ANÁLISE PREDITIVA DE DADOS EM UMA INSTITUIÇÃO DE ABRIGO DE ANIMAIS ABANDONADOS

Caroline Lourenço Alves
Rafael Joseph Pagliuca dos Santos
Thiago Henrique Martinelli
Thomio Watanabe

24 de junho de 2016

Algoritmos de Aprendizado de Máquina (SCC5871-4)

Introdução

Introdução ao problema

Problema de classificação do Austin Animal Center apresentado no website Kaggle[1]:

Predizer o destino de um dado animal abrigado pela entidade, a partir de algumas informações conhecidas sobre ele.

[1] URL: `https://www.kaggle.com/c/shelter-animal-outcomes`.

Atributos

Classes: Adoção, morte, sacrifício (eutanásia), retorno ao dono e transferência.

Atributos de entrada: AnimalID, Name, DateTime, AnimalType, SexuponOutcome, AgeuponOutcome, Breed, Color

Atributos

Classes: Adoção, morte, sacrifício (eutanásia), retorno ao dono e transferência.

Atributos de entrada: ~~AnimalID~~, Name, ~~DateTime~~, AnimalType, SexuponOutcome*, AgeuponOutcome, Breed, Color

* Deve ser utilizado com cautela.

Pré-processamento

Pré-processamento

- AgeUponOutcome: Strings convertidas para dias (numérico), e depois valores agrupados por binning
- DateTime: Separado em dia e hora (numéricos).
- Name: Mantidos 35 nomes mais comuns + ”Outros”
- Breed: Mantidas 85 nomes mais comuns + ”Outros”
- Color: Mantidas 79 cores mais comuns + ”Outros”

Pré-processamento

Codificação para atributos categóricos: *one-hot*.

Algumas análises foram feitas com codificação numérica ordenada.

Classificadores

Classificadores

- Random Forest
- Naive Bayes
- Redes Neurais (MLP)
- SVM

Foi gerado um modelo para cada par classificador-classe, totalizando 20 modelos.

Resultados

Tabela: Perda logarítmica para cada pares classe/classificador.

Classe/Classificador	Random Forest	SVM	MLP	Naive Bayes
Adoption	1.078	7.996	7.928	11.019
Died	0.935	5.844	3.507	16.481
Euthanasia	0.829	7.848	7.056	20.605
Return to owner	0.815	5.909	5.918	14.695
Transfer	0.973	7.098	6.926	16.786

Análise dos resultados

- Naive Bayes: Performance ruim pode ser explicada devido à utilização de codificação *one-hot*, que gera um número muito grande de pseudo-atributos que não satisfazem a hipótese de independência estatística.
- Random Forest: Seleção intrínseca de atributos reduz de forma eficiente a dimensionalidade do problema, produzindo os melhores resultados.

Considerações finais

Foram seguidas as etapas do KDD e obtidos modelos que solucionam o problema proposta.

Os resultados obtidos pelo melhor modelo, do Random Forest, foram submetidos ao Kaggle. Classificação: 875 (resultado: 2.14512).

Código em Python (scikit-learn + pandas + numpy): `https://github.com/Mogi-Bertioga/shelter-animal-outcomes`

Obrigado.

Referências bibliográficas


Breiman, L.:
Random forests.
Machine Learning **45**(1) (Oct. 2001) 5–32


Ho, T.K.:
Random decision forests.
In: Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on. Volume 1. (Aug 1995) 278–282 vol.1


Ho, T.K.:
The random subspace method for constructing decision forests.
IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(8) (Aug 1998) 832–844


Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.:
From data mining to knowledge discovery in databases.