

PC センサデータマイニングによる HDD 故障予兆検知技術

Failure Prognosis Technology for HDDs in Notebook PCs Using Data Mining

西川 武一郎

中津川 実

■ NISHIKAWA Takeichiro

■ NAKATSUGAWA Minoru

東芝は、ノートパソコン (PC) に搭載した“東芝PCヘルスマニタ”により、ユーザーの承諾を得てバッテリー劣化や冷却性能など多岐にわたって稼働データの収集を行っており、既に220万台以上のデータを蓄積している。

今回ハードディスクドライブ (HDD) の故障予兆検知モデルを構築し、収集データを活用したHDDの故障予兆検知エンジンを開発した。このエンジンで判定した結果、HDDの故障確率は“予兆なし”と診断されたグループに比べて“予兆あり”のグループは67倍になることを確認した。予兆ありと診断されたHDDに対して適切な対策を打つことで、ノートPCのデータ保全をより確実に実施することが可能になる。

Toshiba has developed the Toshiba PC Health Monitor, which monitors various sensor data in its notebook PCs via the Internet with users' approval and alerts the users to significant system conditions including battery degradation and system cooling information prior to the possible occurrence of problems. We have already collected a broad range of operating logs from more than 2.2 million notebook PCs around the world.

We have now developed a failure prognosis engine for hard disk drives (HDDs) to diagnose HDDs in notebook PCs using collected sensor data, based on a failure prognosis model established by analyzing the relationships between these logs and failure data to predict the peculiarities of a failing HDD. We conducted evaluation experiments and confirmed that the group of HDD users alerted by this engine had a 67 times higher failure probability compared with the non-alerted HDD users group. This engine allows users to take adequate measures to prevent data loss once they become aware of a sign of a failing HDD.

1 まえがき

東芝は、ノートPCの品質を向上させるために、製造から出荷までのデータや戻入品の修理などのデータを活用して品質管理に取り組んできた。近年、これに加えて、インターネットを介して市場における稼働情報を収集できるようになったことから、インターネット経由で製品を見守り、状況に応じて適切なアクションを取るなどにより、製品価値の最大化が図れるようになってきた。

当社製ノートPCに搭載されている東芝PCヘルスマニタ⁽¹⁾は、この取組みの第一歩である。東芝PCヘルスマニタは、消費電力や、冷却システム、3D (3次元) センサなど、PCシステムの機能を監視し、その状態をメッセージなどで知らせる機能である。個人情報を含まないPCシステムの稼働データだけを収集しており、ユーザーが承諾した場合、インターネットを介してサーバに定期的送信される。当社は、現在220万台を超えるノートPCの稼働履歴を蓄積しており、このデータから、PCの稼働状況に基づく分析が可能である。

今回、収集データを活用した様々なサービスの可能性の中から、HDDの故障予兆検知を実現することを目標とした。HDDの故障はノートPCの故障の主要因の一つで、故障した場合、ユーザーへの影響がもっとも大きいデータ消失につながる可能性がある。HDDには、S.M.A.R.T. (Self-Monitoring,

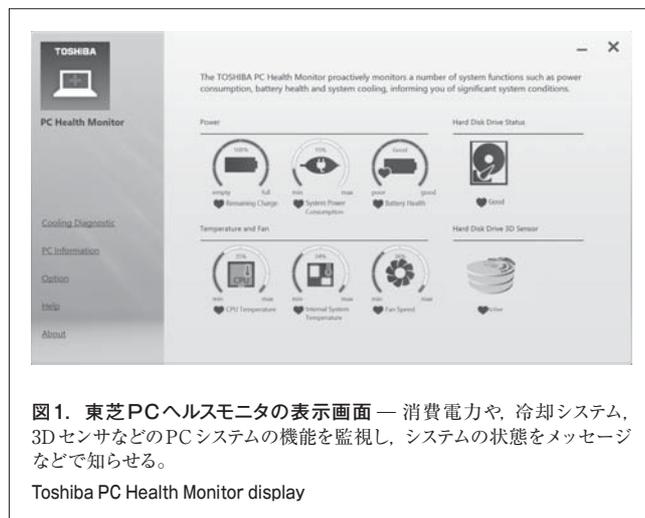


図1. 東芝PCヘルスマニタの表示画面 — 消費電力や、冷却システム、3DセンサなどのPCシステムの機能を監視し、システムの状態をメッセージなどで知らせる。

Toshiba PC Health Monitor display

Analysis and Reporting Technology) と呼ばれる自己診断機能が搭載されているが、東芝PCヘルスマニタではこの機能を活用してデータを蓄積している。更に当社では、HDD故障時の修理データも蓄積しており、これらのデータを用いてHDD故障予兆検知モデルを構築することで、HDDの故障前の状況を知ることが可能になる。

ここでは、東芝PCヘルスマニタの概要、HDD故障予兆検知モデルの構築、及びそれを活用したHDD故障予兆検知エンジンによるサービスについて述べる。

2 東芝PCヘルスマニタの概要

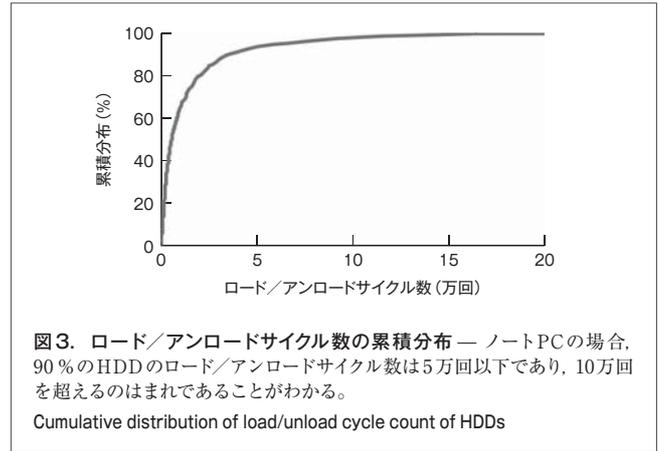
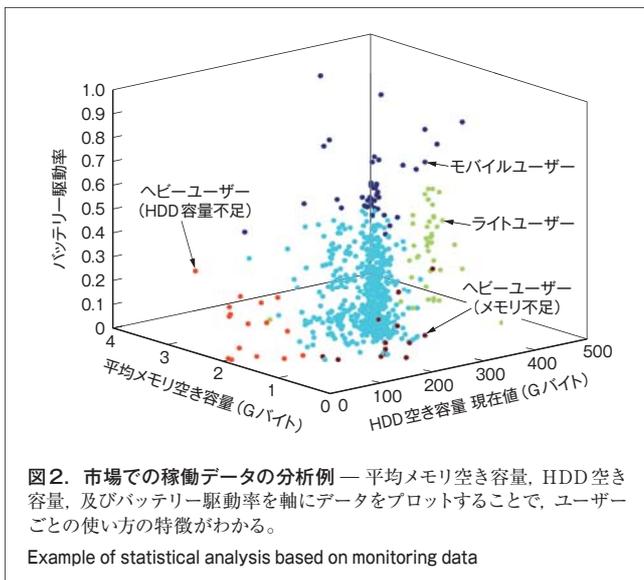
ノートPCに搭載されたPCヘルスマニタは、図1に示すように、稼働中のPCシステムの情報をユーザーに提示できる。ユーザーは、この画面からバッテリーの劣化状況を確認したり、冷却性能診断ツールを呼び出したりできる。

また、ユーザーの承諾が得られた場合には、稼働中のノートPCの情報を蓄積し、定期的にサーバにアップロードする機能も備えている。具体的な収集項目は、AC（交流）アダプタの接続状況や、電流、温度、HDD空き容量、メモリ空き容量、ファンスピード、電源ボタン長押しによる強制終了、HDDのS.M.A.R.T.属性（以下、SMARTデータと呼ぶ）など多岐にわたる。これらのデータは暗号化したうえでサーバに定期的に送信され、既に220万台を超えるノートPCの稼働データを蓄積している。

これらのデータを活用して、当社製品のある機種についてユーザーの使い方を分析した例を図2に示す。このように、市場における稼働データを分析することで、出荷したノートPCがどのように使われているかを確認できる。

3 HDD故障の傾向

企業向けのサーバ⁽²⁾では、多くの場合RAID（Redundant Array of Independent (Inexpensive) Disks）を構築することで冗長性を持たせ、HDDが故障してもデータを復旧できるようにしている。一方ノートPCでは、ほとんどの場合RAIDを構築することはなく、バックアップすら取らないことも多く、HDD故障時にデータを失うおそれがある。また、キーボードや、スピーカ、ファンなど振動を発生させる部品と隣接してHDDが配置されており、電車やバスの中で使用されたり、ユー



ザーが落下させ衝撃を与えたりすることもある。このような厳しい環境で使用されるため、一般にノートPC搭載のHDDでは、振動や衝撃による故障が多いと言われている。このため、同一品質でもユーザーの扱い方によってHDDの故障確率が大きく変わる。

またサーバとは異なり、ノートPCではユーザーのHDDへのアクセス頻度は小さいことから、経年劣化はあまり問題にならないと考えられる。実際、ロード/アンロードサイクル数を評価した結果は図3のようになる。これは、HDDの内部で磁気ヘッドが磁気ディスク表面から待機場所に退避し、その後磁気ディスク表面に戻った回数である。一般に、ロード/アンロードサイクルの保証回数は60万回程度だが、ノートPCのほとんどのユーザーは10万回も使用していないことが図3からわかり、摩耗故障が問題となるほどHDDを使用していないと言える。

このことから、今回対象とするノートPCに搭載されたHDDのほとんどは、偶発故障期間にあると考えられる。この時期の故障の予兆を検知するには、振動や衝撃に起因した時系列データの変化を捉えることが重要である。また、そのような変化から予兆を察知し、故障前にバックアップを取ったり、早めのHDD交換を行ったりすることで、被害を抑えることが可能である。

4 HDD故障予兆検知モデルの構築

PCヘルスマニタでは、ノートPCごとに毎日、表1に示すSMARTデータを収集し蓄積している。そこで当社は、これらのSMARTデータを活用して、故障確率を算出するモデルを構築した。

SMARTデータの各項目が1日ごとに計測された時系列データとして蓄積されているため、過去100日分の時系列データを使うと約2,000個の数値データになる。これらの数値の時系列変化パターンを、故障との相関が高い説明変数に変換し

表 1. SMARTデータ一覧

List of attributes of self-monitoring, analysis, and reporting technology (S.M.A.R.T.) for HDDs

ID	属性
1	RAW READ ERROR RATE
2	THROUGHPUT PERFORMANCE
3	SPIN-UP TIME
4	START/STOP COUNT
5	REALLOCATED SECTORS COUNT
7	SEEK ERROR RATE
8	SEEK TIME PERFORMANCE
9	POWER-ON HOURS COUNT
10	SPIN RETRY COUNT
12	DEVICE POWER CYCLE COUNT
191	SHOCK SENSE COUNT
192	POWER-OFF RETRACT COUNT
193	LOAD/UNLOAD CYCLE COUNT
194	TEMPERATURE PRESENT
196	REALLOCATED EVENT COUNT
197	CURRENT PENDING SECTOR COUNT
198	OFF LINE SCAN UNCORRECTABLE SECTOR COUNT
199	CRC ERROR RATE
220	DISK SHIFT
222	LOAD HOURS
223	LOAD RETRY COUNT
224	LOAD FRICTION
226	LOAD IN TIME
240	WRITE HEAD

ID: 識別番号
CRC: Cyclic Redundancy Check

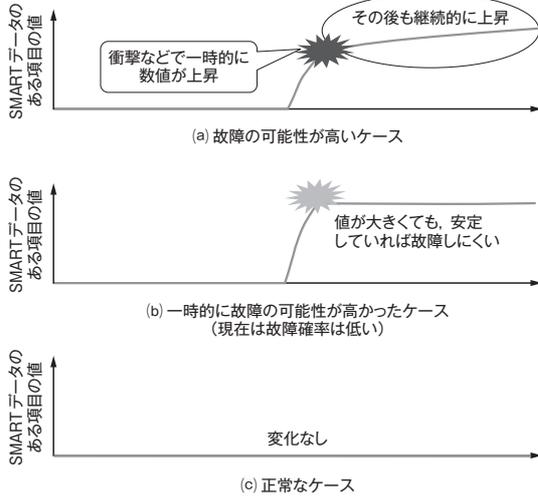


図 4. SMARTデータの時系列変化の例 — SMARTデータの典型的な時系列変化で、異常発生後の値の変化によって故障確率が異なる。
Example of time-series variations in S.M.A.R.T. attribute value

て、故障予兆検知モデルの構築に活用する。説明変数の組合せは、次に述べるデータマイニングアルゴリズムで決定する。

SMARTデータのある項目について、その時間変化を図4

1. 初期値として $F_0 = 0$ とする。
2. $i=1, 2, \dots, N$ において、以下のコスト関数 E を最小にする指標 x_i と弱学習器 $f_i(x_i)$ のパラメータ (a_i, s_i, u_i) を求めることを繰り返す。

$$E = \sum_j w_j (y_j - f_i(x_{ij}))^2$$

ここで、 x_{ij} は j 番目の HDD の i 番目の説明変数とする。

$$\text{また、 } y_j = -\frac{c_j - P_j}{P_j(1 - P_j)}, w_j = P_j(1 - P_j) \text{ である。}$$

c_j は j 番目の HDD が検査不合格のときに 1、合格のときに 0 となる変数である。

3. E を最小にする説明変数 x_i を求めた後、以下の式で F, P を更新する。

$$F_i = F_{i-1} + f_i(x_i)$$

$$P = \frac{1}{1 + \exp(F_i)}$$

4. 引き続き、 $i+1$ 番目のステップへ進む。

図 5. LogitBoost アルゴリズムの概要 — 故障確率モデルの作成には LogitBoost アルゴリズムによる学習を用いた。
Outline of LogitBoost algorithm

に模式的に示す。振動や衝撃などの原因で数値が急上昇した時点で故障確率は上昇するが、その後変化がなければ状態は安定しており故障確率が比較的低いと考えられる。そこで、SMARTデータの値そのもののほか、値の変化を表わす説明変数を多数用意した。今回、抽出する変化期間の長さを変えるなどして、756 個の説明変数を定義した。

故障確率モデルを作成するために、作成した説明変数を成分とする入力変数ベクトル \vec{x}_j を j 番目の HDD に対して定義し、学習用データ $\{(\vec{x}_j, c_j) | j \in \Omega\}$ を用意する。ここで、 c_j は j 番目の HDD が検査不合格のときに 1、合格のときに 0 となる変数である。

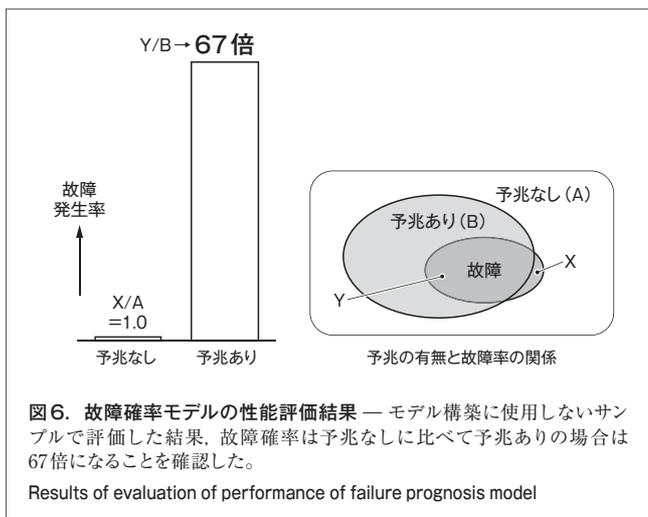
作成した学習データから故障確率モデルを作成するためのアルゴリズムには、LogitBoost⁽³⁾を採用した(図5)。

LogitBoostで作成された学習データにおける故障確率 P は式(1)のように複数の弱学習器 $f_i(x_i)$ の和で表される。

$$P(x_j) = \frac{1}{1 + \exp(f_i(x_{ij}) + \dots + f_k(x_{kj}))} \quad (1)$$

ここで、 $f_i(x_i)$ は式(2)で定義される二分木である。

$$\begin{aligned} f_i(x_i) &= s_i & x_i < a_i \\ f_i(x_i) &= u_i & x_i \geq a_i \end{aligned} \quad (2)$$

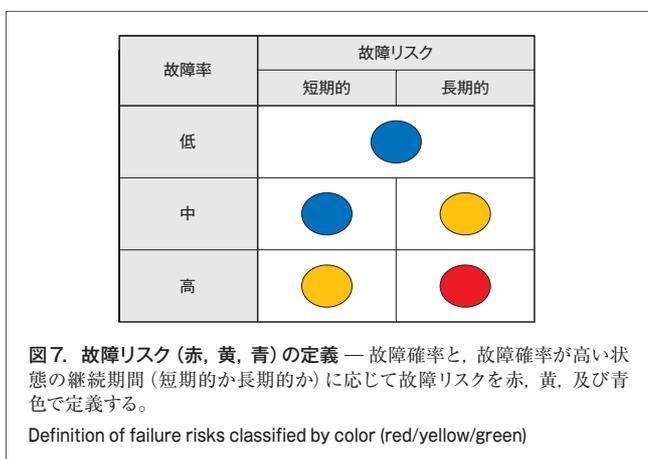


計算された確率値に対してしきい値を一つ決めることで、予兆あり、予兆なしの二つに分けることもできれば、しきい値を二つ決めて、赤、黄、青などの3段階に分けることも可能である。適切にしきい値を決め、テストサンプルをベースにHDD故障予兆検知モデルの性能を推定した結果、予兆ありのサンプルは、予兆なしのサンプルに比べて、市場で故障する確率が67倍となることを確認した(図6)。

5 HDD故障予兆検知エンジンによるサービス提供

HDD故障予兆検知モデルを活用して、ユーザー向けのサービス提供を行うため、故障リスクに応じて信号の色(赤、黄、青)を表示できるHDD故障予兆検知エンジンを開発した。

SMARTデータの時系列変化に伴い、信号が黄(又は赤)になった後に短期間で青に戻るケースがある。このような場合、ユーザーはすぐに対策を打たなければならないのか、対策を打つのを少し待ったほうが良いのか判断に困ることが懸念されるため、信号が頻繁に変わることを防ぐ必要がある。このことから、図7のように故障確率の高さと故障確率が高い



状態の継続期間を元に、信号の色(赤、黄、青)を定義した。この処理により、黄(又は赤)が出現する頻度を大幅に削減できることを確認し、ユーザーへの適切な情報提供を可能にした。

6 あとがき

今回開発したHDD故障予兆検知エンジンは、企業向けの商品“東芝スマートクライアントマネージャー”⁽⁴⁾に搭載された。これは、IT(情報技術)マネージャーが企業内のPCを管理するための商品で、管理対象となるPCの稼働状態を一覧表示で確認できる。また、信号の色に応じて、HDDを早めに交換したり、データのバックアップ頻度を上げたりすることで、問題発生を未然に防ぐことが可能になる。

今後、PCヘルスマニタを活用して他の部品故障の予兆検知や診断を行うなど、更なるサービス検討を行うとともに、PCだけでなく当社製の他の商品にもこの技術を適用していく予定である。

文献

- (1) 儘田 徹. ノートPCの故障予兆技術. 東芝レビュー. 64, 5, 2009, p.64-65.
- (2) 樋渡 仁 他. クラウドコンピューティング環境におけるHDD故障率モデルの提案. 情報処理学会デジタルプラクティス. 2, 3, 2011, p.227-234.
- (3) Friedman, J. et al. Additive logistic regression: a statistical view of boosting. Ann. Statist. 28, 2, 2000, p.337-407.
- (4) 東芝. “東芝スマートクライアントマネージャー(TSCM)”. <http://dynabook.com/pc/catalog/service/1402tscm/index_j.htm>, (参照2014-06-11).



西川 武一郎 NISHIKAWA Takeichiro, Ph.D.
 研究開発センター システム技術ラボラトリー研究主幹, 博士(理学)。
 データマイニング・リスクマネジメント分野の研究・開発に従事。
 日本オペレーションズ・リサーチ学会, 日本品質管理学会会員。
 System Engineering Lab.



中津川 実 NAKATSUGAWA Minoru, Ph.D.
 研究開発センター システム技術ラボラトリー研究主務, 博士(工学)。
 データマイニングの研究・開発に従事。情報処理学会, 人工知能学会, 日本航空宇宙学会会員。
 System Engineering Lab.