Current Development in Theory and Applications of Wavelets
Volume 4, Number 1, 2010, Pages 39-62

ENTROPY MEASURE AND ENERGY MAP USING DICTIONARY PROJECTION PURSUIT IN PATTERN RECOGNITION

R. TAFRESHI[1], F. SASSANI[2], H. AHMADI[3a] and G. DUMONT[3]

[1]Department of Mechanical Engineering
Texas A&M University at Qatar
Education City, P. O. Box 23874, Doha, Qatar
e-mail: rtafreshi@tamu.edu

[2]Department of Mechanical Engineering
University of British Columbia
Vancouver, BC, V6T 1Z4, Canada
e-mail: sassani@mech.ubc.ca

[3]Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, BC, V6T 1Z2, Vancouver, BC, Canada
e-mail: noubari@ece.ubc.ca; guyd@ece.ubc.ca

[a]Department of Electrical and Computer Engineering
University of Tehran, Iran

Abstract

Using the relative entropy criterion, we present a wavelet-based feature extraction algorithm for classification and extend its applications to the projection pursuit technique. The proposed method is applicable to a

---

2010 Mathematics Subject Classification: 42C40.

Keywords and phrases: entropy measure, pattern recognition, wavelet packet.

Communicated by Todar Ganchev

Received March 1, 2010

broad range of pattern recognition problems. In this paper, we apply it to the analysis of cylinder-head vibration data for engine fault detection and diagnosis. The usefulness of the singular value decomposition method in the study and the assessment of classification results is illustrated by analyzing the coefficient matrix, constructed from projecting the data onto the selected bases in determining the extent of correlation among coefficients.

1. Introduction

Condition monitoring for fault detection and prevention is now an integral part of the operation of many industrial processes and machinery [18]. It plays an important role in maintaining quality standards and safety, increasing productivity, and reducing cost. Unlike traditional reactive maintenance practices, condition monitoring and predictive maintenance attempt to prevent unnecessary plant shutdowns and thus reduce machine downtime and improve reliability.

Two distinct approaches have been used to diagnose faults in machinery such as internal combustion engines. In the first approach, a mathematical model of the specific engine or its component under investigation is developed and based on the measurements made of system using the model a search for causes of change in engine performance is conducted. In the second approach, the given engine or a set of its components is considered as a black box. Then, through observations and processing of appropriate sensory data, such as cylinder pressure, cylinder block vibration, exhaust gas temperature, and acoustic emission fault(s) are traced and detected. For a large number of engines in operation in vehicles and industrial settings built-in technologies are not available, therefore external sensing is inevitable. This paper focuses on the latter approach in which vibration data is used for the detection of selected malfunctions in reciprocating internal combustion engines.

The objective here is to develop effective data-driven methodologies for fault detection and diagnosis. The main application is the detection and characterization of combustion-related faults in reciprocating engines, such as knocks, improper ignition timing, loose intake and exhaust valves, and improper valve clearances.

The use of vibration data has become popular in a wide range of fault diagnosis applications, including the detection of knock [3, 20, 23, 30], valve clearance, and gas leakage in both intake and exhaust valves [9], as well as the detection of drift in ignition timings [25, 26]. In [27] we introduced a novel wavelet-based methodology in feature extraction for selecting best basis from a library of orthogonal bases that best discriminates one class of data from others. Then, using this method we modified a discriminatory algorithm referred to as *local discriminant basis* (LDB) [22]. This paper presents another variation of this method used in classification problems referred to as *dictionary projection pursuit* (DPP). The paper also compares the application results of LDB and DPP methods.

The structure of the paper is as follows. After a brief description of pattern recognition and classification schemes we review relative entropy as a useful discriminant measure. Then, a particular multivariate analysis scheme and a well-known technique known as the projection pursuit (PP) are described, followed by the description of DPP as a variation of PP. Description of a normalization scheme to improve DPP is followed by the introduction of cross-data entropy (CDP) approach. Using this approach, a modified form of DPP, referred to as mutual DPP, is used to analyze the vibration data. The classification results of the mutual DPP are compared with the original DPP, then further analyzed using the singular value decomposition approach.

2. Pattern Recognition

For pattern recognition and classification applications, it is highly desirable to identify as few features as possible while extracting as much information from the signal as possible. There is a pressing need to reduce the dimensionality of raw data by extracting a limited number of features that best preserve the useful information. There are many reasons why feature reduction is essential, including the reduction of computational cost, noise reduction, increased robustness, and the more rapid training of classifiers. There may also be a high mutual correlation among the selected features, which can increase computational complexity without any gain in the accuracy of the classification. Furthermore, such a correlation dilutes the information, which is detrimental to the classification results.

In classification problems, not only do we look for features that contain non-superfluous information but we also seek information that can separate classes from each other as distinctly as possible. This type of information is referred to as "discriminantal information". One can justifiably state that it is the superfluous information that introduces complexity in classification tasks. The main objective in feature extraction and classification problems is to find a suitable transformation or a coordinate system that by projecting the signal onto the coordinate directions, we obtain high discriminatory information that reside on a few axes while other axes contain insignificant information.

A linear projection from R^n to R^m is a linear map *B* represented as an $n \times m$ matrix:

$$Z = B^T X, \quad X \in R^{n \times l}, \quad Z \in R^{m \times l} \tag{1}$$

which transforms the *n*-dimensional data set *X* (consisting of *l* data in each column) into an *m*-dimensional space; *Z* is the *m*-dimensional transformed data set. Suppose b_i is the *n*-dimensional column vectors of matrix $B = [b_1\ b_2 \cdots b_m]$. If b_i 's are orthogonal to each other, the projection is orthogonal, and orthonormal if the vector norms are unity. When $m = 1,$ then *B* is a one-dimensional projection, and *Z* is a scalar, referred to as the *projection score*.

Figure 1 shows the main stages of classification in which *X* is the input signal, *Y*, its corresponding class label (e.g., *faulty* or *healthy* condition), and *F*, feature space, which is the discriminant subspace of the reduced dimension $(m < n)$. The maps $f : X \to F$ and $g : F \to Y$ are called *feature extractor* and *classifier*, respectively. It is computationally more efficient to analyze the data in a discriminant subspace of the lower dimension. The classification goal is to determine which class a given data *X* belongs to by constructing a feature space *F* that provides the highest discriminant information among all classes.

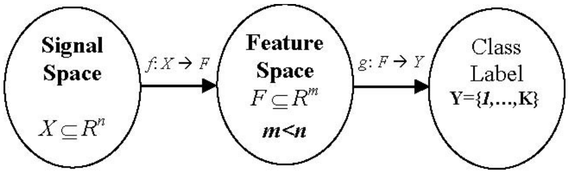

Figure 1. Main stages in classification.

2.1. Analysis of vibration data using wavelets

The vibration signals of a machine always carry information about the state of machine operation and the dynamic behavior of various machine events such as combustion. They can be used to identify faults in machine operation. Vibration signals in internal combustion engines are characterized by transient time behavior and are often corrupted by significant noise content. Wavelets are considered to be highly suitable for the analysis of transient signals to extract features that are used in fault detection problems.

In the next section a measure of distance between two or more classes of data referred to as *discriminant measure* is introduced.

3. Discriminant Measure

The principal objective in a classification problem is to develop measures that are capable of discriminating between different classes as distinctly as possible. The accuracy of the classification results is highly influenced by the extent of class separation in feature space generated by the chosen discriminating measure. A discriminant measure, in general, is designed to evaluate the statistical distance between different classes. The choice of discriminant measure depends on the application at hand. Different researchers have used different discriminant measures in various applications [16, 22, 28]. The approach used in this work is based on *relative entropy* as a measure for discriminating different classes as defined below.

In a two-class case, suppose that $\mathbf{s}^{(l)} = \{s_i^{(l)}\}_{i=1}^{n}$ for $l = 1, 2$ are two non-negative sequences satisfying:

$$\sum_i s_i^{(1)} = \sum_i s_i^{(2)} = 1. \tag{2}$$

The *symmetric relative entropy* for the two classes is then defined as:

$$D(\mathbf{s}^{(1)}, \mathbf{s}^{(2)}) = \sum_{i=1}^{n} \left(s_i^{(1)} \log \frac{s_i^{(1)}}{s_i^{(2)}} + s_i^{(2)} \log \frac{s_i^{(2)}}{s_i^{(1)}} \right) \tag{3}$$

assume that

$$\log 0 = -\infty,\ \log(s_i/0) = +\infty \ \text{ for } s_i > 0, \text{ and } 0.(\pm\infty) = 0. \tag{4}$$

Using only the first term in the right-hand side of (3) the *relative entropy* is then defined as:

$$D(\mathbf{s}^{(1)}, \mathbf{s}^{(2)}) = \sum_{i=1}^{n} s_i^{(1)} \log \frac{s_i^{(1)}}{s_i^{(2)}}. \tag{5}$$

Lemma. *Equation* (5) *is always non-negative and will be zero if the sequences* $\mathbf{s}^{(1)}$ *and* $\mathbf{s}^{(2)}$ *are identical.*

Proof [7]**.** Recalling the elementary inequality for real numbers

$$\log x \le x - 1, \tag{6}$$

with equality if and only if $x = 1$. Then, considering conditions (2) we have

$$\sum_i s_i^{(1)} \log \frac{s_i^{(2)}}{s_i^{(1)}} \le \sum_i s_i^{(1)} \left(\frac{s_i^{(2)}}{s_i^{(1)}} - 1 \right) = \sum_i s_i^{(2)} - \sum_i s_i^{(1)} = 0. \tag{7}$$

Therefore, $\sum_i s_i^{(1)} \log \frac{s_i^{(1)}}{s_i^{(2)}} \ge 0$. The equality holds if and only if $s_i^{(2)} = s_i^{(1)}$, for all i.

Note. Conditions (2) need not be satisfied for a symmetric relative entropy in order to have a non-negative D because the right-hand side of Equation (7) in a symmetric case results in the cancellation of the terms:

$$\left(\sum_i s_i^{(1)} - \sum_i s_i^{(2)} \right) + \left(\sum_i s_i^{(2)} - \sum_i s_i^{(1)} \right) = 0. \tag{8}$$

As can be seen from the above lemma, if two random variables have the same distributions, discriminant measure D will be zero. In classification applications, we are interested in those features that can separate the distribution associated with each class; therefore, one should look for those features that maximize D.

The aim is to find appropriate features that provide discriminant classification results. By projecting a set of data belonging to different classes (vibration signals in our application) onto a set of basis (coordinates) and using squared value of each coefficient as density values $\mathbf{s}^{(l)}$, in different classes, discriminant measure D is derived. In a classification problem, the objective is to identify those bases (from a dictionary of bases) that maximize D. In this manner, the distribution of each class can be transformed into a disjoint distribution with the least overlap with other distributions.

In general, for a classification problem having L classes $(\text{usually } L \le 4)$, one can use a simple approach:

$$D(\{\mathbf{s}^{(l)}\}_{l=1}^{L}) \equiv \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} D(\mathbf{s}^{(i)}, \mathbf{s}^{(j)}). \tag{9}$$

For $L > 4$ the method introduced in [29] is often utilized.

4. Multivariate Analysis

To identify the underlying, though unspecified, structure of a given data set, often a visual representation such as a histogram or a scatterplot is utilized [8]. This can be easily done for low (one, two, or even three) dimensional data; however comprehensive visual tools for higher dimensions are not available.

Classical multivariate analysis provides a powerful tool for gaining insight into the underlying features of the phenomenon or the system that produced the data. These tools include a set of useful summary statistics (such as mean and covariance) as well as correlational structure of the data. In the following section, projection pursuit algorithm used for revealing “interesting” structure of data is introduced after which an extension of PP as a fast version of PP is presented.

5. Projection Pursuit Approach

The projection pursuit (PP) is a method for exploratory analysis of multivariate data sets, which extracts linear projections of data to view them in a lower dimension; often onto a plane or a line. Selection of projection directions is done in a manner that a certain criterion function or *projection index is maximized*. Friedman and Tukey [4] first used the term “projection pursuit”, but the main idea was initially introduced by Kruskal [13]. Projection pursuit seeks a set of projections that are “interesting”, in the sense of their deviation from Gaussian distribution [10]. PP is basically a method for revealing clusters among data.

There are several projection indices, among them, Friedman [5] proposed an index which is the mean-squared difference between the projection score distribution and the Gaussian distribution. This is referred to as the least structured density and measures non-normality in the main body of the distribution (rather than in its entirety). Friedman’s projection index basically measures the departure from normality. Jones and Sibson [12] and Huber [10] set the PP idea in a more structured form and expanded it in a practical implementation. Their approach involves an optimization process that starts at different random positions using the entropy concept from the information theory as the projection index to maximize the divergence of projected data from Gaussian distribution.

By employing a suitable projection index, PP technique can reveal an inherent structure or clusters in a given data. This was then used in a supervised [1, 14, 15], and unsupervised classification scheme of high dimensional data [2, 11], in detecting and classifying images [19] and in feature extraction of acoustic spectra [21].

The use of PP has been limited due to its high computational complexity. Dictionary projection pursuit, described next, attempts to resolve this shortcoming by employing wavelet packet decomposition during the search process of PP as an extension of PP resulting in a computationally more efficient algorithm.

6. Dictionary Projection Pursuit

Rutledge [21] proposed a method that searches for a set of basis functions from a dictionary of redundant wavelet packets in accordance with an orthogonality criterion, in contrast to optimizing a criterion that is done in standard PP. The search is performed in m sets of iterations, where m is the required number of bases, and is decided upon empirically. In each iteration, they use a one-dimensional version of the projection pursuit method (which means $m = 1$ in Equation 1) to find the interesting features of acoustic waveforms. If A is a matrix consisting of all bases of a dictionary such as wavelet packets (Figure 2), then the first base is chosen from dictionary A according to a criterion described below. The dictionary is sometimes called *redundant*, since there are more than one set of basis functions which can span n-dimensional space. The one-dimensional version of projection pursuit is repeated m times where a procedure such as the one in matching pursuit [17] is applied until a set of bases $B = [b_1 \; b_2 \cdots b_m]$ is selected. Then, the data are projected onto the selected basis functions to find an ‘interesting view’ of the data. This is done by the linear projection $Z = B^T X$, where X is the data set and Z is the transformed data in the space of reduced dimension. The algorithm, so-called dictionary projection pursuit (DPP), is a greedy approach in the sense that in each iteration the structure chosen in the previous iteration is eliminated from the data to obtain high projection coefficient values.

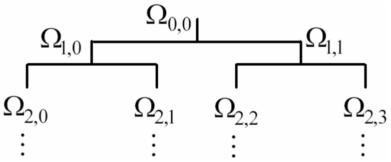

Figure 2. Decomposition tree in a wavelet packet.

To find a set of basis functions *B* that contain desired characteristic information, a weight *w* is assigned to each basis function in the wavelet packet dictionary. The weight $w\ (0 < w < 1)$ is a measure of linear independence of the selected basis from all of the previously selected basis functions. For the initial iteration, the weight is set as a vector of ones at the beginning of the procedure and updated in every iteration. The weight vector is then modified in such a way that each selected basis is orthogonal to the subspace of previously selected basis functions, resulting in a set of orthogonal basis at the final stage of the algorithm. Here is the complete algorithmic procedure:

Step 1. Find wavelet packet coefficients of each training data set in different classes and record them as a set of matrices of size $n \times (\log n + 1)$, where *n* is the signal dimension (call them *map*).

Step 2. Calculate the density (energy) of each packet by squaring each element in the wavelet packet matrices $(map.\text{^}2$ in Matlab notation) to obtain the energy map of each training data.

Step 3. Sum all of the matrices in Step 2 for each class. Divide them by the number of training data in each class (N_l) to determine normalized total energy of training data (energy map) in each class:

$$C_l(j, k, m) \equiv \sum_{i=1}^{N_l} (\mathbf{b}_{j,k,m}^T \cdot \mathbf{x}_i^{(l)})^2 / N_l, \quad \text{for } l = 1, \ldots, L \tag{10}$$

where triple indices (j, k, m) are scale, translation, and oscillation indices of wavelet packet decomposition, respectively. C_l, the energy map of class *l*, is a table which can be rearranged in a matrix form (call it *e-map*). At this stage, there are *L* matrices.

Step 4. Determine the relative entropy of *e-map* (call it *ent_map*) by applying Equation (5) or the symmetric version Equation (3).

Step 5. Repeat the following *m* times to find a set of orthogonal basis:

- Find the basis *b* corresponding to wavelet packet indices associated with $maxarg(w.*ent\_map(:))$, which is the maximum of element-by-element multiplication of vector *w* and vector form of *ent_map*.
- Compute the part of basis *b* which is orthogonal to basis functions already selected (call it residual), and normalize it to unity.
- Starting from left, enter and save the residual of *b* in a matrix as a new column vector.
- Compute *coef*, wavelet packet coefficients of the new basis function (the residual), and accumulate the energy of coefficients:
  $$coef\_e\_sum = coef\_e\_sum + coef.\hat{}2.$$
- Update *w* as $w = 1 - coef\_e\_sum.$

In DPP, the projection index, which is the entropy of the normalized sum energy of wavelet packet coefficients of the data set, is found at the beginning of the algorithm only once, contrary to PP in which the projection index must be calculated in each iteration. This is the key feature of the algorithm, which makes it faster compared to the PP approach.

The normalization in Step 3 will be inconsequential if the number of training data in each class is the same. In Step 5, if *n* bases are selected, then a complete orthogonal basis is found. Then, another method such as *principal component analysis* [8, 24] can be applied for further dimensional reduction. Step 5 is very similar to matching pursuit algorithm [17].

The above algorithm is a one-dimensional projection in the sense that matrix *B* is replaced by a single vector *b*. In each iteration, a basis function *b* is selected and added to the previously selected basis functions in the form of an expanding transpose matrix B^T. In this sense, the final projection is not one-dimensional, but a multi-dimensional projection.

DPP is still a time-consuming algorithm even though is computationally less expensive than the original PP method. The computational cost is $O(m\ n \log n)$, where *m* and *n* are number of selected basis functions and signal size, respectively. If a complete set of basis functions is required, then the computational cost will be $O(n^2 \log n)$, which is relatively high.

In the next section some drawbacks of DPP are highlighted.

6.1. Shortcomings of DPP

Recalling the definition of entropy in Equations (3) and (5), entropy calculations require that each entry to belong to a probability density function (pdf). However, DPP uses the relative entropy of normalized sum of the coefficient energies of all training data in *L* classes. It does not normalize the values of *ent_map* to unity; instead, normalization is done by the number of training data in each class. (It is worth noting that normalizing *ent_map* to unity resolves the above problem; however, it adds another technical glitch. Since we are comparing numbers rather than sequences, normalization to unity means that every element in the *ent_map* matrix must be one, which is trivial.)

As a result, and in accordance with the proof of *lemma* discussed in Section 3, relative entropy will not necessarily be non-negative. Consequently, relative entropy as used in DPP, does not represent a *theoretically* acceptable measure for the separation of different distributions, nevertheless, it may still be considered as a viable measure (albeit not a robust one) for comparing different data and for the selection of wavelets for discriminatory classification.

Because of its “symmetric” property, the relative entropy measure (Equation 3), results in a non-negative value regardless of the sum of the sequence being equal to unity or not. (Please refer to *Note* in Section 3). Still, the symmetric version cannot provide a robust measure. In Section 7 a method for resolving this problem is presented.

Next, we introduce a modification to the normalization scheme used in DPP.

6.2. Class-based normalization

In DPP, coefficients are normalized as defined by Equation (10), where the normalization is basically the average of the sum of the squared coefficients in each class. Consequently, the total energy of coefficients in each class is divided by the number of training data in that class. Under this scheme, normalization is effectively scaling down the signal energy in each class. In the special case, where the number of training data is the same for all classes, energy values are scaled down by the same proportion which corresponds to a uniform scaling of the entropy map values, thus there will be no relative changes in the final outcome. In the proposed approach, normalization as used in Step 3 of the DPP algorithm, is modified. Under the new approach, a class-based normalization is used in which each class is considered separately where the sum of squared *coefficient values* of different wavelet packet nodes, is adjusted by the sum squared values of *all the training data* in that class as:

$$\sum_{i=1}^{N_l} (\mathbf{b}_{j,k,m}^T \cdot \mathbf{x}_i^{(l)})^2 / \sum_{i=1}^{N_l} \| \mathbf{x}_i^{(l)} \|^2 , \tag{11}$$

where N_l is the number of training data in class *l*. Under normalization defined by Equation (11), different classes are normalized with respect to their own factors, resulting in further class differentiation during feature extraction stage and an improved accuracy in the classification stage.

To examine the effectiveness of this modification, numerous trial runs were carried out in which the signal-energy normalization scheme was applied to experimental data from a Ricardo Hydra internal combustion research engine. In order to ascertain and generalize the effectiveness of the proposed method, a wide range of data analysis using different wavelets was planned and performed. These included the use of 32 different analyzing wavelets from the family of orthogonal, biorthogonal, symmetric as well as selected wavelets from Battle-Lemarie spline functions as follows:

1-Haar, 2-Beylkin, 3-Coiflet1, 4-Coiflet2, 5-Coiflet3, 6-Coiflet4, 7-Coiflet5, 8-Daubechies2 (Db2), 9-Db3, 10-Db4, 11-Db5, 12-Db6, 13-Db7, 14-Db8, 15-Db9, 16-Db10, 17-Db20, 18-Db40, 19-Db45, 20-Bior22, 21-Bior31, 22-Bior68, 23-Symmlet4 (Sym4), 24-Sym5, 25-Sym6, 26-Sym7, 27-Sym8, 28-Sym9, 29-Sym10, 30-Vaidyanathan, 31-Battle3, 32-Battle5.

Figure 3 shows the classification results of DPP using the two normalization schemes, in which the horizontal axis indexed from 1 to 32 corresponds to the numbers used above to list the analyzing wavelets. For the majority of wavelets the proposed normalization scheme produced superior performance. For example, by applying Coiflet1 as the analyzing wavelet, misclassification rate was reduced from 4% with the N_l-normalization to 0.5% with the modified normalization scheme (analyzing wavelet number 3 in Figure 3 – “Average”). This is a considerable improvement. In fact, with class-based normalization of signals, additional separation of classes is induced.

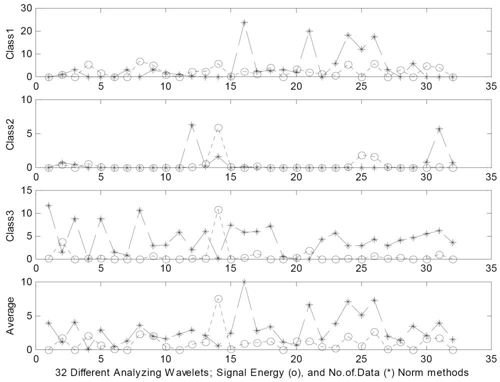

Figure 3. Classification percentage error using DPP under two normalization schemes, with 32 different analyzing wavelets listed above.

In the next section, a novel approach for using relative entropy in the construction of the energy map is proposed.

7. The Cross-data Entropy Approach

The use of entropy measure for feature extraction and classification requires that:

(1) Entries to Equation (5) for the evaluation of relative entropy be all non-negative,

(2) Relative entropy among different classes in each node (in every base of the dictionary in the case of DPP), i.e., the outcome of Equation (5), be also non-negative.

By considering sum of energy of coefficients as the entries to the relative entropy measure, we satisfy the condition of having non-negative sequences as the first condition of using relative entropy. However, the use of relative entropy of the above sum in each class, as prescribed by DPP and LDB methods, does not guarantee condition (2) to be satisfied, i.e., we may not have non-negative relative entropies for each and every data at all times. (Refer to Subsection 6.1 and [27]).

Another consideration for the use of relative entropy measure is the requirement that the sequence constitutes a pdf. On the contrary, in DPP approach, relative entropy is applied to single scalars in all levels of the decomposition process not to a pdf.

To resolve these shortcomings, an approach is proposed here, in which training data are used to generate the required sequence of numbers for proper application and evaluation of entropy. It is proposed that in constructing the entropy measure, instead of using the sum of coefficient energies of all training data in each class, and at each node, as used in DPP described in Section 6, we consider all of coefficients for the evaluation of the entropy measure. As a result, the role of every single data is taken into account in the sense that the relative entropies of each element in the wavelet packet matrix are used to determine the appropriate bases. Under this approach we deviate from the concept of "averaging of data" as is the case in DPP method. Two advantages are gained using the proposed scheme:

(1) Averaging of all training data as used in DPP and LDB methods essentially utilizes the *first order* statistics only. By not involving a second order statistics, such as standards deviation, the dispersion of data is masked. This is considered a limitation of the DPP and LDB methods. The proposed scheme eliminates this limitation by using all training data where coefficients are obtained and used for each and every training data.

(2) In the proposed algorithm, each coefficient is evaluated for all training data and thus at all nodes including the last level of wavelet packet tree, evaluation of entropy is carried-out on a sequence of scalars rather than on a single scalar. The scheme can then be interpreted as a cross-data entropy evaluation or cross-data energy map approach. Since we still use relative entropy, discriminatory bases will be derived as before. Under this scheme the relative entropy of distributions of the coefficients in different classes is taken into account, that is, discriminant information of every data (mutual discrimination among all data) is considered. For this reason, we refer to this method as a *cross-data entropy or mutual-based approach*. The cross-data entropy approach alleviates the shortcoming of standard relative entropy measure used in DPP and LDB methods.

In the following sections, formalization of the extended versions of DPP method, hence referred to as *mutual dictionary projection pursuit* (MDPP), is given. We define the following notations before describing the methods.

Let "*map*" as previously be the wavelet packet coefficients of each training data $\mathbf{x}_i$, for $i = 1, \ldots, N$, which can be illustrated by a set of N matrices of size $n \times (\log_2 n + 1)$, where n is the signal length and N is the number of training data. Let $C^{i=1:N}(j, k, m) \equiv (\mathbf{b}_{j,k,m}^T \cdot \mathbf{x}_{i=1:N})^2$ be the energy map of each training data derived by squaring each element of the *map* matrices, where $C^{1:N}$ is used to denote N energy map matrices, each of the size of *map*. (Matrices $C^{1:N}$ can also be viewed as a 3D-array, *e-map*, of size $n \times (\log_2 n + 1) \times N$.)

Recall that N_l is the number of training data in class l, where $N = \sum_{l=1}^{L} N_l$ is the total number of training data in all classes. If $C_l^{1:N_l}$ are energy maps belonging to each training data in class l, then $[C_l^{1:N_l}(j, k, m)]$ can be defined as a vector consisting of N_l number of element (j, k, m) of $C_l^{1:N_l}$: $[C_l^{1:N_l}(j, k, m)]$ $= [C_l^1(j, k, m), \ldots, C_l^{N_l}(j, k, m)]^T$ for $l = 1, \ldots, L$.

Similarly, we can think of $C_l^{1:N_l}$ as a 3D-array $e\text{-}map_l$.

The process used in the MDDP is described next.

7.1. Mutual dictionary projection pursuit

Consider a time-frequency dictionary such as wavelet packet transform. For a training data set consisting of L classes of signals $\{\{\mathbf{x}_i^{(l)}\}_{i=1}^{N_l}\}_{l=1}^{L}$, MDPP can be implemented by induction on scale j, as follows:

Step 1. Expand each training signal into a dictionary of orthogonal bases (*map* matrices) to obtain coefficients.

Step 2. Calculate the energy map of the coefficients, $C^{1:N}$, composed of squared values of each element of *map* matrices.

Step 3. Normalize matrices $C^{1:N}$.

Step 4. Find the discriminant power (by applying Equation 3 or 5) amongst L vectors $[C_l^{1:N_l}(j, k, m)]_{l=1:L}$ for $j = 0, 1, \ldots, J$, $k = 0, 1, \ldots, 2^j - 1$, and $m = 0, 1, \ldots, 2^{n_0 - j} - 1$, where n_0 is the maximum level of wavelet packet signal decomposition, with $n_0 = \log_2 n \geq J$ and J is the number of the decomposition levels considered for signal analysis.

Call the resultant matrix *ent_map*.

Step 5. Apply Step 5 as outlined in Section 6.

It should be emphasized that the mutual-based approach is not founded on the “sum” of energy map of data in each class as the case is in DPP and LDB. Instead, by employing the energy maps of “every” data sequence in each class, it finds a set of values at every base of the wavelet packet dictionary that “truly” represents the discriminant power of each data.

Computational efficiency of MDPP is similar to DPP. Expectedly, the new method has more data storage requirements than DPP since energy map of each training data must be saved for the evaluation of the entire relative entropy map.

8. Experimental Setup

To evaluate the effectiveness of the algorithm for the classification of different operating conditions in an engine, Ricardo Hydra, a single cylinder spark ignition research engine, was used for conducting a set of experiments and test runs. The engine operates on both gasoline and natural gas modes, but only natural gas mode operation with the compression ratio of 9.26:1 was used. The engine speed and throttling were set at 1500 RPM and 100% (wide open), respectively. Different machine operations with three relative air/fuel ratios namely stoichiometric $(\lambda = 1)$, fuel-lean $(\lambda = 1.5)$, and fuel-rich $(\lambda = 0.8)$ mixtures, each with normal, advance, and retard spark timing were examined. One pressure sensor with 12.5 KHz and two accelerometers with 25 KHz and 12.5 KHz sampling rates were used to measure cylinder pressure and simultaneous vibrations at two positions on the cylinder head in vertical and horizontal directions. Analysis of horizontal vibration data indicated that they did not carry useful information about the combustion event; therefore, only vertical vibration data were used in the runs. A rotational encoder was used to monitor engine speed to determine the starting point of each cycle. In this paper, data belonging to $\lambda = 1$ were utilized. To examine the effect of combustion on cylinder head vibration and pressure signals, data were also collected with engine running on motoring mode, i.e., driven externally and with no ignition/combustion. More information about data preparation details can be found in [25].

The objective of the experiment was to collect acceleration data at the cylinder head position with three different ignition timings of –23 (normal), –33 (advance), and –10 (retard) degrees under stoichiometric conditions $(\lambda = 1)$ assimilating healthy and faulty conditions. The numbers denote ignition timings measured as angles before top dead center.

In the following, selected results using MDPP scheme and the analysis of its performance are presented.

9. Experimental Setup

To assess the effectiveness of the proposed MDPP algorithm and to compare it with MLDB, as well as against DPP and LDB, the algorithms are applied on Ricardo Hydra test data. The normalization method introduced in Subsection 6.2 is also utilized.

In the next section, classification results of the algorithm on a set of machine data are presented. In the classification stage a neural network classifier is used. We examined 12 different backpropagation training algorithms each with one 5-node hidden layer, and Levenberg-Marquardt algorithm was found to provide the best results with respect to both high classification accuracy and computational efficiency.

9.1. MDPP classification

Applying MDPP algorithm on Ricardo Hydra test data with Coiflet1 wavelet resulted in selecting the bases that are plotted in Figure 4 (only the first 8 bases have been shown). The corresponding wavelet packet indices are shown below. The three rows correspond to scale, oscillation, and translation indices, respectively.

4	5	5	5	4	5	4	3
3	6	8	0	2	9	3	0
3	0	0	3	6	0	4	2

It is worth noting that not all of the bases selected by DPP belong to wavelet packet (WP) dictionary. DPP chooses the first base from WP dictionary (which is optimal according to the criterion described in Step 5). The bases selected during the rest of the iterations are components of WP bases that are orthogonal to the bases selected in the previous stages; they do not necessarily belong to the dictionary. Consequently, only the first base is certainly a wavelet packet base, the remaining bases are components of packet that are orthogonal to the previously selected bases. Further observation of the selected wavelets indicates similarity between the selected bases and corresponding wavelet packet bases that gradually decreases as we move to later iterations. In other words, as the number of selected basis increases (the number of columns in matrix *B*) the similarity of packet base and the corresponding selected base is reduced. This can be explained noting the fact that during the evaluation of the residuals in Step 5, only a segment of packet basis that is orthogonal to the previously selected basis is chosen.

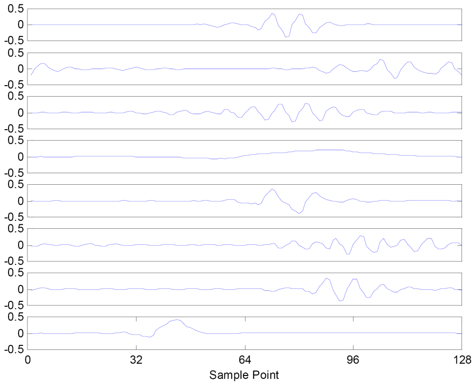

Figure 4. The first 8 bases selected by MDPP using Coiflet1.

Several runs showed that an increase in the number of iterations and bases yields a relatively moderate enhancement in the classification results but escalating the computational cost.

The classification results using MDPP and DPP are found to be close with a maximum classification error of 6% for most of the analyzing wavelets. The proximity of classification results in DPP and MDPP shows that the relative entropy map does not play a significant role in DPP. As we can see from Step 5 of DPP algorithm, the main purpose of using relative entropy is to find the energy map matrix *ent_map* and to determine the weight w. In this context, first order statistics (mean) is considered to be sufficient and there is no need for involving probability density function of features.

To assess the accuracy of the classification results we use the singular value decomposition of the coefficient matrix.

9.2. Analysis of coefficient correlation using singular values

Singular value decomposition (SVD) provides a useful technique in multivariate data analysis and statistical pattern recognition [8]. SVD technique is widely used for evaluating the correlation among experimental data composed of p sets where each set is a sequence of length q. Then, data can be expressed as a $p \times q$ matrix B. Singular values (SV) of matrix B are the eigenvalues of the correlation matrix $B^T B$ ranked from high to low.

We used SVD of the coefficient matrix that is derived by projecting data onto the selected bases to determine the extent to which the feature variables, i.e., the coefficients, are correlated.

The rate of decline in SV (Figure 5), which is the rate of drop from first to second and second to third, and so on, is an important parameter in statistical analysis. For a matrix with large rank, usually the decay of the first few SVs is of interest; where rank is defined as the maximum number of linearly independent rows (columns) [6]. In our case, that the coefficient matrix is of size $96 \times 4,$ there are four SVs (the rank is only four); therefore, the decay from first to second SV is of great importance.

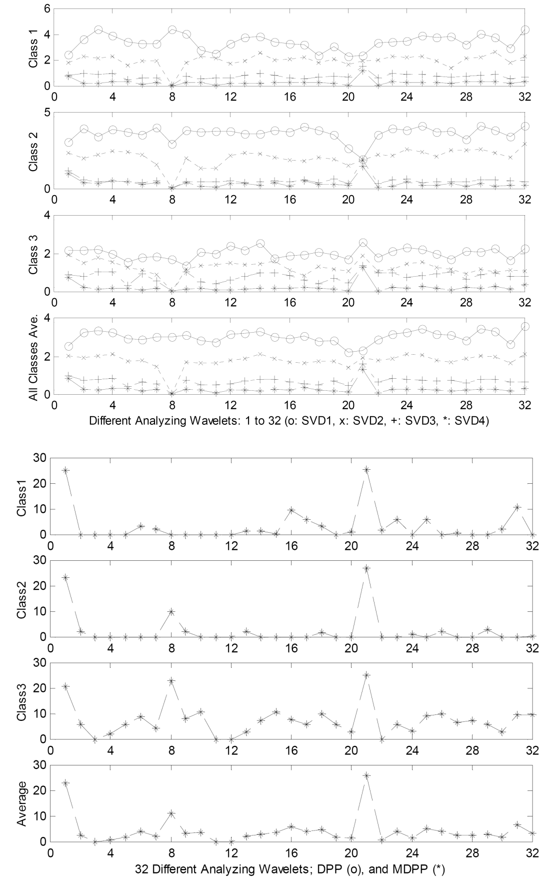

Figure 5. Singular values of coefficient matrix corresponding to the first 4 bases selected by MDPP for 32 analyzing wavelets, along with consequent classification results.

To obtain acceptable classification results, decay of SV of coefficient matrix must be neither too large nor too small. Large SV decay rate indicates that the useful information of coefficients is in one direction only (i.e., the direction of the eigenvector corresponding to that SV); therefore, the rest of selected directions, which are the bases found by the algorithm, contain redundant information, i.e., they are correlated with the first direction. On the other hand, choosing one direction to represent a multi-dimensional data is rarely a reasonable approach, specially noting that dimension reduction has already been implemented in the wavelet coefficient domain. Such cases show that selected bases used to derive the coefficients, do not contain all the essential information about the system performance. Under this condition, we conclude that the algorithm has found a set of bases where only one of them provides useful information, the rest of the bases are mostly redundant. As a result, a coefficient matrix with very high decline in SVI is not actually desirable. Similarly, having low SV decay means that all of bases (directions) carry more or less the same information content since there is a high correlation among them.

To support the above argument, in Figure 5 we have shown four singular values of coefficient matrix corresponding to the first four bases selected by MDPP when applied on Ricardo Hydra test data. The figure demonstrates the SVs in each class and the average for all classes, along with the corresponding classification results. The horizontal axis numbered from 1 to 32 corresponds to the same analyzing wavelets given in Subsection 6.2. Db2 analyzing wavelet (number 8 in Figure 5) maintains a coefficient matrix with large SV decay in all of the classes, which leads to a relatively large classification error of over 10%. Conversely, Bior 3.1 wavelet (number 21) produces a very low SV decay with mutually very close values, but still attains a high classification error of over 25%. Haar wavelet is also in the same category. Wavelets other than these three extreme cases have almost the same SV pattern with favourable decay rate and produce an acceptable classification result-typically with less than 7% error. With highly variable data and the performance errors observed with other classification methods, 7% is considered to be a relatively low and acceptable error.

Another interesting observation made is that SVs in different classes, as shown in Figure 5, follows a discernible pattern in which SVs of different classes are quite distinct from each other. For instance, the first singular values in each class associated with Coiflet1 (wavelet number 3) are 4.5, 3.6, and 2.2 for classes 1, 2, and 3, respectively, which show a difference of at least 20% among different classes.

This is considered an important aspect of the algorithm, which extracts the information that is discriminatory and makes different classes distinguishable from one another.

9.3. Application of different analyzing wavelets

To attain a comprehensive view of the effect of using different analyzing wavelets in classification results, several wavelets from various wavelet families, such as Daubechies, Coiflet, Symlet, and biorthogonal were used and tested in multiple runs. By examining Figure 5 and our experience on other SVD graphs related to various data sets, we conclude that the use of different wavelets has no significant influence on the correlation structure of the coefficients. For this reason, the classification errors for most of the wavelets are almost the same with only minor differences. We can then postulate that the improved classification is due to the algorithm is achieved.

9.4. LDB versus DPP

DPP and LDB methods were introduced and examined in the previous sections. During numerous simulations conducted while developing and examining MDPP and MLDB algorithms, a closer observation of the performance of LDB and DPP algorithms revealed that often the outcome of LDB and DPP methods were the same. For instance, three out of four of the selected bases in these two algorithms were the same when Coiflet1 was used as an analyzing wavelet on Ricardo Hydra data set. This can be traced to rather similar process that LDB and DPP use in searching for the best set of basis. Both search methods are based on the first order statistics, in which the sum-squared coefficients in each class are used for the construction of the relative entropy measure. However, there are some differences the details of which were explained in the present paper. Classification results with different analyzing wavelets are also very similar for DPP and LDB, and the differences are within a few percentage points.

10. Conclusions

The goal of projection pursuit for multivariate data analysis is to find low-dimensional projections, such as one or two dimensions, that provide the most revealing views of the full-dimensional data. In each iteration, DPP finds the component of the selected basis that is orthogonal to the hyper-plane spanned by the previously selected basis functions. In this manner, an orthogonal set of basis is obtained. In this paper, the usefulness of DPP in classification applications was shown.

It was also shown that standard DPP suffers from a technical deficiency in applying relative entropy on coefficients. To overcome this shortcoming, MDPP algorithm was developed as a new method for fault classification.

The classification results were influenced by the selection of appropriate bases. The question dealt here was how one could associate the accuracy of classification to the selected bases. In this respect, the accuracy of the classification results was related to the correlation of coefficient matrix, constructed from projecting the data onto the selected bases, by using singular value decomposition. In the assessment process the relevance and the meaning of various rates of decay of SVs were interpreted. It was shown that a feature matrix (here the coefficient matrix) with neither very high SV decay (where except the very few first SVs, the rest of selected SVs do not carry much information) nor very low SV decay (where there is high correlation among SVs) is desirable.

This paper dealt with the analysis of cylinder-head vibration data for engine fault detection and diagnosis. A novel method, referred to as mutual or cross-data entropy approach, was then presented. Using this approach, two wavelet-based methods namely DPP and LDB for feature selection and classification were modified. DPP and LDB were then, compared with each other and against the cross-data entropy approach. As stated, a close examination of the DPP and LDB methods reveals that their interpretation of entropy is non-standard and this poses certain technical glitches. In both methods, relative entropy is applied on the sequences of numbers that do not constitute a probability density function (pdf). The proposed method overcomes these shortcomings.

Acknowledgment

The authors wish to thank Dr. Howard Malm of REM Technology, British Columbia, Canada, for collaborating in this project by providing data.

References

[1] D. G. Calò, Gaussian mixture model classification: a projection pursuit approach, Comput. Statist. Data Anal. 52(1) (2007), 471-482.

[2] S. S. Chiang, C. I. Chang and I. W. Ginsberg, Unsupervised target detection in hyperspectral images using projection pursuit, IEEE Trans. Geosci. Remote Sensing 39(7) (2001), 1380-1391.

[3] M. M. Ettefagh, M. H. Sadeghi, V. Pirouzpanah and H. Arjmandi Tash, Knock detection in spark ignition engines by vibration analysis of cylinder block: a parametric modeling approach, Mechanical Systems and Signal Processing 22(6) (2008), 1495-1514.

[4] J. H. Friedman and J. W. Tukey, A projection pursuit algorithm for exploratory data analysis, IEEE Trans. Comput. 23 (1974), 881-889.

[5] J. H. Friedman, Exploratory projection pursuit, J. Amer. Statist. Assoc. 82(397) (1987), 249-266.

[6] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, Inc., 1972.

[7] R. M. Gray, Entropy and Information Theory, Springer-Verlag, 1990.

[8] W. Härdle and L. Simar, Applied Multivariate Statistical Analysis, Springer-Verlag, Berlin, Heidelberg, 2003.

[9] Q. Huang, Y. Liu, H. Liu and L. Cao, A new vibration diagnosis method based on the neural network and wavelet analysis, SAE Technical Paper 2003-01-0363, 2003.

[10] P. J. Huber, Projection pursuit (with discussion), Ann. Statist. 13(2) (1985), 435-525.

[11] L. O. Jiménez-Rodríguez, E. Arzuaga-Cruz and M. Vélez-Reyes, Unsupervised linear feature-extraction methods and their effects in the classification of high-dimensional data, IEEE Trans. Geosci. Remote Sensing 45(2) (2007), 469-483.

[12] M. C. Jones and R. Sibson, What is projection pursuit? (with discussion), J. Royal Statist. Assoc. 150(1) (1987), 1-36.

[13] J. B. Kruskal, Towards a practical method which help uncover the structure of a set of observations by finding the line transformation which optimizes a new index of condensation, R. C. Milton and J. A. Nelder, eds., Statistical Computation, Academic Press, New York, 1969, pp. 427-440.

[14] B.-C. Kuo and D. A. Landgrebe, Hyperspectral data classification using nonparametric weighted feature extraction, International Geoscience and Remote Sensing Symposium, Toronto, Canada, June 24-28, 2002.

[15] H. Lin and L. M. Bruce, Parametric projection pursuit for dimensionality reduction of hyperspectral data, Proceedings of IEEE International Geoscience and Remote Sensing Symposium 6(21-25) (2003), 3483-3485.

[16] B. Liu and S. F. Ling, On the selection of informative wavelets for machinery diagnosis, Mech. Syst. Signal Process. 13(1) (1999).

[17] S. Mallat and Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Processing 41 (1993), 3397-3415.

[18] H. Malm, Fundamentals of Reciprocating Machinery Analysis, REM Technology, Inc., 1994.

[19] J. A. Malpica, J. G. Rejas and M. C. Alonso, A projection pursuit algorithm for anomaly detection in hyperspectral imagery, J. Pattern Recognition 41(11) (2008), 3313-3327.

[20] F. Millo and C. V. Ferraro, Knock in SI engines: a comparison between different techniques for detection and control, SAE Technical Paper No. 982477, 1998.

[21] G. Rutledge, Dictionary project pursuit: a wavelet packet technique for waveform feature extraction, Ph.D. Dissertation, University of Victoria, 2001.

[22] N. Saito and R. R. Coifman, Local discriminant bases and their applications, J. Math. Imaging Vision 5(4) (1995), 337-358.

[23] B. Samimy, G. Rizzoni, A. M. Sayeed and D. L. Jones, Design of training data-based quadratic detectors with application to mechanical systems, Proc. of ICASSP-96, Atlanta, GA, 1996.

[24] D. W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, Wiley, New York, 1992.

[25] R. Tafreshi, F. Sassani, H. Ahmadi and G. Dumont, Local discriminant bases in machine fault diagnosis using vibration signals, J. Integrated Computer-Aided Eng. 12(2) (2005).

[26] R. Tafreshi, H. Ahmadi, F. Sassani and G. Dumont, Malfunction detection in multi-cylinder engines using wavelet packet dictionary, SAE Noise and Vibration Conference and Exhibition, 2005.

[27] R. Tafreshi, F. Sassani, H. Ahmadi and G. Dumont, An approach for the construction of entropy measure and energy map in machine fault diagnosis, ASME J. Vibration Acoustics 131(2) (2009).

[28] R. Tafreshi, Feature extraction using wavelet analysis with application to machine fault diagnosis, Ph.D. Dissertation, The University of British Columbia, 2005.

[29] S. Watanabe and T. Kaminuma, Recent developments of the minimum entropy algorithm, Proc. 9th Int. Conf. Pattern Recogn., IEEE, New York, 1988, pp. 536-540.

[30] G. T. Zheng and P. D. McFadden, A time-frequency distribution for analysis of signal with transient components and its application to vibration analysis, Trans. ASME 121 (1999), 328-333.