サポートベクタマシン分類とクラスタリング解析を用いた災害地名と自然特性の関係性評価

1210090　高橋頼良[1]

[1]高知工科大学　システム工学群　建築・都市デザイン専攻

E-mail 210090z@ugs.kochi-tech.ac.jp

日本は古くから自然災害の危険性を示唆する意味の込められた災害地名が多く受け継がれている.しかし,時代の流れ,市町村合併等により,本来の地名の意味を解釈することが困難になってきている.本研究は,サポートベクタマシン(SVM)分類とクラスタリング解析を用いて災害地名と災害箇所の関係性を評価し,災害地名の防災への利用促進を目的とする.

高知県高岡郡四万十町を対象に,標高,傾斜,曲率,尾根谷度を用いて SVM 分類を行い,地すべり箇所を抽出した結果,抽出箇所と災害地名との明確な関係性は確認出来ず,クラスタリング解析に関しても同様に関係性は見いだせなかった.今後は,特徴量,各パラメータの見直し等による精度向上,分類するクラスタ数の増減や比較手法の再検討等を行ったうえで,地名との関係性を改めて評価することが必要である.

Key Words : *CS 立体図, 尾根谷度, 小字*

1. 序論

1－1. 背景

古くから災害が多発していた日本には,災害を示唆するような名前を付けられた「災害地名」と呼ばれるものがある.例えば,「梅（ウメ)」や「栗（クリ)」といった植物地名はそれぞれ「埋める」「刳る」といった意味を持っていると考えられており,これらが含まれる地名は危険地帯の可能性があるとされる.東北大の花岡[1)]は先の東日本大震災の被災地でもある岩手県,宮城県を対に地名語尾と自然災害リスクとの間に一定の関係があることを明らかにした.高知工科大の石川[2)]は,機械学習により四国全域を対象とした地名と自然災害に関する研究を行い,機械学習によって抽出された災害分布と地名との間に明確な関連性を確認することはできなかったが,「古くから伝わる地名はその土地の文化や自然を表しているものが多い」ということを改めて確認している.

1－2. 目的

本研究は「災害地名」の存在を改めて確認しつつ,より高い精度で地名と自然特性の関連性を評価する.そしてその信憑性を再認識し,地名の防災への利用を促進することを目標とし,対象地域の地名と災害箇所の関係性を考察する.防災の指標に地名を組み込むことで,対象地域の住民一人一人がより身近に自然災害の存在を感じ取ることができるようになり,防災意識の向上につながるとともに,地域ごとの具体的な自然災害への対策が可能になると考える.

1－3. 本研究の概要と独自性

本研究は,土地の特徴が最もよく表現されていると言われる地名の小字・俗称データが多く整備されている高知県高岡郡四万十町を対象とし,サポートベクタマシン(SVM)による分類とクラスタリング解析の２つの手法で,自然災害と地名との関係性を分

析した.今回は精度向上,詳細化,を目指し,研究対象を地すべりと地すべりに関する災害地名に着目して研究を行った.地すべり地形図を教師データとし,様々な地形的特徴を特徴量として分類することで,現在確認されている地すべりエリアに加え,未確認の箇所や災害が発生する可能性のある地域も抽出できると考えた.本研究では標高,傾斜,曲率,地上開度と地下開度との関係から傾斜の値と独立した尾根か谷かを示す尾根谷度,地盤変形に植生が影響しうることを踏まえ対象地域の植生データ,これらを特徴量として検討した.加えて,機械学習の精度向上につながるとされるハイパーパラメータを,チューニング法の１つであるグリッドサーチを用いて最適化した.もう 1 つの手法としてクラスタリング解析を行った.クラスタリング解析とはデータ間の類似度に基づきデータをグループ分けする手法で,その中でも,非階層クラスタリングの k-means 法を用いた.k-means 法は分析するデータ量が多くても問題がなく,計算量の軽量化も図れるため手法として採用した.分類後,それぞれに含まれる災害地名読みの有無を確認,その結果とクラスタの特徴を照らし合わすことで関係性を評価した.

2. 使用データ

2−1. 地名データ

地名データは四万十町の小字データ（四万十町地名辞典[4]）を使用した.収録データはポインタデータで,四万十町の税務所管課からの土地台帳の基礎となるデータと,市町村の全図や地籍図を基に作成されたデータである.また, 地名本来の意味を解釈するため,小字・俗称の「読み」に着目する[3].

2−2. SVM 分類･クラスタリングに用いたデータ

SVM 分類及びクラスタリング解析に用いたデータは以下のとおりである.

１）教師データ

・地すべり：1:50,000 地すべり地形分布図
（取得元：防災科学技術研究所）

２）特徴量

・数値標高モデル（取得元：国土地理院）
・傾斜データ（QGIS 内の SAGA により DEM から算出）
・曲率データ（QGIS 内の SAGA により DEM から算出）
・尾根谷度（QGIS 内の SAGA により DEM から算出）
・植生データ
（取得元：環境省自然環境局生物多様性センター）

3. 研究手法

3−1. SVM 分類の手法

四万十町を対象に SVM による分類により地すべり地形の抽出を行った.Python の機械学習ライブラリ scikit-learn を用いた.教師データとした地すべり地形分布図は,地すべり箇所,地すべり箇所と推定される箇所,地すべり箇所かどうか判別できない山体・小丘,の３つに分類されている.今回は,発生箇所かどうかのデータが必要なためこれらを,クラス１：地すべり箇所,クラス０：未発生箇所,に分類した.地すべり箇所,地すべり箇所と推定される箇所をクラス１とし,判別できない箇所は未発生箇所としてクラス０に分類した.分類の際の特徴量には,標高データ,傾斜データ,曲率データ,及び,尾根谷度,植生データを用いた.しかしそれぞれの生データは,次元によって数値の単位がそろっておらず,あるいは大きさが極端に異なる.そのため,そのままデータを使用すると各次元を見比べた際にそれぞれの関係が分かりにくく,機械学習の際に誤差が生じてしまう原因にもなりうる.そこで,標準化を行い,次元のスケールを統一したそれぞれのデータを用いた.加えて,SVM を使用する際に,誤分類を許容する指標のコストパラメータ：C と RBF カーネルのパラメータ：γの最適な値をグリッドサーチによって求め,その値を用いて機械学習を行った.また,分類器のトレーニングデータを取得した位置だが,地すべり地形の分布が少なく,１カ所では不足と考え,２０００ｍ×２０００ｍで地すべり地形が多く含まれている上位５つの小範囲をトレーニングデータとした.

3－2. クラスタリング解析の手法

四万十町の地名のポイントデータ,標高データ,傾斜データ,曲率データ,尾根谷度を用い,非階層クラスタリングの k-means 法を用いて分類を行い,分析を行った.四万十町の地名ポイントからそれぞれ 30m のバッファを生成し,そのバッファに含まれる標高,傾斜,曲率,尾根谷度,の平均,最大,最小を QGIS の地域統計を用いて集計し,その情報を属性としてポイントデータに加えた.その後,加えた変数を標準化,k-means 法によるクラスタリングを行い,分類された結果と元の地名データを結合.分類された各グループの地形的特徴をグラフにて確認しつつ,各グループ内の災害地名の有無,関係性の評価を行った.

4. 分類結果と考察

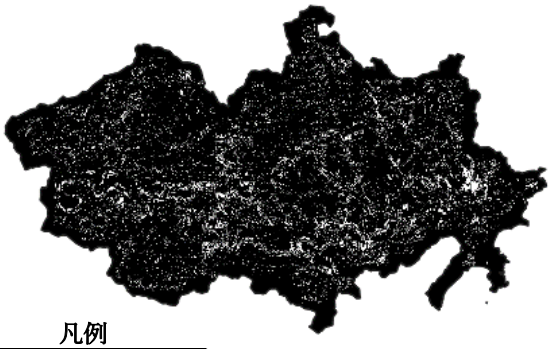
4－1. SVM 分類の結果と考察

まず,四万十町内の 5000m×5000mの範囲をテスト範囲とし,特徴量の最良の組み合わせを求めるため,再現率と適合率の調和平均である F 値を指標として用いて分類結果を比較した.その結果を表‐1 に示す.

表‐1　F値による比較

	地すべり	
	特徴量	F値
ver01	標高、傾斜、曲率	0.715
ver02	標高、傾斜、曲率、尾根谷度	0.958
ver03	標高、傾斜、曲率、植生	0.885
ver04	標高、傾斜、曲率、尾根谷度、植生	0.879

結果から,ver02 の組み合わせを採用し,そのモデルを四万十町全域に適用した.しかし,5000ｍ×5000m のテスト範囲ではうまく抽出できたが,四万十町全域に適用すると地すべり箇所としてうまく分類されず,災害発生箇所がほぼ 0 という結果が出た.そのため,訓練データと各テストデータの特徴量のヒストグラムを作成し比較したところ,テスト範囲が広範囲になるにつれて訓練データとの差が大きくなることがわかった.そのため,四万十町を 1000m×1000ｍの細かいグリッドで分割し,各グリッドに対してそれぞれ分類を行った.分類結果より,地すべり箇所と判断されたメッシュを図‐1 に示す.抽出結果として,F 値の高いものを用いたが,地すべり箇所であるにも関わらずクラス 0（未発生箇所）と判別される箇所が多くあり,クラス 1（地すべり箇所）と判別された箇所が多すぎるように見受けられた.未確認の地すべり箇所の可能性も考えられるため,一概に間違っているとは言えない.今後はさらなる特徴量の見直し,グリッドサーチ以外の手法によるパラメータの最適化等を行うなど,精度向上を図りつつ現地調査なども視野にいれる必要があると考える.



図‐1　機械学習による地すべり箇所の抽出結果

4－2. クラスタリングの結果と考察

10 個のクラスタに分類した後,各クラスタにおいて標高,傾斜,曲率,尾根谷度の傾向を確認するためそれぞれをグラフに表した.結果として,それぞれのクラスタに大きな特徴は見られなかった.それぞれの変数を標準化してクラスタリングを行ったが,それによりそれぞれのクラスタの特徴がわかりにくくなってしまった可能性が考えられる.また,今回はクラスタ数を 10 個と指定したが,クラスタ数を増減させることでクラスタごとの特徴が顕著にみられるかもしれない.今後はクラスタ数を増減させて複数のパターンを試すことに加え,標準化後の変数の比較方法の検討,k-means 法以外のクラスタリング法での分類等が必要だと考える.

5. 自然災害と地名との関係性評価

SVM による分類の結果,クラスタリング解析の結果から,地名との関係性をそれぞれ評価した.小川[3)]の著書から,地すべり地形に多く見られる災害地名の読みを選定し評価に用いた.評価手法は以下の２通りである.

手法１）災害地名の読みが含まれる地名と含まれない地名ポイントからそれぞれ 100m のバッファを生成し,そのバッファに含まれる災害箇所のメッシュ数の平均,標準偏差を計算し比較する.

手法２）クラスタリング分析によって分類した各クラスタに含まれる地名を確認し,災害地名の読みが多く含まれたクラスタの有無,そのクラスタの地形的特徴を確認し関係性を評価する.

表 - 2　地すべり地形に多く見られる災害地名読み

アス	アズ	アブ	ウサ	ウシ	ウト	ウド	ウメ	オシ	カキ
カギ	カケ	ガケ	カネ	カノ	カヤ	クス	クリ	コヤ	スキ
スギ	ズキ	スク	スグ	ホカ	ホキ	ボケ	スケ	スゲ	

6. 結果と考察

表 - 3,表 - 4 に分析結果を示す. 石川[2)]が用いた地名情報は主に大字を扱ったデータであったが,本研究では四万十町に着目することで,大字より地形の特徴がよく表現されているといわれる小字の詳細なデータを多く用いた.そのため,評価手法１で石川[2)]が 1km バッファを用いていたものを,四万十町の豊富な小字データを活かし,バッファを 100m にすることで災害地名読みが含まれる地名と含まれない地名をより部分的に比較できると考えた.しかし,石川[2)]の四国全域での結果と大差はなかった.地すべり箇所抽出の段階で教師データに無い未知の地すべり箇所がかなり多く抽出されたこと,教師データでは発生箇所となっているが抽出結果では未発生箇所となっている箇所があったことが原因と考える.加えて,四万十町は小字データが多くあるがその分土地面積が小さく,教師データの地すべり分布のパターンが不足していた可能性も原因の１つと考えられる.

評価２では各クラスタに含まれる災害地名読みの割合に大差はなかった.ただ,各クラスタに分類された地名数に差があったため,クラスタ数の増減により結果が変化する可能性も考えられる.

表 - 3　評価手法１の結果

	地すべり箇所（個）	標準偏差
災害地名読みあり	2.155	2.705
災害地名読みなし	2.037	2.443

表 - 4　評価手法２の結果

	含まれる災害地名（読み）の数｛割合｝
ID = 0	190（約9.4%）
1	11（約5.3%）
2	30（約9.1%）
3	86（約11.2%）
4	13（約5.9%）
5	83（約10.7%）
6	52（約7.5%）
7	68（約7.1%）
8	34（約5.9%）
9	4（約7.7%）

7. まとめ

地名と自然災害の関係性をより高い精度で再評価することで,地名の防災への利用を促進することを目的としたが,SVM による分類によって抽出された四万十町の地すべり箇所の分布と小字との明確な関連性は確認できず,クラスタ分析に関しても同様に,地名との関連性は確認できなかった.日本各地の地名に含まれるその土地の特徴や危険性を示すメッセージを読み解き今後に活かすためにもさらなる研究が必要だと考える.

8. 参考文献

１）花岡和聖：小地域地名の語尾と自然災害リスクの関係性評価,歴史都市防災論文集 Vol.9,pp.57-64,2015

２）石川惠大：GIS を用いた地名と自然災害の関係性評価,高知工科大学,社会気象工学研究室,pp.1-4,2020

３）小川豊：あぶない地名,三一書房,2012

４）四万十町地名辞典：https://www.shimanto-chimei.com