2B-4

ＰＣセンサデータマイニングによるＨＤＤ故障予兆検出

中津川 実　　西川 武一郎
（株）東芝 研究開発センター

1. はじめに　HDD の故障予兆は検出可能か

ハードディスクドライブ（HDD）が故障すると、重要なデータが失われてしまう。バックアップの実施や RAID 構築により、データ損失リスクを軽減することは可能である。しかし、バックアップをほとんど実施しないユーザは、多く存在していると言われている。データ損失を未然に防止するためには、HDD の状態を監視して故障の予兆を検出し、ユーザに注意喚起することが重要である。

HDD には S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology) と呼ばれる、障害の早期発見・故障の予測を目的とした機能が搭載されている。Pinheiro ら[1] は、Google 社のサーバーの大量の HDD を対象に、S.M.A.R.T.の各項目の値と HDD 故障との関係についての調査を行い、いくつかの S.M.A.R.T.項目でエラーが発生した場合は、その後の故障率が有意に高くなることを示した。また少数の HDD のデータで故障予兆モデルを構築した例として Hamerly[2]、Murray [3]の研究がある。しかし、多数のノート PC での分析事例は我々の知る限り存在しない。本稿では、ノート PC に搭載されている HDD の稼動情報と修理情報を大量に収集して、データマイニングにより HDD 故障予兆検出モデルを構築した研究開発事例について述べる。そして、HDD の故障予兆を高精度で検出可能であることを示す。

2. HDD 故障予兆モデル構築に用いたデータ

本章では、HDD 故障予兆検出モデルの構築に用いたデータについて説明する。HDD の稼動情報は「東芝 PC ヘルスモニタ」を用いて収集した。東芝のノート PC には、センサ情報や各種ログ情報を活用して PC システムをモニタリングするソフトウェア「東芝 PC ヘルスモニタ」が搭載されている（図 1）。ユーザの許可がある場合、モニタリングしたデータはネットワーク経由で収集され、データベースに蓄積される[4]。本稿執筆時点では、166 万台のデータが収集されている。本稿ではこのうち S.M.A.R.T.情報を活用した。

一方、PC 修理センターでは故障が疑われる PC を診断して故障箇所を特定し、修理サービスを行っている。本稿では修理履歴データを用いて、稼働中の HDD が故障したかどうかを把握した。稼動データと修理データの関係を、図 2 に示した。なお、Pinheiro ら[1]も指摘しているが、HDD の故障は定義が難しい。ユーザとベンダーで故障と考える基準が異なることも指摘されている。本稿では PC 修理センターで HDD が修理・交換された場合を HDD 故障と見なす。

HDD failure prediction by PC sensor data mining
† Minoru Nakatsugawa, Takeichiro Nishikawa,
Corporate Research & Development Center, TOSHIBA

図 1　東芝 PC ヘルスモニタ

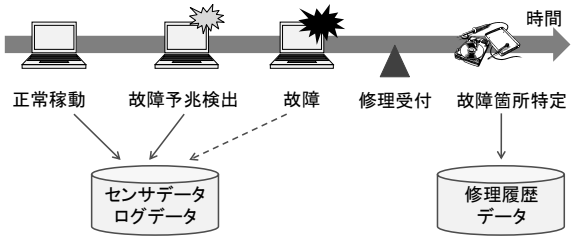

図 2　センサ・ログデータと修理データの関係

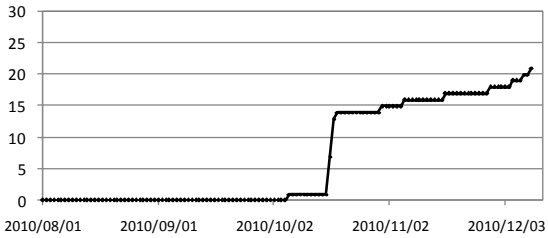

図 3　故障 HDD の Reallocated Sector Count の時系列変化

3. HDD 故障予兆モデル構築

特徴量計算

故障予兆を検知するうえで、徐々に劣化が進行している、急激に悪化した、安定している、というような時系列変化の様子をとらえて判断することが重要である。図 3 に典型的な S.M.A.R.T.の時系列変化を示した。将来の故障発生を予測する際の説明変数の候補として S.M.A.R.T.の時系列変化パターンの種類に応じた特徴量を約 700 種類作成した。

故障／正常判別モデル構築

(X_i, Y_i)を各ドライブのデータとする。$X_i = [x_1, \ldots, x_{Ni}]$ はドライブ i の時刻 1 から N_iまでの時系列 S.M.A.R.T.データ、$Y_i \in \{0,1\}$はドライブ i の故障(1)・正常(0)の別である。HDD 一台ごとに特徴量の計算を、最新時刻 N_i におけるデータで行う（S_i としよう）。故障予兆モデルは、S_i から Y_i を予測するモデルである。正常／故障（Y）の

二値判別モデルを構築するアルゴリズムとして、本稿ではブースティング（LogitBoost）を用いた。

モデル評価

図 4 は、HDD 故障予兆モデルの ROC 曲線である。縦軸は TP/(TP+FN)、横軸は FP/(FP+TN)である。10-fold cross validation による評価結果を示した。故障 HDD のうち 83.3%は故障前に予兆検出することが可能である。また、正常 HDD のうち 90.8%は正常と判定される。

	故障予兆あり	故障予兆なし
故障 HDD	true positive (TP)	false negative (FN)
正常 HDD	false positive (FP)	true negative (TN)

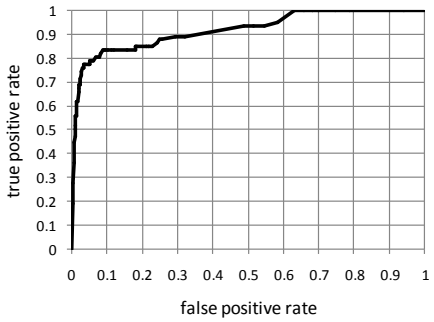

図 4　HDD 故障予兆モデルの ROC 曲線

4. 故障予兆検出に基づくデータバックアップ

バックアップを定期的に行っていないユーザが多い理由に、適切なバックアップ頻度がわからない、ということが挙げられる。バックアップの実施は PC に負荷がかかり、ユーザの手間もかかることから、なるべく少ない方がよい。そこで、故障予兆に応じたバックアップを実施することで、不要なバックアップの手間をかけずにデータ損失の削減効果が得られる。あるいは、故障予兆有無に応じてバックアップ頻度を変更することで、データ損失量を増やさずにバックアップ頻度を削減することが可能である。図 4 に示したモデル精度の条件下でシミュレーション評価を行ったところ、故障予兆発生後にバックアップ頻度を 毎月→毎日 に切り替えることで、毎週バックアップ実施する場合と比較して同程度のデータ損失量（平均 11%減）でバックアップ頻度を平均 61%削減（52→20 回）することが可能である（図 5）。

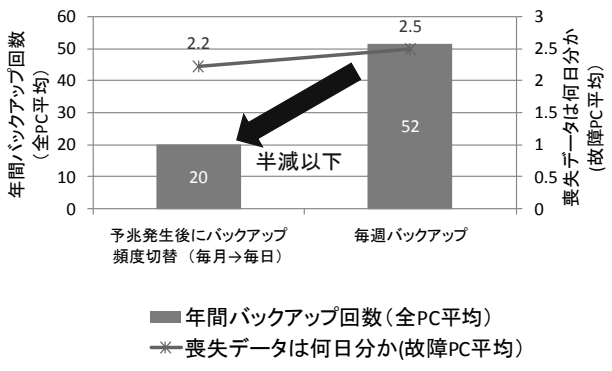

図 5　故障予兆検出に応じたバックアップの実施

ここで、データ損失量は最終バックアップから故障までの経過時間に比例すると仮定して、経過日数で表した。さらに、詳細は講演資料に掲載するが、データ損失コスト（データ損失量に比例すると仮定）とバックアップ実施コスト（バックアップ実施回数に比例すると仮定）の和を最小とするような、最適バックアップ間隔を求めることも可能である。

5. おわりに

本稿では、PC センサデータの時系列変化に関する特徴量を説明変数とし、PC 修理データを教師信号としたブースティングによる HDD 故障予兆検出モデルの構築を示した。これまでメーカーが提供できるサービスは、修理センターに持ち込まれた PC の修理が中心であった。しかし、稼動データを活用することによって、故障発生前の情報提供など新しい価値を提供できる（図 6）。実運用を通じて今後、提案手法のさらなる効果検証と技術改善を行っていく。提案手法は過去の HDD のデータを使用して性能検証を行ったものであり、新たな HDD でも同様の精度であるか、検証していく。

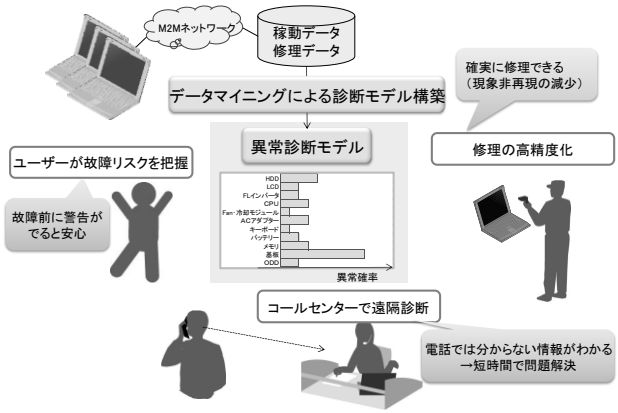

図 6　稼動データの活用によるサービスの実現

6. 参考文献

[1] Eduardo Pinheiro, Wolf-Dietrich Weber and Luiz André Barroso, "Failure Trends in a Large Disk Drive Population", Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), 2007

[2] G. Hamerly, C. Elkan, "Bayesian approaches to failure prediction for disk drives," Proc. 18th ICML, pp.202-209, 2001.

[3] J. Murray, G. Hughes, and K. Kreutz-Delgado, "Machine learning methods for predicting failures in hard drives : a multiple-instance application," JMLR Vol.6, pp.783-816, 2005.

[4] 西川武一郎, 原貫三, "市場品質の監視による早期対策からプロアクティブな品質保全とサービスへ", 東芝レビュー, Vol.64, No.8, 2009.