

教師なし学習による無線品質の自動クラスタリング

Automatic Radio Quality Clustering with Unsupervised Learning

鈴木信雄*¹
Nobuo SUZUKI

津田和彦*²
Kazuhiko TSUDA

*¹ KDDI 研究所
KDDI R&D Laboratories Inc.

*² 筑波大学大学院
University of Tsukuba

Smart devices such as smartphones mount the wireless communication function by standard features. Although this function is necessary in modern society, the wireless devices have some serious radio quality problems. They include connection impossibly and insufficient transmission speed. It takes long time and much human resources to solve such problems with analyzing each wireless device. Machine learning technologies are utilized in data mining issues and provide possible solutions for a lot of problems. It especially is able to reduce the cost of collecting right data and analysis because the unsupervised learning doesn't need right data. This study proposes the method identifies wireless devices which have radio quality problems by classifying the devices with unsupervised learning technology from radio information of Wireless LAN. The evaluation experiment on three kinds of unsupervised learning technology with selecting features of radio quality worsen is also described. Gaussian mixture model showed the highest precision accuracy.

1. はじめに

LTE や無線 LAN などの無線通信はスマートフォンに標準搭載され、現代の情報通信社会では不可欠の機能となっている。便利な無線通信であるが、場所や時刻によって、接続ができない、伝送速度が十分にでないなどの問題が発生している[内藤 2016]。このような問題を解決するためには、問題現象の把握が重要であるが、無線通信が使われている現場では、多くの種類と膨大な数の端末が利用されている。そのため、1台1台の端末の問題を解析するには多くの時間と工数がかかってしまう。

一方で、近年では機械学習のデータマイニングなどへの活用が進んでおり、多くの課題に一定の解決策を提供している。特に、教師なし学習は正解のデータを必要としないため、実際のシステム構築においては、正解データの収集と解析の負担を減らすことができる[Witten 2011]。本研究では、無線 LAN の電波情報から教師なし学習を用いて、同じ無線通信品質の問題を持つ端末を分類し、問題が発生している端末を自動的に特定する手法の初期検討を行ったので報告する。本検討では、品質悪化の現象を表現する素性を選択し、3種類の教師なし学習手法を比較評価した。その結果、混合分布モデル法が最も高い精度を得られることがわかった。

2. 無線通信品質の推定手法

無線通信は空間に放出される電波により通信が行なわれる。そのため、場所や時刻によって電波伝搬環境が刻々と変化し、それに伴って無線通信の品質も変化する。図 1 に無線 LAN における屋外の店舗に設置されたアクセスポイント近くでの受信電波強度の測定結果を示す。この図で示したように、受信電波強度は大きく変動し、一瞥しただけではその品質が良好なのか、正常に通信が行なわれているかを判別することは困難である。

これまで、無線通信の問題を解析するためには、個々の端末で送受信される無線通信のデータを収集し、受信電波強度や再送パケットなどの時系列解析を行っていた。しかし、このような

手法では多くの工数をかけて問題事象を再現し、特定の端末上で通信を試行させる必要がある。実際の現場において、問題の端末を特定し、その場で解析を行うことができれば問題解析の省力化を図ることができる。

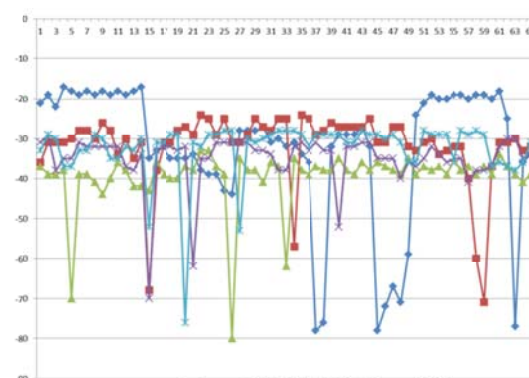


図 1. 受信電波強度の変化

無線通信の品質を推定する技術には多くの研究事例がある。樫原らは、無線品質の劣化を考慮したハンドオーバーを目的として MAC 層の再送回数を無線品質の判断に利用している[Kashihara 2007]。熊谷らは、上りスループットの値から下りに必要なスループットを推定し特定端末の優先順位を制御する手法を提案している[熊谷 2013]。関口らは、セル内の無線リソース割り当ての最適化を目的として、トラフィック増加量、接続端末数、伝送レート、パケットサイズから無線品質を推定し、セルの収容限界への影響を示している[関口 2013]。金子らは、無線メディアの最適な選択を目的として、NAV 時間の間隔を用いた Hurst 指数により無線品質を推定している[金子 2007]。これらの研究は、個々の無線パラメータの変動を時系列に解析することで無線品質を推定する方法であり、パラメータ間の関係が考慮されていない。パラメータ間の関連性を解析すると、パラメータの数が揃えるに従って組合せが多くなり計算コストの増大を招く。そのため、本研究では無線パラメータ間の関連性を考慮する手法として機械学習による推定手法を検討した。

3. 教師なし学習を使った無線品質のクラスタリング手法

機械学習手法には大きく分けて教師あり学習と教師なし学習がある。教師あり学習では、教師データによる入力と出力の正解ペアを用意することで新しい入力データに対して正しい出力を得ることができる。教師なし学習は、教師データではなく入力データのみが与えられ、何らかの基準により最適な出力を得る手法である。教師なし学習は、教師データを用意する必要が無いことからシステム構築と運用コストが比較的安く抑えることができる。本研究では、教師なし学習の複数のクラスタリング手法を用いて無線品質を分類することにより、問題が発生している端末を特定する手法を比較検討した。

クラスタリング手法にはこれまで多くの手法が提案されている[神嶋 2003]。その中でも、本研究では k-means 法、ファジイクラスタリング法、混合分布モデル法の 3 種類を比較評価した。

- k-means 法

本手法は、分割最適化法の 1 つであり、最も代表的なクラスタリング手法である。k-means 法では、クラスタの重心を代表点 c_1, c_2, \dots, c_k に設定し、式(1) の評価関数を最小化するようなクラスタへデータを割り当てる。

$$\sum_{i=1}^K \sum_{x \in C_i} \|x - c_i\|^2 \quad \dots (1)$$

- ファジイクラスタリング法

本手法は、ファジイ k-means 法とも呼ばれ、k-means 法におけるクラスタとデータの間データがそのクラスタに属する度合 g_{ik} を設定し、評価関数においてその度合を考慮する。クラスタ c_k を式(2)によって算出する。

$$c_{kj} = \frac{\sum_{i=1}^N (g_{ik})^m x_{ij}}{\sum_{i=1}^N (g_{ik})^m} \quad \dots (2)$$

- 混合分布モデル法

複数ある確率モデルの c 番目でデータ x が発生する確率密度関数を $f(x)$ とする。この確率の線形和を式(3)で表し混合分布と呼ぶ。混合分布モデル法は、観測データが与えられた時にこの $f(x)$ を求める。具体的には、混合分布モデルの構成要素を 2 つの正規分布と仮定し、式(4)のパラメータを推定する。観測データ x_1, x_2, \dots, x_N からパラメータ θ を求めるには最尤法を使い式(5)の $L(\theta)$ を算出する。

$$f(x) = \sum_{c=1}^K \alpha_c f_c(x) \quad \dots (3)$$

$$\theta = (\alpha_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \quad \dots (4)$$

$$L(\theta) = \sum_{i=1}^N \log \left\{ \sum_{c=1}^K \alpha_c f_c(x_i) \right\} \quad \dots (5)$$

4. 評価

4.1 素性の選択

一般的に、無線通信の問題は次の 3 種類に大きく分けることができる。①接続不可、②伝送速度不足、③頻りに切断である。これらの問題を解析するためには、接続に関する管理フレーム、コンテンツのデータフレーム、電波環境を反映するための情報を利用する必要がある。選択した結果の素性を表 1 に示す。無線品質を推定する研究では、受信電波強度を用いて電波環境を評価することが多い。しかし、今回は無線通信のキャプチャを用いるため、正確に端末での受信電波強度を取得することができない。通常、受信電波強度はデータレートに反映されるため、データレートを選択した。

表 1. 無線品質の素性

素性	説明
MAC アドレス	端末を特定する識別子
フレーム数	収集した端末毎のフレーム数
Authentication/ DeAuthentication フレーム数	認証フレームの数
Association/Reassociation フレーム数	接続フレームの数
Data フレーム数	データフレームの数
データレート	通信能力を示す伝送速度

4.2 評価手順

まず、無線 LAN において接続不可の問題が発生している状況の端末 2 台を用意した。これらの端末に対して無線通信のキャプチャを行った。両端末は静止状態であり、60 秒間のキャプチャを 50 回繰り返した。取得されたキャプチャデータには、対象の 2 端末以外にも周辺に存在する多くの端末のデータも同時に取得された。図 2 にサンプル回数毎の取得された端末数を示す。図からわかるように、接続端末数が時刻によって大きく変動しており、最小で 191 台から最大で 1,239 台までの端末が同時に通信していたことがわかる。

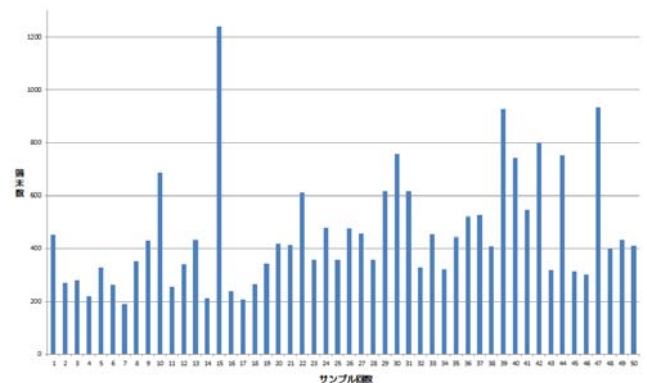


図 2. サンプル回数毎の端末数

次に、キャプチャデータから素性を抽出し、ベクトル化した上で標準偏差を使った正規化を行った。このベクトルを用いて先に述べた 3 種類のクラスタリング処理を実行した。実行環境には R を用いた。ファジイクラスタリングは e1071 パッケージを使い[Wien 2015]、混合分布モデルは mclust パッケージを使った[Fraley 2006]。k-means 法における繰り返し回数は 10、アルゴリ

ズムは Hartigan Wang を用いた。混合分布モデル法における混合分布の型には EII を用いた。

クラスタリング手法の評価は、図 2 に示すような多数の端末の中で同一の問題が発生している当該 2 台の端末が同じクラスタに含まれる割合を評価することで行った。つまり、クラスタリングの結果、当該 2 台の端末が同じクラスタに含まれる場合を正解として正解率を算出した。式(6)に正解率の定義を示す。

$$\text{正解率} = \frac{\text{対象端末 2 台が同じクラスタに含まれた回数}}{\text{総サンプル回数}} \dots (6)$$

4.3 評価結果

それぞれのクラスタリング手法におけるクラスタ数と正解率の関係を図 3 に示す。図の中で、“k-means”は k-means 法，“c-means”はファジイクラスタリング法，“mclust”は混合分布モデル法を指す。全体的には混合分布モデル法が最も性能が高い。次に正解率が高い手法は k-means 法、最後にファジイクラスタリング法と続く。混合分布モデル法が最も高い性能であるのは、無線キャプチャから得られた観測データが比較的正規分布に近い特性を持つためと思われる。

今回は分類されるクラスタ数を 1 から 10 まで変化させて評価を行った。当然ながら、クラスタ数が多いほどクラスタリングの性能は下がっている。先に述べた無線通信の 3 種類の問題に正常な通信状態のクラスタを加えて合計 4 クラスタ程度が無線通信の問題を分類するには最適なクラスタ数であると考えられる。そのため、図 3 のグラフからクラスタ数 4 では混合分布モデル法が k-means 法よりわずかに高い性能を示している。

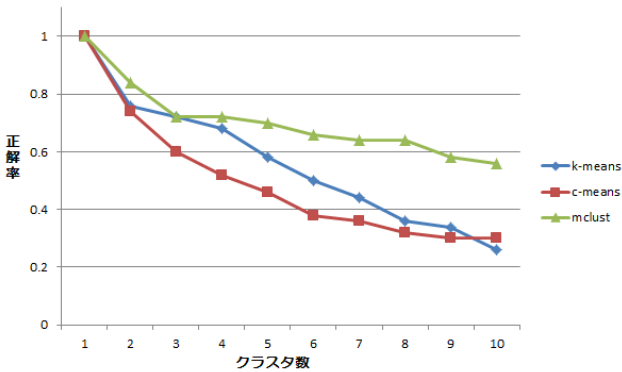


図 3. 各クラスタリング手法における正解率

5. おわりに

本稿では、無線端末の無線通信における問題を早期に解析するために、教師なし学習のクラスタリング手法を使って問題発生端末を特定する手法を評価した。実際の無線通信の問題から、クラスタリングに入力する素性は、接続処理に関する無線フレーム、データフレーム、電波環境に起因するデータを定義した。実際にこれらの観測データを用いて 3 種類のクラスタリング手法を評価したところ、問題の分類に最適なクラスタ数 4 において混合分布モデル法が最も高い性能を示した。これらから、無線通信の問題が発生している端末を特定するためには、混合分布モデルによるクラスタリング手法が適していると考えられる。

一方で課題も見えてきた。今回の評価では、ある 1 ヶ所のある時間帯における結果であり、いつでも観測データが混合分布モデルに有利な正規分布に近いとは限らない。電子レンジなどの干渉源が頻繁に動作しているような状況を含めた評価が必要である。さらに、今回は端末が静止状態で観測を行った。しかし、屋外の公衆無線 LAN などを使う利用者は歩行している場合なども考えられる。そのため、移動状態の端末における評価も必要である。次に、今回は無線のキャプチャ時間を 60 秒間としたが、実際の解析現場では時間がかかり過ぎである。問題発生頻度が高い場合には多数の端末が対象となり、短時間の特定が必要である。そのためには、利用する素性に端末に蓄積された過去の無線通信情報を加えることにより、問題発生時点で原因までが瞬時にわかるような工夫を検討する必要がある。

参考文献

[内藤 2016] 内藤他: モバイルネットワーク(未来へつなぐデジタルシリーズ), 共立出版, 2016.

[Witten 2011] I.Witten et. al.: Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2011.

[Kashihara 2007] S.Kashihara and Y.Oie: Handover management based on the number of data frame retransmissions for VoWLANs, Elsevier Computer Communications, Vol.30, No.17, pp.3257-3269, 2007.

[熊谷 2013] 熊谷他: 無線 LAN における NAV 期間の動的制御を用いた高優先度メディアアクセス制御, 信学技法 AN2012-73, 2013.

[関口 2013] 関口他: 周波数共有環境におけるアプリケーション運用品質向上に向けた無線 LAN の達成 QoS 予測手法, 信学技法 NS2013-47, 2013

[金子 2007] 金子他: コグニティブ無線における無線環境予測の可能性についての一検討, 2007 年電子情報通信学会総合大会 B-17-33, 2007

[神島 2003] 神島, データマイニング分野のクラスタリング手法, 人工知能学会誌 Vol.18, No.1, 2003.

[Fraleley 2006] C.Fraleley et. al.: MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering, Technical Report no.504, Department of Statistics, University of Washington, 2006.

[Wien 2015] T.Wien: Package e1071: Misc Functions of the Department of Statistics, Probability Theory Group, 2015