Avaliação estatística sobre seleção de características para categorização de texto

Rogério C. P. Fragoso*, Lucas F. Melo*, Saulo C. R. P. Sobrinho*
Universidade Federal de Pernambuco (UFPE), Centro de Informática (CIn)
Av. Jornalista Anibal Fernandes s/n, Cidade Universitária 50740-560, Recife, PE, Brazil
rcpf@cin.ufpe.br, lfm2@cin.ufpe.br, scrps@cin.ufpe.br

***Abstract*—ESCREVER RESUMO**

I. Introdução

Algoritmos são sequências de instruções bem definidas, porém, suas implementações podem apresentar comportamentos difíceis de serem previstos. Seja pelo uso de geração de números pseudo-aleatórios que geram um comportamento não determinístico inerente ao código, ou pelo uso de linguagens de alto nível cuja tradução para linguagem de máquina passa por otimizações e diferentes interações com a arquitetura de destino na qual o algoritmo é executado. Sendo assim, a performance de algoritmos de computação em geral é não-determinística e ao analisar comparativamente o desempenho destes algoritmos é necessário levar em consideração que estamos diante de uma amostra aleatória que representa estas performances.

Este trabalho realiza comparação entre algoritmos de categorização de texto mostrando como determinar e aplicar testes estatísticos adequados para este cenário.

A. Fundamentação teórica

Nesta seção, apresentar o problema de seleção de características, categorização de textos e os conceitos básicos necessários para o entendimento do trabalho.

Esta seção apresentou conceitos básicos de categorização de textos e seleção de características. O restante do trabalho é organizado como segue: A Seção II apresenta o objetivo do presente trabalho. Na Seção III são detalhadas as configurações dos experimentos, incluindo descrição da base de dados, os algoritmos de interesse e as hipóteses a serem verificadas sobre os dados. A Seção IV demonstra os procedimentos estatísticos realizados no trabalho. Finalmente, a Seção V apresenta as conclusões do trabalho.

II. Objetivo

O objetivo do presente trabalho é avaliar e comparar o desempenho de quatro métodos de seleção de características usados em categorização de texto. Esta análise comparativa permite determinar se os algoritmos possuem desempenhos significativamente diferentes e, em caso positivo, determinar quais deles se destacam dos demais.

III. Experimentos

Esta seção descreve as configurações dos experimentos realizados para gerar o conjunto de dados sobre o qual a análise será realizada.

A. Base de dados

Neste trabalho a base de dados *Reuters 10* foi utilizada. Esta base de dados é um subconjunto da coleção *Reuters-21578* [1], que é uma das bases mais utilizadas em trabalhos de categorização de texto. A base é composta por documentos coletados do *Reuters newswire* de 1987 e apresenta 135 categorias. Entretanto, neste trabalho foi adotado um subconjunto composto pelas 10 maiores categorias da base. O subconjunto *Reuters 10* contém 9.980 documentos e seu vocabulário abarca 10.987 termos. A base de dados *Reuters 10* também é bastante utilizada em trabalhos de categorização de texto [1]–[3].

A distribuição dos documentos é bastante desbalanceada, apresentando categorias representando desde 2,3% até 39% do tamanho total da base. Nesta base foram aplicados os seguintes procedimentos de pré-processamento: remoção termos com duas ou menos letras, remoção de *stopwords* e *stemming*, com o algoritmo *Iterated Lovins Stemmer* [4].

Vale salientar que para os propósitos da disciplina o uso de repositórios como base para análise não é permitido, porém, a análise comparativa deste trabalho será realizada sobre o desempenho dos algoritmos (tendo a base citada como entrada) e não sobre características da base em si.

B. Metodologia

Conforme mencionado na Seção II, estamos interessados na comparação de algoritmos de seleção de características para categorização de texto. Neste trabalho, a comparação é realizada entre os algoritmos *Maximum Features per Document* (MFD), *Maximum Features per Document-Reduced* (MFDR), *Category-dependent Maximum Features per Document-Reduced* (cMFDR) e *Automatic Feature Subsets Analyzer* (ASFA). A avaliação dos desempenhos dos métodos foi realizada utilizando o algoritmo classificador *Naïve Bayes Multinomial* [5] e a base de dados *Reuters 10*. Deste modo, a base de dados é pre-processada utilizando-se para cada um dos quatro algoritmos

[1] Disponível em http://disi.unitn.it/moschitti/corpora.htm.

de seleção de carcaterísticas, gerando, assim, quatro versões da base original. Em seguida, o classificador *Naïve Bayes Multinomial* é treinado e testado com cada uma destas quatro versões.

A validação cruzada estratificada foi utilizada como método para estimativa de desempenho. Esta técnica é adotada para avaliar a capacidade de generalização de um modelo a partir de um conjunto de dados. Neste trabalho utilizou-se a variação validação cruzada estratificada com *10 folds*, na qual a base de dados $\mathcal{D}$ é particionada em 10 subconjuntos (*folds*), de tamanhos semelhantes, mantendo a proporção de documentos por categorias equivalente à proporção encontrada no conjunto original. Então, são construídos 10 classificadores, cada um utilizando uma parcela dos *folds* para treinamento e outra parcela para realizar o teste do mesmo, de modo a gerar diferentes combinações dos *folds*. A avaliação final é dada pela média das medidas obtidas em cada uma das 10 execuções [6]. Nos experimentos realizados com os métodos MFD, MFDR e cMFDR, nove partições foram utilizadas para treinamento e uma partição foi utilizada para teste. O método AFSA requer uma porção dos dados para configuração de seus parâmetreos. Assim, os experimentos executados com AFSA utilizaram oito partições para treinamento, uma para configuração de parâmetros/validação e uma para teste.

Assim, temos dez medidas de desempenho para cada um dos quatro métodos de seleção de características avaliados. Estes dados de desempenho correspondem às entradas para as análises estatísticas realizadas neste trabalho.

A medida de desempenho utilizada nos experimentos foi *Micro-F1*. Seu cálculo é dado pela Eq. 1.

$$\mathcal{F}1 = \frac{2\text{x}\mathcal{P}\text{x}\mathcal{R}}{\mathcal{P}+\mathcal{R}}, \tag{1}$$

onde $\mathcal{P}$ é uma medida chamada precisão e $\mathcal{R}$ é cobertura [1]. As fórmulas para calcular a precisão $\mathcal{P}$ e a cobertura $\mathcal{R}$ são exibidas a seguir.

$$\mathcal{P} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C}(TP_j + FP_j)} \tag{2}$$

$$\mathcal{R} = \frac{\sum_{j=1}^{C} TP_j}{\sum_{j=1}^{C}(TP_j + FN_j)} \tag{3}$$

TP_j é a quantidade de instâncias corretamente rotuladas como pertencentes à categoria c_j, FP_j é a quantidade de instâncias incorretamente rotuladas como pertencentes à categoria c_j e FN_j é a quantidade de instâncias incorretamente rotuladas como não pertencentes à categoria c_j.

IV. Análise estatística

A. Estatística descritiva

Uma boa prática ao iniciar uma análise de conjunto de dados, e que é sugerida por muitos autores, é o uso de técnicas de estatística descritiva para ganhar intuições iniciais sobre o conjunto de interesse [7].

Para ter uma indicação sobre os tipos de testes que podem ser executados sobre os dados, é interessante verificar se as distribuições que geram os dados aparentam normalidade. A suposição de normalidade é útil pois se esta for plausível, podemos aplicar testes paramétricos sobre os dados. Testes paramétricos possuem maior poder estatístico que permite conclusões mais fortes do que os equivalentes não-paramétricos.

Para verificar a normalidade, começamos utilizando ferramentas visuais. Histogramas permitem visualizar a distribuição das amostras e consequentemente intuir sobre a distribuição da população geradora. As Figuras 5 a 8 apresentam os histogramas das amostras.

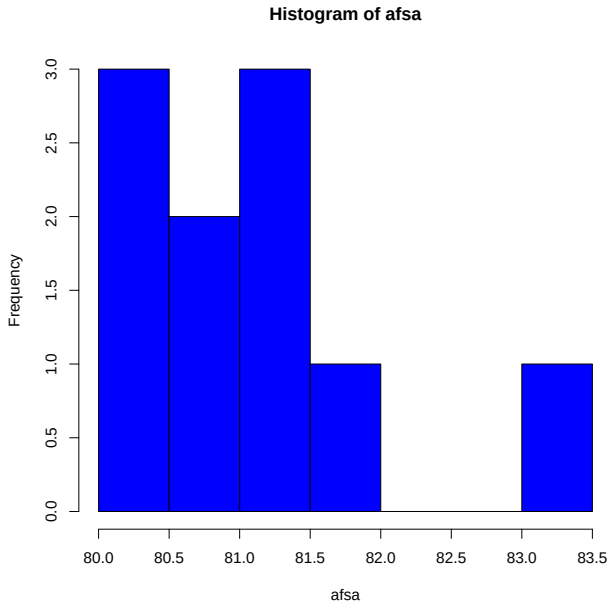


Fig. 1. Histograma dos ... do algoritmo AFSA (10 amostras)

A julgar pelos histogramas apresentados, não temos motivo para acreditar que tais amostras sejam provenientes de normais. Para juntar mais evidências, continuaremos a análise com mais recursos.

Outra ferramenta visual útil é o plot de probabilidade normal ou *normplots* que permite verificar o quão bem os dados podem ser ajustados por uma distribuição normal. Quando os dados são plotados dessa forma, uma linha reta indica uma distribuição normal ideal, sendo essa linha plotada para ser usada como referência. Num cenário real, devido à presença de ruídos, não se espera que os dados se adequem perfeitamente à reta, porém, espera-se que se os dados forem provenientes de uma distribuição normal, estes sejam bem aproximados pela reta de referência e estejam próximos a ela. As Figuras **??** a **??** mostram os normplots de cada distribuição.

ISSUE: Os pontos não aparecem quando o latex importa o pdf Oo Podemos observar que os dados não se aproximam tão bem de uma linha reta.

Por último ainda recorremos aos plot de caixa para verificar em particular a simetria dos dados. ...

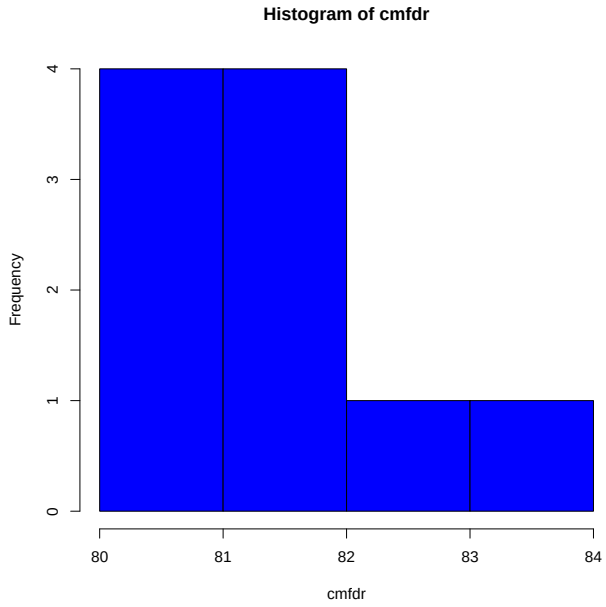


Fig. 2. Escrever uma descrição.

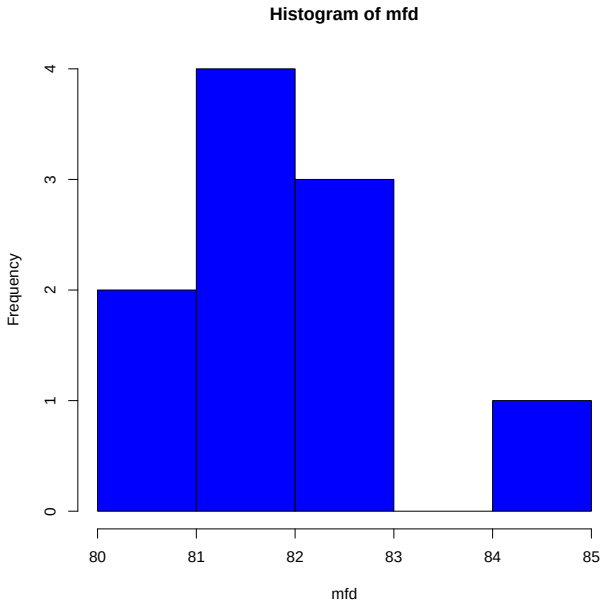


Fig. 4. Escrever uma descrição.

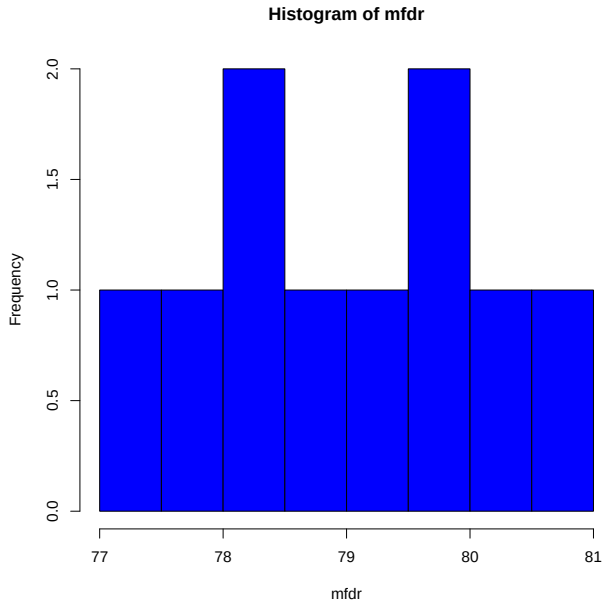


Fig. 3. Escrever uma descrição.

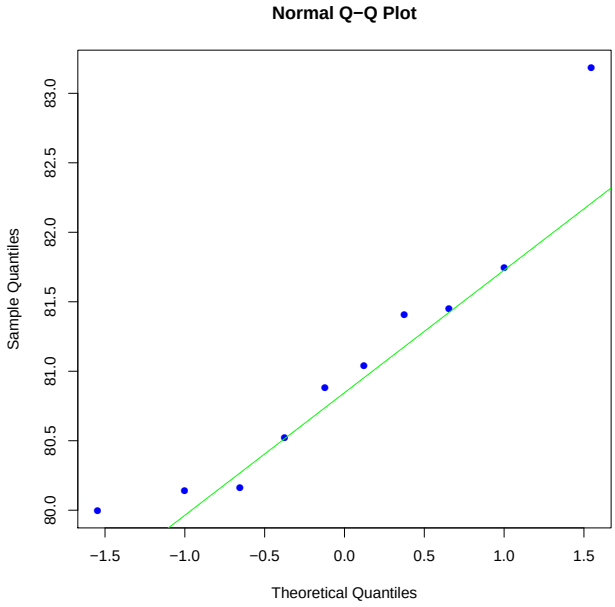


Fig. 5. Normplot dos ... do algoritmo AFSA (10 amostras)

Para confirmar nossa intuição de que as amostras não são normalmente distribuídas, recorremos a testes de hipótese utilizados para verificação de normalidade. Os testes usados foram o de Shapiro-Wilk [8] e ...

As amostras pequenas se mostraram um fator limitante da análise...

FALAR SOBRE OS RESULTADOS Quais aparentam ser normais (média aprox. igual à mediana)

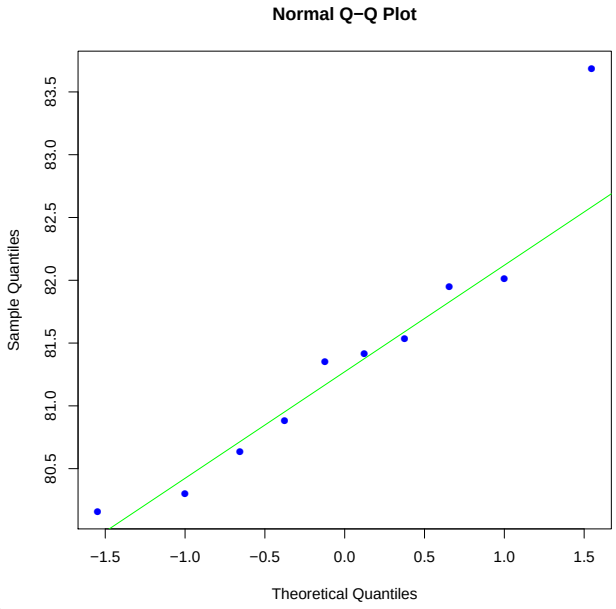
A Figura 9 apresenta uma.....

FAZER UMA CONCLUSÃO DA SEÇÃO Falar das impressões sobre a normalidade dos dados com bases nos histogramas, boxplot e medidas estatísticas.....

B. *Testes de aderência*

Dado que os resultados visuais não foram conclusivos o suficiente, especialmente pelo fato de os conjuntos de dados serem pequenos, se torna interessante realizar um teste para verificação da normalidade dos dados...

Escrever uma motivação para o uso dos testes de



*2*2

Fig. 6. Escrever uma descrição.

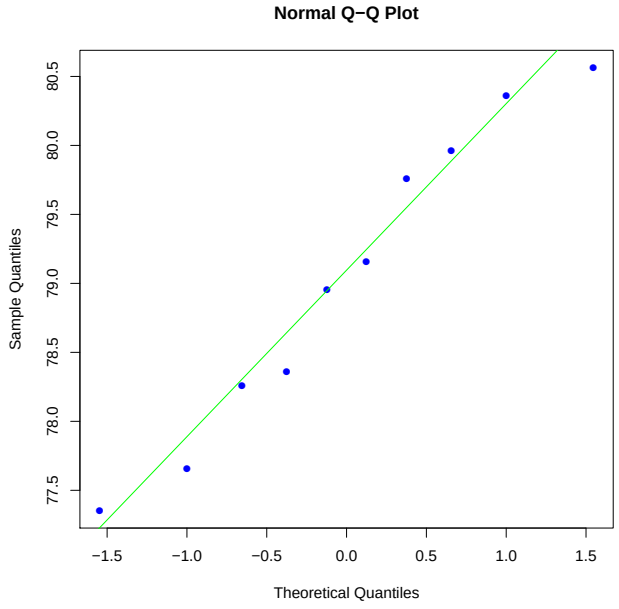


Fig. 7. Escrever uma descrição.

aderência e uma breve explicação de como funcionam

A Tabela II exibe os resultados dos testes Shapiro-Wilk e Kolmogrov-Smirnov para os quatro métodos.

Escrever sobre os resultados do test Shapiro-Wilk

Escrever sobre os resultados do test Kolmogrov-Smirnov

C. *Testes de hipóteses*

Ainda não executei os testes de hipóteses

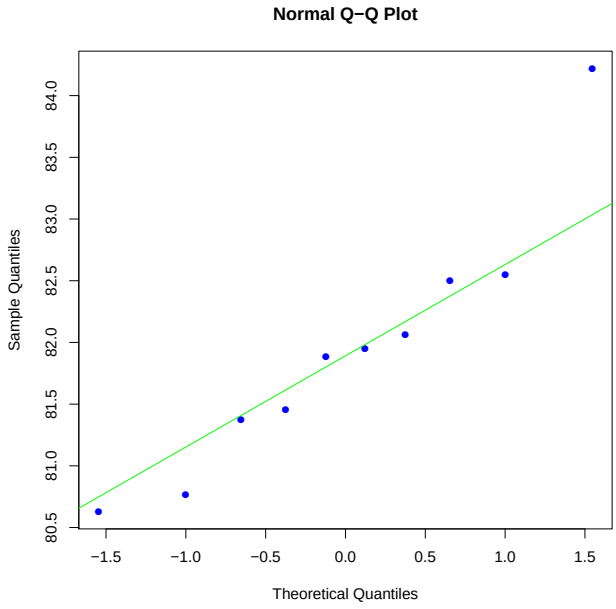


Fig. 8. Escrever uma descrição.

TABLE I. ESTATÍSTICA DESCRITIVA

Método	Média	Mediana	Desv. Padrão
AFSA	81.05262	80.9594	0.9634384
cMFDR	81.39266	81.38469	1.027409
MFDR	79.03808	79.05812	1.119577
MFD	81.93872	81.91752	1.029598

V. CONCLUSÃO

Escrever conclusão

REFERENCES

[1] Y. Chang, S. Chen, and C. Liau, "Multilabel text categorization based on a new linear classifier learning method and a category-sensitive refinement method," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1948–1953, 2008.

[2] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naive Bayes," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5432–5435, 2009.

[3] J. Yang, Y. Liu, Z. Liu, X. Zhu, and X. Zhang, "A new feature selection algorithm based on binomial hypothesis testing for spam filtering," *Knowledge-Based Systems*, vol. 24, no. 6, pp. 904–914, 2011.

[4] J. B. Lovins, *Development of a stemming algorithm*. MIT Information Processing Group, Electronic Systems Laboratory Cambridge, 1968.

[5] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Madison, WI, 1998, pp. 41–48.

[6] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Stanford, CA, 1995, pp. 1137–1145.

[7] D. C. Montgomery and G. C. Runger, *Applied statistics and probability for engineers*. John Wiley & Sons, 2010.

[8] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

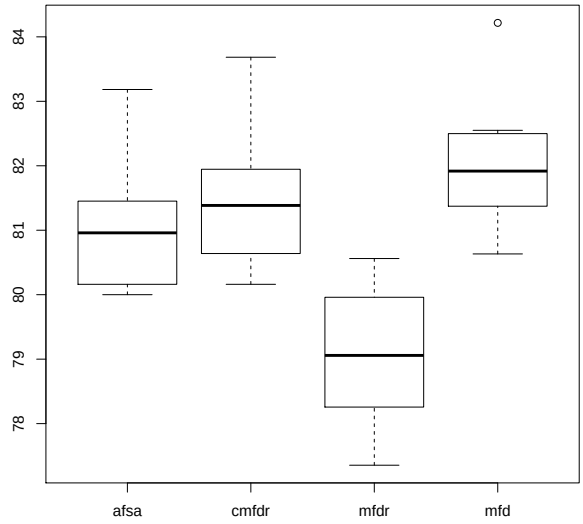


Fig. 9. Escrever uma descrição.

TABLE II. RESULTADOS DOS TESTES DE ADERÊNCIA

	p-value	
	Shapiro-Wilk	Kolmogrov-Smirnov
AFSA	0.2353	0.1818
cMFDR	0.3107	0.1818
MFDR	0.6773	0.1818
MFD	0.3597	0.1818