

rankSVM を用いた不均衡な訓練データからの機械学習と地震被害検知への応用

山梨大学 正会員 ○宮本 崇

山梨大学 非会員 丸山 諒

1. はじめに

土木工学上における機械学習の重要な適用先の一つに、構造物の変状や災害時の都市の被害など、対象事物の異常を発見・検知するものがある。このような異常検知のタスクは、機械学習の枠組みでは教師あり学習による分類問題として捉えられることが多い。

分類問題に対応する機械学習モデルの訓練には、一般には各カテゴリに属するデータの数均等な訓練データを用いることが望ましいが、異常検知に類するタスクでは正常データに比べて異常データの取得が難しいことが多い。カテゴリ間でデータ量が均一でないデータは不均衡データと呼ばれ、不均衡データをそのまま訓練データに用いることは機械学習モデルの深刻な性能低下につながるため、その対処は実用上重要となる。

データ量が不均衡となる異常検知の問題に対しては、正常データの間引きによりデータ数を調整するアンダーサンプリングの手法や、正常データのみから学習を行う 1 クラス分類として問題を捉える手法が用いられる。これらの手法は、それぞれ正常データの一部や異常データをモデルの学習時に取捨するものと言えるが、そのようなデータ数の調整を行わず不均衡データをそのまま学習に利用できる機械学習モデルとして rankSVM の利用が近年に提案されている¹⁾。そこで、本研究では rankSVM の性能を数値実験によって確認すると共に、著者の研究する人工衛星撮影画像からの地震被害検知の問題に応用し、その有用性を検証する。

2. rankSVM の概要

rankSVM は、ウェブ検索結果の最適化などを行うランク学習と呼ばれるタスクに対応する機械学習モデルである。正常データと異常データの 2 クラス分類問題への rankSVM の応用は、次のように定式化される。

今、正常データ $N=\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ と、異常データ $P=\{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_m\}$ が与えられるとき、この 2 クラスの分類問題は任意の $\mathbf{x}_i$ と $\mathbf{x}'_j$ に対して以下の式が成り立つような評価関数 $f(\cdot)$ によって解くことができる。

$$f(\mathbf{x}_i) > f(\mathbf{x}'_j) \quad (1)$$

rankSVM は、任意のカーネルに対応する特徴変換写像 φ と線形変換によって

$$f(\mathbf{x}) = \mathbf{w} \cdot \varphi(\mathbf{x}) \quad (2)$$

として評価関数 $f(\cdot)$ を構成したうえで、以下の最適化問題を解くことによって $\mathbf{w}$ の最適値 $\mathbf{w}_{\text{opt}}$ を得る。

$$\mathbf{w}_{\text{opt}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{(i,j)} l(\mathbf{w} \cdot \varphi(\mathbf{x}_i) - \mathbf{w} \cdot \varphi(\mathbf{x}'_j)) + \frac{1}{2} \|\mathbf{w}\|^2 \quad (3)$$

$$\text{where } l(t) = (\max\{0, 1 - t\})^2 \quad (4)$$

(3)式の解として得られる rankSVM は、次のような 2 つの望ましい性質を有している。まず、rankSVM におけるモデルの学習を表す(3)式では、正常データ $\mathbf{x}_i$ と異常データ $\mathbf{x}'_j$ が常にペアとなって、訓練時には均等な頻度で現れる。したがって、特定のカテゴリのデータが訓練時に多く用いられることで起こる学習の偏りが原理的に生じない。このため、訓練データが不均衡である場合でもデータの取捨を行うことなく、容易に取得可能な多量の正常データや少数の異常データの情報を有効に活用することができる。

また、(3)式の最適化は、分類性能の評価指標の 1 つである ROC-AUC の最大化と等価とであることが知られている²⁾。高い ROC-AUC を有する機械学習モデルは、低い空振り率 (False Positive Rate) の下で高い異常検知率 (True Positive Rate) を実現する、異常検知を行う機械学習モデルとして重要な性質を有している。

キーワード 不均衡データ, 災害異常検知, 機械学習, rankSVM

連絡先 〒400-8511 山梨県甲府市武田 4-3-11 山梨大学 宮本 崇 tmiyamamoto@yamanashi.ac.jp

3. 数値実験

不均衡データに対する rankSVM の分類性能を確認するために、MNIST データセットにおける数字の 4 に対応する画像データと 9 に対応する画像データからなる不均衡な訓練データを表 1 のように構成する。この訓練データセットを用いて学習を行った分類器の性能を、同表に示すテストデータを用いて検証した。

テストデータに対する rankSVM の性能を、2 クラス分類の機械学習モデルであるニューラルネットワーク (NN) と SVM, 1 クラス分類の機械学習モデルである One Class SVM と比較した結果を図 1 に示す。図には、アンダーサンプリングによる訓練データ数の調整を行った場合と行わなかった場合の各モデルの性能を示している。rankSVM の性能が総じて高いことが分かるほか、rankSVM を利用した場合には、アンダーサンプリングを行ってデータ数を調整したケースよりも、不均衡データをそのまま学習に利用したケースの方が高い分類性能を有している。すなわち、多量の正常データと少数の異常データの情報を学習に利用することで、データを取捨するよりも高い性能を得られることが分かる。

表 1 数値実験に用いたデータセットの構成

	“4”を表す画像	“9”を表す画像
訓練データ数	100	5
テストデータ数	100	100

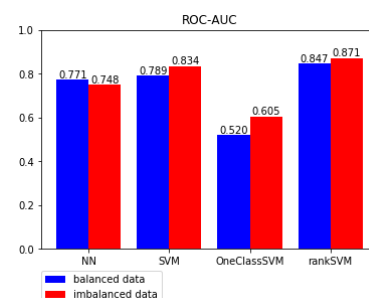


図 1 ROC-AUC による性能の比較 (MNIST)

4. 地震被害検知への応用

次に、著者らの研究する人工衛星撮影画像から住宅構造物の被害検知を行う問題に対して、rankSVM の適用性を検証した。光学衛星による広域の撮影画像から住宅 1 棟単位の撮影部を切り出した小画像片に対して、全壊・無被害の 2 クラス分類を行う問題を考える。このとき、図 2 に示すように、著者の既往の研究で構築した学習済みの畳み込みニューラルネットワークにおいて全結合層を SVM による分類器に置換し、表 2 に示す不均衡データから分類器の訓練と性能の検証を行った。分類器には、前節と同様に 3 種類の SVM を使用し、アンダーサンプリングによりデータ数が均衡になるよう調整した場合と、不均衡データをそのまま学習に利用した場合の 2 つのケースを比較した。

図 3 は、ROC-AUC の観点から各分類器の性能を比較したものである。rankSVM を用いて不均衡データからの学習を行うことにより、わずかながら置換前の NN よりも分類性能が向上しており、実用上の分類タスクにおいても rankSVM の適用性が高いことが示唆される。

表 2 地震被害検知における画像データの構成

	倒壊住宅	非倒壊住宅
訓練データ数	3,880	1,304
テストデータ数	163	163

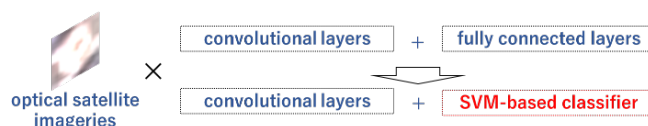


図 2 深層学習と SVM を併用した機械学習モデル

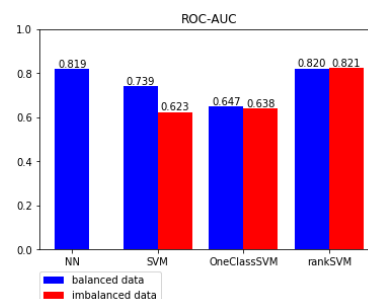


図 3 ROC-AUC による性能の比較 (衛星画像)

5. おわりに

本稿では不均衡データに対する分類モデルとして提案されている rankSVM の有用性を、著者らの研究する人工衛星撮影画像から地震被害の検知を行う問題の中で検証した。rankSVM の利用により、わずかな異常データと多量の正常データが得られる場合でも有効にデータの活用ができると期待される。

参考文献

- 1) R. Cruz et al.: Tackling Class Imbalance with Ranking, 2016 IJCNN, pp.2182-2187, 2016.
- 2) O. Chapelle et al: Efficient Algorithms for Ranking with SVMs, Information Retrieval, 13(3), pp.201-215, 2010.