Nguyen, Thi Thanh Sang; **Do, Pham Minh Thu**

Classification optimization for training a large dataset with naïve Bayes. (English)

Zbl 1444.68159
J. Comb. Optim. 40, No. 1, 141-169 (2020).

Summary: Book classification is very popular in digital libraries. Book rating prediction is crucial to improve the care of readers. The commonly used techniques are decision tree, Naïve Bayes (NB), neural networks, etc. Moreover, mining book data depends on feature selection, data pre-processing, and data preparation. This paper proposes the solutions of knowledge representation optimization as well as feature selection to enhance book classification and point out appropriate classification algorithms. Several experiments have been conducted and it has been found that NB could provide best prediction results. The accuracy and performance of NB can be improved and outperform other classification algorithms by applying appropriate strategies of feature selections, data type selection as well as data transformation.

MSC:

68T05 Learning and adaptive systems in artificial intelligence
62H30 Classification and discrimination; cluster analysis (statistical aspects)

Keywords:

data mining; naïve Bayes; word embedding; feature selection

Software:

AdaBoost.MH; C4.5; GloVe; WEKA; word2vec

Full Text: DOI

References:

[1] Amatriain, X.; Jaimes, A.; Oliver, N.; Pujol, JM; Ricci, F.; Rokach, L.; Shapira, B.; Kantor, PB, Data mining methods for recommender systems, Recommender systems handbook, 39-71 (2011), Boston: Springer, Boston

[2] Frank, E.; Hall, MA; Witten, IH, The WEKA workbench. Online appendix for "data mining: practical machine learning tools and techniques (2016), Burlington: Morgan Kaufmann, Burlington

[3] Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proceedings of the thirteenth international conference on international conference on machine learning, Bari, Italy. Morgan Kaufmann Publishers Inc, pp. 148-156

[4] Freund, Y.; Schapire, RE, A decision-theoretic generalization of on-line learning and an application to boosting, J Comput Syst Sci, 55, 1, 119-139 (1997) · Zbl 0880.68103

[5] Han, J.; Kamber, M.; Pei, J.; Han, J.; Kamber, M.; Pei, J., 2—Getting to know your data, Data mining, 39-82 (2012), Boston: Morgan Kaufmann, Boston

[6] Han, J.; Kamber, M.; Pei, J.; Han, J.; Kamber, M.; Pei, J., 3—Data preprocessing, Data mining, 83-124 (2012), Boston: Morgan Kaufmann, Boston

[7] Han, J.; Kamber, M.; Pei, J.; Han, J.; Kamber, M.; Pei, J., 9—Classification: advanced methods, Data mining, 393-442 (2012), Boston: Morgan Kaufmann, Boston

[8] Han, J.; Kamber, M.; Pei, J., Data mining (2012), Boston: Morgan Kaufmann, Boston

[9] Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Proceedings of the 31st international conference on international conference on machine learning, vol 32. Beijing, China, JMLR.org: II-1188-II-1196

[10] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. arXiv:1310.4546

[11] Nguyen TTS (2019) Model-based book recommender systems using Naive Bayes enhanced with optimal feature selection. In: Proceedings of the 2019 8th international conference on software and computer applications, Penang, Malaysia. ACM, pp 217-222

[12] Novakovic J (2010) The impact of feature selection on the accuracy of Naive Bayes classifier. In: 18th telecommunications forum TELFOR, Serbia, Belgrade

[13] Pennington J, Socher R, Manning CD (2014) GloVe: global vectors for word representation. In: Proceedings of the 2014

conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, Association for Computational Linguistics

[14] Quinlan, JR, C4.5: programs for machine learning (1993), Boston: Morgan Kaufmann Publishers Inc, Boston

[15] Ratanamahatana, C.; Gunopulos, D., Feature selection for the naive bayesian classifier using decision trees, Appl Artif Intell, 17, 5-6, 475-487 (2003)

[16] Refaeilzadeh, P.; Tang, L.; Liu, H.; Liu, L.; Özsu, MT, Cross-validation, Encyclopedia of database systems, 532-538 (2009), Boston: Springer, Boston

[17] Shi, H.; Liu, Y., Naïve Bayes vs. support vector machine: resilience to missing data (2011), Berlin: Springer, Berlin

[18] Taheri, S.; Mammadov, M., Learning the Naive Bayes classifier with optimization models, Int J Appl Math Comput Sci, 23, 4, 787-795 (2013) · Zbl 1284.93218

[19] Tin Kam H (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition

[20] Witten, IH; Frank, E.; Hall, MA; Witten, IH; Frank, E.; Hall, MA, Chapter 2—Input: concepts, instances, and attributes, Data mining: practical machine learning tools and techniques, 39-60 (2011), Boston: Morgan Kaufmann, Boston

[21] Witten, IH; Frank, E.; Hall, MA; Witten, IH; Frank, E.; Hall, MA, Chapter 5—Credibility: evaluating what's been learned, Data mining: practical machine learning tools and techniques, 147-187 (2011), Boston: Morgan Kaufmann, Boston

[22] Witten, IH; Frank, E.; Hall, MA; Witten, IH; Frank, E.; Hall, MA, Chapter 7 - Data Transformations, Data mining: practical machine learning tools and techniques, 305-349 (2011), Boston: Morgan Kaufmann, Boston

[23] Xhemali, D.; Hinde, CJ; Stone, RG, Naïve Bayes vs. decision trees vs. neural networks in the classification of training web pages, Int J Comput Sci Issues, 4, 16-23 (2009)

[24] Xu, W.; Jiang, L.; Yu, L., An attribute value frequency-based instance weighting filter for Naive Bayes, J Exp Theor Artif Intell, 31, 1-12 (2018)

[25] Yu, H.; Liu, L.; Özsu, MT, Support vector machine, Encyclopedia of database systems, 2890-2892 (2009), Boston: Springer, Boston