Proceedings of the ASME 2011 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2011
August 28-31, 2011, Washington, DC, USA

DETC2011-47822

MUTUAL INFORMATION BASED FEATURE SELECTION FROM DATA DRIVEN AND MODEL BASED TECHNIQUES FOR FAULT DETECTION IN ROLLING ELEMENT BEARINGS

Karthik Kappaganthu
Cummins, Inc.
Columbus, Indiana 47201
Email: karthik.kappaganthu@cummins.com

C. Nataraj*
Department of Mechanical Engineering
Villanova University
Villanova, Pennsylvania 19085
Email: nataraj@villanova.edu

ABSTRACT

This paper proposes a novel technique combining data-driven and model-based techniques to significantly improve the performance in bearing fault diagnostics. Features that provide best classification performance for the given data are selected from a combined set of data driven and model based features. Some of the common data driven techniques from time, frequency and time-frequency domain are considered. For model based feature extraction, recently developed cross-sample entropy is used. The ranking and performance of each of these feature sets are studied, when used independently and when used together. Mutual information based technique is used for ranking and selection of the optimal feature set. Using this method, the contribution to performance and redundancy of each of the data driven features and model based features can be studied. This method can be used to design an effective diagnostic system for bearing fault detection.

INTRODUCTION

Rotating systems are amongst the most common machinery in the industry and rolling element bearings are the key components in many high speed rotating systems. Rolling element bearings are the load carrying members of a rotating system and are also one of the prominent sources of nonlinearity. In this current competitive atmosphere it is necessary to maintain the machines in the proper conditions using condition based diagnostics methods. With the present focus in diagnostics shifting to use model and data to develop better diagnostics. Bearing fault detection is usually formulated as a classification problem. The performance of the classification depends on a variety of factors including the quality of the data, the type of classifier, the feature extraction techniques etc. In this research we try to improve the performance of bearing fault detection by improving the quality of features that are input to the classifier by providing information extracted using models.

Feature extraction for bearing fault detection using many signal processing techniques has been studied extensively and numerous methods have been developed. Most of the feature extraction techniques for bearing fault detection are data driven, which are obtained by applying signal processing algorithms in time [1], frequency [2] and time-frequency [3–5] domains. Another type of feature extraction techniques that make use of a model are called model based feature extraction techniques [6,7]. In this research we use a measure of entropy to extract model based features.

An important issue is the use of data driven techniques versus model based techniques. Data driven techniques are useful in incorporating the machine specific information. However, they fail to incorporate the existing knowledge of the system available from its governing laws and are limited by the available data.

*Address all correspondence to this author.

Model based techniques on the other hand not only parameterize the system but can also be used to parameterize the defects so that they can perform well over a range of operating conditions but are limited by the accuracy of the model. If a right combination of model based techniques and data driven techniques are used, the performance might improve.

In order to address these issues, careful feature selection needs to be performed to develop a good bearing fault detection algorithm. Further, the ability to rank features based on their performance will also provide useful insights into the system, model and the techniques. Although there are many studies on feature selection in general, there are only a few that analyze data driven features and model based features for identifying bearing defects quantitatively.

In this study, feature ranking and selection is used as a means to efficiently combine the information from both data and models optimally. A mutual information based method for feature ranking followed by a classifier for feature selection is proposed to generate a set of features obtained using some of the popular data driven feature extraction techniques and a model based technique to identify defects in bearings. Given some sample data, this feature set has the highest information content as well as the right set of features for maximum accuracy.

The aim of this study is to understand quantitatively, the interaction of the various features and their effect on classification performance. Further, we illustrate a methodology that can be sufficiently generalized with less human expertise and can also be used to evaluate the performance of newer model based and data driven feature extraction techniques for bearing fault detection. Although there are many studies on feature selection there are only a few that analyze bearing defects quantitatively.

In [8], the authors use decision trees to select features that provide good performance using a proximal support vector machine. The authors considered time domain features like skewness, kurtosis etc for bearing fault detection. In [9], Principal Component Analysis (PCA) was used to develop a set of features that improved the performance of both supervised and unsupervised learning machines. In this study the features were generated using time domain features like skewness, kurtosis etc.; frequency domain techniques like amplitude fault frequencies; and time-frequency domain techniques like wavelet transforms. In both these studies the outputs of the algorithms were feature sets that provide the best performance for the given data set but there was no ordering and comparison of the features. Some other general computational intelligence based algorithms that can be used for feature selection are [10, 11] and are not the focus of this paper.

The information theoretic approach [12] that is used in this research to determine the optimal set of features quantifies the quality of features as the mutual information content between features and the state of the bearing (faulty or healthy). Mutual information is a statistical measure that correlates different random variables [13]. It can be calculated from the probability distribution between the random variables [14]. Mutual information can be used to compare the features with each other and rank them accordingly. Mutual information based feature selection has been used in computer security, face recognition and biomedicine [15, 16]. We are not aware of any prior study that uses mutual information to select and compare features for bearing fault diagnostics.

The advantage of using this information theoretic approach is that it is independent of the classifier used. Also, among the features, some of these might have similar information among themselves. Hence, using such features together increases the redundancy, uncertainty and degrades the performance. Information theoretic approach addresses this important issue of interaction of features with each other for classification purposes and provides a set that performs better cumulatively. It also provides an ordered set of features that can be used to rank features and increase the efficiency of classification.

Thus, using modeling and mutual information, this paper addresses three important issues. First, it illustrates a model based feature extraction technique; second it integrates these model based features with some of the commonly used data driven features to obtain an optimal feature set for bearing fault classification; and third it provides guidelines about the effectiveness of model based features and data driven features over a range of operating conditions.

The model based features used in the paper are derived from a bearing defect model using cross-sample entropy. Cross-sample entropy is an extension of Approximate entropy [17] which is an estimate of kolmogorov entropy. These techniques are complexity measures that capture the information creation in a time series [18]. Cross-sample entropy captures the match between parts of two signals; greater the similarity in the two signals, lower the cross-sample entropy. Morphology, approximate entropy based techniques are being used increasingly in bio-medical signals and for fault detection purposes in machines.

In the medical field these techniques have been efficiently used to analyze electroencephalogram (EEG) and cardiotocographic (CTG) signals [19–21]. In the analysis of machines these measures have been used to diagnose gearboxes [22] and rolling bearing defects [23–25]. [26] used similar methods to extract envelopes for impulsive-type periodic systems. [27] performed detailed study on approximate entropy for detecting degradation in signals and demonstrated it to detect severity of defects in rolling element bearings.

The measures used in the studies mentioned are data driven and are highly dependent on the operating conditions, because of which it is difficult to generalize these techniques. Cross-sample entropy overcomes this problem as it is evaluated relative to a model. The changes in operating conditions can easily be accommodated by updating the model. Another advantage of cross-sample entropy over approximate entropy is that it has a

lesser bias and is more consistent. Detailed information can be found in [28].

The data driven features used in this paper are from time (skewness, kurtosis) [1], frequency (FFT, envelope spectrum) [2, 5, 29] and time-frequency (discrete wavelet transform) [3–5] domains. Please note that the data driven features considered here are not exhaustive. We have chosen these because these are the most basic techniques that are fundamentally different from each other. There are clearly other techniques which are variations and hybrids of these techniques. There are many other novel signal processing techniques based on nonlinear signal processing techniques [26, 30], demodulation [31–34], empirical mode decomposition [35–37]. These are not a part of this study; however, the proposed methodology could certainly be used to analyze and compare these techniques.

The feature selection methodology is explained in the next section. The experimental setup and the data used in this study are discussed in the third section. The various feature extraction techniques used in this paper are explained in the fourth section followed by a section on discussion on the results of feature selection using mutual information.

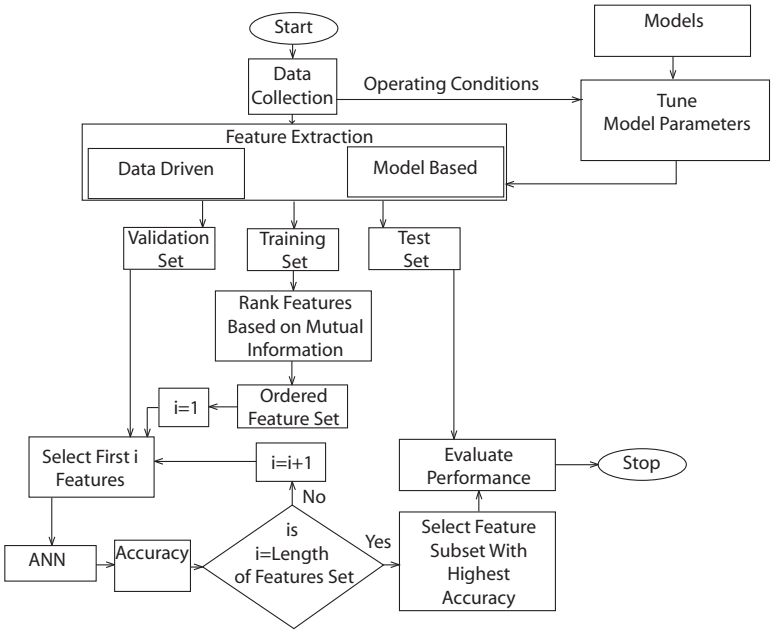

FIGURE 1. Algorithm for feature selection

METHODOLOGY

The flowchart of the process is provided in the Fig. 1. The first step is data collection; vibration data is collected from a system with a faulty bearing and a defect-free bearing over a span of rotating speeds, load and used for training, validation and testing of the algorithm. The faulty bearing has either a localized small or large outer race defect. From the data, various data driven and model based features are extracted. The operating condition parameters of the model namely speed and load are estimated based on the collected data and the corresponding models are used to extract the model based features. Next, a greedy search algorithm is used to rank the features based on the mutual information. Greedy search algorithm is a popular sequential search technique used in statistical research [13].

Now, the validation set is used to extract an optimal feature subset for classification using Artificial Neural Network (ANN) as the classifier. The feature subset selection is performed incrementally using the ordered feature set obtained in the previous stage. The subset with the best ANN classification performance is the optimal solution. This optimal feature subset is then used to test the performance using the test set data. Mutual information based ranking and feature selection are explained in the following sub-sections.

Feature Ranking

As explained earlier the feature ranking is based on mutual information. Estimation of mutual information $I(x;c)$ between a set of features (x) and the class c from a given set of data is explained in Appendix A. The feature selection process using mutual information is an optimization problem which seeks to find a set S which is a subset of set X containing all the features, which maximizes the information content $I(s;c)$ between the data and the health state of the bearing. If the size of S is equal to size of X then the solution to the optimization problem will be an ordered set of features.

The optimization problem can be solved using the greedy search technique. In the first step of this technique, set S is initialized to an empty set and a feature pool set defined as F is initialized to X. Next, S is populated iteratively with a feature from the feature pool such that it maximizes $I(x;c)$ at each stage. The selected feature is then removed from the feature pool. This process is continued until the feature pool is empty.

The algorithm for ranking can be summarized as follows.

1. From the data, find $p(c_k)$ and $H(c_k)$, $k = 1,\ 2,\ 3,\ ...,\ N_C$.
2. Set $S = \{\}$, $F = X$.
3. While F is not an empty set, DO
   (a) Set $i = 1$, Start Loop 1
   (b) Append the i^{th} element of F to S, i.e. $S_i = \{S,\ F_i\}$.
   (c) Set $j = 1$, Start Loop 2
   (d) Using Eqn. 8 find $I(x_j, c)$.
   (e) Using Eqn. 6 find $I(x_i, x_j)$.
   (f) If reached the end of S_i End Loop 2, else increment $j \rightarrow j+1$ and go to Step d.
   (g) Estimate mutual information of set S_i, $I(S_i, c)$ using Eqn. 9.
   (h) If reached the end of F End Loop 1, else increment

$i \rightarrow i+1$ and go to Step b.

(i) Find the element x_i^* corresponding to Maximum $I(S_i, c)$.

(j) Append x_i^* to S and remove it from F.

4. END WHILE
5. The final set S is the ordered feature set.

Feature Selection

The aim of this stage is to extract an optimal subset S_{opt} from the ordered feature set S obtained in the previous stage. The criterion for optimization is to achieve the least classification error using as few features as possible. The validation data is used to train an ANN and the classification accuracy is the measure of the classification. The algorithm for this is as follows.

1. Initialize $i = 1$ and S_i=$S(1)$.
2. Start Loop
3. Train an ANN using S_i and evaluate classification accuracy a_i.
4. If $S_i = S$ Stop Loop and proceed to step 6, else continue.
5. Increment $i \rightarrow i+1$ and $S_i \rightarrow \{S_i,\ S(i+1)\}$ and proceed to step 3.
6. From a_i find i^* corresponding to acceptable accuracy and optimal set size.
7. Obtain S_{opt} as S_{i^*}.

EXPERIMENTAL SETUP

All the experimental data was collected on a 'Machine Fault Simulator (MFS)' [38]. It is a test rig (Fig. 4) with a rotating shaft on two ball bearings. The shaft and the motor are connected using a flexible coupling to minimize misalignment effects. The shaft is loaded using a bearing loader and balancing disks. The different parts of the system can be conveniently assembled and disassembled. The bearings are placed in the bearing casing and can easily be replaced. The bearing parameters for the system used are given in Table 1. The system was loaded with a 5 kg mass. The signals from the MFS were collected using accelerometers placed on the bearing casing; once with a defect-free bearing, later with a bearing with a small outer race defect and finally with a bearing with a large outer race defect. Figures 2 and 3 show the bearings with outer race defects. It can be seen that the width of the defect in the first bearing is smaller than the width of the defect in the larger bearing. The defect width in the first case is 30 mil and in the second case is 90 mil.

Data was captured at different operating conditions, namely rotating speed and load. The load was varied by changing the unbalance in the system. Rotating speed can be easily measured and unbalance was estimated using least squares technique. The rotating speed was 1200, 1800 and 2400 r.p.m. Unbalance was varied using balancing disks. Three different sets of unbalance

FIGURE 2. Bearing with a small outer race defect

FIGURE 3. Bearing with a large outer race defect

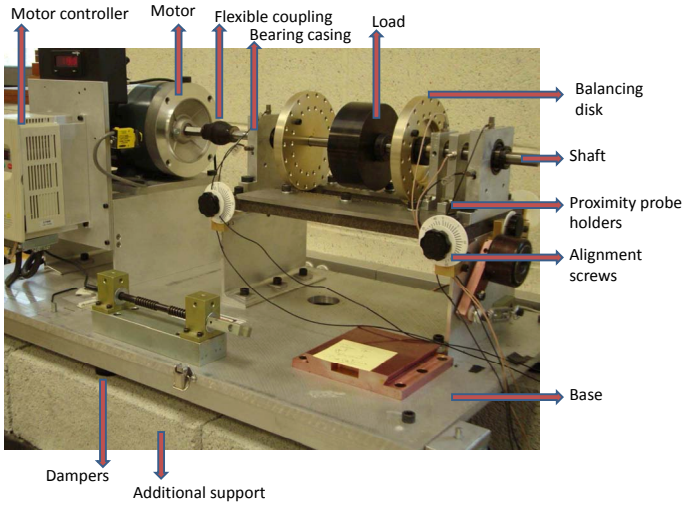

FIGURE 4. Experimental Setup

were induced. In the first case no external unbalance was used, next an unbalance was induced by placing a 5.5 gm screw at a distance of 9 cm. In the third case an external unbalance was induced by placing a 7 gm screw at a distance of 11 cm. Thus, there are nine different sets of operating conditions and at each set hundred independent samples of data were collected at a sampling rate of 32768 Hz. For convenience, data from the system without a defect is labeled '*DF*', data from a system with smaller outer race defect is labeled '*ORDs*' and data from a system with larger outer race defect is labeled '*ORDl*'. The corresponding models are labeled DF_m, $ORDs_m$, and $ORDl_m$ respectively. 50% of samples at each operating condition were used for training, 25% for validation and 25% for testing. Care is taken that data in each set is distributed evenly over the entire operating range.

Parameter	Value
Number of Rolling Elements (N_b)	8
Pitch Diameter (D_m)	1.319 in
Rolling Element Diameter (D_b)	.3125 in
Ball Pass Frequency (ω_{bpf0})	3.052 Ω
Contact Angle	0

TABLE 1. Bearing parameters

FEATURE EXTRACTION

In this section we provide a brief description of the feature extraction techniques we have used for rolling element bearing diagnostics. The list of methods discussed here are not exhaustive, but are representative of some of the basic techniques.

Data Driven Techniques

Among the first feature extraction techniques for rolling element bearing fault detection were the time domain techniques. Rolling element bearings with faults showed higher peak to peak vibration compared to a healthy bearing [1, 2]. The time domain features considered in this study are skewness and kurtosis.

Frequency domain methods are among the most used feature extraction techniques for bearing fault detection. When the rolling element enters a defect, an impulse acts on the casing. These impacts excite the structural resonances. The impulse is exerted at a frequency with which the rolling elements enter the defect. This frequency can be calculated from the geometry of the bearing and rotating speed [39], [40], [41]. Frequency domain techniques use these excitations to detect the defects in the bearing. The frequency component associated with a race defect is called the corresponding race ball pass frequency. The rotation of the cage also produces some frequency components. The other frequency components present in a typical bearing signal are the 1X response usually due to rotating unbalance, its harmonics and sub-harmonics. The presence of harmonics and sub-harmonics indicates nonlinear behavior in general. Note that a system with healthy bearing is also nonlinear in nature.

Fast Fourier Transform is the most common method to extract the frequency components in a signal. The spectrum usually contains a peak at the defect frequency. However, this is not always clearly observable because of slip and masking by other stronger vibrations.

In order to overcome this problem envelope spectrum is used. The impulse's excitations are amplitude modulated and can be recognized as side bands in frequency spectrum and can be seen as peaks in the Envelope Spectrum [29]. To find the envelope spectrum, the signal is band pass filtered around a frequency that the maximum signal to noise ratio, then Hilbert transform is used to find the envelope spectrum. There are many techniques to find the central frequency and range for the band pass filter [42–44]. Because of the simplicity and ease of use we use spectral kurtosis to select this band [45].

Discrete wavelet transforms (DWT) is a method for obtaining the time-frequency information of the signal. These methods are useful to extract the transients in the signal and are hence popular for defect detection. More information about the wavelet transforms can be found in [46]. Some of the recent work on bearing diagnostics using DWT are [3, 4, 47–51].

The relationship between these features and the magnitude of the defect is not straight forward as these features depend nonlinearly on other factors like speed, selection of the frequency band, load etc. The effectiveness of these features in identifying the severity of the defect for the current data can only be determined after feature ranking and selection.

The reconstructed DWT detail signal from a defect-free bearing and bearings with small and large outer race defects are shown in Figs. 5, 6 and 7 respectively. For brevity, signals only between level 1 and level 4 are shown. The defect-free bearing has the least energy content at all the three levels and the small outer race defect bearing signal has slightly more energy. The signal from a bearing with a large outer race defect has higher energy. In general the signals from bearings with defects are peaky and have more energy content. For the purpose of bearing fault detection, the level whose energy is most correlated with the severity of defect needs to be selected. This task is usually performed iteratively by verifying the different DWT signals in the expected frequency range. As will be seen in the next section, the feature ranking and selection algorithm used in this paper is a useful tool to choose the most useful signals.

The data driven features that are considered in the paper using the techniques explained above are listed in Table 3. Please note that the data driven methods discussed here are not exhaustive, but are representative of some of the basic techniques.

It should be noted that among the methods considered here, some of the methods perform well under certain conditions while others perform better at other conditions. For example, FFT features would perform well when the effect of slip is minimal and the defect signals are not masked by other signals, envelope spectrum on the other hand performs well when the band with highest signal to noise level can be efficiently selected. It is often a difficult task to pick the right set of features; much depends on the system and conditions such as speed, support properties, material properties etc., and other operating characteristics of the bearing.

Model Based Features

To extract the model based features the rotor-bearing system is modeled as a rigid rotor on nonlinear bearings. A rigid rotor model is well accepted in rotor dynamics literature, it is reason-

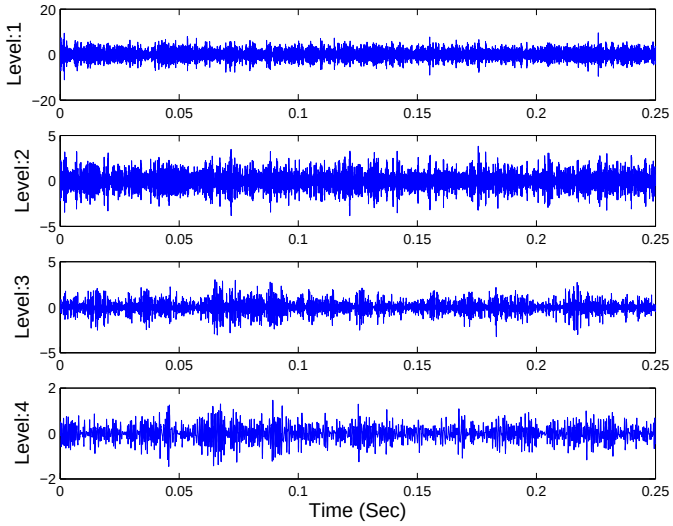

FIGURE 5. DWT (detail) of a defect-free bearing signal

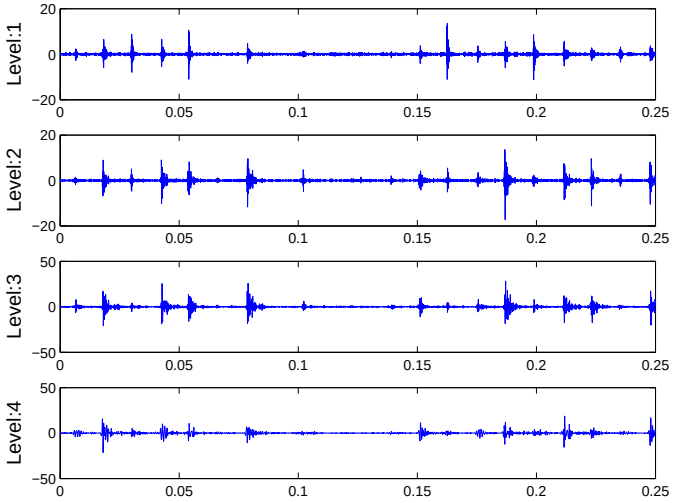

FIGURE 6. DWT (detail) of a bearing signal with a small outer race defect

ably valid for up to the first two critical speeds. The bearings are modeled using Hertzian contact forces and the outer race defects as pits. The bearing stiffness is implicit in the bearing force. The rotor-bearing system schematic is shown in Fig. 8.

The rotor-bearing system has four degrees of freedom $q = [V\ W\ B\ \Gamma]^T$. V, W are the displacement degrees of freedom in y and z directions respectively and B, Γ are the corresponding angular degrees of freedom. The forces acting on the rigid rotor with mass m, inertia I_D and polar moment of inertia I_p are the

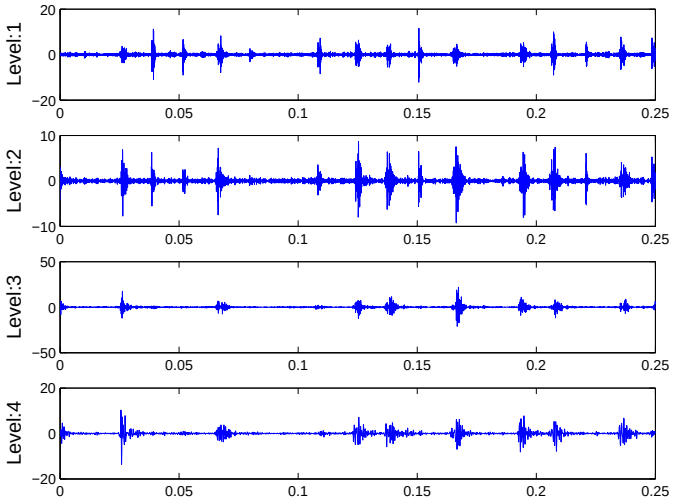

FIGURE 7. DWT (detail) of a bearing signal with a large outer race defect

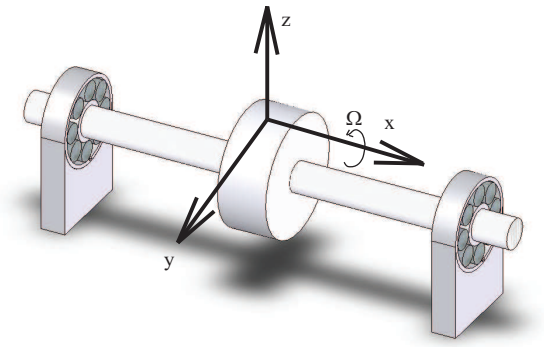

FIGURE 8. The rotor-bearing system

bearing forces (Q_b)and the unbalance forces (Q_u). Using the Lagrangian equation, the equation of motion for the rotor-bearing system is given by Eqn. (1).

$$M\ddot{q} + (C - \Omega G)\dot{q} = Q_b + Q_u. \tag{1}$$

where M and G are the mass and gyroscopic matrices, Ω is the rotating speed of the shaft. The unbalance force Q_u is dependent on the unbalance parameter 'e'. The bearing force Q_b is a function of shaft displacement, bearing geometry, rolling el-

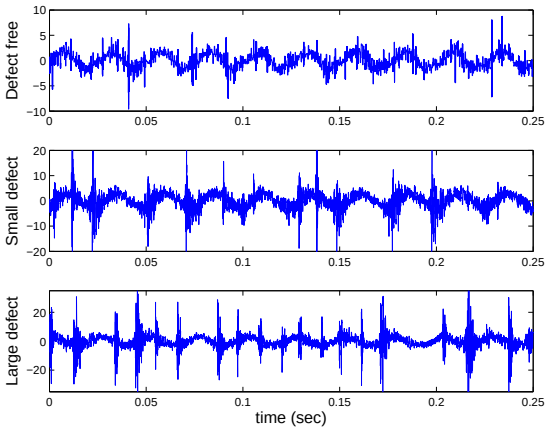

FIGURE 9. Measured bearing vibration signals. (a) Defect free. (b) Small defect. (c) Large defect

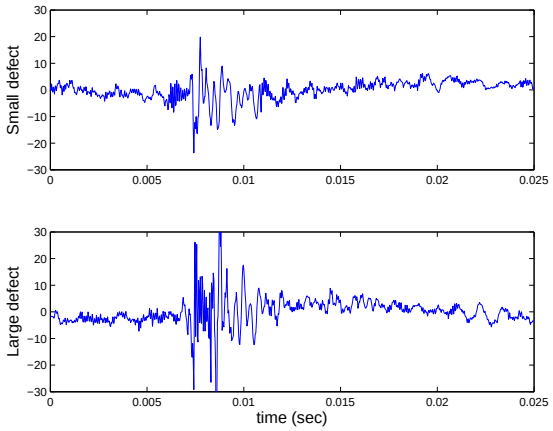

FIGURE 10. Measured bearing vibration signals for one rotation. (a) Defect free. (b) Small defect. (c) Large defect

ement positions and the magnitude and position of the defect. The defects are modeled as pits in the races with some depth and width. The effect of the defect on the bearing is simulated by assuming that the rolling element rotates about the edge of the defect before hitting a point in the defect and then gets out of the defect rotating about the other edge. Because of this change in the rolling element motion, the effective deflection and hence the restoring force on the rotor changes. We do not discuss the derivation of model in this paper, the details of this model can be found in [52]. A sample of the data captured from the experimental setup is shown in Fig. 9. In this figure the first subplot is the vibration signal of defect free system, the second subplot is the signal from a system with the smaller outer race defect and the third subplot is the signal from a system with the larger outer race defect. As is evident from these figures, bearings with outer race defects have impulses which are excited by the rolling element entering the defect. Figure 10 shows the bearing signal near these pulses for each of the defective bearings. The duration, magnitude and shape of these pulses is dependent on imbalance and defect magnitude. These changes in the characteristics of the shape of the signal are used to extract features to identify the severity of defect.

Consider the simulations generated from the models developed above shown in Fig. 11. Figure 12 shows the simulations for one rotation. It can be seen that the simulations and the measurements have a similar pattern which consists of a base signal modulated with impulses whose shape and duration is dependent on imbalance, magnitude of defect etc. The change in the shape is subtle and cannot be identified by observation. Feature extraction techniques are used to capture the differences not directly visible to naked eye. For convenience, simulations of a defect free model, small outer race defect and large outer race defect are labeled DF_m, S_m and L_m respectively.

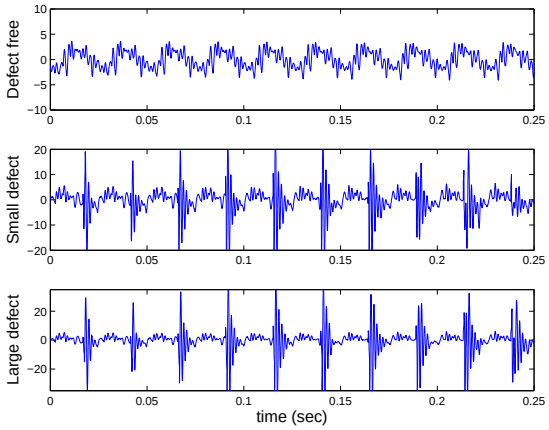

FIGURE 11. Simulated bearing vibration signals. (a) Defect free. (b) Small defect. (c) Large defect

Now, cross-sample entropy is used to extract features that can be used to identify defects using information from both model and data. Cross-sample entropy is an indicator of nonlinearity in a signal which measures the match between various parts of signals; greater the similarity in two signals lower the cross-sample entropy between them. Further, cross-sample entropy calculations need two parameters to be chosen. These are template length m and tolerance r. In this paper cross-sample entropy is calculated at $m = 1, 2$ and $r = 0.05, 0.1$.

Thus for a given signal S, using cross-sample entropy between the signal and the simulations from the three models (Defect free, small defect and large defect), features that compare the closeness of the data to one of the three expected models are extracted. Also, since these features are extracted at various template lengths and tolerance levels that are twelve model based

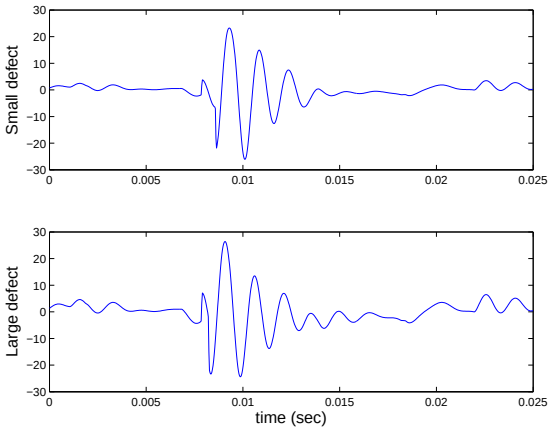

FIGURE 12. Simulated bearing vibration signals. (a) Defect free. (b) Small defect. (c) Large defect

Features
$CrsEn(S,\ DF_m,\ m,\ r);\ m = 1,\ 2;\ r = 0.05,\ 0.1$
$CrsEn(S,\ ORDs_m,\ m,\ r);\ m = 1,\ 2;\ r = 0.05,\ 0.1$
$CrsEn(S,\ ORDl_m,\ m,\ r);\ m = 1,\ 2;\ r = 0.05,\ 0.1$

TABLE 2. Model based features

features for a given signal; these are listed in Table 2.

FEATURE RANKING, SELECTION AND DISCUSSION

In order to understand the effect of model based features and data driven features, feature ranking and selection is performed using the algorithm presented in Section 2, with both model based and data driven features together (all the features listed in Tables 3 and 2). The performance of the algorithm is discussed to provide insights into the performance of the features under the given conditions.

The data from all the different unbalance experiments collected at 1200, 1800 and 2000 r.p.m are used to extract the various features described in the previous section. In this study we are interested in the overall performance of algorithm under all operating conditions, hence feature ranking and selection are performed to obtain best performance with all the data taken together. When the features are input to the feature ranking algorithm, the ordered feature set is shown in Table 4. The top three features are data driven features and the next three features are model based features. It is interesting to note that for a signal, cross-sample entropy with all the three models are grouped together, indicating the presence of essential non-redundant information. The other model based features are redundant or inferior

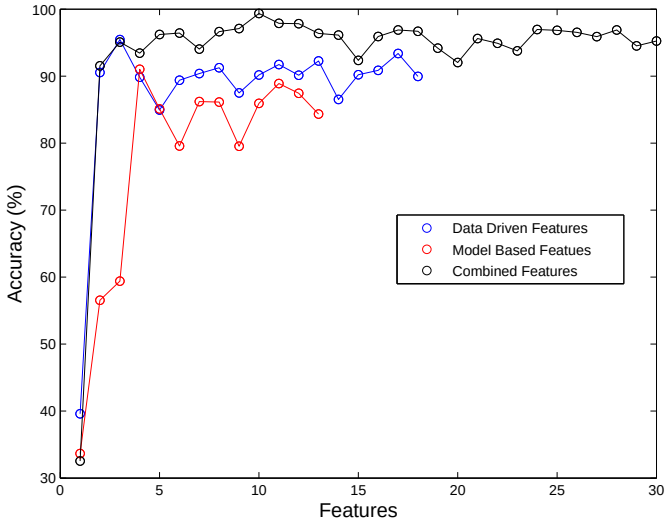

FIGURE 13. Classification Performance

and are present only towards the end of ordered feature set. Also note that the FFT magnitude at rotating speed occurs lower in the order. This could be because the cross-sample entropy based features have implicit information about the speed and unbalance.

To understand the value of the model based features and the significance of the ordering, we now analyze the classification performance and perform feature selection using the algorithm explained in Section 2. Consider the classification performance plot for each of the ordered feature sets as shown in Fig. 13. In this plot the performance of classification at each index is plotted. Performance is defined as the ability of the ANN to identify the severity of the defect from the inputs presented to it.

The performance at each index is evaluated by considering the set of features from the ordered set up to that index. For example, to evaluate the performance at index 10, the first ten features listed in Table 4 are used as input to the classifier. The average accuracy of classification for twenty independent classification runs using an artificial neural network is used as a measure of performance. The artificial neural network has one hidden layer with ten neurons. Tangential sigmoid is used as the transfer function for each of the neurons in the hidden layer.

For the purpose of comparison the feature ranking and selection algorithm is used to obtain the ordered feature set using only data driven features and model based features. From all the possible feature sets from the three ordered feature sets, the best possible performance is obtained when both data driven features and model based features were used. This performance is achieved when the first ten features from the ordered feature set shown in Table 4 are used. Further the performance is significantly increased when model based features were used. The best accuracy using data driven features was 93.4% which improved

to 99.3% when the optimal feature set suggested by the algorithm was used.

CONCLUSION

In this paper we address some of the important issues in identifying outer race defects of different magnitudes in rolling element bearings. We first used cross-sample entropy to extract model based features. These model based features are extracted by comparing the measured signal to simulations of a bearing model with parameterized defects. Then, to determine the performance of these methods and to further improve it by using some of the existing data driven features we performed feature ranking and feature selection using mutual information.

This approach allowed us to study the performance of the various features together as a set, and to determine an optimal feature set that consists of information from both data driven features and model based features. Thus, using this method we were able to optimally combine the information available in both data driven features and model based features.

The performance of this approach was studied on data captured from a rotating system at various speeds and loads. There were nine sets of operating conditions and a total of eighteen data driven features and twelve model based features were used. When both data driven features and model based features were used together the performance of the ordered feature set was excellent. The top ranked features in the ordered feature set consisted of both data driven features and model based features. This indicates that the model based features interact well data driven features and contain significant information that is relevant and not redundant. The performance showcases the importance of information contained in model when the operating conditions vary. This variation in conditions is parameterized and is implicitly available in model based features.

Thus, in this paper we used cross-sample entropy to extract model based features which were derived using models that parameterized defects and operating conditions. These model based features were combined with some of the common data driven techniques to significantly improve the classification performance. From all the multitude of combinations of features that were possible from the features considered, feature ranking using mutual information was performed to efficiently select the optimal feature set containing both data driven features and model based features, that provided the best classification performance.

Domain	Features
Time domain	Skewness, Kurtosis
Frequency domain	FFT magnitude at Ω, $\Omega/2$, 2Ω, 3Ω, ω_{bpfo}, $2\omega_{bpfo}$, $3\omega_{bpfo}$ Envelope magnitude at ω_{bpfo}, $2\omega_{bpfo}$, $3\omega_{bpfo}$
Time-frequency domain	DWT detail signals' energy up to level six

TABLE 3. Data driven features by domain

Rank	Features	Rank	Features
1	Envelope Mag. Ball Pass Freq.	2	DWT Energy: Level 4
3	FFT Mag. 3 Rot. Speed	4	crsEn(S,$ORDl_m$, $m = 2$, $r = .1$)
5	crsEn(S,DF_m, $m = 2$, $r = .1$)	6	crsEn(S,$ORDs_m$, $m = 2$, $r = .1$)
7	Envelope Mag. 3 Ball Pass Freq.	8	DWT Energy: Level 6
9	FFT Mag. 1 Rot. Speed	10	Envelope Mag. 2 Ball Pass Freq.
11	Kurtosis	12	DWT Energy: Level 3
13	FFT Mag. .5 Rot. Speed	14	DWT Energy: Level 1
15	FFT Mag. Ball Pass Freq.	16	crsEn(S,DF_m, $m = 1$, $r = .1$)
17	FFT Mag. 3 Ball Pass Freq.	18	FFT Mag. 2 Rot. Speed
19	DWT Energy: Level 5	20	FFT Mag. 2 Ball Pass Freq.
21	DWT Energy: Level 2	22	crsEn(S,DF_m, $m = 1$, $r = .05$)
23	Skewness	24	crsEn(S,DF_m, $m = 2$, $r = .05$)
25	crsEn(S,$ORDs_m$, $m = 1$, $r = .05$)	26	crsEn(S,$ORDl_m$, $m = 1$, $r = .05$)
27	crsEn(S,$ORDl_m$, $m = 1$, $r = .1$)	29	crsEn(S,$ORDl_m$, $m = 2$, $r = .05$)
29	crsEn(S,$ORDs_m$, $m = 1$, $r = .1$)	30	crsEn(S,$ORDs_m$, $m = 2$, $r = .05$)

TABLE 4. Ordered Combined Features

REFERENCES

[1] Tandon, N., 1994. "A comparision of some vibration parameters for condition monitoring of rolling element bearings". *Measurement,* ***12***, pp. 285–286.

[2] Barkov, A., Barkova, N., and Mitchell, J., 1995. "Condition assessment and life prediction of rolling element bearings". *Sound and Vibration,* ***28***, pp. 10–17.

[3] Cade, I. S., Keogh, P. S., and Sahinkaya, M. N., 2005. "Fault identification in rotor/ magnetic bearing systems using discrete time wavelet coefficients". *IEEE/ ASME Transctions on Mechatronics,* ***10***(6), December 2005, pp. 648–657.

[4] Mori, K., Kasashmi, N., Yoshioka, T., and Ueno, Y., 1996. "Prediction of spalling on ball bearings by applying discrete wavelet transform to vibration signals". *Wear,* ***8***, pp. 195–162.

[5] Ypma, A., 2001. "Learning methods of machine vibration analysis and health monitoring". PhD thesis, Delft University.

[6] Gertler, J., 1993. "Residual generation in model-based fault diagnosis". *Control, theory and advanced technology,* ***9***, pp. 259–285.

[7] Chen, J., and Patton, R., 1999. *Robust Model-Based Fault Diagnosis for Dynamic Systems*. Kluwer Academic Publishers.

[8] Sugumaran, V., Muralidharan, V., and Ramachandran, K. I., 2007. "Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing". *Mechanical Systems and Signal Processing,* ***21***, pp. 930–942.

[9] Malhi, A., and Gao, R. X., 2004. "PCA-based feature selection scheme for machine defect classification". *IEEE Transactions on Instrumentation and Measurement,* ***53***, pp. 1517–1525.

[10] Guo, H., Jack, L. B., and Nandi, A. K., 2004. "Automatic feature extraction for bearing fault detection using genetic programming". *Eighth International Conference on Vibrations in Rotating Machinery - IMechE Conference Transactions,* ***2***, pp. 363–372.

[11] Raymer, M. L., Punch, W. F., Goodman, E. D., Kuhn, L. A., and Jain, A. K., 2000. "Dimensionality reduction using genetic algorithms". *IEEE Transactions on Evolutionary Computation,* ***4***, pp. 164–171.

[12] Peng, H., Long, F., and Ding, C., 2005. "Feature selection based on mutual information : Criteria of max-dependency, max-relevance and min-redundancy". *IEEE Transactions on Pattern Analysis and Machine Intelligence,* ***27***, pp. 1226–1238.

[13] Duda, R. O., Hart, P. E., and Stork, D. G., 2001. *Pattern Classification*. Wiley-Interscience, New York.

[14] Cover, T. M., and Thomas, J. A., 1991. *Elements of Information Theory*. John Wiley and Sons, New York.

[15] Huang, J., Cai, Y., and Xu, X., 2007. "A hybrid genetic algorithm for feature selection wrapper based on mutual information". *Pattern Recognition Letters,* ***28***, pp. 1825–1844.

[16] Yana, Z., Wanga, Z., and Xieb, H., 2008. "The application of mutual information-based feature selection and fuzzy LS-SVM-based classifier in motion classification". *Computer Methods and Programs in Biomedicine,* ***90***, pp. 275–284.

[17] Pincus, S., 1995. "Approximate entropy as a complexity measure". *Chaos,* ***5***, pp. 110–117.

[18] Nayfeh, A. H., and Balachandran, B., 1995. *Applied Nonlinear Dynamics*. Wiley-VCH.

[19] Diambra, L., Bastos, J., and Malta, C., 1999. "Epileptic activity recognition in eeg recording". *Physica A,* ***273***, pp. 495–505.

[20] Acharya, R., Faust, O., Kannathal, N., Chua, T., and Laminarayan, S., 2005. "Non-linear analysis of eeg signals at various sleep stages". *Computer Methods and Programs in Biomedicine,* ***80***, p. 3745.

[21] Signorini, M., Magenes, G., Cerutti, S., and Arduini, D., 2003. "Linear and nonlinear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings,". *IEEE Transactions on Biomedical Engineering,* ***50***, p. 365374.

[22] Jiang, J., and Chen, J., 1999. "The application of correlation dimension in gearbox condition monitoring". *Journal of Sound and Vibration,* ***224***, p. 529541.

[23] Logan, D., and Mathew, J., 1996. "Using the correlation dimension for vibration fault diagnosis of rolling element bearings-i. basic concepts". *Mechanical Systems and Signal Processing,* ***10***, p. 241250.

[24] Logan, D., and Mathew, J., 1996. "Using the correlation dimension for vibration fault diagnosis of rolling element bearings-ii. selection of experimental parameters". *Mechanical Systems and Signal Processing,* ***10***, p. 251264.

[25] Yan, R., and Gao, R., 2004. "Complexity as a measure for machine health evaluation". *IEEE Transactions on Instrumentation and Measurement,* ***53***, p. 12371334.

[26] Nikolaou, N. G., and Antoniadis, I. A., 2007. "Application of morphological operators as envelope extractors for impulsive-type periodic signals". *Mechanical Systems and Signal Processing,* ***17***, pp. 1147–1162.

[27] Yan, R., and Gao, R. X., 2007. "Approximate entropy as a diagnostic tool for machine health monitoring". *Mechanical Systems and Signal processing,* ***21***, pp. 824–839.

[28] Richman, J. S., and Moorman, J. R., 2000. "Physiological time-series analysis using approximate entropy and sample entropy". *American Journal of Physiolgical Heart and Circulatory Physiology,* ***278***, pp. H2039–H2049.

[29] Randall, R. B., and Gao, Y., 1994. "Extraction of modal parameters from the response of power cepstrum". *Journal*

of Sound and Vibration, ***176***, pp. 179–193.

[30] Garimella, P., and Yao, B., 2005. "Robust model-based fault detection using adaptive robust observers". In Proceedings of the 44th IEEE Conference on Deccision and Control and the European Control Conference, pp. 3073–3078.

[31] Hu, X., He, Q., and Wang, H., 2008. "Rolling bearing fault detection based on svd denoising and stft demodulation method". *China Railway Science,* ***29***, pp. 95–100.

[32] Bozchalooi, I. S., and Liang, M., 2009. "Parameter-free bearing fault detection based on maximum likelihood estimation and differentiation". *Measurement Science & Technology,* ***20***, p. 065102.

[33] Altmann, J., and Mathew, J., 2001. "Multiple band-pass autoregressive demodulation for rolling-element bearing fault diagnosis". *Mechanical Systems and Signal Processing,,* ***15***, pp. 963–977.

[34] Randall, R., and Sawalhi, N., 2009. "Signal processing tools for tracking the size of a spall in a rolling element bearing". In IUTAM Symposium on Emerging Trends in Rotor Dynamics.

[35] Dong, H., Qi, K., Chen, X., Zi, Y., He, Z., and Li, B., 2009. "Sifting process of emd and its application in rolling element bearing fault diagnosis". *Journal of Mechanical Science and Technology,* ***23***, p. 2000.

[36] Yu, Y., Dejie, Y., and Junsheng, C., 2006. "A roller bearing fault diagnosis method based on EMD energy entropy and ANN". *Journal of Sound and Vibration,* ***294***, pp. 269–277.

[37] Li, M., and Zhao, P., 2008. "The application of wavelet packet and SVM in rolling bearing fault diagnosis". In IEEE International Conference on Mechatronics and Automation.

[38] http://www.spectraquest.com.

[39] Harris, T. A., 2002. *Rolling Bearing Analysis*, 4 ed. Wiley-Interscience.

[40] Nataraj, C., and Pietrusko, R. G., 2005. "Dynamic response of rigid rotors supported on rolling element bearings with an outer raceway defect". *ASME Conference Proceedings,* ***2005***(47381), pp. 1249–1261.

[41] Harsha, S. P., Sandeep, K., and Prakash, R., 2004. "Nonlinear dynamic behaviors of rolling element bearings due to surface waviness". *Journal of Sound and Vibration,* ***272***(3-5), pp. 557 – 580.

[42] Guo, L., Chen, J., and Li, X., 2009. "Rolling bearing fault classification based on envelope spectrum and support vector machine". *Journal of Vibration and Control,* ***15***, pp. 1349–1363.

[43] Sawalhi, N., and Randall, R. B., June, 2007. "Semi-automated bearing diagnostics - three case studies". In Comadem Conference, Faro, Portugal.

[44] Tse, P. W., Peng, Y. H., and Yam, R., 2001. "Wavelet analysis and envelope detection for rolling element bearing fault diagnosis-their effectiveness and flexibilities". *Transactions of the ASME. Journal of Vibration and Acoustics,* ***123***, pp. 303–313.

[45] Antoni, J., and Randall, R. B., 2006. "The spectral kurtosis: Application to the vibratory surveillance and diagnostics of rotating machines". *Mechanical Systems and Signal Processing,* ***20***, pp. 308–331.

[46] Chan, Y. T., 1995. *Wavelet Basics*. Kluwer Academic Publihsers.

[47] Ocak, H., Loparo, K. A., and Discenzo, F. M., 2007. "Online tracking of bearing wear using wavelet packet decomposition and probabilistic modeling". *Journal of Sound and Vibration,* ***302***, pp. 951–961.

[48] Pan, Y., Chen, J., and Guo, L., 2009. "Robust bearing performance degradation assessment method based on improved wavelet packet -support vector data descriptions". *Mechanical Systems and Signal Processing,* ***23***, pp. 669–681.

[49] Djebala, A., Ouelaa, N., and Hamzaoui, N., 2008. "Detection of rolling bearing defects using discrete wavelet analysis". *Meccanica,* ***43***(3), pp. 339–348.

[50] Wu, J.-D., and Liu, C.-H., 2008. "Investigation of engine fault diagnosis using discrete wavelet transform and neural network". *Expert Systems and Applications,* ***35***, pp. 1200–1213.

[51] Feng, Y., and Schlindwein, F. S., 2009. "Normalized wavelet packets quantifiers for condition monitoring". *Mechanical Systems and Signal Processing,* ***23***, pp. 712–723.

[52] Kappaganthu, K., Nataraj, C., and Samanta, B., 2009. "Fault parameter identification for model based prognostics of a ball bearing with an outer race defect". *ASME Conference Proceedings,* ***2009***(48982), pp. 1259–1267.

A Mutual Information

Let x_i be the random variable with pdf $p(x_i)$ corresponding to the i^{th} feature. Let C be any classifier that maps the features into N_C classes and c_k is the corresponding random variable with pdf $p(c_k)$, $k = 1,\ 2,\ \ldots,\ N_C$. Note that c_k is a discrete random variable. The entropy and mutual information are the defined as in Eqs. 2 and 3.

$$H(x_i) = -\int p(x_i)\log p(x_i)dx \tag{2}$$

$$I(x_i;c_k) = -\int p(x_i,c_k)\log\frac{p(x_i,c_k)}{p(x_i)p(c_k)}dx \tag{3}$$

Further, the entropy and mutual information are related by Eq. 4.

$$I(x_i;c_k) = H(c_k) - H(c_k|x_k) \tag{4}$$

In order to calculate the mutual information we need to find $p(c_k)$ and $p(c_k|x_i)$ from the data. It is easy to find $p(c_k)$ as it is a discrete random variable. By Bayesian rule we have

$$p(c_k|x_i) = \frac{p(x_i|c_k)p(c_k)}{p(x_i)} \tag{5}$$

The pdf of a continuous random variable x can be calculated from a given data using a Parzen's Window.

$$p(x) = \frac{1}{N}\sum_{i=1}^{N}\phi(x - x_i, h) \tag{6}$$

where, N is the number of samples, h is a parameter that defines the size of the window, x_i are the data points and ϕ is a finite valued non-negative density function called the window function. In this work a Gaussian function is used for ϕ (as is done typically).

Using Eqs. 6 and 7, $p(x_i|c_k)$ can be calculated.

$$p(x_i|c_k) = \frac{1}{N_k}\sum_{i=1}^{N_k}\phi(x - x_{k_i}, h) \tag{7}$$

where, N_k are the number of data points in the k^{th} class and x_{k_i} are the data points belonging to k^{th} class. Using Eqs. 5, 6 and 7 mutual information between a feature and a class can be calculated using Eq. 8.

$$I(x_i, c) = \sum_{k=1}^{N_C} p(c_k)\log p(c_k) - \int \sum_{k=1}^{N_C} p(x_i|c_k)p(c_k)\log p(x_i|c_k) \tag{8}$$

However, in order to calculate the mutual information between a set of features, $x = [x_1\ x_2\ ,...,\ x_n]$ and a class, we would need to calculate the joint pdf $p(x)$ of the feature set and the conditional joint pdf $p(x|c)$. Although it is possible to do this, it is cumbersome and is often inaccurate. A simpler procedure is to use Eq. 9.

$$I(x;c) = \frac{1}{|S|}\sum_{x_i \varepsilon S} I(x_i;c) - \frac{1}{|S-1|^2}\sum_{x_i, x_j \varepsilon S} I(x_i;x_j)$$
$$x = \{x : x \varepsilon S \subset X\} \tag{9}$$

The first part of the right hand side of Eq. 9 is the mean of the mutual information of each of the features and class; it is a measure of relevance of the set S. The second part consists of the information between the features themselves; it is a measure of redundancy of the set S. Using this method it is necessary to only calculate the joint pdf of two features at a time. This method, when used in a sequential search, has similar performance to the actual value [12].