呼気データ解析による疾患判定

情報科学科　小山 裕太郎　　指導教員：作村 諭一

1　はじめに

人は健康維持のために予防措置を行ったり，今の健康状態の把握に努める．健康状態を知るためには，血液検査などの侵襲的な医療行為を伴った健康診断を受診するのが一般的である．しかし，受診者の身体的・心理的な負担は解決すべき問題である．

そこで，近年では非侵襲的な検査方法に関する研究が盛んに行われている．例えば，光学技術を用いた血液検査や唾液を用いた検査などがある．このような非侵襲的検査は，侵襲的検査と比較すれば受診者の負担を軽減できるのは明らかである．

本研究では，呼気ガスを用いて疾患判定を行うことを目的とし，肺がん・NASH (Non-alcoholic steatohepatitis: 非アルコール性脂肪肝炎)・肝がん・歯周炎の 4 つの疾患を扱う．

先行研究により，既に各疾患の判定において有効とするガス成分の抽出が行われた．しかし，アーティファクトな可能性の高いガス成分を含んでいることや，パラメータをデフォルト値のまま扱っているといった改善点が挙げられるため，その改善を行う．また，本研究からの新しい試みも行った．

2　呼気ガスデータ

本研究では，肺がんとそれ以外の疾患とでは，異なる呼気ガスデータを使用する．肺がん患者（n=107）と健常者（n=29）からは 59 種類のガス成分が，NASH 患者（n=67），肝がん患者（n=95），歯周炎患者（n=29），健常者（n=101）からは 16 種のガス成分が検出された．各疾患において，解析で扱うには不適切であるガス成分は除外し，肺がんは 20 種，NASH は 16 種全て，肝がん・歯周炎は 15 種のガス成分を用いる．ここで除外したガス成分には，先行研究では扱っていた（肺がんにおいて）アーティファクトな可能性が高いガス成分も含まれている．

また，両クラス間において不均衡データの問題があるため，解析時にはオーバーサンプリングを用いてデータ数を揃えた．

3　解析手法

疾患判定は，採取をした呼気データを用いて，サンプルを「疾患」か「健常」の 2 クラスのどちらか一方に分類することによって行う．呼気データは，ガス成分量から構成した多次元データであるため，多次元データの 2 値分類問題において高い分離性能を持つことで知られる Support Vector Machine（SVM）[1] を用いる．採取した呼気データと，それに対応する「疾患」と「健常」のラベルを学習データとして SVM に学習させた．

また，SVM はカーネル関数を併用するため，ガウシアンカーネルを採用した．SVM とカーネル関数はパラメータ設定が可能だが，先行研究ではデフォルト値であった．しかし，パラメータチューニングを行えば，サポートベクトルを減らし，分類境界を滑らかにできる．つまり過学習を抑えられる．よって，先行研究よりも過学習を抑えるためにパラメータチューニングも行った．

4　性能評価

SVM によって生成された判別器の汎化性能を評価するため，一個抜き交差検証を用いる．性能評価の指標には，真陽性率と真陰性率による精度を用いた（百分率）．

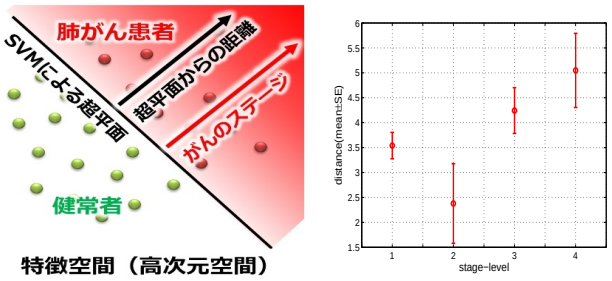
5　結果

先行研究の改善点を改善し解析を行った結果，各疾患において表 1 のような最大精度をもつ判定器を得た．ただし，表 1 の括弧内の数値は，先行研究で疾患判定に有効とされたガス成分での解析結果である．また，SV 数とはサポートベクトルのデータ数の平均と標準偏差を表している．

さらに，本研究からの新しい試みとして，肺がんにおいて，がんステージを推定するという研究も行った．図 1(a) に示すように，特徴空間における 2 クラスを分類する境界（超平面）からデータ点が離れるほどがんステージが増加し，がんが進行しているという，距離とがんステージの関係性の仮説を立てた．ガス 5 種での真陽性率 1 位において図 1(b) のようなステージ毎の距離の平均が得られ，仮説の関係性は十分にあると考えられる．

表 1　各疾患における精度 1 位の判定器（括弧内は先行研究のもの）

疾患	肺がん	NASH	肝がん	歯周炎
真陽性率	96.3% (98.1%)	97.0% (89.6%)	92.6% (77.9%)	69.0% (69.0%)
真陰性率	82.8% (82.8%)	93.1% (80.2%)	94.1% (92.1%)	95.0% (89.1%)
精度	93.4% (94.9%)	94.6% (83.9%)	93.4% (85.2%)	89.2% (84.6%)
成分数	9 種類 (11 種類)	5 種類 (7 種類)	5 種類 (7 種類)	6 種類 (7 種類)
SV 数	58.3±2.35 (122.4±2.6)	32.1±1.41 (119.3±2.2)	50.1±1.53 (133.7±1.8)	52.8±3.08 (142.7±3.0)



(a)　(b)

図 1　(a) 非線形写像による特徴空間での超平面からの距離とがんステージの関係性の仮説の図．(b) 肺がん患者の各ステージにおける，超平面からの距離の平均と標準誤差のグラフ．

6　おわりに

表 1 より，どの疾患においても先行研究よりもサポートベクトルの平均が半分以下となった．つまり過学習を大幅に抑えた上，より少ないガス成分で高精度な疾患判定器を作成できた．

がんステージと距離の関係性は，健常者側のクラスでも同様の考えにより，将来の疾患予測が可能なのではないかと考える．

参考文献

[1] C.M. ビショップ著，元田浩ほか監訳 (2012)『パターン認識と機械学習 下-ベイズ統計学による統計的予測-』丸善出版．