Avaliação de Técnicas de Aprendizado de Máquina Aplicadas a Análise de Crédito

Autores: Jane Thais Soares, Rogério Alves e Honovan Paz
DOI: 10.21528/CBIC2021-150

Aluno: Felipe Israel Corrêa

Análise de Risco de Crédito e Objetivo

Disponibilizar o crédito e diminuir o risco de inadimplência;

Comparar 10 dos mais comuns classificadores;

Base utilizada German Credit, disponível no repositório UCI com informações bancárias reais de 1000 clientes e 24 atributos;

Classificadores

- Naive Bayes;
- Logistic Regression;
- K-nearest-neighbors;
- Support Vector Machine;
- Decision Tree;
  - Bagging Decision Tree;
  - Boosting Decision Tree;
  - Random Forest;
- Neural Network (Multilayer Perceptron);
- **Voting Classification;**

*Bagging, boosting e bootstrap;

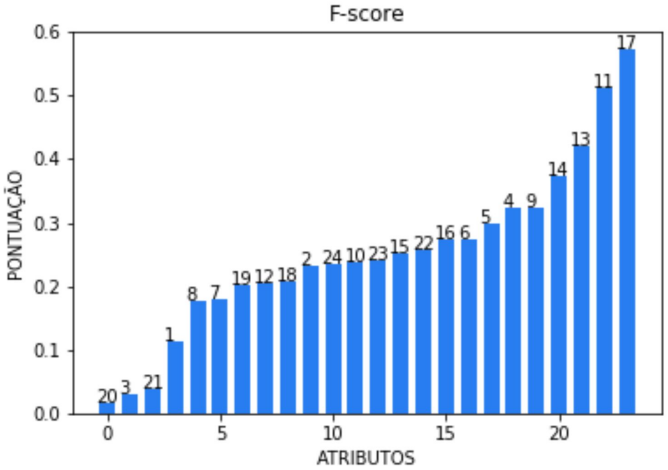
Pré-processamento

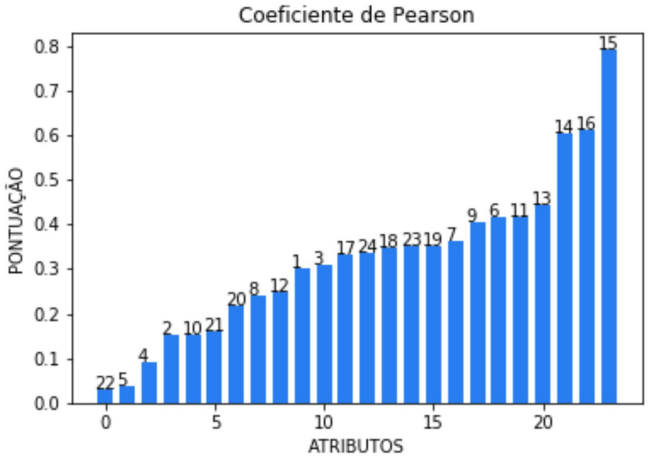
Normalização minmax;

Seleção de características:

- F-score;
- Coeficiente de correlação Pearson;
- Algoritmo Genético;

Seleção de Características





Seleção de Características

ATRIBUTOS SELECIONADOS PELO GA PARA CADA MÉTODO UTILIZADO.

nº	Atributos	KNN	Bayes	R. Logistica	A. decisão	SVM	R. forest	bagging	boosting	Voting	MLP
1	Verificação saldo	1	1	1	0	1	1	1	1	1	1
2	Nº meses do emprestimo	1	1	1	0	1	0	1	0	1	0
3	Historico de credito	1	0	0	1	0	1	0	1	0	0
4	Crédito de imposto mín. alternativo - AMT	0	0	0	1	0	1	1	1	1	1
5	Saldo poupança	1	1	0	1	0	0	1	0	1	0
6	Trabalho Atual	1	0	0	0	0	0	1	0	1	0
7	Sexo	1	0	1	1	0	0	1	0	1	0
8	Tempo de residencia atual	1	1	1	0	1	0	1	0	1	0
9	Propriedade	0	0	0	0	1	1	1	1	1	1
10	Idade	0	1	1	1	0	0	0	0	0	0
11	Outros parcelamentos	0	0	1	0	0	0	1	0	1	0
12	Crédito em outro banco	0	0	0	1	0	1	1	1	1	1
13	Conta individual/conjunta	0	0	1	0	0	1	0	0	0	0
14	Telefone	0	0	1	1	1	1	0	1	0	1
15	Trabalho estrangeiro	1	0	1	1	1	1	1	1	1	1
16	Compra carro novo	0	1	1	1	0	0	1	0	1	0
17	Compra carro usado	0	0	1	1	1	1	1	1	1	1
18	devedor s/ avalista	0	0	1	0	1	1	1	1	1	1
19	Devedor c/ avalista	0	0	0	0	0	1	0	0	0	0
20	Aluguel casa	0	1	1	0	1	0	0	0	0	0
21	Possui casa	1	1	1	0	0	1	0	1	0	1
22	Desempregado	0	1	1	1	0	1	1	1	0	1
23	Trabalho informal	0	0	0	0	1	0	1	1	1	0
24	Trabalho formal	1	0	1	1	1	0	1	1	1	0

NÚMERO DE ATRIBUTOS SELECIONADOS POR CADA MÉTODO DE SELEÇÃO DE CARACTERÍSTICAS.

Modelo	Nº atributos
Sem seleção (todos os modelos)	24
F-score (todos os modelos)	6
Pearson (todos os modelos)	4
F-score + Pearson (todos os modelos)	8
KNN (GA)	10
Bayes (GA)	9
R. Logística (GA)	16
SVM (GA)	11
Árvore de decisão (GA)	12
Random forest (GA)	13
Bagging (GA)	17
Boosting (GA)	13
Voting (GA)	16
MLP (GA)	10

Resultados

RESULTADOS OBTIDOS SEM SELEÇÃO DE CARACTERÍSTICAS

Modelo	parâ-metros	Acurácia (%)	Desvio padrão	Máximo (%)	Mínimo (%)	Tempo médio(s)
KNN	[7]	0,711	0,032	0,719	0,688	0,012
Bayes	-	0,688	0,041	0,740	0,650	**0,003**
R. logística	[8]	0,752	0,033	0,780	0,699	0,010
SVM	[9]	0,738	0,029	0,760	0,700	0,129
A. decisão	[29]	0,699	**0,027**	0,719	0,650	0,004
R. Forest	[8]	0,767	0,037	0,810	**0,740**	3,312
Bagging	[30]	**0,785**	0,036	**0,839**	**0,740**	3,529
Boosting	[31]	0,739	0,034	0,798	0,710	0,827
Voting	-	0,746	0,042	0,800	0709	0,191
MLP	[11]	0,740	0,031	0,794	0,700	0,879

RESULTADOS OBTIDOS ENTRE OS FILTROS UNIVARIADOS.

Modelo	Acurácia		
	F-score	C. Pearson	*F-score* + C. Pearson
KNN	0,631	0,677	0,691
Bayes	0,687	0,645	0,689
R. logística	0,692	0,691	0,730
SVM	0,699	0,692	0,728
A. decisão	0,683	0,680	0,686
R. Forest	0,714	0,714	0,720
Bagging	0,715	**0,716**	0,740
Boosting	0,711	0,715	**0,758**
Voting	**0,721**	0,713	0,729
MLP	0,706	0,701	0,735

Resultados

RESULTADOS OBTIDOS COM SELEÇÃO DE CARACTERÍSTICAS (*F-score* + COEFICIENTE DE PEARSON)

Modelo	parâ-metros	Acurácia (%)	Desvio padrão	Máximo (%)	Mínimo (%)	Tempo médio(s)
KNN	[7]	0,691	0,042	0,720	0,649	**0,001**
Bayes	-	0,689	0,031	0,711	0,644	0,002
R. logística	[8]	0,730	0,041	0,778	0,698	0,004
SVM	[9]	0,728	0,036	0,799	0,701	0,067
A. decisão	[29]	0,686	0,054	0,721	0,630	0,002
R. Forest	[8]	0,731	0,034	0,806	0,701	0,770
Bagging	[30]	0,740	0,043	0,811	0,699	1,760
Boosting	[31]	**0,758**	0,046	**0,825**	**0,711**	0,740
Voting	-	0,729	0,048	0,801	0,699	0,070
MLP	[11]	0,735	**0,029**	0,798	0,703	0,739

RESULTADOS OBTIDOS COM SELEÇÃO DE CARACTERÍSTICAS (ALGORÍTIMO GENÉTICO)

Modelo	parâ-metros	Acurácia (%)	Desvio padrão	Máximo (%)	Mínimo (%)	Tempo médio(s)
KNN	[7]	0.731	0.038	0800	0,670	0,003
Bayes	-	0.728	0,039	0,780	0,650	**0,002**
R. logística	[8]	0.785	0,032	0,839	0,719	0,032
SVM	[9]	0.759	**0,028**	0,810	0,709	0,069
A. decisão	[29]	0.706	0,050	0,770	0,640	0,003
R. Forest	[8]	0.766	0,039	**0.869**	0,729	1,467
Bagging	[30]	**0.792**	0,034	0,819	**0,731**	2,34
Boosting	[31]	0,765	0,038	0,829	0,728	0,807
Voting	-	0,771	0,033	0,801	0,719	0,095
MLP	[11]	0,762	0,044	0,831	0,711	0,369

Conclusão

Classificadores apresentaram resultados condizentes com a literatura;

Bom desempenho de Logistic Regression;

A seleção de características se mostrou bastante efetividade;

Métodos de ensemble apresentaram melhores resultados;