Voronoi Networks and Their Probability of Misclassification

K. Krishna, M. A. L. Thathachar, *Fellow, IEEE,* and K. R. Ramakrishnan

***Abstract*—Nearest neighbor classifiers that use all the training samples for classification require large memory and demand large online testing computation. To reduce the memory requirements and the computation cost, many algorithms have been developed that perform nearest neighbor classification using only a small number of representative samples obtained from the training set. We call the classification model underlying all these algorithms as *Voronoi networks* (Vnets), because these algorithms discretize the feature space into Voronoi regions and assign the samples in each region to a class. In this paper we analyze the generalization capabilities of these networks by bounding the generalization error. The class of problems that can be "efficiently" solved by Vnets is characterized by the extent to which the set of points on the decision boundaries fill the feature space, thus quantifying how efficiently a problem can be solved using Vnets. We show that Vnets asymptotically converge to the Bayes classifier with arbitrarily high probability provided the number of representative samples grow slower than the square root of the number of training samples and also give the optimal growth rate of the number of representative samples. We redo the analysis for decision tree (DT) classifiers and compare them with Vnets. The bias/variance dilemma and the curse of dimensionality with respect to Vnets and DTs are also discussed.**

***Index Terms*—Neural networks, pattern recognition, statistical learning theory.**

I. Introduction

ONE OF THE most popular and simple pattern classifiers is the nearest neighbor classifier (NNC). NNC assigns to an unclassified sample the class label of the nearest sample in the set of training examples. Though there is no training phase, NNC needs sufficient memory to store all the training samples and needs to compute, for every test sample, its distance from each training sample. There is a class of algorithms that overcome this problem. These algorithms find a small number of representative samples using the training set and use these samples to perform nearest neighbor classification. Some of the algorithms in this class find a set of representative samples, which is a proper subset of the training set, such that the classification error over training samples is minimized [1]. The other algorithms obtain the representative samples, which need not be a subset of the training set, by applying an iterative algorithm on the training set, [2]–[7]. The popular learning vector quantization (LVQ) algorithm [3] belongs to this category of algorithms. All these algorithms, though need a training phase, are shown to perform better than NNC in many cases, apart from reducing the computational effort and memory requirements.

All the above-mentioned algorithms share the same classification model, i.e., perform NN classification using some representative samples instead of all the training samples. This classification model is called the *Voronoi network* (Vnet) in this paper because it discretizes the feature space into Voronoi regions [8] and assigns samples in each region to a class. Voronoi regions are formed by a set of points in the feature space. Each Voronoi region contains those points of the space that are closest to a point among points in the set. From now on, the representative samples are referred to as the *Voronoi centers* (Vcenters). In this paper, we analyze the capabilities of Vnets to generalize from a set of training samples to unseen test samples and address some issues in their design pertaining to the solution of pattern classification (PC) problems.

A. A Note on the Name "Voronoi Network"

The considered classification model is referred to by various names which have mostly originated from different learning algorithms. For example, we have LVQ classifier [3], nearest prototype or multiple prototype classifier [6], prototype-based NNC [4], nearest-neighbor-based multilayer perceptron (NN-MLP) [5], etc. One of the most popular among these names is the LVQ classifier. However, LVQ is also used to refer to the class of clustering algorithms (see, for example, [9]) though it was originally used to refer to a class of supervised learning algorithms [3]. In view of the above facts, a need was felt to denote the classification model underlying the above algorithms by an appropriate name. Since the fundamental building block for the classification model is a Voronoi region, the classification model is referred to as Vnet in this paper.

B. Generalization Error of Vnets

Since Bayes decision rule (BDR) is the best map that can be learned, it is proper to judge the generalization capability of a rule by the error between the rule and the BDR, which is referred to as the generalization error (GE) in this paper. GE is decided by two factors in case of Vnets, as in any other neural network. One of the factors is the representational capacity of the network. There could be some error between the BDR and the map v_K which is the best among those generated by Vnets with a fixed number K of Vcenters. To make this error small, the space of maps generated by Vnets must have sufficient power to represent or closely approximate the class of BDRs. This error is called the approximation error (AE). Second, the Vnet learned from a finite number of training samples might be far from the

Manuscript received August 26, 1999; revised July 11, 2000.

K. Krishna was with the the Department of Electrical Engineering, Indian Institute of Science, Bangalore, India. He is now with IBM India Research Laboratories, New Delhi, India.

M. A. L. Thathachar and K. R. Ramakrishnan are with the Department of Electrical Engineering, Indian Institute of Science, Bangalore, India.

Publisher Item Identifier S 1045-9227(00)09847-7.

best Vnet v_K. The reason for this error is the limited information available on the underlying distribution of samples because of finite number of training samples. This error is called the estimation error (EE). Previously, these two issues have been addressed in isolation [10]–[13] as well as together [14]–[16] for some class of networks. In this paper, we address both these issues for Vnets under a common framework as in [16], [14]; however, the analysis in [16], [14] deals with the problem of learning real-valued functions.

The approximation capabilities and the behavior of approximation error of various function approximation schemes, neural networks in specific, have been analyzed in the past (for example, see [10] and [11]). In all the studies the functions are assumed to be real valued and the problem of PC, where the functions map to a finite set of class labels, is considered as a special case [11]. In classical approximation theory, the rate of convergence of AE with respect to the number of parameters of the approximating function is obtained in terms of the degree of smoothness of the function that is to be approximated. In the case of PC problems any such characterization does not appear to have been done in the past. In this paper, the rate of convergence of AE of the maps, that take values from a finite set of symbols, *viz.*, the maps associated with PC problem, has been characterized by the extent to which decision boundaries fill the domain of the maps. One of the contributions of the present work is that the above-mentioned property is captured by a function called the P-*coarse index function* whose value at infinity is the *generalized fractal dimension* under a given probability measure P. The approximation error made by Vnets is bounded in terms of this function exploiting the nice geometric properties of Voronoi regions. This analysis, first of its kind to the best of our knowledge, gives insights into the sources of errors made by Vnets.

Blumer *et al.* [17] analyzed the estimation error in learning subsets of $\Re^d$ using the results of uniform convergence of empirical process from [18], [19] under the framework of Valiant's [20] probably approximately correct (PAC) learning. Haussler [13] generalized the analysis to learning real valued functions and Ben-David *et al.* [21] to $\{0, \ldots, N\}$-valued functions. In this paper, a bound on the estimation error has been derived by considering a class of $\{0, 1\}$-valued functions that is derived from the class of maps generated by Vnets using the results from [13]. Elsewhere [7], we have also derived a bound on estimation error by bounding Ψ-dimension of the class of maps generated by Vnets using the results from [21] and found that the former approach gave better results.

The main theorem of this paper gives a bound on the generalization error of Vnets obtained by separately bounding the components corresponding to AE and EE. As stated earlier, this result characterizes the approximating capabilities of maps taking values from a finite set of class labels that is used to find the rate of decrease of AE of Vnets. Furthermore, it gives a sufficient condition on the relative growth of the number of Vcenters in Vnet with respect to the number of training samples for the asymptotic convergence of the probability of misclassification to the Bayes error, thus proving the asymptotic convergence of Vnets to a corresponding BDR under fairly large class of probability measures. As a corollary, the optimal rate of growth of the number of Vcenters is derived so as to minimize GE. The theorem also sheds light on the tradeoff between the number of Vcenters and the number of training samples which has been referred to as the bias-variance dilemma in [15]. The phenomenon of "curse of dimensionality" in Vnets is also analyzed.

Decision tree (DT) is another popular classification model [22], [23] that shares a structure similar to that of Vnet. DT is a binary tree where each nonleaf node contains a split rule that decides whether a sample belonging to the left or right subtree, and each leaf node is assigned to a class. The bound obtained for the Vnets can be rederived for DTs too. A bound on GE is obtained for DTs and is compared with that obtained for Vnets. This kind of analysis suggests a basis for analytical comparison of any two different classification models. Moreover, it appears that, the observations made with respect to Vnets are applicable to other classification models too.

This paper is organized as follows. The following section develops Vnet classification model. The generalization error in Vnets is formulated in Section III. Section IV contains the main result *viz.*, a bound on GE. Proofs of most of the results stated in this section are relegated to Appendix for clarity in presentation. Various implications of the main result are discussed in Section V. The bound on GE for DTs is rederived and is compared with that of Vnets in Section VI. Section VII makes some comments on and discusses some possible extensions of the presented results. Finally, the paper ends with a summary in Section VIII.

II. Voronoi Networks

The task of pattern recognition is to learn a mapping from the feature space to the set of class labels. The learning takes place with respect to a set of training samples.[1] The mapping that is learned should be able to map unknown samples (samples that are not present in the training set) into their correct class labels. The ultimate aim of learning is to find a map that minimizes the average misclassification over the feature space. This is a well studied topic in pattern recognition literature [24]. The map that has the least average misclassification error is the BDR. Vnet approximates Bayes decision regions by Vornoi regions.

Vnet discretizes the feature space into disjoint convex sets, viz., Voronoi regions and assigns each region to a class label. Let $X \subseteq \Re^d$ denote the feature space and $Y = \{l_1, l_2, \ldots, l_N\}$ the set of class labels. Each Voronoi region is specified by a point *viz.*, the Vcenter, in X. Let $\{c_i \in X, i = 1, 2, \ldots, K\}$ be a set of Vcenters. Then, Vnet is defined as

$$v(\boldsymbol{c}, \boldsymbol{a}, x) \triangleq \sum_{i=1}^{K} a_i \chi_{S_i}(x), \qquad a_i \in Y \tag{1}$$

where $\boldsymbol{c} = [c_1, c_2, \ldots, c_K]$, $\boldsymbol{a} = [a_1, a_2, \ldots, a_K]$, $a_i \in Y$, $\chi_C(\cdot)$ is the indicator function of set C and S_i are as defined below. Let

$$\breve{S}_i \triangleq \{x\colon \|x - c_i\| < \|x - c_j\|,\ j \neq i,\ \forall j\},$$

and

$$\overline{S}_i \triangleq \{x\colon x \in \partial\breve{S}_i \cap \partial\breve{S}_j,\ j > i,\ \forall j\}$$

[1]The word "samples" is used to refer to a pair of a feature vector and its class label as well as to a feature vector alone. The particular reference would be clear from the context of usage.

where $||\cdot||$ is the Euclidean norm and ∂S is the boundary of the set S, i.e., if x is at an equidistant from two or more c_i then x is assigned to the set with the least index. Then define

$$S_i \triangleq \breve{S}_i \cup \overline{S}_i. \tag{2}$$

S_i are convex sets. Observe that $\{S_i\}$ partition X into K disjoint convex sets[2] i.e., $S_i \cap S_j = \emptyset$ if $i \neq j$, and $\cup S_i = X$.

A. Learning Vnets

Learning an input–output map using a Vnet involves finding the Vcenters $\{c_1, c_2, \ldots, c_K\}$ from a set of m training samples, $D_m = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, such that the misclassification error of the resulting map over the training set is minimized. It may be noted that NNC is an instance of Vnet, where $K = m$.

The empirical risk or the average misclassification error over D_m of a Vnet with K Vcenters specified by $\boldsymbol{c}$ and $\boldsymbol{a}$ is given by

$$\mathcal{E}(\boldsymbol{c}, \boldsymbol{a}) = \frac{1}{m} \sum_{i=0}^{m} d_Y(v(\boldsymbol{c}, \boldsymbol{a}, x_i), y_i) \tag{3}$$

where $v(\boldsymbol{c}, \boldsymbol{a}, \cdot)$ is as defined in (1) and d_Y is a discrete metric on Y given by

$$d_Y(a, b) = \begin{cases} 1 & \text{if } a \neq b, \\ 0 & \text{if } a = b \end{cases} \qquad \forall\, a, b \in Y. \tag{4}$$

The learning problem is to find the pair $(\boldsymbol{c}^*, \boldsymbol{a}^*)$ that minimizes $\mathcal{E}(\boldsymbol{c}, \boldsymbol{a})$. It turns that for a fixed $\boldsymbol{c}$, the vector $\boldsymbol{a}$, that minimizes $\mathcal{E}(\boldsymbol{c}, \cdot)$, is uniquely determined. So, the problem of finding $(\boldsymbol{c}^*, \boldsymbol{a}^*)$, in effect, boils down to finding $\boldsymbol{c}^*$ alone.

As mentioned in the previous section, there exist many algorithms that try to find c^*. As in case of any classification model, there exists a tradeoff between the average misclassification over training set and the probability of misclassification which reflects the generalization capabilities of Vnets. For example, one can easily show using a simple example that, a network with $K = m$ has a higher probability of misclassification than an optimal net with $K \ll m$, when K is properly chosen [7]. Thus the problem of finding K in Vnets is related to well-studied structural risk minimization in neural networks [18]. The results presented in this paper throw some light on this issue.

III. Problem Formulation

Let $X \subseteq \Re^d$ be the feature space and $Y = \{l_1, l_2, \ldots, l_N\}$ be the finite set of class labels. Denote $X \times Y$ as Z and (x, y) as z. Let $\mathcal{S}_X$ be a σ-field of subsets of X and $\mathcal{S}_Y$ be the power set of Y. Let $\mathcal{S}_Z = \mathcal{S}_X \times \mathcal{S}_Y$ be the smallest σ-algebra in $X \times Y$ which contains every set of the form $A \times B$, $A \in \mathcal{S}_X$, and $B \in \mathcal{S}_Y$. Let P be a probability measure on $(Z, \mathcal{S}_Z)$. Let P_X, and P_Y be the probability measures induced by P on $(X, \mathcal{S}_X)$, and $(Y, \mathcal{S}_Y)$, respectively. Let v be a $\mathcal{S}_X$ measurable map from X into Y. Then the average misclassification error, referred to as *expected risk,* incurred using v is given by

$$R[v] = \int_Z d_Y(v(x), y) P(dz).$$

[2] $\overline{S}_i$ has been defined to make the definition precise. If the underlying probability measure is absolutely continuous with respect to Lebesgue measure, then these $\overline{S}_i$'s can be ignored.

The aim of learning is to find a v that minimizes $R[v]$. Let

$$v^*(x) = \sum_{i=1}^{N} l_i \chi_{B_i}(x) \tag{5}$$

be a BDR that minimizes $R[v]$ i.e., $v^* = \arg\min_v R[v]$, where B_i are appropriately defined. Since we are interested in analyzing how best the learned rule approaches a BDR, the class of BDR that can be "efficiently" approximated by the rules generated by Vnets are characterized, instead of characterizing the space of probability measures.

Since P is unknown, $R[v]$ is unknown. Only source of information available is that of random samples drawn according to P. Let $D_m = \{(x_i, y_i)\}_{i=1}^m$ be the set of m training samples randomly drawn according to $P(x, y)$. Using this data set, the expected risk can be approximated by the *empirical risk* $\hat{R}_m$

$$\hat{R}_m[v] \triangleq \frac{1}{m} \sum_{i=1}^{m} d_Y(v(x_i), y_i). \tag{6}$$

A common strategy is to estimate the BDR as the mapping that minimizes the empirical risk.

Remark III.1: It is to be noted that, under fairly general assumptions, the expected risk converges in probability to the empirical risk for each given v, and not for all v simultaneously. Therefore, it is not guaranteed that the minimum of the empirical risk will converge to the minimum of the expected risk as m tends to infinity. Therefore, as pointed out and analyzed in the fundamental work of Vapnik and Chervonenkis [18], [19]the notion of uniform convergence in probability has to be introduced and it will be discussed later.

The standard practice in pattern recognition using neural networks is that the number of parameters of the network is fixed and $\hat{R}_m[v]$ is minimized over the parameter space, i.e., over the space of maps generated by networks with a fixed number of parameters. In the case of Vnets, the number of parameters is proportional to the number of Voronoi regions or the number of hidden units. Let $\mathcal{V}_K$ be the space of maps generated by Vnets with K hidden units, i.e.,

$$\mathcal{V}_K \triangleq \left\{ v\colon v(x) = \sum_{i=1}^{K} a_i \chi_{S_i}(x),\ a_i \in Y \right.$$
$$\left. S_i \text{ as in (2), for } \boldsymbol{c} \in X^K \right\}. \tag{7}$$

[From now on, a Vnet is denoted by $v(\cdot)$ omitting the variables $\boldsymbol{c}$ and $\boldsymbol{a}$.] Then, $\hat{R}_m[v]$ is minimized over $\mathcal{V}_K$ i.e., v^* is approximated by the function $v_{K,m}$ defined as

$$v_{K,m} \triangleq \arg\min_{v \in \mathcal{V}_K} \hat{R}_m[v]. \tag{8}$$

Assuming that the problem of learning $v_{K,m}$ from D_m, *viz.*, minimization of $\hat{R}_m[v]$ over $\mathcal{V}_K$, is solved, we are interested in finding out the "goodness" of $v_{K,m}$, apart from finding out the class of problems that can be solved "efficiently" using Vnets. The problem of finding $v_{K,m}$ is addressed by various learning algorithms mentioned in Section I. The goodness of $v_{K,m}$ is measured in terms of its expected risk $R[v_{K,m}]$, which is equal to the probability of misclassification. Since v^* is the best possible mapping that minimizes the probability of misclassification, $R[v^*] \leq R[v_{K,m}]$. The above issues are analyzed below

by bounding the difference $|R[v_{K,m}]-R[v^*]|$, which is referred to as the generalization error (GE).

Remark III.2: One expects $v_{K,m}$ to become a better and better estimate as K and m go to infinity. In fact, when m increases, the estimate of the expected risk $\hat{R}_m[v]$ and hence the estimator $v_{K,m}$ improves. But, when K increases, the number of parameters to be estimated increases with the same amount of data, and hence the estimate of the expected risk deteriorates. In this situation, the estimate can be improved by seeking more data. Therefore, K and m have to grow as a function of each other for convergence to occur. The relation between K and m for the convergence is derived as a corollary of the main theorem of this paper.

IV. Bound on Generalization Error

There are two factors involved in the approximation of v^* by $v_{K,m}$, viz., how accurately v^* can be approximated by the maps with K Voronoi regions, and given a data of size m, how accurately $v_{K,m}$ approximates the closest function among $\mathcal{V}_K$ to v^* with respect to the metric $d_{Y,P}$ [defined below in (10)] for a given P. To study these factors, let v_K be the mapping with K Vcenters having the least expected risk, i.e.,

$$v_K = \arg\min_{v\in\mathcal{V}_K} R[v].$$

Then it follows from the definitions that $R[v^*] \le R[v_K] \le R[v_{K,m}]$, and

$$\begin{aligned}|R[v_{K,m}]-R[v^*]| &= |R[v_K]-R[v^*]| + |R[v_{K,m}]-R[v_K]|. \end{aligned} \tag{9}$$

The first term in the right-hand side (RHS) of (9) is referred to as the AE since, it is the error in approximating v^* by the functions in $\mathcal{V}_K$. The second term is referred to as the EE since, $v_{K,m}$ uses m observations, D_m, to estimate the K Voronoi centers associated with v_K by minimizing the empirical risk $\hat{R}_m$. We bound the generalization error by bounding these two terms separately.

A. Bound on Approximation Error

Notice that the approximation error does not depend on D_m but, it depends on v^*. Therefore, to bound the approximation error, the space $\mathcal{U}$ of BDRs for which approximation error asymptotically goes to zero with K is characterized. In particular, our interest is in such a $\mathcal{U}$ for which the following expression holds:

$$\forall\, v\in\mathcal{U},\ \exists\, v_K\in\mathcal{V}_K\ \ni\colon |R[v]-R[v_K]| < \xi(K)$$

where $\xi(K)$ goes to zero faster than $1/K^t$, for some $t>0$, as K goes to infinity.

This problem is analyzed using a metric defined on the space $\mathcal{U}$ of maps from X into Y. For a given probability measure P on $(X, \mathcal{S}_X)$, the discrete metric d_Y, defined on Y (4), can be extended to a pseudometric $d_{Y,P}$ on $\mathcal{U}$ as follows:

$$\begin{aligned} d_{Y,P}(v_1, v_2) &= \int_X d_Y(v_1(x), v_2(x))P(dx) \\ &\qquad \forall\, v_1, v_2 \in \mathcal{U}. \end{aligned} \tag{10}$$

It may be noted that

$$|R[v]-R[v_K]| \le d_{Y,P}(v, v_K). \tag{11}$$

Therefore, in this section, $d_{Y,P}(v, v_K)$ is bounded to bound $|R[v]-R[v_K]|$ which depends on v especially, the *P-coarse index function* of v. Before formally deriving the bound, the intuition behind defining P-coarse index function of a mapping is described below.

Any two neighboring Vcenters generate a hyper planar decision boundary. Since these hyper planes can efficiently approximate any surface which is smooth enough, it is intuitive that BDR with smooth decision boundary surfaces can be efficiently approximated by Vnets. On the other hand, considering the contribution by a Voronoi region to the average misclassification error, if a Voronoi region is a proper subset of B_i [see (5)] for some i, then the error made in this region is zero. Therefore, the number of such Voronoi regions, if each region is equally probable, should reflect how close is the mapping generated by the Vnet to the BDR that is being approximated and hence the approximation error. The relation between this number and the smoothness of the decision boundary surfaces is brought out in Section V.

Definition IV.1: Let v be a measurable function from $X \mapsto Y$. Then, $B_i^{(v)}$ denote the preimages of v corresponding to the class l_i, i.e., $B_i^{(v)} \triangleq \{x\colon v(x) = l_i\}$. Let P be a probability measure on $(X, \mathcal{S}_X)$. Then a set $A\in\mathcal{S}_X$ is said to be *P-pure* with respect to v if $P(A\cap B_i^{(v)}) = P(A)$ for some i. A is said to be *P-impure* if A is not P-pure.

Let $\mathcal{C}_K = \{C_1, \ldots, C_K\}$ be a set of hypercubes formed by $(n_1+\ldots+n_d)$ hyperplanes such that n_i hyperplanes are perpendicular to ith axis and $\prod_{i=1}^d (n_i+1) = K$. If P is absolutely continuous with respect to Lebesgue measure μ (denoted by $P \ll \mu$), then C_i are chosen such that $P(C_i) \le 1/K$. Otherwise, C_i are chosen such that probability of each hypercube is less than $1/K$ except those containing points of positive probability measure greater than $1/K$. If a hypercube contains a point of probability θ $(>1/K)$ then the probability of such a hypercube is utmost $\theta+(1/K)$. Then, the minimum number of P-impure sets in $\mathcal{C}_K$, minimum over all possible $\mathcal{C}_K$, represented by $\Lambda_{(P,v)}(K)$, is called the *P-coarse index function* of v. □

Theorem IV.1: Let P be a probability measure on $(X, \mathcal{S}_X)$. Then for all measurable $v\colon X\mapsto Y$, there exist a $v_K\in\mathcal{V}_K$ such that

$$d_{Y,P}(v, v_K) \le \frac{\Lambda_{(P,v)}(K)}{K}.$$

Proof: This theorem is proved by constructing a v_K, for a given v, that satisfies the above bound. For some v and $v_K \in \mathcal{V}_K$, $v_K(x) = \sum_{i=1}^K a_i \chi_{S_i}(x)$

$$\begin{aligned} d_{Y,P}(v, v_K) &= \int_X d_Y(v(x), v_K(x))P(dx) \\ &= \sum_{i=1}^K \int_{X\cap S_i} d_Y(v(x), a_i)P(dx) \end{aligned}$$

since $v_K = a_i$ on S_i. Let

$$\eta_i = \int_{X\cap S_i} d_Y(v(x), a_i)P(dx).$$

then, $\eta_i = \sum_{j=1}^{N} d_Y(l_j, a_i) P(B_j^{(v)} \cap S_i)$. Therefore, if $i' = \arg\max_j P(B_j^{(v)} \cap S_i)$ then, $a_i = l_{i'}$ minimizes η_i. It is to be noted here that if S_i is a P-pure set with respect to v, ($P(S_i \cap B_j^{(v)}) = 0$ for all j except for $j = i'$), then $\eta_i = 0$. Therefore

$$d_{Y,P}(v, v_K) \leq \sum_{i:\, S_i \text{ is } P\text{-impure set}} \eta_i.$$

Now, S_i are appropriately chosen to bound the error. For a given K, consider a $\mathcal{C}_K$ containing $\Lambda_{(P,v)}(K)$ P-impure sets with respect to v. Since these hypercubes can be obtained from K Vcenters, choose $S_i = C_i \in \mathcal{C}_K$; for $i = 1, 2, \ldots, K$. Now observe that if $P \ll \mu$ then, η_i can be bounded by $((N-1)/N)\ \Pr(S_i)$ because, $B_j^{(v)}$ partition S_i and in the worst case, $\Pr(B_1^{(v)} \cap S_i) = \ldots = \Pr(B_N^{(v)} \cap S_i)$. If P is not absolutely continuous with respect to μ and S_i is a P-impure set containing a point with probability measure greater than $1/K$, then $\eta_i < 1/K$. From the above construction and observations, we get

$$d_{Y,P}(v, v_K) \leq \frac{\Lambda_{(P,v)}(K)}{K}$$

thus proving the theorem. ∎

The above theorem, because of (11), implies that for a given P and v, there exists a v_K in $\mathcal{V}_K$ such that

$$|R[v] - R[v_K]| \leq \frac{\Lambda_{(P,v)}(K)}{K}. \tag{12}$$

Therefore, the rate at which this bound approaches zero as K tends to infinity, depends on $\Lambda_{(P,v)}(K)$, which in turn depends on the underlying P and v. We relate $\Lambda_{(P,v)}(K)$ to the generalized fractal dimension (defined later) of the set of points on the decision boundaries of v and characterize the rate of decrease of the bound in terms of this dimension in Section V.

B. Bound on Estimation Error

Since $R[v_{K,m}]$ is a random variable, $|R[v_{K,m}] - R[v_K]|$ is bounded only in a probabilistic sense, i.e., with probability $1-\delta$

$$|R[v_{K,m}] - R[v_K]| \leq \omega(m, K, \delta), \qquad \forall v_K \in \mathcal{V}_K. \tag{13}$$

Let

$$q(m, \varepsilon) \triangleq \sup_{v_K \in \mathcal{V}_K,\, P \in \mathcal{P}} \Pr\{(x, y) \in (X \times Y)^m: |R[v_{K,m}] - R[v_K]| > \varepsilon\}$$

Then, EE can be bounded if $q(m, \varepsilon)$ is bounded. We bound $q(m, \varepsilon)$ and consequently bound $\omega(m, K, \delta)$ using results from statistical learning theory [25].

One expects $v_{K,m}$ to converge to v_K as one gets more and more data. This happens if $\hat{R}_m[\cdot]$ converges to $R[\cdot]$ uniformly in probability [18]. The following lemma gives a relationship between the estimation error and $|\hat{R}_m[\cdot] - R[\cdot]|$.

Proposition IV.2: [25] If $|\hat{R}_m[v] - R[v]| \leq \omega, \forall v \in \mathcal{V}_K$ then, $|R[v_{K,m}] - R[v_K]| < 2\omega$.

Thus EE is bounded if the difference between the empirical risk and true risk is bounded uniformly over $\mathcal{V}_K$. This problem is converted to that of analyzing the convergence of empirical estimate of expectation of a class of random variables to their true values below. Let $\mathcal{A}$ be the class of $\{0, 1\}$-valued functions defined on $X \times Y$ by the metric d_Y, viz.,

$$\mathcal{A} = \{a: X \times Y \to \{0, 1\},\ a(x, y) = d_Y(v(x), y),\ v \in \mathcal{V}_K\}. \tag{14}$$

Then for any $a \in \mathcal{A}$, it may be noted that the empirical expectation of a based on D_m, $\hat{E}_m[a] = \hat{R}_m[v]$, and the true expectation of a under the probability distribution $P(x, y)$, $E[a] = R[v]$. Bounding the error between empirical and true expectations of a class of random variables is a well-studied topic [19]. Let

$$r(m, \varepsilon) = \sup_{a \in \mathcal{A},\, P \in \mathcal{P}} \Pr\{z \in Z^m: |\hat{E}_m[a] - E[a]| > \varepsilon\}. \tag{15}$$

Then the following proposition directly follows from the definitions of $r(m, \varepsilon)$ and $q(m, \varepsilon)$ and from Proposition IV.2.

Proposition IV.3: If $r(m, \varepsilon/2) \leq \delta$ then, $q(m, \varepsilon) \leq \delta$.

Therefore, it is enough to bound $r(m, \varepsilon/2)$ in order to bound $q(m, \varepsilon)$. Let

$$\mathcal{C} = \{C \subseteq Z: C = \{z: a(z) = 1\},\ a \in \mathcal{A}\}. \tag{16}$$

Then, the convergence of $\hat{E}_m[a]$ to $E[a]$ is same as the convergence of empirical probability to true probability of the corresponding subset of Z. It may be noted that $r(m, \varepsilon)$ can be bounded using the VC-dimension of $\mathcal{C}$ [19]. Ben-David *et al.* [21] gave a relation between the VC-dimension of $\mathcal{C}$ and the Ψ-dimension of $\mathcal{A}$. Therefore $r(m, \varepsilon)$ can be bounded by bounding Ψ-dimension of $\mathcal{A}$. $r(m, \varepsilon)$ can also be bounded, considering $\mathcal{A}$ as a family of real valued functions and using the pseudodimension of $\mathcal{A}$ as given in [13]. We have derived the bounds using both the techniques [7] and found that the results obtained using the latter approach are better than those obtained using the former. The derivation of the bound using the pseudodimension of $\mathcal{A}$ is given below which follows along similar line to the one given in [16]. Some definitions and results that are useful in the derivation of bound on $r(m, \varepsilon)$ are given in Appendix A.

Consider the family of functions $\mathcal{A}$ defined in (14). Then Theorem A.4 implies

$$\Pr\Big(\exists\, a \in \mathcal{A}: |\hat{E}_m[a] - E[a]| > \varepsilon\Big) \leq 4E\Big[\mathcal{N}\Big(\frac{\varepsilon}{8}, \mathcal{A}_{D_m}, d_{a1}\Big)\Big] \exp\Big(-\frac{\varepsilon^2 m}{32}\Big) \tag{17}$$

where $\mathcal{A}_{D_m}$ is the D_m-restriction of $\mathcal{A}$ given by $\mathcal{A}_{D_m} = \{(a(z_1), \ldots, a(z_m)): a \in \mathcal{A}\}$, d_{a1} the average l_1 metric, and $\mathcal{N}(\varepsilon, \mathcal{A}_{D_m}, d_{a1})$ the ε-covering number of $\mathcal{A}_{D_m}$ as defined in Appendix A. The following lemma, whose proof has been relegated to Appendix B for continuity of presentation, relates the metric capacities of $\mathcal{A}$ ($\mathcal{C}(\varepsilon, \mathcal{A}, d_{a1})$), and $\mathcal{V}_K$ ($\mathcal{C}(\varepsilon, \mathcal{V}_K, d_Y)$), and the expectation of the covering number of $\mathcal{A}_{D_m}$.

Lemma IV.4: $E[\mathcal{N}(\varepsilon, \mathcal{A}_{D_m}, d_{a1})] \leq \mathcal{C}(\varepsilon, \mathcal{A}, d_{a1}) \leq \mathcal{C}(\varepsilon, \mathcal{V}_K, d_Y)$.

Using the above lemma and taking supremum on both sides of (17) over all P, the $r(m, \varepsilon)$ defined in (15) can be bounded as

$$r(m, \varepsilon) \leq 4\mathcal{C}\left(\frac{\varepsilon}{8}, \mathcal{V}_K, d_Y\right) \exp\left(-\frac{\varepsilon^2 m}{32}\right). \quad (18)$$

To find a bound on $\mathcal{C}(\varepsilon, \mathcal{V}_K, d_Y)$, $\mathcal{V}_K$ is decomposed as the composition of two classes of functions. Let $\mathcal{H}_1$ be the class of functions from $\Re^k$ to $\overline{\boldsymbol{e}}$, generated by K-hidden neurons of Vnet where, $\overline{\boldsymbol{e}} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\}$, and $\boldsymbol{e}_i$ is the K-dimensional unit vector in the ith dimension i.e., ith component of $\boldsymbol{e}_i$ is 1 and the rest of the components are zeros. Then

$$\mathcal{H}_1 = \{h_1(\boldsymbol{x}) = (\chi_{S_1}(\boldsymbol{x}), \ldots, \chi_{S_K}(\boldsymbol{x}))\} \quad (19)$$

where $\{S_i\}$ are defined as in (2). Let $\mathcal{H}_2$ be the class of functions, from $\overline{\boldsymbol{e}}$ to Y, generated by the single output neuron of the network. Note that $|\mathcal{H}_2| \leq N^K$. Then, $\mathcal{V}_K = \{h_2 \circ h_1 \colon h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$. The following lemma bounds $\mathcal{C}(\varepsilon, \mathcal{V}_K, d_Y)$ in terms of the metric capacity of $\mathcal{H}_1$. The proof of this lemma is given in Appendix B.

Lemma IV.5:

$$\mathcal{C}(\varepsilon, \mathcal{V}_K, d_Y) \leq N^K \mathcal{C}\left(\frac{2\varepsilon}{K}, \mathcal{H}_1, d_{a1}\right).$$

Since $\mathcal{H}_1$ is a vector of index functions of some class of subsets, some results from Appendix A can be used to derive a bound on $\mathcal{C}(\varepsilon, \mathcal{H}_1, d_{a1})$, thus bounding $\mathcal{C}(\varepsilon, \mathcal{V}_K, d_Y)$.

Theorem IV.6:

$$\mathcal{C}(\varepsilon, \mathcal{V}_K, d_Y) \leq (2N)^K \left(\frac{Ke}{\varepsilon} \ln \frac{Ke}{\varepsilon}\right)^{K\eta}$$

where $\eta = 2(d+1)(K-1)\log(3(K-1))$.

Proof: Let $\mathcal{S}$ be the class of subsets of the form S_i defined in Section II. Let $\mathcal{G} = \{\chi_S(x) \colon S \in \mathcal{S}\}$, then $\mathcal{H}_1 = \mathcal{G}_K$ where, $\mathcal{G}_K = \{(g_1(x), \ldots, g_K(x)) \colon g_i \in \mathcal{G}, 1 \leq i \leq K\}$. Then, from Theorem A.5, we get, $\mathcal{C}(\varepsilon, \mathcal{H}_1, d_{a1}) \leq (\mathcal{C}(\varepsilon, \mathcal{G}, d_{a1}))^K$. Because $\mathcal{G}$ is a space of the indicator functions, $\mathbf{dim}_P(\mathcal{G}) = \mathbf{dim}_{\mathrm{VC}}(\mathcal{S})$. Since $\mathcal{S}$ is a class of subsets where each subset is an intersection of at most $K-1$ half spaces, Lemmas A.1 and A.2 implies $\mathbf{dim}_P(\mathcal{G}) \leq \eta$. Substituting this in (34) for $\mathcal{G}$, we get $\mathcal{C}(\varepsilon, \mathcal{G}, d_{a1}) < 2((2e/\varepsilon)\ln(2e/\varepsilon))^\eta$. Therefore, because of Theorem A.5, $\mathcal{C}(\varepsilon, \mathcal{H}_1, d_{a1}) \leq 2^K((2e/\varepsilon)\ln(2e/\varepsilon))^{K\eta}$. The above result follows from Lemma IV.5 on substitution of the bound for $\mathcal{C}(\varepsilon, \mathcal{H}_1, d_{a1})$. ■

Theorem IV.7: With probability $(1-\delta)$, the following bound holds:[3]

$$|R[v_{K,m}] - R[v_K]| \leq O\left(\left[\frac{K\ln(N) + dK^2\ln(K)\ln(Km) + \ln\left(\frac{1}{\delta}\right)}{m}\right]^{1/2}\right). \quad (20)$$

Proof: For η defined in Theorem IV.6, substituting the bound for $\mathcal{C}(\varepsilon, \mathcal{V}_K, d_Y)$ in (18) we get

$$r(m, \varepsilon) \leq 4(2N)^K \left(\frac{8Ke}{\varepsilon} \ln \frac{8Ke}{\varepsilon}\right)^{K\eta} \exp\left(-\frac{\varepsilon^2 m}{32}\right).$$

From Proposition IV.3, $q(m, \varepsilon) \leq \delta$ if

$$4(2N)^K \left(\frac{16Ke}{\varepsilon} \ln \frac{16Ke}{\varepsilon}\right)^{K\eta} \exp\left(-\frac{\varepsilon^2 m}{128}\right) \leq \delta \quad (21)$$

This implies that $\Pr(\forall v \in \mathcal{V}_K \colon |R[v_{K,m}] - R[v_K]| \leq \varepsilon) \geq (1-\delta)$ if, the above inequality (21) is satisfied. Therefore, the following inequality should be satisfied:

$$\varepsilon^2 m \geq 128\Big(K\ln(2N) + 2K\eta(\ln(16Ke) - \ln(\varepsilon)) + \ln\left(\frac{4}{\delta}\right)\Big). \quad (22)$$

It is later shown that the above inequality is satisfied if (23), shown at the bottom of the page, is true. Let $p = 128[K\ln(2N) + 2K\eta(\ln(16Ke) + (1/2)\ln(m)) + \ln(4/\delta)]$. Since $p \geq 1$, the inequality (22) is satisfied for the above value of ε; thus proving the theorem. ■

C. The Main Theorem

Theorem IV.8: Let P be a probability measure on $(Z, \mathcal{S}_Z)$. Let v^* be a BDR corresponding to P and $\Lambda_{(P_X, v^*)}(K)$ be the P_X-coarse index function of v^*. Let $D_m = \{(x_i, y_i) \colon i =$

[3]In general, these bounds are known to give very large numbers if different variables in the bounds are substituted with their values. However, they are qualitatively tight in the sense that there exist some distributions for which the bound is attained. Therefore, in this paper, the bound is found only in the order notation.

$$\varepsilon = \left[\frac{128\left[K\ln(2N) + 2K\eta(\ln(16Ke) + \frac{1}{2}\ln(m)) + \ln\left(\frac{4}{\delta}\right)\right]}{m}\right]^{1/2}. \quad (23)$$

$1, \ldots, m\}$ be a set of independent training examples drawn according to P. Let $v_{K,m}$ be as defined in (8). Then, with probability $(1-\delta)$

$$|R[v^*] - R[v_{K,m}]| \leq \frac{\Lambda_{(P_X, v^*)}(K)}{K} + O\left(\left[\frac{K\ln(N) + dK^2\ln(K)\ln(Km) + \ln\left(\frac{1}{\delta}\right)}{m}\right]^{1/2}\right). \tag{24}$$

Proof: The theorem follows from (9), Theorem IV.1 and Theorem IV.7. ■

V. Implications of the Theorem

In this section, various implications of the main theorem of this paper are discussed.

A. Convergence of Vnet to Bayes Decision Rule

Under the assumption that AE asymptotically converges to zero with K, which is justified in the next section, the discussion in the sequel implies that, if $K = m^r$ and $r < 1/2$ then, the bound on GE in (24) asymptotically goes to zero. Substituting $K = m^r$ in the expression for $\omega(m, K, \delta)$ [refer (13)] in (20) and letting m tend to infinity, we get

$$\lim_{m\to\infty} \omega(m, K, \delta) = \lim_{m\to\infty}\left(\frac{m^r\ln(N) + dm^{2r}r(r+1)\ln^2(m) - \ln(\delta)}{m}\right) = \lim_{m\to\infty} m^{2r-1}\ln^2(m) = 0.$$

Therefore, Vnet asymptotically converges to a BDR with arbitrary high probability if the number of Voronoi regions grow slower than the square root of the number of training samples.

B. How Efficiently can Vnets Solve a Problem?

Efficiency of learning a given pattern recognition task, v^*, from a finite set of training samples can be measured by the rate of convergence of $R[v_{K,m}]$ to $R[v^*]$. The larger the rate the more the efficiency. It follows from (24) that, for a fixed number of Vcenters K, as the number of training samples m increases, the bound on EE monotonically goes to zero where as the bound on AE remains constant since it is independent of m. Therefore, the rate of convergence of $R[v_{K,m}]$ to $R[v^*]$ also depends on the rate of decrease of the bound on AE, and hence AE, to zero. The only parameter that depends on the problem under consideration *viz.*, $\Lambda_{(P_X, v^*)}(\cdot)$ (refer Definition IV.1), in the bound on GE is due to the bound on AE and the bound on EE is independent of v^*. So, $\Lambda_{(p_X, v^*)}(K)$ is analyzed to find the efficiency of learning by Vnets. This analysis also gives an insight into the approximation capabilities of Vnets.

Definition V.1: Let P be a probability measure on $(X, \mathcal{S}_X)$ and $v = \sum_{i=1}^{N} c_i\chi_{B_i}(x)$ be a measurable function. Then define

$$B_v \triangleq \{x\text{: } B_\varepsilon(x) \text{ is } P\text{-impure}, \ \forall\varepsilon > 0\}$$

where $B_\varepsilon(x)$ is the ε-ball around x in X. Let

$$d_{(P,v)}(K) \triangleq \begin{cases} 0 & \text{if } \Lambda_{(P,v)}(K) = 0, \\ \dfrac{d\ln(\Lambda_{(P,v)}(K))}{\ln(K)} & \text{otherwise.} \end{cases} \tag{25}$$

Then, $\limsup_{K\to\infty} d_{(P,v)}(K)$, if the limit exists, is called the *generalized fractal dimension* (GFD) of B_v with respect to P, denoted by $P\text{-}\mathbf{dim}_{GF}(B_v)$. □

GFD is called so because, it is a generalization of the *fractal dimension*, which is same as the *upper entropy index* defined in [26]. The fractal dimension is defined on the compact sets using the Lebesgue measure. Here, GFD is defined with regard to v and P. However, the above definition can be appropriately modified for defining the dimension on any arbitrary sets with regard to a given P. In this modified definition, $P\text{-}\mathbf{dim}_{GF}(A)$ is equal to the fractal dimension of A, when P has a compact support and is uniform.

Writing the bound on AE in terms of $d_{(P,v)}(K)$, we get

$$|R[v] - R[v_K]| < \frac{1}{K^{((d-d_{(P,v)}(K))/d)}}. \tag{26}$$

It may be observed that $d_{(P,v)}(K)$ always lies in the interval $[0, d]$. The above inequality (26) says that the nearer $d_{(P,v)}(K)$ stays to d the slower is the rate of decrease of the bound on AE with K.

To analyze further, assume P has a compact support and is uniform. Then higher the filling property of B_v more the value of $d_{(P,v)}(K)$ and hence more the difficulty to approximate v by Vnets. For a given B_v, if P's distribution over the regions neighboring the points of B_v is less, then it makes $d_{(P,v)}(K)$ small making the rate of decrease of bound large. In conclusion, as it should be expected, the approximation of a mapping v by Vnets depends on the space filling property of B_v and the probability distribution over the regions around it.

C. Optimal Growth Rate of K

It is shown above that for $R[v_{K,m}]$ to converge to $R[v^*]$ the number of Voronoi regions should grow slower than the square root of the number of training samples. Then the question arises; does there exist an optimal growth rate of K with respect to m? By the optimal growth, we mean the relation $K^*(m)$ between K and m such that for a fixed number of training samples m, the Vnet with $K^*(m)$ number of Vcenters minimizes the bound on GE. The existence of such a $K^*(m)$ is quite intuitive from

Remark III.2. This is reflected in the bound on GE (24). To get an explicit expression for $K^*(m)$, the minimum of the bound on GE, which is assumed to be close to the minimum of GE, is found by solving the following equation:

$$\frac{\partial}{\partial K}\left[\frac{1}{K^{1-(d_K/d)}} + \left(\frac{K\ln(N)+dK^2\ln(K)\ln(Km)-\ln(\delta)}{m}\right)^{1/2}\right]=0$$

where $d_K = d_{(P_X, v^*)}(K)$. Applying the derivative and neglecting the less significant terms we get

$$\frac{d-d_K}{d}\,\frac{1}{K^{2-(d_K/d)}} = \frac{(Kd^3(\ln(m)\,\ln(K))^3)^{1/2}}{\sqrt{m}}.$$

After some algebra we get that the estimate of $K^*(m)$, for large m, is proportional to

$$\left(\frac{(d-d_K)^2 m}{d^5\,\ln^5(m)}\right)^{(d/(8d-2d_K))}. \tag{27}$$

The above expression (27) suggests one more interesting behavior of the optimal growth rate of K. When d_K is more, i.e., the problem under consideration is more difficult (refer the previous sub-sections), the optimal growth rate of K is higher. That is, the number of Vcenters in Vnets can grow faster as the number of training samples increase, when solving more difficult problems. This is indeed desirable, because larger Ks make the AE small.

D. Dependence on the Dimension of X

It is known from the theory of linear and nonlinear widths [27] that if the target function has d variables and a degree of smoothness s, one should not expect to find that an approximation error goes to zero faster than $O(k^{-(s/d)})$, when the number of parameters that have to be estimated is proportional to k. Here the degree of smoothness s is a measure of how constrained the class of functions are; for example, the number of derivatives that are uniformly bounded or the number of derivatives that are integrable or square integrable. Therefore, from the classical approximation theory, one expects that unless certain constraints are imposed on the class of functions to be approximated, the rate of convergence will dramatically slow down as the dimension d increases, showing the phenomenon known as the "curse of dimensionality."

Let $d_K = d_{(P_X, v^*)}(K)$. The term d_K reflects the frequency of variation of v^* on its domain and the term $(d-d_K)$ in (26), in a way, characterizes its smoothness. In case of Vnets, AE goes to zero as fast as $O(K^{-(d-d_K)/d})$. Therefore, the above conclusions from the classical approximation theory are valid in case of Vnets also with $(d-d_K)$ as smoothness parameter.

It is interesting to note the following observation at this point. The expression in (27) suggests that when $d_K \approx d-1$, the optimal growth rate of K increases from 1/8 to 1/6 as $d/(6d+2)$. That is, K can be increased at a faster rate for higher dimensions maintaining the optimal performance. Nevertheless, the increase in the optimal rate of K cannot compensate AE because of the $1/d$ term in the exponent of K in the bound AE. In fact, for higher dimensions the optimal growth rate of K is almost independent of the dimension of X because, the optimal rate reaches very close to 1/6 even for moderate values of K.

E. Tradeoff Between K and m

Theorem IV.8 suggests that there exist various choices of K and m for a given specifications of confidence and accuracy parameters. This is due to the tradeoff between AE and EE. When a large number of centers and less number of training samples are chosen, the resulting EE could be high, but this can be compensated by low AE and vice versa, in order to keep GE constant. If the availability of the training samples is expensive then, one could use more number of centers with the same bound on GE. Or, if the network size is expensive one could use large number of training samples. Thus the economics of this tradeoff would yield a suitable values for K and m to solve the pattern recognition problem with the required accuracy and confidence. Geman *et al.*, termed this trade off as bias/variance dilemma [15].

F. Sample Complexities

Finally, we consider the bounds on sample complexity of Vnets. As defined in Section IV-B, sample complexity reflects the number of samples sufficient to learn a mapping with a given accuracy and confidence requirements. From the value of η defined in Theorem IV.6, the expression in (22) implies that $q(m, \varepsilon) < \delta$ if

$$m(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon^2}\left(K\,\ln(N) + dK^2\,\ln^2(K)\,\ln(1/\varepsilon) + \ln(1/\delta)\right)\right).$$

VI. Bound on GE of Decision Trees

The classification model that shares a similar structure, but not the same, with Vnets is the DT classifier [22]. A DT is a binary tree. Each nonleaf node in the tree is associated with a split rule that decides whether a given pattern will go to the left or right subtree. A class label is assigned to each leaf node implying that all the patterns landing in that leaf node are to be classified into that class. When the feature space is a subset of $\Re^d$, then the class of split rules considered could be the class of hyper surfaces that are either parallel or oblique to the axes. The DT's using the former class of split rules are called Axes-parallel DTs and those using the later are oblique DTs. There are many learning algorithms for these DTs (for example, [28]). Better and better learning algorithms are still being discovered.

The learnability of this class of classifiers has been analyzed and bounds on sample complexities have been found by bounding the VC-dimension of DTs which are used for two-class pattern classification problems [29], [30]. The analysis presented in Section IV can be easily reworked to the class of maps generated by DTs. A brief sketch of the proof of the theorem for DTs is given below. As expected, the approximating behavior of DTs is quite similar to that of Vnets

except that the number of Vcenters is replaced by the number of leaf nodes. To find a bound on EE, DT is considered as a combination of l (number of leaf nodes) number of indicator functions where the set associated with each indicator function is an intersection of n (the depth of the tree) half spaces. Let $\mathcal{V}_{n,l}$ be the class of mappings generated by DTs for a fixed n and l. Then, the $v_{K,m}$ and v_K corresponding to DTs are

$$v_{n,l,m} \triangleq \arg \min_{v \in \mathcal{V}_{n,l}} \hat{R}_m[v] \tag{28}$$

where $\hat{R}_m[\cdot]$ is as defined in (6), and

$$v_{n,l} \triangleq \arg \min_{v \in \mathcal{V}_{n,l}} R[v],$$

where $R[\cdot]$ is the average misclassification error.

Theorem VI.1: Let P be a probability measure on $(Z, \mathcal{S}_Z)$. Let v^* be a BDR corresponding to P and $\Lambda_{(P_X, v^*)}(l)$ be the P_X-coarse index function of v^*. Let $D_m = \{(x_i, y_i)\colon i = 1, \ldots, m\}$ be a set of independent training examples drawn according to P. Let $\mathcal{V}_{n,l}$ be as defined in (28). Then, with probability $(1-\delta)$

$$\begin{aligned} |R[v^*] - R[v_{n,l,m}]| &\leq \frac{\Lambda_{(P_X,v^*)}(l)}{l} \\ &\quad + O\left(\left[\frac{l\ln(N) + dln\ln(n)\ln(lm) - \ln(\delta)}{m}\right]^{1/2}\right). \end{aligned} \tag{29}$$

A Sketch of Proof: The first part of the above inequality follows from the definition of P_X-coarse index function of v^*. To get the second part, we need to bound $\mathcal{C}(\varepsilon, \mathcal{V}_{n,l}, d_Y)$ because, following similar arguments as in Section IV-B:

$$r(m, \varepsilon) \leq 4\mathcal{C}\left(\frac{\varepsilon}{8}, \mathcal{V}_{n,l}, d_Y\right) \exp\left(-\frac{\varepsilon^2 m}{32}\right). \tag{30}$$

Since $\mathcal{V}_{n,l}$ is a combination l indicator functions each is associated with an intersection of n half spaces, Theorem IV.6 implies that

$$\mathcal{C}(\varepsilon, \mathcal{V}_{n,l}, d_Y) \leq (2N)^l \left(\frac{4le}{\varepsilon} \ln \frac{4le}{\varepsilon}\right)^{l\eta}$$

where $\eta = 2(d+1)(n-1)\log(3(n-1))$. Substituting this bound in (30) and after some algebra we get the second part of the inequality in the statement of the theorem. □

From the above theorem, one can say that the efficiency of solving a problem, the dependence of GE on the dimension of X, and the tradeoff between l and m in case of DTs are the same as in Vnets. One can analyze the optimal growth rate of the depth n of a tree and the number l of leaf nodes in a tree. It is observed that the optimal rate of growth of n is a function of $\ln(m)$. This is to be expected because the number of subsets one can get in a tree of depth n is 2^n. But, the optimal rate of growth of l is almost similar to that of K; it differs only in the rate. Following the same procedure as in the previous section, we get that the number $l^*(m)$ of leaf nodes should be proportional to

$$\left(\frac{(d-d_l)^2 m}{d^3 \ln^2(m)}\right)^{(d/(6d-2d_l))}$$

where $d_l = d_{(P_X, v^*)}(l)$, for the bound on GE to be optimal. This rate is faster than that of K in Vnets [refer (27)]. The difference is obvious because each element in $\mathcal{S}$ is assumed to be an intersection of $K-1$ half spaces whereas each subset in case of DT is assumed to be an intersection of n half spaces and the derived bounds are worst case bounds.

VII. Discussion

In this section, we make some comments on the nature of results obtained along with some possible extensions of the results.

The main result of this paper has been developed on a framework similar to Valiant's PAC framework for learning and along the lines of those given in [16]. However, the result is not entirely distribution free. The bound on AE is dependent on the underlying probability measure P in terms of the properties of a corresponding BDR, v^* *viz.*, P_X-coarse index function of v^*. Moreover, in this paper, we have started from an approximation scheme and found the space of problems that can be efficiently solved by the scheme as opposed to the other way round viz., from a class of problems to an optimal class of approximation schemes.

As previously mentioned, the bounds obtained using the PAC framework are known to be very loose. Therefore, no attempt has been made in this paper to find the exact numerical bounds. Rather, these bounds have been used to analyze the behavior of GE with respect to various parameters in terms of bounds on the growth of K with respect to m, etc.

In this paper, we have assumed the existence of an algorithm that finds $v_{K,m}$, which minimizes $\hat{R}_m[v]$, for a K and a training set D_m. We guess that the problem of finding $v_{K,m}$ is NP-hard since its counterpart in unsupervised learning, viz., clustering, is NP-hard and the present problem is very similar to clustering. (A formal verification of this conjecture needs to be done.) Recently, an algorithm based on evolutionary algorithms has been proposed and shown to asymptotically converge to $v_{K,m}$ with probability one [7]. This algorithm may converge to a suboptimal if run for a finite time, which is the case in reality. Nevertheless, it is empirically shown to converge to a better optimum than that obtained by LVQ3 algorithm [3].

The behavior of generalization error of the networks with sigmoidal units [11] and of the RBF networks [16] have been analyzed considering them as function approximation schemes. It appears that finding a bound on GE for these networks when used as pattern classifiers is difficult. First, the class of decision regions that can be obtained using the functions generated by these networks as discriminating functions could be very difficult to characterize. However, the conclusions arrived from the analysis presented in Section IV-A should, in general, be applicable to these networks too. It is interesting to investigate whether the techniques similar to the one given in Section IV-B

can be used in bounding the EE of these networks. The problem may arise when calculating the metric entropy of the space of mappings in the last stage of the network where the vectors of real numbers are mapped to class labels.

One has to address two interrelated issues in comparing two classification models. One issue is the average misclassification error and second, the time taken to arrive at a particular instance of the model. The first issue has been analyzed in case of Vnets and DTs in the present paper. The analysis could be extended to MLPs in which each perceptron uses a hard-limiting function; because the fundamental structure in the associated decision regions is a hyperplane. It is interesting if a similar analysis can be done on other types of MLPs and RBF networks that are popularly used in PC models. However, it appears to be difficult to access the computation needed to arrive at a globally optimal classifier with existing tools in computation complexity. If one succeeds in precisely finding the computation requirements of learning these networks, then these classification models can be compared purely on an analytical basis and the comparison then holds for a wide class of PC problems irrespective of the specific learning algorithm.

VIII. Summary

We have considered the classification model that discretizes the feature space into Voronoi regions and assigns samples in each region to a class label. We have called this model as "Vnets." The error between a rule and a BDR corresponding to the problem under consideration has been considered as the generalization error. GE has been split into AE and EE and these factors have been individually bounded, thus bounding GE. We have analyzed the approximating behavior of Vnets, in turn their efficiency in solving a problem from the bound on AE. We have shown that the efficiency of Vnet depends on the extent to which Bayes decision boundaries fill the feature space. The bound on EE has been derived using two techniques and the relative merits of these two techniques have been discussed. Putting together the bounds on AE and EE, we have shown that Vnets converge to the Bayes classifier with arbitrarily high probability provided the number of Vcenters grow slower than the square root of the number of training samples. All these results have been reworked for DTs. The optimal growth rate of the number of Voronoi centers for optimum GE has been calculated for Vnets as well as DTs. The main result of this paper also quantitatively explains the bias/variance dilemma and the curse of dimensionality that is generally observed in classification models.

Appendix I
Some Definitions and Results

The analysis of the process of learning a concept can be done by analyzing the uniform convergence of the associated empirical process as shown in Section IV-B for a class of concepts. The classes of concepts that are of interest in this paper are the class $\mathcal{A}$ of subsets of X, and the class $\mathcal{F}$ of real valued functions. In each of the above classes, the uniform convergence of the empirical process depends on the growth function associated with the concept class. The growth function in turn depends on a number (dimension) that is specific to a concept class. In this section, we define these dimensions for the above mentioned two concept classes and give the relation between this dimension and the property of uniform convergence for the class of real valued functions in a series of theorems and lemmas. Readers are referred to [25] and [31] for more details.

A. VC Dimension of Subsets of X

The following notation is used in this section. Let $(X, \mathcal{S})$ be a measurable space and let $\mathcal{C} \subseteq \mathcal{S}$ be a collection of subsets of X.

Definition A.1: A subset $K \subseteq X$ is said to be *shattered* by $\mathcal{C}$ if for every partition of K into disjoint subsets K_1 and K_2, there exists $C \in \mathcal{C}$ such that $K_1 \subseteq C$ and $K_2 \cap C = \emptyset$. Then the Vapnik–Chervonenkis (VC) dimension of $\mathcal{C}$, $\mathbf{dim}_{\mathrm{VC}}(\mathcal{C})$, is the maximum cardinality of any subset of X shattered by $\mathcal{C}$, or ∞ if arbitrary large subsets can be shattered. □

The following two lemmas give the VC dimension of the class of half spaces in $\Re^d$ and that of the class of subsets of X which are intersections of n elements belong to a concept class of a given VC dimension. These results are useful in deriving the bounds for Vnets.

Lemma A.1: ([17, Lemma 3.2.3]) Let $C \subseteq 2^X$ be a concept class of finite VC dimension, *viz*, $\mathbf{dim}_{\mathrm{VC}}(C) = d$, $1 \leq d < \infty$. For all $n \geq 1$, let $C_n = \{\bigcap_{i=1}^{n} c_i \colon c_i \in C, 1 \leq i \leq n\}$ or $C_n = \{\bigcup_{i=1}^{n} c_i \colon c_i \in C, 1 \leq i \leq n\}$. Then $\forall n \geq 1$, $\mathbf{dim}_{\mathrm{VC}}(C_n) \leq 2dn \log(3n)$. □

Lemma A.2: ([32]) Let $\mathcal{G}$ be the class of half spaces in $\Re^k$ then, $\mathbf{dim}_{\mathrm{VC}}(\mathcal{G}) = k + 1$. □

B. Learning the Class $\mathcal{F}$ of Real Valued Functions

Definition A.2: For $x \in \Re$ let, $\mathrm{sign}(x) = 1$ if $x > 0$ else $\mathrm{sign}(x) = 0$. For $\overline{x} = (x_1, \ldots, x_k) \in \Re^k$ let, $\mathrm{sign}(\overline{x}) = (\mathrm{sign}(x_1), \ldots, \mathrm{sign}(x_k))$, and for $T \subset \Re^k$ let, $\mathrm{sign}(T) = \{\mathrm{sign}(\overline{x}) \colon \overline{x} \in T\}$. Any set $T \subset \Re^k$ is said to be *full* if there exists $\overline{x} \in \Re^k$ such that $\mathrm{sign}(T + \overline{x}) = \{0, 1\}^k$, where $T + \overline{x} = \{\overline{y} + \overline{x} \colon \overline{y} \in T\}$.

Let $\mathcal{H}$ be a family of functions from a set X into $\Re$. Any sequence $\overline{x} = (x_1, \ldots, x_k)$ is said to be *shattered* by $\mathcal{H}$, if its $\overline{x}$-restriction, $\mathcal{H}|_{\overline{x}} = \{(h(x_1), \ldots, h(x_k)) \colon h \in \mathcal{H}\}$, is full. The largest k such that there exists a sequence of k points which are shattered by $\mathcal{H}$ is said to be the *pseudodimension* of $\mathcal{H}$ denoted by $\mathbf{dim}_P(\mathcal{H})$. If arbitrarily long finite sequences are shattered, then $\mathbf{dim}_P(\mathcal{H})$ is infinity. □

Definition A.3: Let A be a subset of a metric space Ω with metric d_M. Then a set $\Gamma \subset \Omega$ is called an *ε-cover* of A with respect to the metric d_M if for every $a \in A$, there exists some $t \in \Gamma$ satisfying $d_M(a, t) \leq \varepsilon$. The size of the smallest ε-cover is called *covering number* of A and is denoted by $\mathcal{N}(\varepsilon, A, d_M)$. A set $\Gamma \subset \Omega$ is called an *ε-separated* if for all $\alpha, \beta \in \Gamma$, $d_M(\alpha, \beta) > \varepsilon$. The *packing number* $\mathcal{M}(\varepsilon, A, d_M)$ of a set A is the size of the largest ε-separated subset of A. □

Let d_{a1} be a metric on $\Re^k$ defined as $\forall \alpha, \beta \in \Re^k$, $d_{a1}(\alpha, \beta) = 1/k \sum_{i=1}^{k} |\alpha_i - \beta_i|$ ($a1$ is used in the subscript to represent average l_1 metric as in [25]). Let $d_{a1,P}$ be a

pseudometric on a space $\mathcal{H}$ of functions from X to $\Re^k$, induced by a probability distribution P on X, defined as

$$d_{a1,P}(h_1, h_2) = \int_X d_{a1}(h_1(x), h_2(x))P(x)\,dx.$$

Theorem A.3: ([13]) Let $\mathcal{H}$ be a family of functions from a set X into $[0, M]$, where $\mathbf{dim}_P(\mathcal{H}) = d$ for some $1 \le d < \infty$. Let P be a probability distribution on X. Then for all $0 < \varepsilon \le M$

$$\mathcal{M}(\varepsilon, \mathcal{H}, d_{a1,P}) < 2\left(\frac{2eM}{\varepsilon}\ln\frac{2eM}{\varepsilon}\right)^d. \tag{31}$$

□

Theorem A.4: ([13], [25]) Let $\mathcal{H}$ be a family of functions from a set X into [0, 1]. Let D_m be a sequence of m training examples randomly drawn according to a distribution P. Then, for all $\varepsilon > 0$,

$$\Pr(\exists h \in \mathcal{H}: |\hat{E}_m[h] - E[h]| > \varepsilon) \le 4E\left[\mathcal{N}\left(\frac{\varepsilon}{8}, \mathcal{H}|_{D_m}, d_{a1}\right)\right]\exp\left(-\frac{\varepsilon^2 m}{32}\right). \tag{32}$$

Where $\mathcal{H}|_{D_m}$ is the restriction of $\mathcal{H}$ to the data set D_m, that is: $\mathcal{H}|_{D_m} \triangleq \{(h(z_1), \ldots, h(z_m)): h \in \mathcal{H}\}$. □

The original version of the above theorem is given in [13]; an improved version, with less conservative bounds, is given in [25]. $E[\mathcal{N}((\varepsilon/8), \mathcal{H}|_{D_m}, d_{a1})]$ depends on the probability distribution P, since D_m is a random set that depends on P. To get a bound that is independent of P, a quantity called metric capacity is defined and the error is bounded in terms of this.

Definition A.4: The *metric capacity* of $\mathcal{H}$ is defined as

$$\mathcal{C}(\varepsilon, \mathcal{H}, d_{a1}) \triangleq \sup_P\{\mathcal{N}(\varepsilon, \mathcal{H}, d_{a1,P})\} \tag{33}$$

where the supremum is taken over all the probability distributions defined on Z. □

Since the right-hand side of (31) is independent of P and $\mathcal{N}(\varepsilon, \mathcal{H}, d_{a1,P}) \le \mathcal{M}(\varepsilon, \mathcal{H}, d_{a1,P})$ for all P, we have, under the same hypothesis of Theorem A.3

$$\mathcal{C}(\varepsilon, \mathcal{H}, d_{a1}) < 2\left(\frac{2eM}{\varepsilon}\ln\frac{2eM}{\varepsilon}\right)^d. \tag{34}$$

We use the following theorem in Theorem IV.6.

Theorem A.5: ([13]) Let $\mathcal{G}_k = \{(g_1(x), \ldots, g_k(x)): g_i \in \mathcal{G}, 1 \le i \le k\}$, where $\mathcal{G}$ is a class of functions from X to $\Re$. Then $\mathcal{C}(\varepsilon, \mathcal{G}_k, d_{a1}) \le (\mathcal{C}(\varepsilon, \mathcal{G}, d_{a1}))^k$. □

APPENDIX II
PROOFS OF LEMMA IV.4 AND LEMMA IV.5

Proofs of Lemma IV.4: Given a sequence D_m in Z, sampled according to the distribution P, define P_{D_m} as the empirical distribution on Z, i.e., $P_{D_m}(z) = (1/m)\sum_{i=1}^m \delta(z - z_i)$, where δ is the Dirac delta function and z_i is the ith element of D_m. Let $a_1, a_2 \in \mathcal{A}$, then $d_{a1,P_{D_m}}(a_1, a_2) = d_{a1}((a_1(z_1), \ldots, a_1(z_l)), (a_2(z_1), \ldots, a_2(z_l)))$. Therefore, $\mathcal{N}(\varepsilon, \mathcal{A}_{D_m}, d_{a1}) = \mathcal{N}(\varepsilon, \mathcal{A}, d_{a1,P_{D_m}})$ because of the isometry. Hence the first inequality follows from (33) and by taking expectation on both sides with respect to P.

Fix a distribution P on Z. Let P_X, P_Y be the marginal distribution with respect to X and Y, respectively. Let K be an ε-cover for $\mathcal{V}_K$ with respect to the metric d_{Y,P_X} such that $|K| = \mathcal{N}(\varepsilon, \mathcal{V}_K, d_{Y,P_X})$. Then, it is easy to show that $H(K) = \{d_Y(f(x), y): f \in K\}$ is an ε-cover for $\mathcal{A}$ with respect to the metric $d_{a1,P}$. Then, $\mathcal{N}(\varepsilon, \mathcal{A}, d_{a^1,P}) \le |H(K)| = \mathcal{N}(\varepsilon, \mathcal{V}_K, d_{Y,P_X}) < \mathcal{C}(\varepsilon, \mathcal{V}_K, d_Y)$. Taking supremum over all P we get the second inequality in the lemma. □

Proof of Lemma IV.5: Fix a distribution P on $\Re^k$. Let K be $2\varepsilon/K$-cover for $\mathcal{H}_1$ with respect to the metric $d_{a1,P}$ such that $|K| = \mathcal{N}((2\varepsilon/K), \mathcal{H}_1, d_{a1,P})$. Then, $\mathcal{H}(K) = \{h_2 \circ h: h \in K, h_2 \in \mathcal{H}_2\}$ can be shown to be an εp-cover for $\mathcal{V}_K$ with respect to $d_{Y,P}$. Hence

$$\mathcal{N}(\varepsilon p, \mathcal{V}_K, d_{Y,P}) \le |\mathcal{H}(K)| = \sum_{f\in K}|\mathcal{H}_2| \le \sum_{f\in K} N^K \le N^K\mathcal{N}\left(\frac{2\varepsilon}{K}, \mathcal{H}_1, d_{a1,P}\right).$$

Since this holds for any P, taking supremum over all P, the lemma is proved. □

REFERENCES

[1] B. V. Dasarathy, Ed., *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, CA: IEEE Comput. Soc. Press, 1991.

[2] C.-L. Chang, "Finding prototypes for nearest neighbor classifiers," *IEEE Trans. Comput.*, vol. C-23, pp. 1179–1184, Nov. 1974.

[3] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, pp. 1464–1480, Sept. 1990.

[4] C. Decaestecker, "Finding prototypes for nearest neighbor classification by means of gradient descent and deterministic annealing," *Pattern Recognition*, vol. 30, no. 2, pp. 281–288, Feb. 1997.

[5] Q. Zhao, "Stable on-line evolutionary learning of NN-MLP," *IEEE Trans. Neural Networks*, vol. 8, pp. 1371–1378, Nov 1997.

[6] J. C. Bezdek, T. R. Reichherzer, G. S. Lim, and Y. Attikiouzel, "Multiple-prototype classifier design," *IEEE Trans. Syst., Man, Cybern. C*, vol. 28, pp. 67–79, Feb. 1998.

[7] K. Krishna, "Hybrid evolutionary algorithms for supervised and unsupervised learning," Ph.D. dissertation, Dept. Elect. Eng., Indian Inst. Sci., Bangalore, India, Aug. 1998.

[8] G. Voronoi, "Recherches sur les paralleloedres primitives," *J. Reine Angew. Math.*, vol. 134, pp. 198–287, 1908.

[9] N. R. Pal, J. C. Bezdek, and E. C. K. Tsao, "Generalized clustering networks and Kohonen's self-organizing scheme," *IEEE Trans. Neural Networks*, vol. 4, pp. 549–557, July 1993.

[10] G. Cybenko, "Approximation by superposition of sigmoidal functions," *Math. Contr., Syst, Signals*, vol. 2, no. 4, pp. 303–314, 1989.

[11] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," Dept. Statist., Univ. Illinois Urbana-Champaign, Champaign, IL, Tech. Rep. 58, Mar. 1991.

[12] E. Baum and D. Haussler, "What size net gives valid generalization?," *Neural Comput.*, vol. 1, no. 1, pp. 151–160, 1989.

[13] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. Comput.*, vol. 100, pp. 78–150, 1992.

[14] H. White, "Connectionist nonparametric regression: Multilayer perceptron can learn arbitrary mappings," *Neural Networks*, vol. 3, pp. 535–549, 1990.
[15] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, pp. 1–58, 1992.
[16] P. Niyogi and F. Girosi, "On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions," Artificial Intell. Lab., Massachusetts Inst. Technol., Cambridge, MA, A. I. Memo 1467, Feb. 1994.
[17] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik–Chervonenkis dimension," *J. Assoc. Comput. Machinery*, vol. 36, no. 4, pp. 926–965, Oct. 1989.
[18] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*. Berlin, Germany: Springer-Verlag, 1982.
[19] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probability and its Applications*, vol. 12, no. 2, pp. 264–280, 1971.
[20] L. G. Valiant, "A theory of the learnable," *Commun. Assoc. Comput. Machinery*, vol. 27, no. 11, pp. 1134–1143, 1984.
[21] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. M. Long, "Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions," *J. Comput. Syst. Sci.*, vol. 50, pp. 74–86, 1995.
[22] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
[23] J. R. Quinlan, "Decision trees and decision making," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, pp. 339–346, May 1968.
[24] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[25] M. Vidyasagar, *A Theory of Learning and Generalization With Applications to Neural Networks and Control Systems*. Berlin, Germany: Springer-Verlag, 1996.
[26] G. A. Edgar, *Measure, Topology and Fractal Geometry*. New York: Springer-Verlag, 1990.
[27] G. G. Lorentz, *Approximation of Functions*. New York: Chelsea, 1986.
[28] S. K. Murthy, S. Kasif, and S. Salzberg, "A system for induction of oblique decision trees," *J. Artificial Intell. Res.*, vol. 2, pp. 1–32, 1994.
[29] D. Haussler, "Quantifying inductive bias: AI learning algorithms and Valiant's learning framework," *Artificial Intell.*, vol. 36, pp. 177–221, 1988.
[30] A. Ehrenfeucht and D. Haussler, "Learning decision trees from random examples," *Inform. Comput.*, vol. 83, no. 3, pp. 247–251, 1989.
[31] B. K. Natarajan, *Machine Learning: A Theoretical Approach*. San Mateo, CA: Morgan Kaufmann, 1991.
[32] R. S. Wenocur and R. M. Dudley, "Some special Vapnik–Chervonenkis classes," *Discrete Math.*, vol. 33, pp. 313–318, 1981.

K. Krishna received the B.Sc. degree in physics from Nagarjuna University, India, in 1986. He received the M.E. and Ph.D. degrees in electrical engineering from the Indian Institute of Science, Bangalore, India, in 1993 and 1999, respectively.

He has been a Research Staff Member in the IBM India Research Laboratory, New Delhi, India, since December 1998. His present research interests include machine learning, statistical learning theory, evolutionary algorithms, and data and web mining.

Dr. Krishna received the Alfred Hay Medal for the best graduating student in electrical engineering in 1993 from the Indian Institute of Science.

M. A. L. Thathachar (SM'79–F'91) received the B.E. degree in electrical engineering from the University of Mysore, India, in 1959, the M.E. degree in power engineering, and the Ph.D. degree in control systems from the Indian Institute of Science, Bangalore, India, in 1961 and 1968, respectively.

He was a Member of the Faculty of Indian Institute of Technology, Madras, from 1961 to 1964. Since 1964, he has been with Indian Institute of Science, Bangalore, where currently he is a Professor in the Department of Electrical Engineering. He has been a Visiting Professor at Yale University, New Haven, CT, Michigan State University, East Lansing, Concordia University, Montreal, PQ, Canada, and the National University of Singapore. His current research interests include learning automata, neural networks, and fuzzy systems.

Dr. Thathachar is a Fellow of Indian National Science Academy, the Indian Academy of Sciences, and the Indian National Academy of Engineering.

K. R. Ramakrishnan received the B.S., M.S., and Ph.D. degrees from in the Department of Electrical Engineering, Indian Institute of Science, Bangalore, India.

He is currently an Associate Professor in the same department. His research interests include signal processing, computer vision, medical imaging, and multimedia.