theta_0) = \frac{n-1}{n}\frac{\rho_{n-1}}{\rho_n}\int_0^{\theta_0}(\sin\theta)^{n-2}\,d\theta \le \frac{n-1}{n}\frac{\rho_{n-1}}{\rho_n}\theta_0(\sin\theta_0)^{n-2}. \tag{26}$$

But

$$\frac{\rho_{n-1}}{\rho_n} = \frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\pi^{-1/2} \le \frac{\Gamma\left(\frac{n+3}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\pi^{-1/2} = \left(\frac{n}{2}+1\right)\pi^{-1/2} \tag{27}$$

and therefore $\eta \to 0$ as $n \to \infty$ for any θ_0, $0 \le \theta_0 < \pi/2$.

Returning to our original problem, it is obvious that as $n \to \infty$, the probability of finding a nearest neighbor of P in the shell region corresponding to the dome generated by the intersection with $S_Q(R'_k)$ becomes equal to the ratio $\eta(\theta_0)$, where θ_0 is determined by

$$\lim_{n\to\infty}\theta_0 = \lim_{n\to\infty} 2\arcsin\frac{R'_k}{2R_1} = \frac{\pi}{3} \tag{28}$$

hence this probability goes to zero, as n goes to infinity. Thus

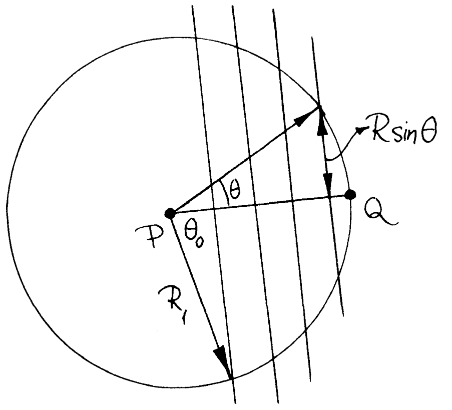

Fig. 5. Hypersphere in n-dimensional space.

TABLE I

n	$E(R_1)$	$\sigma^2(R_1)$	$E(R_5)$	$\sigma^2(R_5)$	θ_0(rads)	$\eta(\theta_0)$
10	0.866	1.086×10^2	1.059	2.435×10^3	1.429	0.3402
15	1.029	7.092×10^3	1.179	1.349×10^3	1.286	0.1464
20	1.168	5.247×10^3	1.294	9.174×10^4	1.221	0.0644
25	1.292	4.156×10^3	1.402	6.898×10^4	1.184	0.0288
30	1.403	3.437×10^3	1.503	5.508×10^4	1.160	0.0131

any two points, even if they are nearest neighbors, have distinct sets of k-NN's, themselves excepted. Note that this holds even for two points closer than $E(R_1)$ since $\lim_{n \to \infty} \theta_0 < \pi/2$.

In order to illustrate the behavior of η, we computed θ_0 as 2 arcsin $E(R'_5)/(2E(R_1))$ and $\eta(\theta_0)$ by (26) for n between 10 and 30 (Table I). It is clear from the very low values of σ^2 that using expected radii to evaluate θ_0 is indeed justified.

Consider the edited NN rule. Suppose that a point x is to be classified. Before editing, let x_α and x_β be two of its nearest neighbors. By the precedent discussion, with high probability neither x_α nor x_β is a k-NN of the other. If so, they are further apart than P and Q in Fig. 4 and can be expected to have no common k-nearest neighbors. Hence, with high probability they are tested independently in the editing procedure.

A question of major interest is the dimensionality required to make the independence assumption accurate. The results of Penrod and Wagner [6] differ in performance by only about six percent from that obtained by assuming independence. Yet for this case nearest neighbors have an inordinate number of common k-nearest neighbors. Thus it would be reasonable to expect the independence assumption to be highly accurate (perhaps within one percent) for an $\eta(\theta_0) = 0.1$ that is achieved with a dimensionality of about 15.

References

[1] T. M. Cover and P. E. Hart, "Nearest neighbor patter classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21–27, 1967.

[2] N. G. De Bruijn, *Asymptotic Methods in Analysis*, 3rd ed. Amsterdam: North Holland, 1970.

[3] P. A. Devijver and J. Kittler, "On the edited nearest neighbor rule," in *Proc. 5th Int. Conf. Patt. Recog.*, Miami Beach, FL. 1980, pp. 72–80.

[4] S. Karlin and H. M. Taylor, *A Second Course in Stochastic Processes*, New York: Academic, 1981.

[5] J. Koplowitz and T. A. Brown, "On the relation of performance to editing in nearest neighbor rules," *Patt. Recog.*, vol. 13, pp. 251–255, 1981.

[6] C. S. Penrod and J. J. Wagner, "Another look at the edited nearest neighbor rule," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-7, pp. 92–94, 1977.

[7] D. M. Y. Sommerville, *An Introduction to the Geometry of N Dimensions*. London: Methuen & Co., 1929.

[8] I. Tomek, "A generalization of the k-NN rule," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-6, pp. 121–126, 1976.

[9] D. L. Wilson, "Asymptotic properties of NN rules using edited data," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-2, pp. 408–421, 1972.

Information Energy of a Fuzzy Event and a Partition of Fuzzy Events

LEANDRO PARDO

Abstract—**In order to define a measure of the information processed by a fuzzy event and by a partition of fuzzy events, the "information energy" provided by a fuzzy event and partition of fuzzy events is considered. This measure integrates the statistical uncertainty resulting from the occurrence of events and the uncertainty of meaning of events that is expressed by the membership function. The functional information energy is formally similar to the Onicescu's information energy, which used an analogy to kinetic energy from mechanics, although it is conceptually different.**

I. Introduction

Let X be an arbitrary set and $(X, \mathscr{A}, P)$ a probability space, where $\mathscr{A}$ is a σ-field of subsets X and P a probability on $\mathscr{A}$. A fuzzy event in X is a fuzzy set X^j on X whose membership function f_X^j is Borel measurable. The lattice of all fuzzy events defined on X is denoted by $\Gamma(X)$.

K. Okuda, T. Tanaka, and K. Asai [3], in analogy with Shannon's entropy of information theory, studied a measure of the uncertainty processed by a fuzzy event X^j. This measure integrates the uncertainty resulting from the occurrence of elements of X on the one hand, the degree of fuzziness of X^j and on the other hand. This concept was extended by K. Kuriyama [2] in order to define the amount uncertainty processed by a fuzzy partition of fuzzy events.

O. Onicescu [4] introduced the concept of information energy as an alternative way of building an information theory. This concept can be interpreted as a measure of our information concerning a random variable. L. Pardo [5] and A. Theodorescu [6] justified Onicescu's final observation of "*L'energie informationelle peut servir, aussi bien que l'entropie comme fondament d'une thèorie de l'information*".

In order to define the information processed by a fuzzy event and a fuzzy partition of fuzzy events in a fuzzy setting, a function in the class of fuzzy events and a function in the class of a fuzzy partition of fuzzy events is defined. These functionals are called the "information energy of a fuzzy event" and the "information energy of a partition of fuzzy events", respectively. These concepts are quite different from those of Onicescu's information theory because they integrate the information before carrying out an experiment with values for X, and because the meaning of fuzzy events in each element of a partition A, which is expressed by the membership function, is uncertain.

II. Information Energy

In this section several definitions are established that are needed in later sections.

Definition 1: The information energy contained in the fuzzy event X^j is defined by

$$W(X^j) = P(X^j)^2 + P(\bar{X}^j)^2$$

where $\bar{X}^j$ is the complement set of X^j and

$$P(X^j) = \int_X f_X^j(x)\, dP(x).$$

Example: Suppose that a machine produces a defective item with probability p $(0 < p < 1)$ and produces a nondefective item probability $1 - p$. Furthermore, 15 items produced by the machine are selected at random and inspected, and the outcomes for these 15 items are independents. The information energy contained by the fuzzy event "approximately 15 of the items are

Manuscript received August 11, 1983; revised July 1984.

The author is with the Departamento de Estadistica e I.O., Facultad de Matemàticas, Universidad Complutense de Madrid, Madrid-3, Spain.

0018-9472/85/0100-0139$01.00 ©1985 IEEE