データサイエンス　第 2 回レポート課題

工学系研究科　社会基盤学専攻　修士 1 年　37-176011　梶原裕希(Name:37176011)

1．タスクの詳細

●タスク名：Human Activity Recognition Using Smartphones Sensor Data

2．試した方法、性能を向上させるためのアイデア・工夫点など

●手法選択の経緯

(1) 与えられた教師データの特徴把握

配布された X_train.csv、y_train.csv の特徴として、「サンプルデータ数が十分に多い」、「目的変数であるカテゴリが 3 つ以上存在する」の 2 つが挙げられる。前者については、表 1 の通り、全体で 5000 以上のデータ、カテゴリ別でも 690 以上のデータが保証されている。よって、今回は、データ数が多い場合にも効率的な学習が可能であり、マルチクラス分類が容易である**ランダムフォレスト**を用いて分類を行うことを検討する。

表 1：データ数

Category	Number
1	868
2	747
3	691
4	871
5	951
6	952
Total	5080

(2) 次元削減による各行動カテゴリの分布調査

ランダムフォレストの欠点は、複雑なデータを用いると汎化性能が落ちることである。そこで、今回使用するデータの複雑さを評価するために、説明変数の次元削減を行い各行動カテゴリの分布を見た。その結果を図 1～図 7 に示す。図 1 によれば、各データの分布は動的行動カテゴリ(WALKING, WALKING UPSTAIRS, WALKING DOWNSTAIRS)と静的行動カテゴリ(SITTING, STANDING, LAYING)に大きく分けられることが窺える。また、図 2~図 7 によれば、各データはカテゴリごとに一定のまとまりをもって分布していることが分かる。このことから、サンプルデータの説明変数の値は、カテゴリごとに一定の傾向を示している点でさほど複雑ではないと判断できる。

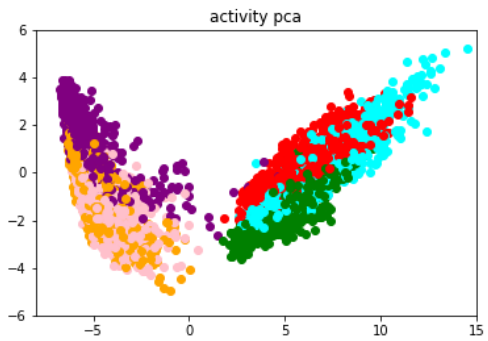

図 1：　次元削減結果と各カテゴリの分布

【図 1 凡例】

赤色：WALKING

緑色：WALKING UPSTAIRS

水色：WALKING DOWNSTAIRS

橙色：SITTING

桃色：STANDING

紫色：LAYING

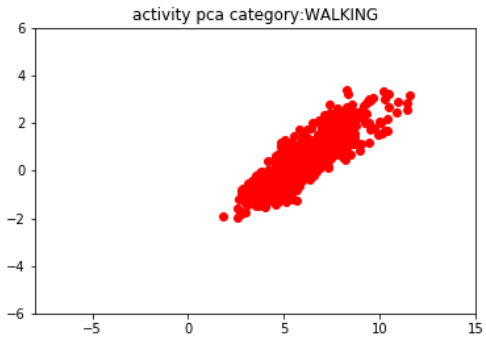

図 2：「WALKING」の分布

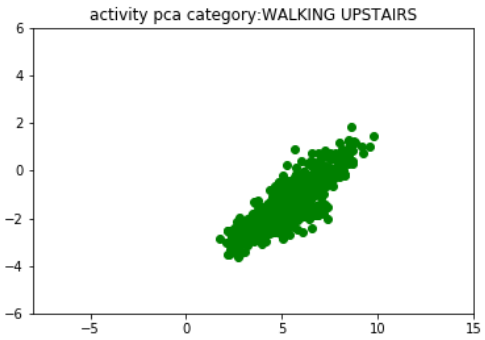

図 3：「WALKING UPSTAIRS」の分布

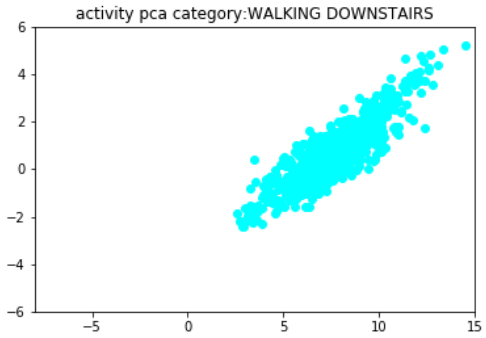

図 4：「WALKING　DOWNSTAIRS」の分布

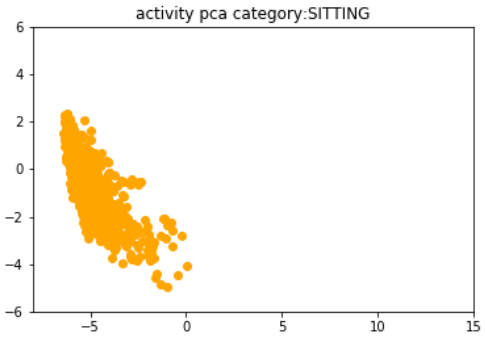

図 5：「SITTING」の分布

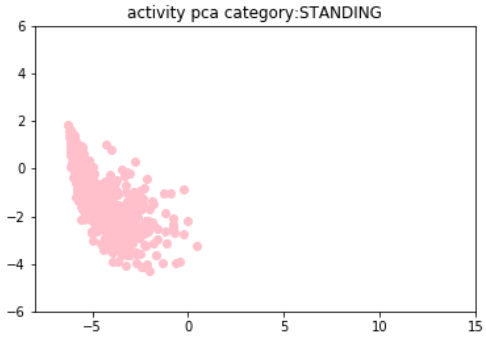

図 6：「STANDING」の分布

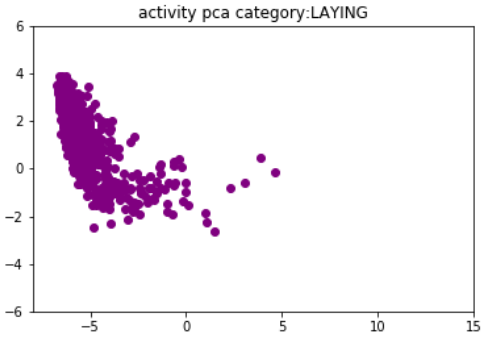

図 7：「LAYING」の分布

(3) ランダムフォレストとその他の手法の分類正確性比較

今回のデータセットの場合、ランダムフォレストがその他の手法に対して優位であることを検証するために、ランダムフォレストと k 最近傍法、SVM の 3 手法の訓練誤差とテスト誤差を比較した。具体的には、3 手法とも 5080 のデータのうち 4000 のデータを学習に用い、残りのデータのうち最初の 100 のデータをテストデータとして、訓練データとテストデータについて分類の正解率を求めた。その結果を表 2 に示す。これによれば、k 最近傍法と SVM と比較した場合、ランダムフォレストの方がより正確に分類を行っていると言える。

表 2：3 手法の訓練誤差とテスト誤差

Method	knn	svm	randomforest
Training accuracy	1.00	1.00	1.00
Validation accuracy	0.87	0.91	0.95

●性能を向上させるためのアイデア・工夫点など

(1) 訓練データ数の調整

ランダムフォレストの性能を向上させるために、学習に用いる訓練データ数の調整を行った。訓練データ数を 100 から 5000 まで 100 刻みで変化させ、残りのデータをテストデータとした時のテスト分類正解率を計算した。図 8 にその結果を示す。この図によれば、訓練データ数を増加させても、汎化性能が大きく落ちることはなく、正確な予測を行うことに成功していることが窺える。今回は、最も正解率の高かった訓練データ数 5000 のモデルを用いることにした。

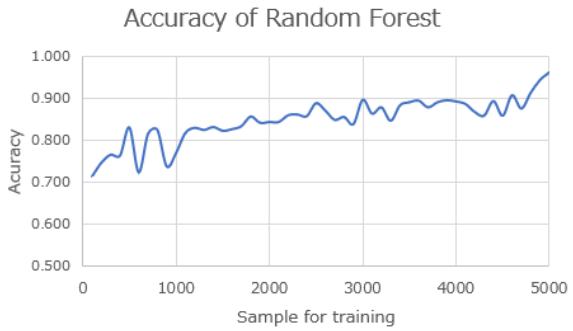

図 8：訓練データ数を変化させた時の分類正解率

3. 結果と考察

●結果

(1) 最終的に選択された手法

手法：ランダムフォレスト

訓練データ数：5000

テストデータ数：80

(2) Test error と Validation error

Test error：0.9994, Validation error：0.9625

(3) スコアリングサーバに提出した結果

Accuracy：0.88444444444444

●考察

(1) 各説明変数の重要度

ランダムフォレストの利点の一つは、分析後に目的変数を予測する上での各説明変数の重要度を把握できることである。得られた全説明変数の重要度の値を降順に表したものが図 9 である。このうち、特に重要度の高かった上位 4 つの説明変数を表 3 に示した。図 9 によれば、561 の説明変数のうち、目的変数に大きな影響を与えているのはせいぜい 100 程度であることが分かる。また、表 3 から、目的変数に比較的関連の深い説明変数は、降順に「X 方向の重力加速度の平方和を値で割ったもの」「X 方向の重力加速度の最小値」「X ベクトルと重力平均ベクトルとの間の角度」「Y 方向の体の加加速度の中央絶対偏差」であることが把握できる。

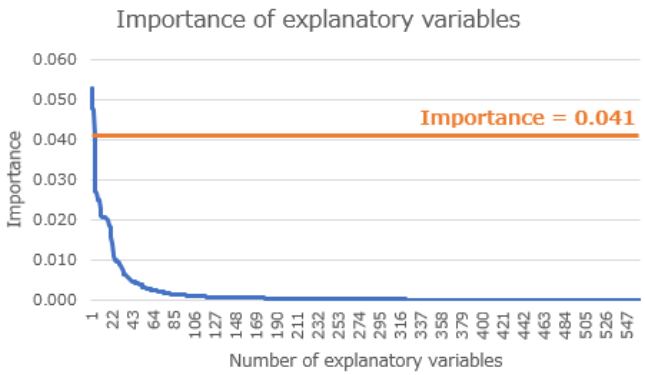

図 9：各説明変数の重要度(降順)

表 3：重要度の高い説明変数

No	Detail	Importance
57	tGravityAcc-energy()-X	0.05274845829361
53	tGravityAcc-min()-X	0.04778437232463
559	angle(X,gravityMean)	0.04769271702271
88	tBodyAccJerk-mad()-Y	0.04125934624571

(2) 今後の課題

今回の分析にはランダムフォレストを用いた。この手法で更に精度を高めるためには、今回考慮しなかったパラメータである「作成する決定木の個数」「作成する決定木の深さ」「サンプリングする目的変数の個数」を調整することが必要であると考えられる。

4. 講義の感想、要望など

・講義資料が充実しており、予習・復習の際に大変役立ちました。今後も python を使った分析を行う際に参照していきたいです。

・前回のレポート課題の振り返りにもありましたが、この講義の内容は初学者には少し難しいかなと思います。