R 言語を用いた決定木構築に関する実験レポート

5419080 関一樹
日本大学文理学部情報科学科

概要

本実験レポートでは, まず最初に決定木に関する知識及び手法について説明を行う. その後, 与えられたデータに対し決定木の構築を行う. 最後に得られた結果に関して考察を行う.

1 目的

本実験レポートは,R 言語を用いたデータの分析を通じ, 決定木の考え方を確認・復習すること, および具体的な操作を習得することを目的とする.

2 理論

通常, データを分類する目的の一つとして未来の事象を予測する目的があるが, データを分類する手法は多々存在する. その中でも本実験レポートでは決定木と呼ばれる手法を採択し, データの分類及び予測を行う.

決定木とは, データを段階的に分類を行い入力データを分類する手法である. 決定木の特徴として, 段階的に分類を行う事より分類結果の直感的な理解が可能である利点がある. しかし一方で, 分類を行う際に 1 つの変数しか考慮出来ない点がある.

2.1 分割基準

前述した通り, 決定木はデータを段階的に分割するが, 分割を評価する手法として, 情報利得とジニ係数を用いた分割指標がある.

2.1.1 情報利得

情報利得とは, 分割の前後での平均情報量について着目する手法である. 情報量とは, ある集合内の事象 E が滅多に発生しない事象なのか否かについて着目した値であり, 下式で算出される.

$$log_2 \frac{1}{P_D(E)} = -log_2 P_D(E)$$

上式より確率の逆数を対数に取っていることから, 滅多に起きない事象の際, 得られる情報量が大きくなることが読み取れる. この時の平均情報量は下式で表せされる.

$$H(D) = \sum_{c \in C} P_D(c) log_2 \frac{1}{P_D(E)} \left(= -\sum_{c \in C} P(c) log_2 P_D(E) \right)$$

前述した通り, 情報利得は, 分割前の平均情報量から分割後の各平均情報量の重み付き平均の差から得ることができる. 即ち, あるデータ集合 D について属性 A の情報利得は下式で表られる.

$$H(D) - \sum_{a \in A} P_D(a) H(D_a)$$

2.1.2 ジニ係数

ジニ係数とは, あるデータ集合 D のクラス分布について着目する手法である. 特に, 集合 D について取り出した 2 事例が同じクラスに属さない確率について着目する. 即ち, 違うクラスである確率が小さい時ジニ係数も小さくなり, 良い分割が行えている評価を得られる. また,2 事例が同じクラスに属さない確率は,2 事例が同じクラスに属する確率の余事象であることから確率は下式で表される.

$$G(D) = 1 - \sum_{c \in C} P_D(c) * P_D(c) \left(= 1 - \sum_{c \in C} P_D(c)^2 \right)$$

以上より, ジニ係数は異なるクラスである確率の期待値より求められることから, 属性 A に関するジニ係数は以下の通りになる.

$$\sum_{a \in A} P_D(a) G(D_a)$$

以上より, 上記の評価指標を用いることでより高精度の決定木の構築が可能となった. しかし, 決定木構築の終了条件の一つとして, 分割後のデータが同一クラスに属する場合構築が終了するが, ここでデータの中に外れ値が存在する場合を考える. すると, 外れ値を分類するためにデータの分割が行われるが, 決定木全体としては冗長な決定木が構築され過学習する恐れがある. そこで, 枝刈りと呼ばれるある基準を設定することで汎化能力の向上を行う手法がある. 枝刈りには大きく事前枝刈り, 事後枝刈りの 2 種に分けられる.

2.1.3 事前枝刈り

事前枝刈りとは, データの最低事例数を定めることで決定木の過学習を軽減する手法のことである.

2.1.4 事後枝刈り

事後枝刈りとは, 決定木が生成された後に決定木の簡略化を行う手法のことである. また, 事後枝刈りについて各手法が存在するが, 本実験レポートではコスト複雑度枝刈りと呼ばれる手法を採択した.

コスト複雑度枝刈りとは, 決定木の精度と複雑度について着目する手法である. ノード t での誤分類数を $M(t)$, データ数を N, 葉の個数を $|\tilde{T}|$ とすると下式の通りに表される.

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$
$$R(T) = \sum_{t \in \tilde{T}} R(t) \frac{M(t)}{N}$$

上式ではでは R(T) で部分木の精度,$\alpha|\tilde{T}|$ で部分木の複雑度について算出していることから, 一般に値が小さいほど, 高精度かつ簡潔な決定木である評価を得られる. また, 得られた値が大きい場合, 部分木が不正確または複雑である評価を得られる.

3　実験手法

本実験レポートでは UCI Machine Learning Repository[1] より入手した zoo.csv データについて, 情報利得及びジニ係数を用いたクラス Type に関する決定木構築を行う.zoo データの各属性については付録に記載した.

また構築した決定木の評価指標として χ^2 検定を採択した. 有意水準 α は 0.05 とし, 帰無仮説を「予測クラスと実際のクラスは独立である」とし, 対立仮説を「予測クラスと実際のクラスは独立でない」とし, 評価を行う. また分割基準や事前枝刈りや事後枝刈りの有無による結果に対して考察を行う. (本実験レポートでは χ^2 検定の際,0 除算を回避する手法として, クラス予測の際に 10^{-10} の値を加算する.)

4　実験結果・考察

4.1　情報利得

まず情報利得を用いた決定木構築を行う. 以下に R を用いて決定木構築を行なった結果を示す.

- 分割基準及び枝刈りなしの場合

	zoo.predict						
	amphibian	bird	fish	insect	invertebrate	mammal	reptile
amphibian	0	0	0	1	0	0	0
bird	0	6	0	0	0	0	0
fish	0	0	3	0	0	0	0
insect	0	0	0	2	0	0	0
invertebrate	0	0	0	4	0	0	0
mammal	0	0	0	0	0	5	0
reptile	0	0	0	4	0	0	0

この時の χ^2 検定の p 値が $0.0001475 < 0.05$ より, 帰無仮説を棄却し, 対立仮説を採択する. また, この時の精度は 64% である.

- 分割基準あり, 枝刈りなしの場合

	zoo.test.predict						
	amphibian	bird	fish	insect	invertebrate	mammal	reptile
amphibian	0	0	1	0	0	0	0
bird	0	6	0	0	0	0	0
fish	0	0	3	0	0	0	0
insect	0	0	0	2	0	0	0
invertebrate	0	0	0	4	0	0	0
mammal	0	0	0	0	0	5	0

reptile	0	0	3	1	0	0	0

この時の χ^2 検定の p 値が $0.0006447 < 0.05$ より, 帰無仮説を棄却し, 対立仮説を採択する. また, この時の精度は 64% である.

- 分割基準なし, 枝刈りありの場合

	zoo.test.predict						
	amphibian	bird	fish	insect	invertebrate	mammal	reptile
amphibian	0	0	0	1	0	0	0
bird	0	6	0	0	0	0	0
fish	0	0	3	0	0	0	0
insect	0	0	0	2	0	0	0
invertebrate	0	0	0	2	2	0	0
mammal	0	0	0	0	0	5	0
reptile	3	0	0	1	0	0	0

この時の χ^2 検定の p 値が $8.75\text{e-}08 < 0.05$ より, 帰無仮説を棄却し, 対立仮説を採択する. また, この時の精度は 72% である.

- 分割基準及び枝刈りありの場合

	zoo.test.predict						
	amphibian	bird	fish	insect	invertebrate	mammal	reptile
amphibian	1	0	0	0	0	0	0
bird	0	6	0	0	0	0	0
fish	0	0	3	0	0	0	0
insect	0	0	0	1	1	0	0
invertebrate	0	0	0	0	4	0	0
mammal	0	0	0	0	0	5	0
reptile	2	0	1	0	1	0	0

この時の χ^2 検定の p 値が $4.094\text{e-}07 < 0.05$ より, 帰無仮説を棄却し, 対立仮説を採択する. また, この時の精度は 80% である.

以上より, いずれも χ^2 検定より, 予測クラスと実クラスとは独立でない事が得られ, 特に, 分割基準と枝刈りを行なう決定機構築が, 最も精度が高い分類を行う事が得られた. 一方枝刈りを行わなかった 2 つの構築結果について, 双方とも最も精度が低い結果が得られたことより, 情報利得を用いた決定木構築において枝刈りを行うことで汎化能力を高めている事が推測される.

4.2 ジニ係数

次にジニ係数を用いた決定木構築を行う. 以下に R を用いて決定木構築を行なった結果を示す. 分割基準を用いない場合は上記の情報利得にて前述しているため省略する.

- 分割基準あり, 枝刈りなしの場合

	zoo.test.predict						
	amphibian	bird	fish	insect	invertebrate	mammal	reptile
amphibian	0	0	0	1	0	0	0
bird	0	6	0	0	0	0	0
fish	0	0	3	0	0	0	0
insect	0	0	0	2	0	0	0
invertebrate	0	0	0	4	0	0	0
mammal	0	0	0	0	0	5	0
reptile	0	0	0	4	0	0	0

この時の χ^2 検定の p 値が $0.0001475 < 0.05$ より, 帰無仮説を棄却し, 対立仮説を採択する. また, この時の精度は 64% である.

- 分割基準及び枝刈りありの場合

	zoo.test.predict						
	amphibian	bird	fish	insect	invertebrate	mammal	reptile
amphibian	0	0	0	1	0	0	0
bird	0	6	0	0	0	0	0
fish	0	0	3	0	0	0	0
insect	0	0	0	2	0	0	0
invertebrate	0	0	0	2	2	0	0
mammal	0	0	0	0	0	5	0
reptile	3	0	0	1	0	0	0

この時の χ^2 検定の p 値が $8.75\text{e-}08 < 0.05$ より, 帰無仮説を棄却し, 対立仮説を採択する. また, この時の精度は 72% である.

以上よりジニ係数による分割結果を得られたが, いずれも情報利得を用いた決定木構築の結果と酷似している. 考えられる原因としては 2 つ推測できる.1 つ目は, データ数に関する原因である. 理論にて記述したが, データ数が少なくなると決定木は過学習を起こすため, 事前枝刈りで適切な設定が行えていない可能性がある. 加えて, 本実験レポートでは 1:3 の割合でデータを割り当てたが, 前述した通り適切なデータの割り振りではない可能性がある.2 つ目は分割基準についてである. 情報利得及びジニ係数は双方とも確率を用いた指標であるため, 実行結果が酷似してしまった可能性が推測できる.

5　まとめ

本実験レポートでは, 決定木に関する知識の確認と定着を目的とし,zoo.csv データを対象とした計算機実験を行った. その結果, 分割基準及び枝刈りを用いた決定木構築が最も精度が高い事が確認された. 加えて,R 言語を用いた決定木構築に関する一定の経験を得ることができた.

6　付録

本実験レポートで扱った zoo.csv データの各属性は以下の通りである.

データ数:100，**属性数**:17，**欠損地**:**なし**

hair: **毛の有無** (false / true)
feathers: **羽毛の有無** (false / true)
eggs: **卵** (false / true)
milk: **ミルク** (false / true)
airborne: **空中を飛ぶか** (false / true)
aquatic: **水生** (false / true)
predator: **捕食者** (false / true)
toothed: **歯の有無** (false / true)
backbone: **背骨の有無** (false / true)
breathes: **呼吸の有無** (false / true)
venomous: **毒性** (false / true)
fins: **ヒレの有無** (false / true)
legs: **脚の数** (0,2,4,5,6,8)
tail: **尾** (false / true)
domestic: **国内** (false / true)
catsize: **猫の大きさか** (false / true)
type: **タイプ** (1,2,3,4,5,6,7)

各タイプの属性数

1(mammal)　　: 41
2(bird)　　: 20
3(reptile)　　: 5
4(fish)　　: 13
5(amphibian)　　: 3
6(insect)　　: 8
7(invertebrate) : 10

参考文献

[1] e-Stat

http://archive.ics.uci.edu/ml/datasets/zoo