UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXIII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIENTÍFICA E TECNOLÓGICA - 2019

Análise do Comportamento de Diferentes Classificadores a partir de Bases de Dados Completas e Reduzidas

Matheus Galvão Correia[1]; Fabiana Cristina Bertoni[2]

1. Bolsista PIBIC/FAPESB, Graduando em Engenharia de Computação, Universidade Estadual de Feira de Santana, e-mail: matheusgcorreia@gmail.com
2. Orientadora, Departamento de Exatas, Universidade Estadual de Feira de Santana, e-mail: fcbertoni@gmail.com

PALAVRAS-CHAVE: Redução de dados, Classificação de padrões, Algoritmos para classificação de padrões.

INTRODUÇÃO

Sistemas de Classificação de Padrões buscam obter um modelo baseado em um conjunto de exemplos que descrevem uma função desconhecida. Esse modelo é utilizado posteriormente para fornecer o valor de atributos de novos exemplos. A tarefa de classificação é uma função de aprendizado que mapeia bases de dados de entrada em um número finito de classes. Nela, cada exemplo pertence a uma classe, entre um conjunto pré-definido de classes. O objetivo de um algoritmo de classificação é encontrar alguma correlação entre os atributos e uma classe, de modo que o processo de classificação possa usá-la para predizer a classe de um exemplo novo e desconhecido.

As bases de dados utilizadas para aprendizado dos sistemas de classificação de padrões possuem muitas vezes dados redundantes ou que não representam bem o problema, apresentando alta dimensão e grande escala, fatores que afetam a acurácia e o tempo computacional dos algoritmos de classificação. Neste contexto, métodos para reduzir estas bases de dados vêm sendo propostos. A redução de dados pode ser alcançada de diversas maneiras, dentre elas, a Seleção de Características e a Seleção de Instâncias. García-Pedrajas (2013) afirma que, ao selecionar características, reduzimos o número de colunas em uma base de dados e, selecionando instâncias, reduzimos o número de linhas, segundo Tsai (2014).

Os trabalhos iniciais do projeto de pesquisa ao qual este trabalho está vinculado se concentraram em encontrar bons algoritmos para redução de dados. Entretanto, após a análise dos resultados e apresentação destes em eventos científicos, surgiu o questionamento relacionado a estabilidade na acurácia de diferentes tipos de classificadores a partir de bases de dados reduzidas, fazendo surgir a proposta de trabalho ligado a este projeto.

METODOLOGIA

Para efeito didático, a metodologia pode ser dividida em cinco etapas:

1) Teórica, onde foram estudados os problemas de classificação de padrões e realizada uma revisão da literatura, a fim de compreender os algoritmos K-Nearest Neighbor (KNN), Classificadores Bayesianos, Sistemas de Classificação Fuzzy e Redes Neurais, empregados na sua solução;
2) Ferramental, onde foram modelados e implementados os algoritmos de classificação;
3) Experimental, na qual foram realizados os experimentos necessários à verificação da acurácia e do tempo computacional dos algoritmos de classificação, considerando as bases de dados completas e reduzidas;
4) Testes e Análise dos Resultados, na qual os resultados obtidos foram avaliados e discutidos;
5) Divulgação dos Resultados (esta etapa ainda está em andamento). Os resultados obtidos serão apresentados no formato de artigos científicos no SEMIC/UEFS (Seminário de Iniciação Científica da UEFS) e em congressos e/ou periódicos nacionais e internacionais relacionados com a pesquisa.

Para a realização das etapas 2), 3) e 4) foi usado um dos computadores disponíveis no LASIC (Laboratório de Pesquisa em Sistemas Inteligentes e Cognitivos), com um processador INTEL CORE I7-3770 e memória RAM de 8GB. Dentre os repositórios e ferramentas de programação utilizadas estão: a linguagem Java (Paul e Deitel, 2014), o software Weka (Hall et al., 2009), o framework jMetal (Juan e Durillo, 2011) e o repositório de bases de dados *Keel*, disponível em: http://sci2s.ugr.es/keel/datasets.php.

Na etapa experimental, foram utilizadas 37 bases de dados dos mais variados tamanhos. Detalhes relevantes sobre estas bases, como o número de instâncias e características, estão disponíveis na Tabela 1 em anexo. Em todas elas foram aplicados cinco algoritmos de classificação, a saber: KNN, SVM, Redes Bayesianas, Naive Bayes e Rede Neural. Nos experimentos, para cada algoritmo, foi utilizada a abordagem *ten-fold cross validation*. Além disso, cada fold foi testado três vezes, logo, para cada base de dados foram efetuadas 30 execuções. Os resultados apresentados expressam a média destas 30 execuções. Ao final, coletou-se a média de acurácia, coeficiente Kappa (Fleiss, 1981) e tempo de execução. Em seguida, para redução de características, foram aplicados os algoritmos de seleção NSGA-DO, NSGA-II e NSGA-III, propostos por Pimenta (2015), Deb (2002) e Yuan (2014), respectivamente. Após a redução, os mesmos algoritmos de classificação foram executados nas bases reduzidas, da mesma forma que nas bases originais e foram coletados os mesmos dados citados anteriormente. A análise destes resultados é apresentada a seguir.

RESULTADOS

Por conta do grande número de bases utilizadas, os resultados serão apresentados em tabelas sob a perspectiva dos classificadores. A Tabela 1 apresenta os resultados dos algoritmos executados com as bases originais, a Tabela 2 com as bases reduzidas utilizando o NSGA-DO, a Tabela 3 com redução através do NSGA-II e, por fim, a Tabela 4 com as bases otimizadas pelo NSGA-III.

Tabela 1. Classificadores executados nas bases originais

	Acurácia (%)	**Coeficiente Kappa**	**Tempo de execução (ms)**
KNN	79.45	0.60	5,071.76
SVM	79.56	0.58	25,911.86
Rede Bayesiana	78.04	0.61	1,742.72
Naive Bayes	75.42	0.56	1,304.90
Rede Neural	80.85	0.63	245,467.04

Tabela 2. Classificadores executados nas bases reduzidas com NSGA-DO

	Acurácia (%)	**Coeficiente Kappa**	**Tempo de execução (ms)**
KNN	77.16	0.57	3,546.67
SVM	73.53	0.44	5,075.93
Rede Bayesiana	74.66	0.53	1,320.78
Naive Bayes	72.46	0.49	1,039.22
Rede Neural	76.59	0.55	66,383.60

Tabela 3. Classificadores executados nas bases reduzidas com NSGA-II

	Acurácia (%)	**Coeficiente Kappa**	**Tempo de execução (ms)**
KNN	77.13	0.57	3,566.30
SVM	73.60	0.44	5,750.40
Rede Bayesiana	74.52	0.53	1,316.11
Naive Bayes	72.17	0.49	1,078.20
Rede Neural	76.48	0.55	66,955.62

Tabela 4. Classificadores executados nas bases reduzidas com NSGA-III

	Acurácia (%)	**Coeficiente Kappa**	**Tempo de execução (ms)**
KNN	77.18	0.57	3,518.85
SVM	73.78	0.45	5,339.79
Rede Bayesiana	74.69	0.53	1,310.33
Naive Bayes	72.32	0.49	1,039.80
Rede Neural	76.68	0.55	73,130.04

A acurácia é obtida pela razão entre o número de instâncias corretamente classificadas e o total de instâncias de uma base de dados. O coeficiente Kappa indica a confiabilidade do classificador, ou seja, o quanto ele é capaz de atribuir a mesma classe à mesma instância durante a classificação. Valores mais próximos de 1 indicam maior acurácia. Os dados apresentados mostram que, em todos os casos, a acurácia de classificação

obtida com as bases reduzidas decresceu em cerca de 5% (considerando uma média entre as precisões de todos os classificadores), em relação às bases originais. Em contrapartida, o tempo de execução foi consideravelmente diminuído com a redução das bases de dados, em aproximadamente 71,3%, considerando também uma média dos tempos de execução de todos os classificadores.

CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi analisar o comportamento de diferentes tipos de classificadores a partir de bases de dados reduzidas. Os experimentos realizados com 37 bases dados, de diferentes tamanhos, mostram que a eliminação de características irrelevantes é capaz de reduzir satisfatoriamente o custo computacional de execução de todos os classificadores analisados, e portanto, seu tempo de execução, com um impacto mínimo na acurácia.

Ainda não foram realizados experimentos envolvendo sistemas de classificação fuzzy, mas estes se constituem em um dos objetivos a serem ainda investigados.

REFERÊNCIAS

DEB, Kalyanmoy et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation, v. 6, n. 2, p. 182-197, 2002.

HALL M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., and WITTEN, I. H.. 2009. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter. pp. 10-18.

JUAN, A. J. N.; DURILLO, J. 2011. jMetal: A Java framework for multiobjective optimization. Advances in Engineering Software, p. 760–771.

SCHILDT, H. 2011. Java The Complete Reference. McGraw-Hill.

GARCÍA-PEDRAJAS, N.; A. d. HARO-GARCÍA, et al. 2013. A scalable approach to simultaneous evolutionary instance and feature selection. Information Sciences 228: 150-174.

PAUL , D. H. M. D.; DEITEL, J. 2014. Java How to Program. Pearson.

TSAI, C.-F.; Z.-Y. Chen. 2014. Towards high dimensional instance selection: An evolutionary approach. Decision Support Systems 61: 79-92.

YUAN, Yuan; XU, Hua; WANG, Bo. An improved NSGA-III procedure for evolutionary many-objective optimization. In: Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation. ACM, 2014. p. 661-668.

PIMENTA, Adinovam HM; DE ARRUDA CAMARGO, Heloisa. NSGA-DO: Non-Dominated Sorting Genetic Algorithm Distance Oriented. In: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). IEEE, 2015. p. 1-8.

FLEISS J. L. “Statistical methods for rates and proportions”. New York: John Wiley, 1981. p 212-236.

ANEXO

Tabela 1. Bases de Dados

Base de Dados	Nº de instâncias	Nº de características
Abalone	4177	8
Australian	2565	22
Automobile	205	26
Balance	625	4
Banana	5200	3
Bupa	345	7
Cleveland	303	75
COIL 2000	9000	86
Contraceptive	1473	9
CRX	690	15
German	1000	20
Glass	214	10
Haberman	306	3
Heart	267	22
Hepatitis	155	19
Ecoli	336	8
Iris	150	4
Magic	19020	11
Marketing	45211	17
Newthyroid	7200	21
Optdigits	5620	64
Page Blocks	5473	10
Pen-Based	10992	16
Phoneme	528	10

Pima	769	9
Ring	23	4
Satimage	6435	36
Segment	2310	19
Spambase	4601	57
Tae	151	5
Texture	168	148
Thyroid	7200	21
Titanic	887	5
Twonorm	7400	21
Vehicle	946	18
Wine	178	13
Winsconsin	569	32