Avaliação estatística sobre seleção de características para categorização de texto

Rogério C. P. Fragoso*, Saulo C. R. P. Sobrinho*, Lucas F. Melo*
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

***Resumo*—Este trabalho realiza uma análise estatística sobre desempenho de métodos de seleção de características para categorização de texto. Os métodos cMFDR, MFDR e MFD são analisados com objetivo de verificar se a configuração paramétrica automática, fornecida pelo método AFSA, impacta no desempenho de classificação dos métodos. Análises estatísticas não foram conclusivas quanto à aderência das amostras a uma distrbuição normal. Deste modo, foi utilizado o teste de hipóteses não-paramétrico de Wilcoxon (teste de sinais com postos).**

I. Introdução

Um grande problema dos algoritmos de seleção de características é a identificação do número ideal de características a serem selecionadas. Um número pequeno pode acarretar em baixo desempenho de classificação. Um número grande, por outro lado, pode levar a um alto custo computacional. Normalmente este número é encontrado de maneira empírica. Entretanto, problemas de categorização de texto costumam ser de alta dimensionalidade. Assim, o processo de encontrar o número ideal de características a serem selecionadas pode ser lento e custoso.

Os métodos de seleção de características *Maximum f features per Document* (MFD) [3], *Maximum f features per Document-Reduced* (MFDR) [3] e *Class-dependent Maximum f features per Document-Reduced* (cMFDR) [4] introduzem uma abordagem de seleção local de características, o que facilita a parametrização da quantidade de características a serem selecionadas. Em um estudo [3], verificou-se que, com a abordagem tradicional de seleção de características, é necessário avaliar subconjuntos desde 11 características até mais de 5000. Com a abordagem de seleção local, proposta nos métodos supracitados, o melhor desempenho de classificação é encontrado em um espaço de 10 valores para seu único parâmetro (f).

Todavia, ainda assim, esta parametrização necessita de uma interação manual. Assim, dependendo do tipo de aplicação em que será usado o método de seleção de características, esta parametrização pode ter custos indesejáveis. O método *Automatic Feature Subsets Analyzer* (ASFA) [5] foi proposto para tratar deste problema. AFSA realiza automaticamente a parametrização dos métodos cMFDR, MFDR e MFD.

Neste trabalho analisamos os desempenhos dos métodos cMFDR, MFDR e MFD com configuração manual do parâmetro f e com configuração automática deste parâmetro, fornecida pelo método AFSA. Desejamos averiguar se a parametrização automática acarreta em diferenças nos desempenhos dos métodos.

A. Fundamentação teórica

Na abordagem de aprendizagem de máquina, uma instância é representada como um vetor composto por pares de característica e valor. Uma abordagem comum para a representação de textos na forma de vetores de características é a técnica conhecida como *Bag of Words* (BoW) [1]. Nela, um texto é tratado como um conjunto de palavras, sem considerar gramática ou ordem de ocorrência das palavras no texto. Cada palavra do vocabulário da base de dados é considerada uma característica e é associada à frequência desta palavra no documento. Ou seja, o tamanho do vocabulário da base de dados define a dimensionalidade dos vetores. Desta forma, em uma base de dados de tamanho médio, é comum que os vetores de características contenham dezenas de milhares de dimensões [2]. Entretanto, a maior parte destas características é irrelevante ou redundante. A alta dimensionalidade pode tornar a categorização de textos muito dispendiosa em termos de memória e tempo de execução. Adicionalmente, este grande número de características pode impactar negativamente no desempenho de classificação, especialmente em bases de dados com um número pequeno de instâncias em relação ao número de características, fenômeno conhecido como "praga da dimensionalidade". Como muitas das características são irrelevantes para a categorização, estes problemas podem ser tratados através da restrição da quantidade de características do conjunto de dados. Esta abordagem é conhecida como Redução de Dimensionalidade (DR, do inglês *Dimensionality Reduction*).

Uma técnica de DR muito utilizada é a seleção de características. Nesta abordagem, o conjunto final é formado por parte das características do conjunto original. A utilização de métodos de filtragem é a técnica de FS considerada mais adequada para problemas de TC, devido ao custo computacional ser bem mais baixo que o de outras técnicas, como métodos *wrapper*. Métodos de filtragem realizam um ordenamento das características através do uso de algoritmos determinísticos e métricas estatísticas, conhecidas com funções de avaliação de características (FEF, do inglês *Feature Evaluation Function*). Após o ordenamento, uma quantidade, estabelecida pelo usuário, de características é selecionada para a formação do novo subconjunto.

B. Métodos de seleção de características

Maximum f features per Document (MFD) [3], *Maximum f features per Document-Reduced* (MFDR) [3] e *Class-dependent Maximum f features per Document-Reduced* (cMFDR) [4] são métodos de seleção de características para categorização de texto.

MFD realiza a seleção local (por documento) das f características consideradas mais relevantes, de acordo com alguma *feature evaluation function*. O parâmetro f é informado pelo usuário. MFDR introduz um limiar onde apenas documentos considerados relevantes são considerados na seleção de características. cMFDR aperfeiçoa o limiar introduzido por MFDR. Com cMFDR, existe um limiar para cada categoria, o que melhora o desempenho em bases de dados desbalanceadas.

Estes três métodos requerem um valor para o parâmetro f, que indica a quantidade de características a serem selecionadas por instância. A escolha de um bom valor para este parâmetro pode ser um trabalho demorado e exaustivo. Neste contexto, o método *Automatic Feature Subsets Analyzer* (ASFA) [5] foi proposto para ser usado conjuntamente com um dos três métodos: cMFDR, MFDR ou MFD. O objetivo do AFSA é prover, para estes métodos, um valor para o parâmetro f de forma automática.

Esta seção apresentou conceitos básicos de categorização de textos e seleção de características e introduziu os métodos de seleção de características que são avaliados no trabalho. 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 desta pesquisa é verificar se o método AFSA é capaz de prover uma configuração paramétrica, para os métodos de seleção de características cMFDR, MFDR e MFD e , de modo que o desempenho de classificação destes métodos não seja prejudicado. Ou seja, desejamos averiguar se o desempenho destes métodos é impactado quando eles são usados conjuntamente com AFSA. Assim, o desempenho de cada um dos métodos cMFDR, MFDR e MFD é comparado com sua versão combinada com AFSA (AFSA+cMFDR, AFSA+MFDR, AFSA+MFD), com vistas a determinar se os algoritmos apresentam diferenças estatisticamente significativas de desempenho.

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

Para a categorização de texto, foi utilizada a base de dados *Reuters 10*. 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, o subconjunto adotado neste trabalho é composto pelas 10 maiores categorias da base. O base de dados *Reuters 10* contém 9.980 documentos e seu vocabulário abarca 10.987 termos. A base de dados *Reuters 10* também é muito utilizada em trabalhos de categorização de texto [6]–[8].

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: *stemming*, com o algoritmo *Iterated Lovins Stemmer* [9], remoção de termos com duas ou menos letras e remoção de *stopwords*.

Vale salientar que a análise comparativa deste trabalho é realizada sobre o desempenho dos algoritmos, tendo a base citada como entrada, e não sobre características da base em si. O processo de geração das amostras utilizadas na análise estatística é detalhado na Seção III-B.

B. Metodologia

Conforme mencionado na Seção II, esta pesquisa visa realizar uma comparação de desempenho de algoritmos de seleção de características para categorização de texto. Neste trabalho, é feita uma avaliação do desempenho do método AFSA. Para tanto, AFSA é usado em conjunto com cada um dos métodos cMFDR, MFDR e MFD (conforme descrito na Seção I-B). O desempenho de cada um dos métodos configurado manualmente é comparado com o desempenho destes métodos configurados automaticamente por AFSA.

O desempenho de um método de seleção de características pode ser auferido em termos de redução de dimensionalidade (tamanho do vetor final de características), tempo de execução e do desempenho de classificação atingido com o vetor de características resultante do processo de seleção. Neste trabalho, a avaliação dos desempenhos dos métodos levou em conta o desempenho de classificação sobre a base de dados resultante do processo de seleção de características. O algoritmo de aprendizagem de máquina empregado para a avaliação dos métodos foi classificador *Naïve Bayes Multinomial* [10].

A base de dados *Reuters 10* foi pré-processada utilizando cada um dos seis algoritmos de seleção de carcaterísticas (cMFDR, MFDR e MFD e a combinação de cada um destes com AFSA), gerando, assim, seis versões da base original. Em seguida, o classificador *Naïve Bayes Multinomial* foi treinado e testado com cada uma destas seis 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

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

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* [11].

Nos experimentos realizados com os métodos cMFDR, MFDR e MFD, 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. Deste modo, ao final deste processo, temos dez medidas de desempenho para cada um dos seis métodos de seleção de características avaliados. Estes dados de desempenho correspondem às amostras que são 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 de precisão e $\mathcal{R}$ é uma medida de cobertura [6]. 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, a qual é sugerida por muitos autores, é o uso de técnicas de estatística descritiva para se obter intuições iniciais acerca do conjunto de interesse [12].

Para se ter uma indicação sobre os tipos de testes de hipótese que podem ser executados sobre os dados, é interessante verificar se as distribuições que geram as amostras apresentam 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 do que seus equivalentes não-paramétricos, o que nos permite extrair conclusões mais fortes.

Para verificar a normalidade, começamos analisando diretamente os sumários numéricos dos dados. Ao comparar médias e medianas, podemos ter uma idéia da simetria do conjunto em questão. Adicionalmente, verificamo o índice de assimetria para cada amostra. Comumente, uma amostra é considerada simétrica se apresenta índice de assimetria entre -0.5 e 0.5. Amostras com índice de assimetria com valores entre -1 e -0.5 ou entre 0.5 e 1.0 são consideradas levemente assimétricas. Já amostras com índice de assimetria menor que -1 ou maior que 1 são tidas como assimétricas. A Tabela I mostra os valores de média, mediana e assimetria de cada conjunto.

Tabela I. Sumários numéricos dos dados

Método	Média	Mediana	Assimetria
AFSA+cMFDR	81.58	81.50	0.39
cMFDR	81.34	81.30	0.38
AFSA+MFDR	80.16	80.23	0.46
MFDR	80.06	79.94	-0.16
AFSA+MFD	81.72	81.95	-0.64
MFD	81.78	81.95	-0.34

A maioria das amostras apresenta valores de suas médias relativamente próximos aos de suas medianas. Além disso, a maioria das amostras apresentam índices de assimetria entre -0.5 e 0.5. Uma exceção é AFSA+MFD, que apresenta índice de assimetria igual a -0.64. À primeira vista, podemos imaginar que os dados seguem uma distribuição normal. Entretanto, somente esta análise numérica é insuficiente para nos fornecer informações consistentes para subsidiar uma conclusão sobre a normalidade dos dados.

É comum o uso de recursos visuais para analisar a distribuição de uma amostra de dados. Histogramas permitem visualizar a distribuição de frequências das amostras e, consequentemente, intuir sobre a distribuição da população da qual a amostra foi extraída. A Figura 1 apresenta os histogramas dos desempenhos de classificação de cada algoritmo, nas suas versões combinadas com o ASFA e sem o mesmo.

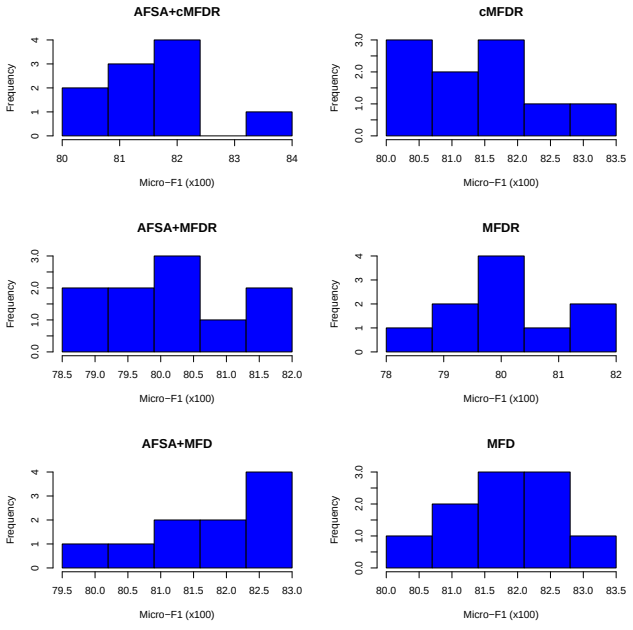

Figura 1. Histograma do desempenho de classificação dos métodos cMFDR, MFDR e MFD e suas versões combinadas com AFSA.

Histogramas de amostras de dados normalmente distribuidos apresentam o formato de um sino, com os valores centrais exibindo maiores frequencias. Analisando os histogramas, não observamos indícios fortes de normalidade em nenhuma das amostras. O histograma da amostra de AFSA+cMFDR exibe uma forma semelhante a de uma distribuição normal. Entretanto, a presença de valores com frequência igual a zero não nos permite concluir que a amostra tem distribuição semelhante à normal. Já para cMFDR, o histograma apresenta uma forte assimetria positiva. O histograma de AFSA+MFDR não guarda semelhanças com uma distribuição normal. O histograma de MFDR apresenta maior frequência nos valores centrais, mas com uma frequência muito alta na cauda direita. A amostra de AFSA+MFD exibe um histograma com forte assimetria negativa. O histograma de MFD apresenta forma análoga a de uma distribuição normal. Notamos uma maior frequência nos valores centrais, porém o histograma apresenta uma leve assimetria negativa. A impressão obtida a partir dos histogramas contradiz a conclusão da observação das médias, medianas e assimetrias das amostras. Sendo assim, continuamos a análise com outros recursos visuais, em busca de evidências mais robustas.

Outra ferramenta visual útil é o gráfico QQ (*QQ Plot*), que permite verificar a semelhança entre duas distribuições de probabilidade. Nesta pesquisa, usamos o gráfico QQ para verificar a semelhança das distribuições de probablidades das amostras com uma distribuição normal. Caso a distribuição da amostra seja semelhante à uma normal, os pontos no gráfico QQ são posicionados sobre a reta $y = x$. Num cenário real, devido à presença de ruídos, não se espera que os dados adequem-se perfeitamente à tal reta. Porém, espera-se que, se os dados forem provenientes de uma distribuição semelhante à normal, estes se aproximem da reta $y = x$. A Figura 2 mostra os Gráficos QQ para cada amostra.

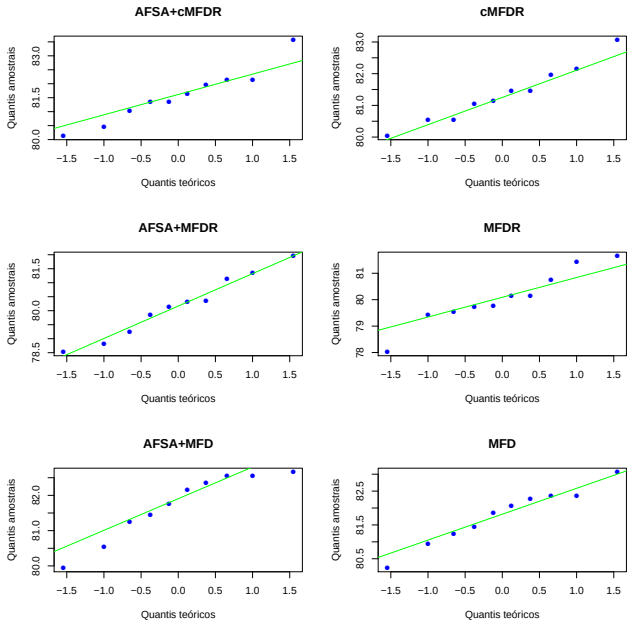

Figura 2. Gráfico QQ dos desempenhos de classificação dos métodos cMFDR, MFDR e MFD e suas versões combinadas com AFSA.

Podemos observar que o ajuste das distribuições de probabilidade das amostras à normal, em geral, não foi tão próximo ao ponto de nos levar a uma conclusão de que as amostras provêm de uma distribuição normal. Nas amostras que representam os resultados dos métodos cMFDR e AFSA+MFDR e MFD, os pontos aproximam razoavelmente a reta $y = x$. Todavia, com estas informações, ainda não temos subsidíos que nos apoiem a adotar testes de hipóteses paramétricos. Cada método que apresenta aparente aderência à normal é comparado com um método que não apresenta indícios de normalidade nos gráficos QQ (AFSA+cMFDR, MFDR e AFSA+MFD).

Ainda recorremos aos diagramas de caixa (*boxplots*) como um último recurso visual. Este tipo de diagrama tem o objetivo de verificar sobretudo a simetria das amostras. Os diagramas de caixa das amostras em análise são apresentados na Figura 3.

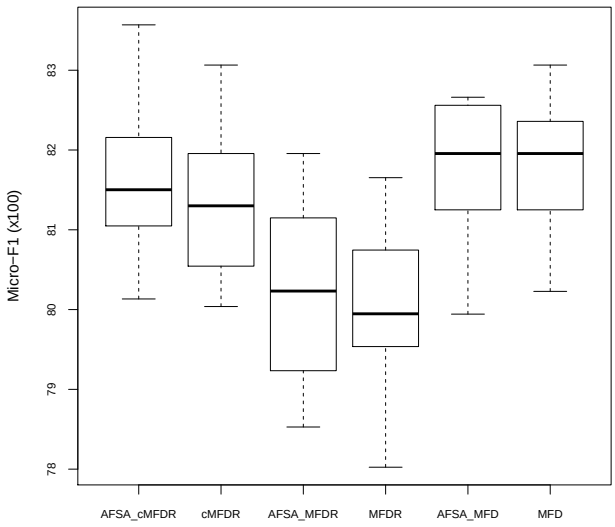

Figura 3. Diagramas de caixa de todos os conjuntos de amostras.

Observando os diagramas, vemos que, com exceção de AFSA+MFDR, os diagramas exibem assimetria, alguns inclusive apresentando uma assimetria acentuada, como é o caso de AFSA+MFD. A simetria é uma característica inerente à distribuição normal. Um indício de que uma amostra provém de uma população com distribuição aproximadamente normal é a simetria, ou seja, quando média, moda e mediana apresentam valores aproximados e todos eles se encontram aproximadamente no centro da distribuição. Como esta característica não é observada nas amostras, temos mais indícios de que não devemos assumir normalidade dos dados.

B. Testes de aderência

Uma maneira mais formal de verificar se as amostras seguem uma distribuição normal é com o uso de testes de aderência. Nesta pesquisa, foram executados os testes Shapiro-Wilk [13], Anderson-Darling [14] e Cramer-von

Mises [15]. Os resultados dos testes para cada amostra são apresentados na Tabela II.

Tabela II. RESULTADOS DOS TESTES DE ADERÊNCIA.

Método	p-value		
	Shapiro-Wilk	Anderson-Darling	Cramer-von Mises
AFSA cMFDR	0.75	0.67	0.74
cMFDR	0.90	0.86	0.86
AFSA MFDR	0.88	0.89	0.87
MFDR	0.61	0.47	0.45
AFSA MFD	0.19	0.25	0.30
MFD	0.88	0.77	0.72

Consideramos que uma amostra não segue uma distribuição normal caso o *p-value* seja menor que 0.05. Deste modo, o resultado dos três testes apontam que, para todas as amostras, não existem evidências suficientes para rejeitar a hipótese nula de que a amostra segue uma distribuição normal, o que corrobora a impressão obtida da análise dos valores das médias, medianas e assimetria das amostras (Tabela I). Porém, é importante salientar que a eficiências de testes de aderência é prejudicada quando as amostras são pequenas, como no caso em análise. Adicionalmente, a observação dos histogramas, gráficos QQ e dos diagramas de caixa dão indícios de que a maioria das amostras não seguem uma distribuição normal. Por estas razões, concluímos que, com os dados disponíveis, não temos segurança em afirmar que as amostras seguem uma distribuição normal. Diante disto, utilizamos teste de hipótese não-paramétrico para verificar a existência de diferenças significativas entre as amostras.

C. Resultados

Depois de avaliar as distribuições das amostras, optamos pelo caminho mais seguro, que consiste na comparação dos pares de métodos por testes não paramétricos. Como estamos realizando comparações entre pares de amostras dependentes, optamos pelo teste de sinais com postos de Wilcoxon [16].

Estabelecemos a hipótese nula de que os desempenhos dos métodos sendo comparados (com e sem AFSA) são iguais. A hipótese alternativa é que os desempenhos são diferentes. Com este conjunto de hipóteses, buscamos averiguar se a configuração paramétrica realizada automaticamente por AFSA apresenta algum impacto no desempenho de classificação dos métodos cMFDR, MFDR e MFD.

Os resultados dos testes são apresentados na Tabela III.

Tabela III. RESULTADOS DOS TESTES DE HIPÓTESES WILCOXON.

Hipótese alternativa	*p-value*
H_1 : AFSA MFD $\neq$ MFD	0.44
H_1 : AFSA MFDR $\neq$ MFDR	0.54
H_1 : AFSA cMFDR $\neq$ cMFDR	0.08

Para este teste, consideramos um nível de significância de 0.05. Diante dos resultados apresentados na Tabela III, concluímos, com uma confiança de 95%, que não há indícios suficientes para rejeitar a hipótese nula. Ou seja, não é possível afirmar que a configuração automática de parâmetros realizada pelo método AFSA impacta no desempenho classificatório dos métodos cMFDR, MFDR e MFD.

V. CONCLUSÃO

Este trabalho fez uma avaliação estatística sobre o desempenho de métodos de seleção de características para categorização de texto. Verificamos se a configuração paramétrica automática afeta o desempenho dos métodos cMFDR, MFDR e MFDR. Pelos experimentos e resultados apresentados, não foi possível rejeitar a hipótese nula de que a configuração paramétrica do AFSA não afeta a performance dos algoritmos de de seleção de características.

Observamos que, para ganhar mais segurança no resultado dos teses, seria necessário colher amostras maiores, ou seja, executar mais vezes os métodos de seleção de características. Isso seria importante especialmente porque testes não paramétricos geralmente necessitam de amostras maiores que seus equivalentes paramétricos para obter as mesmas conclusões. O aumento das amostras poderia inclusive levar à verificação de normalidade nos dadoss, o que alteraria até mesmo o tipo teste de hipóteses empregado. Entretanto, a avaliação dos algoritmos apresentados é computacionalmente custosa, devido ao tamanho das bases de dados e da alta dimensionalidade inerente ao problema de categorização de texto. Dessa forma, a análise inicial precisou ser limitada a um número pequeno de amostras. Ainda assim, a não rejeição dá um indício ao experimentador de que o AFSA não causa impactos no desempenho dos algoritmos, o que seria o comportamento desejado.

REFERÊNCIAS

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] E. Gabrilovich and S. Markovitch, "Text categorization with many redundant features: using aggressive feature selection to make svms competitive with c4. 5," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 41.

[3] R. Pinheiro, G. Cavalcanti, and T. Ren, "Data-driven global-ranking local feature selection methods for text categorization," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1941–1949, 2015.

[4] R. Fragoso, R. Pinheiro, and G. Cavalcanti, "Class-dependent feature selection algorithm for text categorization," in *International Joint Conference on Neural Networks*, 2016.

[5] ——, "A method for automatic determination of the feature vector size for text categorization," in *2016 Brazilian Conference on Intelligent Systems*, 2016.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] T. W. Anderson and D. A. Darling, "A test of goodness of fit," *Journal of the American statistical association*, vol. 49, no. 268, pp. 765–769, 1954.

[15] J. Durbin and M. Knott, "Components of cramer-von mises statistics. i," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 290–307, 1972.

[16] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.