A Multiple Model Prediction Algorithm for CNC Machine Wear PHM

Huimin Chen [1]

[1] *Department of Electrical Engineering, University of New Orleans, New Orleans, LA, 70148, U. S. A.*
hchen2@uno.edu

ABSTRACT

We present a multiple model approach for wear depth estimation of milling machine cutters using dynamometer, accelerometer, and acoustic emission data. The feature selection, initial wear estimation and multiple model fusion components of the proposed algorithm are explained in details and compared with several alternative methods using the training data. The performance evaluation procedure and the resulting scores from the submitted predictions are also discussed.

1. INTRODUCTION

The 2010 PHM data challenge focuses on the remaining useful life (RUL) estimation for cutters of a high speed CNC milling machine using measurements from dynamometer, accelerometer, and acoustic emission sensors (See http://www.phmsociety.org/competition/phm/10). The challenge data set contains six individual cutter records, denoted by c1, ..., c6. Records c1, c4 and c6 are training data while records c2, c3 and c5 are testing data. Each cutter was used repeatedly to cut certain work piece with the spindle speed of 10400 RPM. The wears of three flutes were measured after each cut (in 10^{-3}mm). In addition, 3-component platform dynamometer was used to measure the cutting forces in X, Y, Z directions. Three accelerometers were used to measure the vibrations of the cutting process in X, Y, Z directions. Acoustic emission (AE) sensor was used to measure the acoustic signature (AE-RMS) of the work piece during the cutting process. Prognostic algorithm development with similar equipment setup was reported in (Li et al., 2009). The training data contain 315 cut files for each cutter with the measured time series of forces, vibrations and AE-RMS and the resulting wears of three flutes after each cut. The testing data only contain the force, vibration and AE-RMS measurements for each cut without the wear depth measurement of each flute. The goal is to estimate the maximum number of cuts one can safely make for each testing cutter at a given wear limit. Note that one has to implicitly or explicitly predict the maximum wear of the flutes after each cut without knowing the initial wear of the cutter using only the force, vibration and AE-RMS measurements from consecutive cuts. Upon completion of the competition, the author was invited to submit a paper that fully discloses the algorithm used in 2010 PHM data challenge.

The rest of the paper is organized as follows. In Section 2, we first explain the performance evaluation criterion of the data challenge. Then we discuss feature selection for linear regression model on the additional wear after each cut for all three cutters. Finally, we reveal the need of individual regression model for each cutter and call for a multiple model approach to predict the wear depth of the testing cutter. In Section 3, we present the detailed description of the algorithm applied to the data challenge. The concluding summary is in Section 4.

2. PRELIMINARY ANALYSIS OF THE TRAINING DATA

2.1 Data Challenge Submission and Score Function

Each participant in 2010 PHM data challenge is required to estimate the maximum safe cuts at wear limit of 66, 67, ..., 165 ($\times 10^{-3}$mm) for three testing cutters. However, one does not know the actual wear after each cut and can only use the sensor measurements up to the current cut to infer the cumulative wear (although a participant can use the measurements from all 315 cuts, which is clearly unrealistic in practice). Note that the wear limit is on the maximum wear among three flutes. From the training data, we obtained the maximum safe cuts for the three training cutters in Figure 1. Clearly, c1 and c4 have early jump in the number of maximum cuts before wear depth reaches 100 while c6 has a small jump before wear depth reaches 120. These are important change points where the wear depth of the cutter is small after each cut (and in some cases unnoticeable) so that the number of safe cuts increases abruptly when the wear limit changes incrementally. Note that the initial wears of c1, c4, and c6 are 48.9, 31.4, 62.8, respectively.

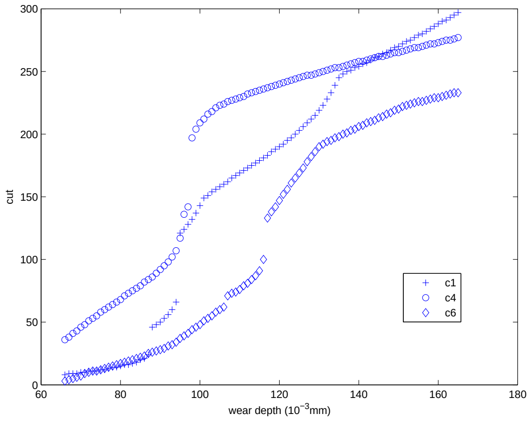

Figure 1. Maximum safe cuts for c1, c4 and c6.

Denote by δ the difference between the estimated number of cuts and the actual number of maximum cuts for a given cutter before it reaches the wear limit. The score function is defined as

$$S(\delta) = \begin{cases} e^{-\delta/10} - 1 & \delta < 0 \\ e^{\delta/4.5} - 1 & \delta \geq 0 \end{cases} \tag{1}$$

The goal is to minimize the sum of scores at the given 100 wear limits for all three testing cutters. Note that over estimate of the cut number has a larger penalty than under estimate. The above exponential penalty function focuses more on penalizing any single bad estimate rather than the average performance among all possible wear limits.

2.2 Feature Extraction and Selection

We first consider a regression model to estimate the additional wear after each cut assuming that the initial wear of the cutter is known. The goal is to identify useful features and to avoid overfitting to the training data. Linear regression was used with the candidate features and their p-values resulting from the regression on c1, c4, and c6 cutters in Table 2.2. A small p-value indicates the corresponding feature is likely to have non-zero regression coefficient, i.e., it should be included in wear depth estimation (Schervish, 1996). Unfortunately, most of the candidate features have relatively small p-values, making feature selection a challenging task with limited training data. Note that one can explore many candidate features including peak, mean, standard deviation, skew, kurtosis of force and vibration along X, Y, Z directions as well as their frequency domain indicators as listed in Table 2.2. Thus it is computationally infeasible to exhaustively enumerate all possible feature subsets with the linear regression model and choose the best one. To expedite the automatic feature selection procedure for any regression model, we do not want to test whether an individual regression coefficient should be zero or not based on p-value, instead, we control the false discovery rate (FDR) among all the selected features. It is an effective method for testing multiple hypotheses simultaneously (Benjamini & Hochberg, 1995).

Maximum force	0.08
Mean force	0.12
Standard deviation of force	0.07
Maximum vibration	0.05
Mean vibration	0.06
Standard deviation of vibration	0.09
Maximum AE-RMS	0.06
Mean AE-RMS	0.08
Standard deviation of AE-RMS	0.15
Peak magnitude of force frequency spectrum below 2000Hz	0.11
Peak magnitude of vibration frequency spectrum below 2000Hz	0.07
Peak magnitude of AE-RMS frequency spectrum above 2000Hz	0.14

Table 1. Candidate features and the corresponding p-values in linear regression

Formally, we consider d candidate features to be possibly included in the linear regressor. A hypothesis H_k describes the index set $\mathcal{I}_k \subseteq \{1, \cdots, d\}$ of the non-zero regression coefficients of $\mathbf{w}$, i.e., the selected feature subset.

H_k: $w_i \neq 0$ if $i \in \mathcal{I}_k$, otherwise $w_i = 0,\ i = 1, \cdots, d$.

We apply the FDR control technique to estimate $\mathcal{I}_k$. Note that the procedure does not require any independence assumption of the test statistics which is important in our case since some candidate features can have strong correlations in the regression model. The feature selection procedure is a step-down test (by successively selecting features) which is more efficient than a step-up test (by successively eliminating non-diagnostic features) when the number of selected features is relatively small compared with d. The procedure starts with the test statistic $T_1, \cdots, T_d$ based on the element-wise estimate $\hat{w}_1, \cdots, \hat{w}_d$. Each test statistic T_i is associated with a p-value, π_i, indicating its statistical significance when $w_i = 0$.

For any user specified FDR level $q \in (0, 1)$, the feature subset is selected by performing the following steps which controls the FDR to be below q (Chen, Bart, & Huang, 2008).

- Order the p-values such that $\pi_{(1)} \leq \cdots \leq \pi_{(d)}$.
- Compute the index $u_i = \min\left(1, \frac{d}{(d-i+1)^2} q\right)$, $i = 1, ..., d$.
- Reject all hypotheses $w_{(j)} = 0$ for $1 \leq j \leq k-1$ where k is the smallest index for which $\pi_{(k)} > u_k$. If no such k exists, then $\mathbf{w} = 0$.

Once the subset $\hat{\mathcal{I}}_k$ is determined, the regression coefficients should be recomputed using only the selected input features. Clearly, the FDR controlled feature selection procedure is much more efficient than finding the optimal feature subset via enumeration. It has been shown that the above procedure does control the FDR at the significance level q (Chen et al., 2008).

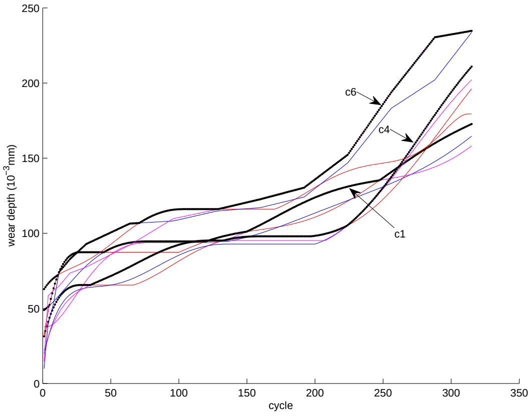

Figure 2. Flute wear after each cut and the maximum wear for cutter c1, c4, c6.

We set q=0.1 and applied linear regression to 68 candidate features. Only 5 features were selected based on the training data. With the selected features, we obtained the regressed additional wear for c1, c4 and c6 and the score being $3 \cdot 10^6$. We also tried to build the linear regression model for the wear of each flute instead of using the maximum wear after each cut. Figure 2 shows the individual flute wear and the maximum wear after each cut for the three training cutters. With the same FDR control level, the resulting score becomes $4 \cdot 10^6$ owing to the regression model fitting to the individual wear of each flute instead of the maximum wear. However, when we applied linear regression to c1 and c4 and the resulting model to estimate the wear of c6, the score becomes $2 \cdot 10^8$. Similar effects were observed when using c1 and c6 to build the model and estimate the wear of c4 as well as using c4 and c6 to estimate the wear of c1. Note that the results are based on the true initial wear of the cutter. It is clear that the linear regression method on the selected features tends to overfit even with fairly restrictive FDR control. The resulting score on the training set is unsatisfactory.[1] From the preliminary data analysis, we concluded that

- A single linear regression model yields large estimation error of the wear depth even with controlled FDR in feature selection.
- Regression on the additional wear after each cut for individual flute does not gain any benefit compared with regression on the maximum wear.
- No small subset of the candidate features can correlate well with the maximum wear after each cut for all training cutters.

3. MULTIPLE MODEL PREDICTION ALGORITHM

3.1 Building Regression Model for Each Cutter

We assume that each cutter has its own model to estimate the additional wear after each cut based on the selected features. In general, we consider a class of models M_1, ..., M_K where model M_i assumes that the observation z is governed by a likelihood functional $f_i(z|\theta_i)$ depending on the unknown parameter θ_i $(i = 1, ..., K)$. The dimension of θ_i is denoted by p_i. Denote by z^n a vector of n independent observations. Given z^n, one wants to find the best model M_i among the K candidates. Existing model selection criteria can be written in a general form (Chen & Huang, 2005)

$$l_j = -\log f_j(z^n|\hat{\theta}_j) + d_j(z^n), \ j = 1, ..., K \qquad (2)$$

being minimized among the K candidates. The first term of l_j uses the best estimate of θ_j to fit the negative log-likelihood function. The second term $d_j(z^n)$ is a penalty function that varies for different model selection criteria.

From our past experience, we applied the minimum description length (MDL) criterion which yields $d_j(z^n)$ =log $\left(\int f_j(z^n|\hat{\theta}_j)dz^n\right)$. It is interpreted as part of the normalized maximum likelihood (NML) (Rissanen, 1996). Under certain regularity conditions, one can use the asymptotic expansion of $d_j(z^n)$ given by

$$d_j(z^n) = \frac{p_j}{2}\log\left(\frac{n}{2\pi}\right) + \log\int |I(\theta_j)|^{1/2} d\theta_j \qquad (3)$$

where $I(\theta_j)$ is the Fisher information matrix given by

$$I(\theta_j) = \lim_{n\to\infty}\frac{1}{n}E\left[-\frac{\partial^2 \log f_j(z^n|\theta_j)}{\partial\theta_j(\partial\theta_j)'}\right] \qquad (4)$$

and the integral in (3) is over an appropriate subset of the parameter space. The MDL criterion intends to minimize the overall code length of a model and the observation described by the model.

As a special case of the above MDL principle, we assume that the additional wear of each cutter is a polynomial function of the selected features of unknown order. Thus for model M_i, the observation equation is $\mathbf{y} = \mathbf{H}_i\theta_i + \mathbf{w}_i$ where $\mathbf{H}_i$ is a known $n \times p_i$ matrix; θ_i is an unknown $p_i \times 1$ vector; and $\mathbf{w} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is the Gaussian noise vector with unknown variance σ^2. The minimum variance unbiased (MVU) estimate of θ_i is $\hat{\theta}_i$ = $(\mathbf{H}_i^T\mathbf{H}_i)^{-1}\mathbf{H}_i^T\mathbf{y}$. The residual sum of squares (RSS) of $\hat{\theta}_i$ is R_i = $||\mathbf{y} - \mathbf{H}_i\hat{\theta}_i||^2$. The MVU estimate of σ^2 is $\hat{\sigma}_i^2$ = $R_i/(n - p_i)$, which is different from the ML estimate given by R_i/n. The MDL based on the NML density minimizes the cost (Rissanen, 1996)

$$\mathrm{MDL}(p_i) = \frac{n}{2}\log\left(\hat{\sigma}_i^2\right) + \frac{p_i+1}{2}\log F_i + L_i \qquad (5)$$

[1] At the time that I registered for 2010 PHM data challenge, the top score on the leader board was already $4 \cdot 10^5$ on testing data without knowing the initial wear on each cutter.

where $F_i = (\mathbf{y}^T\mathbf{y} - R_i)/(p_i\hat{\sigma}_i^2)$ and $L_i = \frac{1}{2}\log\left(\frac{n-p_i}{p_i^3}\right)$. We control the FDR level q=0.1 as before. For cutter c1, 7 features were selected resulting in a score of $3 \cdot 10^3$ from the penalized regression model. Interestingly, 6 features were selected (4 of which are identical to those used for c1) for cutter c4 leading to a score of $2 \cdot 10^3$. 9 features were selected for cutter c6 (including all 6 featured used for c4) leading to a score of $2 \cdot 10^3$. We can see that three training cutters with different regression models have much better accuracy in estimating the wear depth than using a single regression model. However, we can no longer perform cross validation on the resulting models with the training data.

3.2 Multiple Model Fusion

From the Bayesian point of view, selecting a single model to make prediction or interpolation ignores the uncertainty left by finite data as to which model is the correct one. Thus a Bayesian formalism uses all possible models in the model space under consideration when making predictions, with each model weighted by its "posterior" probability being the correct one. The approach is called Bayesian model averaging and widely used in combining different learning algorithms (MacKay, 1992). The major difficulty in applying the Bayesian approach lies in the specification of priors when the candidate models are nested or partially overlapped. To be more specific, assuming that we have the observation vector z^n, the likelihood of a new observation y can be approximated by

$$f(y|z^n) \approx \sum_{i=1}^{K} P(M_i|z^n) f_i(y|\hat{\theta}_i) \tag{6}$$

using multiple models while the single model likelihood is $f_i(y|\hat{\theta}_i)$ if model M_i is selected. Without assigning prior to θ_i, one can not apply Bayes formula to estimate the model probability.

If one does not stick to the formal Bayesian solution, then the penalty l_i in the model selection criterion can be used to estimate the model probability. We apply the following estimator

$$P(M_i|z^n) = e^{-l_i} \Big/ \left(\sum_{j=1}^{K} e^{-l_j}\right) \tag{7}$$

and use MDL for l_i. In our opinion, the model probability can not be interpreted as the posterior since the prior of θ_i is unspecified. It represents the self-assessment of how likely M_i is selected using the criterion l_i. Denote by $\hat{y}_i$ the estimate (prediction or interpolation) of y using M_i. The single model estimate uses $\hat{y}_i$ if M_i is selected as the best model. The multiple model estimate uses

$$\hat{y} = \sum_{i=1}^{K} P(M_i|z^n)\hat{y}_i \tag{8}$$

and a subset of the K models can be identified based on a predetermined minimum model probability. Note that the outcome provided via multiple model fusion is valid for any model set, not limited to linear models. The essence of (7) is to approximate the Bayesian evidence of each model without any dependence on the unknown parameter θ_i so that all data can be used for the inference purpose, i.e., parameter estimation (Chen & Huang, 2005). If we know the initial wear of the testing cutter, then we can apply the above prediction method with three regression models obtained from the training data.

3.3 Initial Wear Estimate and RUL Prediction

Typically, the initial wear of the testing cutter should be known to the algorithm developer. Then for each cut, one collects the sensor measurements and sequentially predicts the additional wear after the cut. However, this information was unavailable during the competition. To estimate the initial wear, we made a hypothesis that the standard deviation of the high frequency peak from AE-RMS measurements correlates with the cutter's initial wear. This also has a statistical support with relatively small p-value from the linear regression model. We extrapolate from the three training cutters and obtained the estimated initial wears for c2, c3, c5 being 60, 55, 45, respectively. Note that we have used the AE-RMS measurements from the testing cutter of the first 15 cuts, which seems reasonable in practice.

With the estimated initial wear and the three regression models learned from the training data, we apply the multiple model algorithm via sequentially estimating the model probabilities after each cut by processing new sensor measurements and combining the individual predictions using (8). The resulting prediction of the maximum safe cuts is shown in Figure 3 where one can see that the prediction looks like a weighted average of the wears made by c1, c4, and c6. The first submission (with alias UNO-PHM) to 2010 PHM data challenge had a score of $9 \cdot 10^5$, which was among those top performance teams on the leader board.

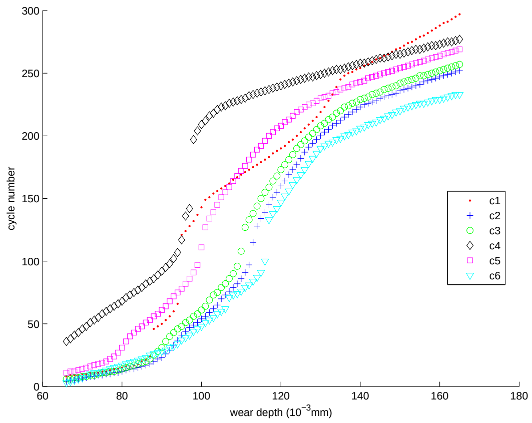

Figure 3. First submission to PHM data challenge for cutter c2, c3, c5 (maximum safe cuts of c1, c4, c6 included for comparison).

4. CONCLUSION

We provided detailed description of the multiple model prediction algorithm with automatic feature selection and initial wear estimation submitted to 2010 PHM data challenge. The final submission ranked #2 among professional and student participants and the method is applicable to other data driven PHM problems.

ACKNOWLEDGMENTS

The author wants to express his gratitude to Dr. Neil Eklund for organizing the 2010 PHM data challenge and inviting him to the special session, to Dr. Kai Goebel for hosting his visit to NASA Ames Research Center, arranging stimulating discussions with his team members on various prognostic problems, and to the ijPHM editor, Dr. Abhinav Saxena, as well as anonymous reviewers for providing constructive comments on an earlier version of this paper.

He is also grateful to the research sponsorship in part by ARO through grant W911NF-08-1-0409, ONR-DEPSCoR through grant N00014-09-1-1169, and Louisiana Board of Regents through NASA EPSCoR DART-2 and NSF(2010)-LINK-50.

NOMENCLATURE

δ	estimation error of maximum safe cuts
d	number of candidate features
w_i	regression coefficient of feature i
π_i	p-value of feature i
y	observation or quantity to be estimated
$\hat{y}$	estimate of y
M_i	statistical model i
θ_i	unknown parameter associated with model i
f_i	likelihood function of model i
l_i	penalized log-likelihood function of model i
z^n	n independent observations

REFERENCES

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, *B57(1)*, 289-300.

Chen, H., Bart, H., & Huang, S. (2008). Integrated Feature Selection and Clustering for Taxonomic Problems within Fish Species Complexes. *Journal of Multimedia*, *3(3)*, 10-17.

Chen, H., & Huang, S. (2005). A Comparative Study on Model Selection and Multiple Model Fusion. In *International Conference on Information Fusion.*

Li, X., Lim, B., Zhou, J., Huang, S., Phua, S., Shaw, K., et al. (2009). Fuzzy Neural Network Modelling for Tool Wear Estimation in Dry Milling Operation. In *Annual Conference of the Prognostics and Health Management Society.*

MacKay, D. (1992). Bayesian Interpolation. *Neural Computation*, *4*, 415-447.

Rissanen, J. (1996). Fisher Information and Stochastic Complexity. *IEEE Trans. on Information Theory*, *42*, 40-47.

Schervish, M. (1996). P Values: What They Are and What They Are Not. *The American Statistician*, *50(3)*, 203-206.

Huimin Chen received the B.E. and M.E. degrees from Department of Automation, Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, in 2002, all in electrical engineering. He was a post doctorate research associate at Physics and Astronomy Department, University of California, Los Angeles, and a visiting researcher with the Department of Electrical and Computer Engineering, Carnegie Mellon University from July 2002 where his research focus was on weak signal detection for single electron spin microscopy. He joined the Department of Electrical Engineering, University of New Orleans in Jan. 2003 and is currently an Associate Professor. His research interests are in general areas of signal processing, estimation theory, and information theory with applications to target detection and target tracking.