

# Short Papers

## Marginal Maximum Entropy Partitioning Yields Asymptotically Consistent Probability Density Functions

Tom Chau, *Member, IEEE*

**Abstract**—The marginal maximum entropy criterion has been used to guide recursive partitioning of a continuous sample space. Although the criterion has been successfully applied in pattern discovery applications, its theoretical justification has not been clearly addressed. In this paper, it is shown that the basic marginal maximum entropy partitioning algorithm yields asymptotically consistent density estimates. This result supports the use of the marginal maximum entropy criterion in pattern discovery and implies that an optimal classifier can be constructed.

**Index Terms**—Marginal maximum entropy, recursive partitioning, pattern discovery, asymptotic optimal classification.

### 1 INTRODUCTION

THE marginal maximum entropy (MME) criterion for recursive partitioning has been exploited in missing data analysis [1], the automatic detection of relational information [2], knowledge discovery in databases [3] and, more recently, in multivariate data mining [4]. In these applications, the utility of the MME-driven recursive partitioning lies in the estimation of probability density functions (pdf) from data. These pdfs are subsequently employed in pattern classification. While asymptotically consistent probability density estimates have been validated for several recursive partitioning schemes [5], [6], [7], [8], [9], a like result has not been established for the MME criterion. This paper will argue that under mild assumptions about the underlying data distribution, MME-driven recursive partitioning furnishes consistent probability density estimates, i.e., estimates with asymptotically unbiased mean and vanishing variance.

The remainder of the paper is organized as follows: In Section 2, the marginal maximum entropy criterion is stated and the basic recursive partitioning algorithm is paraphrased. Section 3 comprises the bulk of the paper and details the arguments towards asymptotic consistency. The implications of this theoretical result on pattern classification are discussed in Section 4.

### 2 THE MME CRITERION AND RECURSIVE PARTITIONING

The following definition from probability theory will apply throughout the discussion:

**Definition 1 (Event).** *An event in  $\mathbb{R}^d$  is a Borel set.*

A Borel set [10] is a subset of the  $d$ -dimensional continuous sample space  $\mathbb{R}^d$  belonging to a Borel field of  $\mathbb{R}^d$ . For the purpose of this

• The author is with Bloorview MacMillan Centre and the Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, Ontario, Canada. E-mail: tkchau@ieee.org.

Manuscript received 21 June 2000; revised 30 Oct. 2000; accepted 20 Dec. 2000.

Recommended for acceptance by I. Sethi.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112308.

discussion, an event can be thought of as a hyperrectangle in  $\mathbb{R}^d$ . Indeed, the two terms will be used interchangeably.

Partitioning a data set involves the segmentation of observations into disjoint groups or events. The objective of partitioning is to describe the data probabilistically, that is, to estimate a probability for each event. Event probabilities are estimated as,

$$\hat{P}_j = \frac{n_j}{N}, \quad (1)$$

where  $N$  is the total number of data points under consideration and  $n_j$  is the number of data points within  $E_j$ , or the cardinality of the set  $E_j$ . Similarly, the probability density of an event is commonly estimated by,

$$\hat{p}_j = \frac{n_j}{Nv_j}, \quad (2)$$

where  $v_j$  is the volume of the event or hyperrectangle. These estimates are well-established in histogram-based density estimation [11] and have been prevalent in automatic pattern recognition applications [12].

Clearly, a unique partition of the sample space does not exist and numerous partitioning criteria, such as maximum Kolmogorov-Smirnoff distance [8], [6], [13], minimum within-partition sum-of-squares [14], and minimum misclassification cost [5], have been proposed. A recent review of partitioning criteria is found in Murthy [15]. The marginal maximum entropy criterion [3], [16] is derived from information theory and prescribes yet another method of partitioning the sample space. According to this criterion, the subspace of interest is segmented in a way which *approximately* maximizes the overall entropy [17],  $H$ , of the partition. If the subspace is partitioned into  $J$  events, then the entropy of the partition is expressed as,

$$H = - \sum_{j=1}^J P_j \log P_j, \quad (3)$$

where  $P_j$  is the probability associated with event  $E_j$ . The approximate nature of the method is due to the fact that partitioning is not performed in the full dimensionality, but rather *marginally* along each dimension. It has been shown experimentally that this is a reasonable approximation [16] for a number of distributions. The marginal maximum entropy partitioning criterion is summarized below.

**Definition 2 (Marginal Maximum Entropy (MME) Criterion).** *To approximately maximize the entropy of a  $d$ -dimensional partition,  $d \geq 2$ , maximize the entropy of each one-dimensional (marginal) partition.*

It is well-known that the entropy of a partition (3) is maximized when all probabilities are equalized [17], [18]. Consequently, the equalization of probabilities in (3) translates into an equalization of event frequencies. Therefore, marginal maximization of entropy involves the segmentation of individual axes into 1-dimensional intervals such that each interval contains an approximately equal number of samples [16].

The MME criterion has been used in conjunction with recursive partitioning to estimate the probability density of a data set. The algorithm draws upon recursive partitioning [5], [8], [16] and hierarchical discretization [3], [2]. Suppose we want to partition a subspace  $\Omega \subset \mathbb{R}^d$ . At a given level of recursion, if we choose a

partition size  $Q$  where  $Q \geq 2$ , the  $d$ -dimensional subspace of interest will be segmented into  $Q^d$  hyperrectangles or events. The basic algorithm is paraphrased below.

### Simple Recursive Partitioning with MME Criterion

Choose a partition size  $Q_0, Q_0 \geq 2$

Call Partition( $\Omega, Q_0$ )

*Procedure* Partition( $S, Q$ )

Partition  $S$  into  $Q^d$  events according  
to the MME criterion.

For each event  $E_j, j = 1 \dots Q^d$

If  $n_j < \xi$ ,

Estimate probability density as  $\hat{p}_j = \frac{n_j}{Nv_j}$

Else

Choose a partition size  $Q_j$

Call Partition( $E_j, Q_j$ )

End if

End For

End Partition

In the above,  $\xi$  is the threshold sample size below which further partitioning is not warranted. The minimum requisite sample size for statistical testing in contingency tables [19], [20], [21] has served as a guide in determining this threshold [4]. The choice of partition size  $Q_j$  is governed by the trade-off between maximizing statistical significance of the resulting events while minimizing computational complexity. A crude upper bound for  $Q_j$  is  $(\frac{N}{\xi})^{1/d}$ . In other words, choose  $Q_j$  such that each  $d$ -dimensional event has at least the minimum requisite number of data points for statistical testing. For the purpose of this paper, we do not need to be concerned with the exact method of determination of these two parameters,  $\xi$  and  $Q_j$ . It is important to note that, while the MME criterion equalizes frequencies within intervals, the lengths of the intervals are not uniform. Herein lies the potential to estimate any general distribution of data. Extensions to this basic algorithm can be found in [2], [3] and, more recently, [4].

### 3 ASYMPTOTIC CONSISTENCY

The fundamental measure of correctness of a classification rule is its unconditional error rate,  $R_n$ , in the limit of an infinite training sample. A rule is Bayes risk consistent or efficient if it satisfies the following asymptotic condition:

$$\lim_{n \rightarrow \infty} R_n = R^*, \quad (4)$$

where  $R^*$  is the Bayes error rate or Bayes probability of error. This is the optimal error rate achievable and can occur only if the class conditional densities are completely specified. It will be argued that, in the presence of large samples, a Bayes risk consistent classifier can be constructed via marginal maximum entropy partitioning. The critical component of this argument is to demonstrate that MME-driven recursive partitioning furnishes consistent class density estimates. The proof will proceed in three stages. First, the large sample behavior of the recursive algorithm is characterized (Proposition 1). Second, asymptotic unbiasedness of the density estimate is argued (Proposition 2). Third, vanishing variance in the limit is shown to be true (Proposition 3). For notational convenience, the event indicator function  $I_j(\mathbf{x})$  is defined.

**Definition 3 (Event Indicator Function).** *The indicator function for the event,  $E_j$ , is defined as,*

$$I_j(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in E_j \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Here,  $\mathbf{x} \in \mathbb{R}^d$  is a point in the  $d$ -dimensional sample space. We now characterize the large sample behavior of the hyperrectangles obtained from recursive partitioning.

**Proposition 1 (Large Sample Behavior).** *Suppose MME-driven partitioning is recursively applied to a sample space  $\Omega \subset \mathbb{R}^d$  with  $N$  data points. If  $N \rightarrow \infty$  then,*

1. *The maximum number of allowable recursions,  $r_{max}$ , increases at the rate  $\ln N$ .*
2. *The ratio  $\frac{n_r}{N} \rightarrow 0$ , where  $n_r$  is the number of points in the event containing  $\mathbf{x}$  at the  $r$ th recursion.*
3. *The hyperrectangle volume vanishes, i.e.,  $V_N \rightarrow 0$ , but  $NV_N \rightarrow \infty$ .*
4. *For each data point  $\mathbf{x}$ , there will be an event center  $\mathbf{c}$  such that,  $|\mathbf{c} - \mathbf{x}| < \epsilon, 0 < \epsilon \ll 1$ .*

**Proof.** To prove the first item, we find a lower bound on the maximum number of recursions  $r_{max}$ . Let  $\xi$  represent the minimum allowable number of points in an event and let  $Q_i$  be the partition size at the  $i$ th level of recursion. Recall that  $N$  is the total number of points under consideration. For one particular subspace, we may write,

$$\xi = \prod_{i=1}^{r_{max}} \left[ \left( \frac{1}{Q_i^d} \right) \right] N. \quad (6)$$

The above equation is a direct result of the marginal frequency equalization property of the marginal maximum entropy criterion. Let  $Q_{min} = \min_i Q_i$ . We can then form an upper bound to the product,

$$\frac{1}{Q_1} \frac{1}{Q_2} \dots \frac{1}{Q_{r_{max}}} \leq \left( \frac{1}{Q_{min}} \right)^{r_{max}}. \quad (7)$$

Using this result in (6) and taking logarithms yields,

$$\ln \xi \leq \ln N + dr_{max} \ln \left( \frac{1}{Q_{min}} \right). \quad (8)$$

Rearranging, we arrive at a lower bound for  $r$ ,

$$r_{max} \geq A \ln N - B, \quad (9)$$

where  $A = \frac{1}{d \ln Q_{min}}$  and  $B = A \ln \xi$ . As  $A$  and  $B$  are constants with respect to  $N$ , then  $r_{max}$  grows approximately as  $\ln N$ . Therefore,  $r_{max} \rightarrow \infty$  as  $N \rightarrow \infty$ .  $\square$

Consider a sequence of nested events,  $E_k, k = 1 \dots r$ , produced by recursive partitioning. Here,  $k$  denotes the level of recursion. Consider a point  $\mathbf{x}$  contained in the nested events, i.e.,  $\mathbf{x} \in \mathbf{E}_r \subseteq \mathbf{E}_{r-1} \subseteq \dots \subseteq \mathbf{E}_1$ . Let  $n_r$  be the number of points in  $E_r$ . By the frequency equalization property, we have, for fixed  $N$ ,

$$n_r = \frac{n_{r-1}}{Q_r^d} = \frac{n_{r-2}}{Q_r^d Q_{r-1}^d} = \dots = \frac{N}{\prod_{i=1}^r Q_i^d}. \quad (10)$$

Rearranging,

$$\frac{n_r}{N} = \frac{1}{\prod_{i=1}^r Q_i^d} \quad \text{where } Q_i > 1 \forall i. \quad (11)$$

We know that  $r \rightarrow \infty$  and, thus,  $\frac{n_r}{N} \rightarrow 0$  as  $N \rightarrow \infty$ .

To prove that the volume decreases to 0, we note that in a particular subspace,

$$V_N = V_{Global} \prod_{i=1}^r \frac{1}{Q_i^d} \quad (12)$$

where  $V_{Global}$  is the constant, finite, global volume of the space occupied by the data. As  $Q_i > 1$  for all  $i$  and  $r \rightarrow \infty$  as  $N \rightarrow \infty$ , therefore,  $\lim_{N \rightarrow \infty} V_N = 0$ . However, we need to show that  $V_N$  decreases slower than  $1/N$ . Recall the crude upper bound for the partition size,  $Q_i \leq (\frac{n_i}{\xi})^{1/d}$ , where  $\xi$  denotes the minimum number of points required for statistical testing and  $n_i$  is the number of points in an event  $E_i$  at the  $i$ th level of recursion. Thus, the volume can be loosely bound as follows:

$$V_N \geq V_{Global} \prod_{i=1}^{r(N)} \frac{\xi}{n_i}, \quad (13)$$

where we have written  $r(N)$  to indicate that  $r$  is a function of  $N$ . Multiplying both sides by  $N$ , we obtain,

$$NV_N \geq V_{Global} \xi \prod_{i=1}^{r(N)} \frac{N}{n_i}. \quad (14)$$

However, we know that  $\lim_{N \rightarrow \infty} \frac{n_i}{N} = 0$  and, therefore, we conclude that the right-hand side of the above inequality increases to infinity, implying that  $NV_N \rightarrow \infty$  as  $N \rightarrow \infty$ .

Again, consider the sequence of events  $E_k$ ,  $k = 1 \dots r$ , that contain  $\mathbf{x}$ . Let  $\mathbf{c}_k$  be the centers of  $E_k$ . Let  $h_{kj}$  be the length of the  $j$ th dimension of  $E_k$ ,

$$h_{kj} = h_{1j} \prod_{i=1}^{r-1} \frac{1}{Q_i}, \quad Q_i > 1, \quad (15)$$

where  $h_{1j}$  is the original length, prior to any recursion. Define the diagonal of the event  $E_k$  as  $diag(E_k) = \sqrt{\sum_{j=1}^d h_{kj}^2}$ . For  $\mathbf{x} \in \mathbf{E}_k$ ,

$$\|\mathbf{x} - \mathbf{c}_k\| \leq diag(E_k), \quad (16)$$

where  $\|\cdot\|$  is the Euclidean norm. Since  $Q_i > 1$ ,  $h_{kj} \rightarrow 0$  and, therefore,  $diag(E_k) \rightarrow 0$  as  $r \rightarrow \infty$ . We conclude then that there exists a positive integer  $R$  such that for  $r > R$  recursions and  $0 < \epsilon \ll 1$ ,  $\|\mathbf{x} - \mathbf{c}_r\| < \epsilon$ . In other words, the center of the event  $\mathbf{c}_r$  can become arbitrarily close to  $\mathbf{x}$ . Having characterized the large sample behavior of the recursive partitioning algorithm, we are now ready to determine the asymptotic properties of the pdf estimate.

**Proposition 2 (Unbiased Estimate).** *Suppose we have a sample space  $\Omega \in \mathbb{R}^d$  from which we independently draw  $N$  samples  $\mathbf{x}$  according to the probability law  $f(\mathbf{x})$ . The pdf estimate,*

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^K I_i(\mathbf{x}) \frac{n_{E_i}}{NV_{E_i}} \quad (17)$$

*based upon the final set of events,  $\{E_i\}$ ,  $i = 1 \dots K$ , of a sequence of recursive maximum entropy partitions of  $\Omega$  is asymptotically unbiased, i.e., for  $\mathbf{x} \in \Omega$ ,*

$$E[\hat{f}(\mathbf{x})] \rightarrow f(\mathbf{x}) \quad \text{as } N \rightarrow \infty \quad (18)$$

*Here,  $E[\cdot]$  denotes expectation over different random samplings.*

**Proof.** For a test point  $\mathbf{x}_{test}$  contained in the event  $E_i$  of the final partition,

$$E[\hat{f}(\mathbf{x}_{test})] = E[\hat{f}_{E_i}] \quad (19)$$

$$= E\left[\frac{n_i}{NV_i}\right] \quad (20)$$

$$= \frac{P}{V_i}. \quad (21)$$

In the last equation, we exploited the fact that the number of points  $n_i$  falling within the event  $E_i$  is binomially distributed

with probability  $P = \int_{E_i} f(\mathbf{x}) d\mathbf{x}$ , i.e.,  $n_i \sim Bin(N, P)$ . Hence,  $E[n_i] = NP$ .

Assuming that  $f(\mathbf{x})$  is continuous and bounded on  $E_i$  and that  $E_i$  is connected, then we may apply the mean value theorem to evaluate  $P$ . In other words, there exists a point  $\mathbf{p}$  in  $E_i$  such that

$$\int_{E_i} f(\mathbf{x}) d\mathbf{x} = f(\mathbf{p}) V(E_i), \quad (22)$$

where  $V(E_i) = \int_{E_i} d\mathbf{x}$ . Employing this result in (21) yields,

$$E[\hat{f}(\mathbf{x}_{test})] = f(\mathbf{p}) \quad \mathbf{p} \in E_i. \quad (23)$$

Now, let  $\mathbf{c}_i$  be the center of  $E_i$ . By Proposition 1, we know that as  $N \rightarrow \infty$ ,

$$\|\mathbf{p} - \mathbf{c}_i\| < \epsilon_1 \quad (24)$$

$$\|\mathbf{x}_{test} - \mathbf{c}_i\| < \epsilon_2, \quad (25)$$

where  $0 < \epsilon_1, \epsilon_2 \ll 1$  and  $\|\cdot\|$  is the Euclidean norm. To relate  $\mathbf{p}$  and  $\mathbf{x}_{test}$ , we make use of the triangle inequality,

$$\|\mathbf{p} - \mathbf{x}_{test}\| \leq \|\mathbf{p} - \mathbf{c}_i\| + \|\mathbf{x}_{test} - \mathbf{c}_i\| \quad (26)$$

$$< \epsilon_1 + \epsilon_2. \quad (27)$$

Hence, as  $N \rightarrow \infty$ ,  $\|\mathbf{p} - \mathbf{x}_{test}\| \rightarrow 0$ , implying that  $\mathbf{p} \rightarrow \mathbf{x}_{test}$ . This leads to the conclusion that as  $N \rightarrow \infty$ ,

$$E[\hat{f}(\mathbf{x}_{test})] \rightarrow f(\mathbf{x}_{test}). \quad (28)$$

□

The unbiasedness of the pdf estimate has now been addressed. To complete the argument for consistent estimation, the variance of the estimate needs to be examined.

**Proposition 3 (Vanishing variance).** *The pdf estimate (17), when applied under the conditions as in Proposition 2, has asymptotically vanishing variance, i.e.,*

$$Var[\hat{f}(\mathbf{x})] \rightarrow 0 \quad \text{as } N \rightarrow \infty. \quad (29)$$

**Proof.** For a test point  $\mathbf{x}$  that falls inside a event  $E_i$ ,

$$Var[\hat{f}(\mathbf{x})] = E[\hat{f}(\mathbf{x})^2] - E[\hat{f}(\mathbf{x})]^2 \quad (30)$$

$$= E\left[\frac{n_i^2}{N^2 V_N^2}\right] - E\left[\frac{n_i}{NV_N}\right]^2 \quad (31)$$

$$= \frac{1}{N^2 V_N^2} \left( E[n_i^2] - E[n_i]^2 \right) \quad (32)$$

$$= \frac{1}{NV_N^2} P(1 - P). \quad (33)$$

In the last inequality, we again have used the fact that  $n_i \sim Bin(N, P)$  and, thus,  $E[n_i^2] - E[n_i]^2 = Var[n_i] = NP(1 - P)$ . As before,  $P$  is the probability of finding a point in  $E_i$ . Again, if we assume that  $f(\mathbf{x})$  is bounded and continuous over  $E_i$  and  $E_i$  is connected, then, we can apply the mean value theorem. Substituting,  $P = f(\mathbf{p}) V_N$ ,  $\mathbf{p} \in E_i$ , in (33), we obtain

$$Var[\hat{f}(\mathbf{x})] = \frac{f(\mathbf{p})}{NV_N} - \frac{f(\mathbf{p})^2}{N}. \quad (34)$$

By Proposition 1, we know  $NV_N \rightarrow \infty$  as  $N \rightarrow \infty$  and, thus, the variance vanishes as  $N \rightarrow \infty$ . □

The preceding propositions have established the asymptotic bias and variance of the pdf estimate. It now remains to show that

the pdf estimate is asymptotically consistent. By Markoff's Theorem [10, p. 212], we know that asymptotic unbiasedness (Proposition 2) and vanishing asymptotic variance (Proposition 3) imply the mean square convergence of  $\hat{f}$  to  $f$ , i.e.,

$$E[(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2] \rightarrow 0 \quad \text{as } N \rightarrow \infty. \quad (35)$$

By Tchebycheff's Inequality [10, p. 113], we know that mean square convergence implies convergence in probability. In other words, the probability density estimate,  $\hat{f}(\mathbf{x})$ , is asymptotically consistent.

#### 4 IMPLICATIONS AND CONCLUSION

When the recursive partitioning process is applied independently for each class, the above results imply that each class density estimate converges in probability to the respective true density. Hence, in the presence of large samples, a Bayes risk consistent classifier can be constructed via recursive partitioning with the MME criterion. Furthermore, unequal prior probabilities can be assigned to each class to form a general classification rule. This result supports the use of the MME criterion in supervised classification problems.

It is important to note that this paper has not addressed the rate of convergence of the pdf estimates to their asymptotic values with respect to increasing sample size. In other words, the established result does not yield any specific insight into the finite sample size behavior of the estimate, information that would impact the practical utility of MME-driven recursive partitioning. Nonetheless, some past studies have shown experimentally that finite sample performance is practically adequate and comparable to that of other established methods [4], [3], [2]. No mention has been made about the rate of degradation in the pdf estimates, in the presence of noise corrupted data. This robustness issue along with finite sample convergence rates are areas requiring further theoretical and experimental investigation.

In this short paper, it has been shown that the marginal maximum entropy criterion, when used in conjunction with recursive partitioning, provides asymptotically consistent pdf estimates. In turn, these pdf estimates can be used to construct a theoretically optimal classifier.

#### ACKNOWLEDGMENTS

The author would like to acknowledge the financial support of the Hospital for Sick Children Foundation, the Bloorview Children's Hospital Foundation, and the Natural Sciences and Engineering Research Council of Canada.

#### REFERENCES

- [1] A.P.A. daSilva, V.H. Quintana, and G.K.H. Pang, "Solving Data Acquisition and Processing Problems in Power Systems Using a Pattern Analysis Approach," *IEE Proc. C*, vol. 138, no. 4, 1991.
- [2] D.K.Y. Chiu, B. Cheung, and A.K.C. Wong, "Information Synthesis Based on Hierarchical Maximum Entropy Discretization," *J. Experimental and Theoretical Artificial Intelligence*, vol. 2, pp. 117-129, 1990.
- [3] A.K.C. Wong, D.K.Y. Chiu, and B. Cheung, "Information Discovery through Hierarchical Maximum Entropy Discretization," *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W.J. Frawley, eds., pp. 125-140, AAAI/MIT Press, 1987.
- [4] T. Chau and A.K.C. Wong, "Pattern Discovery by Residual Analysis and Recursive Partitioning," *IEEE Trans. Data and Knowledge Eng.*, vol. 11, no. 6, pp. 833-852, Nov./Dec. 1999.
- [5] E.G. Henrichon and K.S. Fu, "A Nonparametric Partitioning Procedure for Pattern Classification," *IEEE Trans. Computers*, vol. 18, no. 7, pp. 614-624, July 1969.
- [6] L. Gordon and R.A. Olshen, "Asymptotically Efficient Solutions to the Classification Problem," *The Annals of Statistics*, vol. 6, no. 3, pp. 515-533, 1978.

- [7] W.S. Meisel and D.A. Michalopoulos, "A Partitioning Algorithm with Application in Pattern Classification and the Optimization of Decision Trees," *IEEE Trans. Computers*, vol. 22, no. 1, pp. 93-103, Jan. 1973.
- [8] J.H. Friedman, "A Recursive Partitioning Decision Rule for Nonparametric Classification," *IEEE Trans. Computers*, pp. 404-408, 1977.
- [9] T.W. Anderson, "Some Nonparametric Multivariate Procedures Based on Statistically Equivalent Blocks," *Multivariate Analysis*, P.R. Krishnaiah, ed., pp. 5-28, 1966.
- [10] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.
- [11] D. W. Scott, *Multivariate Density Estimation*. John Wiley and Sons, 1992.
- [12] R.O. Duda and P.H. Hart, *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.
- [13] R.E. Haskell and A. Noui-Mehidi, "Design of Hierarchical Classifiers," *Computing in the 90's: The First Great Lakes Computer Science Conf. Proc.*, pp. 118-124, 1991.
- [14] J.A. Songquist, E.L. Baker, and J.N. Morgan, *Searching for Structure*. Survey Research Center, 1973.
- [15] S.K. Murthy, "Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey," *Data Mining and Knowledge Discovery*, pp. 345-389, 1998.
- [16] M. Lascrain, "On Maximum Entropy Discretization and its Applications in Pattern Recognition," Ph.D. thesis, Systems Design Eng., Univ. of Waterloo, Waterloo, Ontario, Canada, 1983.
- [17] C.E. Shannon, "Mathematical Theory of Communication," *Bell Systems Technical J.*, vol. 27, no. 3, pp. 379-423, 1948.
- [18] E.T. Jaynes, "Information Theory and Statistical Mechanics I," *Physical Rev.*, vol. 106, no. 1, pp. 620-630, 1957.
- [19] H.O. Lancaster, *The Chi-Squared Distribution*. John Wiley and Sons, 1969.
- [20] R.A. Christensen, *Log-Linear Models*. Springer-Verlag, 1990.
- [21] A. Agresti, *Categorical Data Analysis*. New York: Wiley, 1990.