DETECTING CLASS-INDEPENDENT LINEAR RELATIONSHIPS WITHIN AN ARBITRARY SET OF FEATURES

Ashwin Sarma

Naval Undersea Warfare Center, Newport, RI 02841

ABSTRACT

Classifiers for surveillance sonar systems are often designed to operate on large sets of predefined clues, or features. Sometimes the mathematical definitions for these features are poorly known. Other times the designer is not aware that a fixed and class-independent linear (or affine) relationship exists between subsets of features. We discuss a method based on Gram-Schmidt orthogonalization which allows the classifier designer to determine whether subsets of features have such relationships. Certain features can then be shown unnecessary by application of Wozencraft and Jacobs' "Theorem of Irrelevance". An approach is also described to rank features to aid in the selection of an effective subset.

I. INTRODUCTION

The design of classifiers for a surveillance sonar system is often based on a pre-existing set of measurements or "features" considered useful for discrimination. This set can be large and usually includes features that are simple functions of basic physical measurements of the objects to be classified. Features are often defined via linear functions of fundamental measurements or other more basic features. However, these functional relationships are sometimes unknown to the designer or difficult to sort out.

Denote as $\mathbf{Z}$ the raw measurements collected from an object to be classified. Features contained in the vector $\vec{f}(\mathbf{Z})$ are said to be *affinely dependent* on another set of features contained in the vector $\vec{g}(\mathbf{Z})$ if $\vec{f}(\mathbf{Z}) = \vec{b} + A\vec{g}(\mathbf{Z}), \forall\, \mathbf{Z}$. In other words there exists fixed vector $\vec{b}$ and matrix A such that $\vec{f}(\mathbf{Z})$ is given by the above relationship regardless of the object type (class). Examples include:

(1) If $\vec{g}(\mathbf{Z})$ consisted of a single feature such as an estimate of the angular width of an object and $\vec{f}(\mathbf{Z})$ was the cross-range extent measured by a sonar system corresponding to that object at a fixed (non- random) range R then $\vec{g}(\mathbf{Z}) = \mathbf{\Delta}\theta$ and $\vec{f}(\mathbf{Z}) = R\mathbf{\Delta}\theta$. Certain classes of objects may possess greater angular widths than others. However, feature $\vec{f}(\mathbf{Z})$ is completely defined by $\vec{g}(\mathbf{Z})$.

(2) If $\vec{g}(\mathbf{Z})$ and $\vec{f}(\mathbf{Z})$ are estimated probabilities of an object undergoing acceleration or maintaining constant velocity respectively at time t. Such estimates are provided by tracking algorithms commonly used in sonar. Note that $\vec{g}(\mathbf{Z})+\vec{f}(\mathbf{Z}) = 1$. It may be that certain classes of objects perform more maneuvers than others and will exhibit this in terms of higher values of $\vec{g}(\mathbf{Z})$ and lower values of $\vec{f}(\mathbf{Z})$. However $\vec{f}(\mathbf{Z})$ is completely defined by $\vec{g}(\mathbf{Z})$.

This work was supported by the Office of Naval Research (Award Number: N0001404WX20326) and a NUWC in-house research grant.

In such cases the dependence may have been known to the classifier designer, but in real cases with large feature sets the dependence may be unknown or difficult to determine. Furthermore, in real problems, more complicated examples of affine dependence can and do arise in obscure and inadvertent ways.

We are interested in detecting if any such affine relationships are present among features. In this short paper we describe a nonparametric method for detecting affinely dependent features that is based on the linear algebraic structure of the feature space and is independent of the underlying distributions as well as separability among classes. Removal of such dependent features is argued through application of Wozencraft and Jacobs' "Theorem of Irrelevance" [1]. Removal of such features is critical as they add no additional information and unnecessarily increase the dimensionality of the feature space.

"Approximate" affine dependence can also be identified and a ranking of the features is possible. The pre-processing advocated herein is a simple but often overlooked first step in the design of sonar classifiers. The method has proved useful for detecting and removing affinely dependent features present in Navy feature databases.

II. "THEOREM OF IRRELEVANCE"

Denote the class-conditional pdf for class i as $p_{\mathbf{r}}(\rho|i)$ where $\mathbf{r}$ is the random vector of features and ρ is a particular realization of $\mathbf{r}$. Then decompose $\mathbf{r}$ as $\mathbf{r} = [\mathbf{r_1}\ \mathbf{r_2}]^{\mathbf{T}}$. Denote $\rho_{\mathbf{1}}$ and $\rho_{\mathbf{2}}$ as the corresponding realizations of $\mathbf{r_1}$ and $\mathbf{r_2}$. We observe that $p_{\mathbf{r}}(\rho|i) = p_{\mathbf{r_1}}(\rho_{\mathbf{1}}|i)\ p_{\mathbf{r_2}}(\rho_{\mathbf{2}}|i, \mathbf{r_1} = \rho_{\mathbf{1}})$ by Bayes rule. Wozencraft and Jacobs' "Theorem of Irrelevance" states that the optimum classifier/dichotomizer may disregard a vector $\mathbf{r_2}$ if and only if

$$p_{\mathbf{r_2}}(\rho_{\mathbf{2}}|i, \mathbf{r_1} = \rho_{\mathbf{1}}) = p_{\mathbf{r_2}}(\rho_{\mathbf{2}}|\mathbf{r_1} = \rho_{\mathbf{1}}) \qquad (1)$$

In other words $\mathbf{r_2}$ conditioned on $\mathbf{r_1}$ must be statistically independent of i for $\mathbf{r_2}$ to be declared "irrelevant". Another

useful interpretation is that satisfaction of this requirement implies that $\mathbf{r_1}$ is a sufficient statistic for estimating the pdfs $p_{\mathbf{r}}(\rho|i)$, $\forall\, i$, although not necessarily minimal [2], [3], [4].

III. DETECTING AFFINELY DEPENDENT FEATURES

The approach involves construction of a data matrix $\mathbf{X}$ in which *all* available data are included. If C classes are to be distinguished and N_i (1 x D) real-valued vector measurements (i.e. feature vectors) are available for $i = 1, ..., C$, $\mathbf{X}$ is a N x D data matrix where each row is a feature vector and $N = \sum_{i=1}^{C} N_i$. It is assumed that $N > D$, a reasonable assumption as there are commonly more measurements than features. An unsupervised situation in which N unlabeled feature vectors are presented can also be considered.

Denote the D features by $f_1, f_2, \ldots, f_D$. We are interested in affine dependence of the form:

$$f_k = \sum_{j=1, j\neq k}^{D} \alpha_j^k f_j + b_k \tag{2}$$

where b_k need not equal zero. Note that the linear algebraic concept of *linear dependence* among a set of vectors does not accommodate affine dependence. Thus a transformation such as standardization [5] must be performed before standard linear algebraic techniques can be applied. This amounts to estimating the sample mean $\hat{\mu}$ and sample standard deviation $\hat{\sigma}$ of each column of $\mathbf{X}$ and transforming every element x in that column according to $x = (x - \hat{\mu})/\hat{\sigma}$. Denote $\mathbf{X}$ after standardization as $\tilde{\mathbf{X}}$. The transformed features are denoted $\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_D$. Any affine dependence ($b_k \neq 0$ in eqn. (2)) is eliminated, i.e.

$$\tilde{f}_k = \sum_{j=1, j\neq k}^{D} \tilde{\alpha}_j^k \tilde{f}_j \tag{3}$$

(see appendix). It is clear that the elements of $\tilde{\mathbf{X}}$ will be approximately zero mean with unit variance. Thus standardization has the added effect of controlling the dynamic range of the elements of $\mathbf{X}$ and numerically stabilizing Gram-Schmidt orthogonalization.

Gram-Schmidt orthogonalization can be performed on the columns of $\tilde{\mathbf{X}}$ via QR matrix decomposition, i.e $\tilde{\mathbf{X}} = \mathbf{QR}$. The linearly dependent column vectors in $\tilde{\mathbf{X}}$ are marked by zeros along the diagonal of $\mathbf{R}$.

If R_{kk} is the leftmost zero element along the diagonal of $\mathbf{R}$ then $\tilde{\mathbf{x}}_\mathbf{k}$ is linearly dependent on the vectors $\tilde{\mathbf{x}}_\mathbf{1}, \tilde{\mathbf{x}}_\mathbf{2}, \ldots, \tilde{\mathbf{x}}_\mathbf{k-1}$ and there exists a set of coefficients, $\tilde{\alpha}_1^k, \tilde{\alpha}_2^k, \ldots, \tilde{\alpha}_{k-1}^k$ such that $\tilde{\mathbf{x}}_\mathbf{k}$ can be expressed as:

$$\tilde{\mathbf{x}}_\mathbf{k} = \tilde{\alpha}_1^k \tilde{\mathbf{x}}_\mathbf{1} + \tilde{\alpha}_2^k \tilde{\mathbf{x}}_\mathbf{2} + \ldots + \tilde{\alpha}_{k-1}^k \tilde{\mathbf{x}}_\mathbf{k-1} \tag{4}$$

Note that at least one (but not all) $\tilde{\alpha}^k$'s need be nonzero. Since this dependence is enforced over all C classes and for all N measurements, the probability that such a dependence manifested due to chance is effectively zero. It is worth mentioning that the dependence is not linked to the manner in which $\mathbf{X}$ was populated. Specifically a permutation of the rows of $\mathbf{X}$ (equivalently $\tilde{\mathbf{X}}$) will not alter the linear dependence. This can be seen by noting that pre-multiplying $\tilde{\mathbf{x}}_\mathbf{k}$ in (4) by a permutation matrix results in

$$\tilde{\mathbf{u}}_\mathbf{k} = \tilde{\alpha}_1^k \tilde{\mathbf{u}}_\mathbf{1} + \tilde{\alpha}_2^k \tilde{\mathbf{u}}_\mathbf{2} + \ldots + \tilde{\alpha}_{k-1}^k \tilde{\mathbf{u}}_\mathbf{k-1} \tag{5}$$

where $\tilde{\mathbf{u}}_\mathbf{j}$, $j = 1, ..., D$ is the permuted $\tilde{\mathbf{x}}_\mathbf{j}$.

Thus feature f_k must be expressible as

$$f_k = \alpha_1^k f_1 + \alpha_2^k f_2 + \ldots + \alpha_{k-1}^k f_{k-1} + b_k \tag{6}$$

or

$$f_k = \sum_{j=1}^{D} \alpha_j^k f_j + b_k \tag{7}$$

where $\alpha_j^k, j \geq k$ must equal 0. If R_{mm} ($k < m \leq D$) is the next zero element along the diagonal of $\mathbf{R}$ then $\tilde{\mathbf{x}}_\mathbf{m}$ is linearly dependent on the vectors $\tilde{\mathbf{x}}_\mathbf{1}, \tilde{\mathbf{x}}_\mathbf{2}, \ldots, \tilde{\mathbf{x}}_\mathbf{m-1}$ and there exists another set of coefficients, $\tilde{\alpha}_1^m, \tilde{\alpha}_2^m, \ldots, \tilde{\alpha}_{m-1}^m$ such that $\tilde{\mathbf{x}}_\mathbf{m}$ can be expressed as:

$$\tilde{\mathbf{x}}_\mathbf{m} = \tilde{\alpha}_1^m \tilde{\mathbf{x}}_\mathbf{1} + \tilde{\alpha}_2^m \tilde{\mathbf{x}}_\mathbf{2} + \ldots + \tilde{\alpha}_k^m \tilde{\mathbf{x}}_\mathbf{k} + \ldots + \tilde{\alpha}_{m-1}^m \tilde{\mathbf{x}}_\mathbf{m-1} \tag{8}$$

Once again not all $\tilde{\alpha}^m$'s need be nonzero. Furthermore since $\tilde{\mathbf{x}}_\mathbf{k}$ is linearly dependent on $\tilde{\mathbf{x}}_\mathbf{1}, \tilde{\mathbf{x}}_\mathbf{2}, \ldots, \tilde{\mathbf{x}}_\mathbf{k-1}$, $\tilde{\alpha}_k^m$ can be set equal to zero. Thus f_m can be expressed as

$$f_m = \sum_{j=1}^{D} \alpha_j^m f_j + b_m \tag{9}$$

where α_j^m must equal zero for $j = k$ as well as for $j \geq m$.

Therefore the set of features $\mathbf{r_2}$ for which $R_{jj} = 0$ are affinely dependent on the remainder of the features denoted $\mathbf{r_1}$. From eqns. (7) and (9), we can see that $\rho_\mathbf{2}$ is completely determined by $\rho_\mathbf{1}$ independent of class i. Thus eqn. 1 is satisfied and $\mathbf{r_2}$ is irrelevant. It is worth noting that if $N = D$ and $\mathbf{X}$ is full rank then $\tilde{\mathbf{X}}$ has a rank of $D - 1$. Thus this process can be successfully applied only if N is strictly greater than D.

Even if an exact affine dependence existed, errors in machine calculation of $\mathbf{R}$ can lead to non-zero R_{jj}. Higham [6, p. 24,122] has shown that performing the QR decomposition via a sequence of Givens rotations leads to an ultimate relative error that is acceptable and on the level of machine precision u. Here relative error, defined as $\frac{||\mathbf{R}-\hat{\mathbf{R}}||_\mathbf{2}}{||\tilde{\mathbf{X}}||_\mathbf{2}}$ ($= \frac{||\mathbf{R}-\hat{\mathbf{R}}||_\mathbf{F}}{||\tilde{\mathbf{X}}||_\mathbf{F}}$) is the normalized difference between the exact $\mathbf{R}$ and $\hat{\mathbf{R}}$, estimated by the actual decomposition. Note that

$$\frac{||\mathbf{R} - \hat{\mathbf{R}}||_\mathbf{F}}{||\tilde{\mathbf{X}}||_\mathbf{F}} \approx \frac{D\epsilon}{\sqrt{ND}} = \sqrt{\frac{D}{N}}\epsilon = u \tag{10}$$

where ϵ is the average value of $|R_{ij} - \hat{R}_{ij}|$ $(i, j = 1, ..., D)$. Thus if $N \approx 100D$, R_{jj} can be compared to a threshold of ten times machine precision to detect linear dependence.

If certain features are preferred (ex. considered more intuitive or powerful) by the designer, reordering can be performed such that these populate the leftmost columns of $\mathbf{X}$. This step increases the likelihood that such a feature will be retained in the event that it is affinely dependent on other features.

IV. "APPROXIMATELY" AFFINELY DEPENDENT FEATURES

At this point let us assume that features that are exactly affinely dependent have been detected and removed. Operating on the resultant set, the process of Gram-Schmidt orthogonalization of $\tilde{\mathbf{x}}_\mathbf{k}$ amounts to

(1) Determining the least-squares fit of $\tilde{\mathbf{x}}_\mathbf{k}$ in the subspace spanned by $\{\tilde{\mathbf{x}}_\mathbf{1}, ..., \tilde{\mathbf{x}}_{\mathbf{k-1}}\}$. This is equivalent to determining the projection of $\tilde{\mathbf{x}}_\mathbf{k}$ onto this subspace. We denote this projection as $\mathbf{P}\tilde{\mathbf{x}}_\mathbf{k}$ where $\mathbf{P}$ is the projection matrix.

(2) Subtracting the result from $\tilde{\mathbf{x}}_\mathbf{k}$ to form the orthogonal error vector $\mathbf{e}$. Thus $\tilde{\mathbf{x}}_\mathbf{k} = \mathbf{P}\tilde{\mathbf{x}}_\mathbf{k} + \mathbf{e}$. The norm of the error vector $\mathbf{e}$, $||\mathbf{e}||$, is minimized in this process [3, p. 365] but is never zero.

Again we observe that permuting the rows of $\mathbf{X}$ (equivalently $\tilde{\mathbf{X}}$) only permutes the elements of $\mathbf{e}$ but leaves $||\mathbf{e}||$ unchanged.

The relationship of Gram-Schmidt orthogonalization to least-squares estimation is important in that, for $N > D$, we can interpret feature f_k as a sum of A) a fixed (and class-independent) linear function of the previous features and B) a residual. It is possible that some (or all) of the previous features may be useful for classification. It is also possible that the least-squares estimate of f_k in terms of the previous features is useful as well. However this information is implicit in the previous features. Thus it is sufficient to determine the added information captured in the residual. This amounts to a statistical test[1] comparing the scalar values in $\mathbf{e}$ corresponding to class i with those corresponding to class j. Nonparametric tests such as a Chi-squared test (or Kolmogorov-Smirnov test) can be used to compare two samples. Specifically the p-value of the test statistic can be returned. The p-value is the probability of the event that a value of the test statistic greater than or equal to that observed occurs when both samples have a common probability distribution. Standard critical levels of significance are 0.05. Thus if the p-value is less than the critical level we can safely conclude that the distribution governing the two samples is different. A difference between the distributions indicates that the feature may prove useful for classification. It must be stressed that if there are only a few measurements per class, a large p-value need not indicate distributional similarity. However if there are enough measurements such that p-values lower than the critical level are at least possible then if the p-value is greater than the critical level we may conclude f_k is "approximately" affinely dependent on the previous features. Specifically eqn. (1) with $\mathbf{r_2} = f_k$ and $\mathbf{r_1} = [f_1, f_2, \ldots, f_{k-1}]^T$ is satisfied at the reported p-value. Graphical methods such as quantile-quantile plots can be used to corroborate the conclusions.

[1] or set of pairwise tests when more than two classes are considered

The process of feature subset selection involves considering various subsets and, via a pre-chosen separability measure (a.k.a. Filter Method) or a classifier architecture (a.k.a. Wrapper Method), ranking the quality of each subset [7], [8]. As testing every possible subset is usually prohibitive, alternatives that consider only specific subsets are almost always applied. One such approach is as follows. Features can be ranked according to p-value. If it is required that a feature must be discarded, the feature with the *largest* p-value can be selected. The entire process of Gram-Schmidt orthogonalization followed by statistical testing is then repeated if further reduction is required.

V. SUMMARY

Methodologies are provided to detect exact and "approximate" affine relationships within a set of features. They provide insight into feature structure and aid in feature subset selection.

VI. APPENDIX

Assume the affine dependence of eqn. (11) for feature f_k.

$$f_k = \sum_{j=1, j\neq k}^{D} \alpha_j^k f_j + b_k \tag{11}$$

This dependence must be enforced regardless of class. Thus

$$\mathbf{x_k} = \sum_{j=1, j\neq k}^{D} \alpha_j^k \mathbf{x_j} + b_k \mathbf{1_N} \tag{12}$$

where $\mathbf{x_j}$ is the j^{th} column of $\mathbf{X}$ ($j = 1, .., D$) and $\mathbf{1_N}$ is the vector $\underbrace{[1\, 1\, \ldots\, 1]}_{N}^T$.

Denote $\mathbf{x_j}^\mathbf{T}\mathbf{1_N}$, the sum of elements in column j, as S_j. Denote the columns of $\mathbf{X}$ after standardization as $\tilde{\mathbf{x}}_\mathbf{j}$. Thus $\tilde{\mathbf{x}}_\mathbf{j}$ equals

$$\tilde{\mathbf{x}}_\mathbf{j} = \frac{\mathbf{x_j} - \hat{\mu}_\mathbf{j}\mathbf{1_N}}{\hat{\sigma}_\mathbf{j}} \tag{13}$$

where $\hat{\mu}_j = \frac{S_j}{N}$ and $\hat{\sigma}_j^2 = \frac{(\mathbf{x_j} - \hat{\mu}_j\mathbf{1_N})^T(\mathbf{x_j} - \hat{\mu}_j\mathbf{1_N})}{N}$. Rewriting eqn. (12) in terms of $\tilde{\mathbf{x}}_\mathbf{j}$ results in:

$$\hat{\sigma}_k\tilde{\mathbf{x}}_\mathbf{k} + \hat{\mu}_k\mathbf{1_N} = \sum_{j=1, j\neq k}^{D} \alpha_j^k(\hat{\sigma}_j\tilde{\mathbf{x}}_\mathbf{j} + \hat{\mu}_j\mathbf{1_N}) + b_k\mathbf{1_N} \tag{14}$$

Subtracting $\hat{\mu}_k \mathbf{1_N}$ from both sides of (14) yields:

$$\hat{\sigma}_k \tilde{\mathbf{x}}_\mathbf{k} = \sum_{j=1, j\neq k}^{D} \alpha_j^k \hat{\sigma}_j \tilde{\mathbf{x}}_\mathbf{j} + \underbrace{\left[\sum_{j=1, j\neq k}^{D} \alpha_j^k \hat{\mu}_j \mathbf{1_N} + b_k \mathbf{1_N} - \hat{\mu}_\mathbf{k} \mathbf{1_N} \right]}_{\Gamma}$$

It is clear that the vector $\sum_{j=1, j\neq k}^{D} \alpha_j^k \hat{\mu}_j \mathbf{1_N} + b_k \mathbf{1_N}$ and $\hat{\mu}_k \mathbf{1_N}$ can be represented as $c\mathbf{1_N}$ and $d\mathbf{1_N}$ respectively and that that $\Gamma = (c-d)\mathbf{1_N}$. Pre-multiplying both Γ and eqn. (12) by $\mathbf{1_N}^\mathbf{T}$ reveals that $c = d$. Thus

$$\tilde{\mathbf{x}}_\mathbf{k} = \sum_{j=1, j\neq k}^{D} \tilde{\alpha}_j^k \tilde{\mathbf{x}}_\mathbf{j} \tag{15}$$

where $\tilde{\alpha}_j^k$ equals $\alpha_j^k \hat{\sigma}_j / \hat{\sigma}_k$.

VII. REFERENCES

[1] John M. Wozencraft and Irwin M. Jacobs, *Principles of Communication Engineering*, John Wiley and Sons, New York, 1965.

[2] Thomas S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*, Academic Press, New York, 1967.

[3] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*, Addison-Wesley: Reading Massachusetts, 1991.

[4] Edward C. Real, "Feature extraction and sufficient statistics in detection and classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.

[5] R.O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, John Wiley and Sons: New York, second edition, 2001.

[6] Nicholas J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, 1996.

[7] Isabelle Guyon and Andre Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research 3*, pp. 1157–1182, 2003.

[8] K.Z. Mao, "Orthogonal forward selection and backward elimination algorithms for feature subset selection," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 34, no. 1, pp. 629–634, February 2004.