Adaption of Naïve Bayes Classifier in Various Fields

Shengyuan Zhang[1,a,*]

[1]*University of California, Davis, Shields Ave, Davis, CA 95616*
a. sophiazhang217@gmail.com
**corresponding author*

Keywords: naïve Bayes classification, review

Abstract: Despite the directness of the Naive Bayes classifier, it has been one of the most used and powerful classification algorithms. Not only does the model perform better than other more complicated models, but it also requires less training sets than other models. This paper provides a summarization of the implications of the Naive Bayes classifier across many fields of study to see the execution of the Naive Bayes model in nine-teen different circumstances.

1. Introduction

Base on the Bayes' Theorem, the Naive Bayes classifiers finds the probability of an event happening given the plausibility of another event that has already occurred. It is not a single algorithm, but a family of algorithms where all of them share a universal principle, such as every pair of features being classified is independent of each other. The theorem behind the Naïve Bayes classification is straightforward $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, which find the probability of event A happening, given that event B has occurred, under the assumption that all predictors have the same effect on the outcome. This paper explores the uses of Naive Bayes in different fields, from medical, educational, to genic.

2.1. Economic

The Naïve Bayes classification is applied in the economic field by Leng[1] to analyze one of the most significant problems in the area of Chinese economic operation is the lack of methods to organically combine non-timing and timing information. Not until Leng's uses of K order multi Markov chain dynamic naïve Bayesian classifier did dynamic and static information finally combined organically. During the research, Leng analyzed the effect of time lag effect on macroeconomic risk, with data he collected on macroeconomic risk index using the macroeconomic sentiment index, and twelve factors that affecting macroeconomic risks. Through the analysis, Leng successfully proved the classifier is more reliable and practical than other existing methods in the analysis of the time delay effect.

2.2. Computer Science

The classification method is also widely used in computer science field for virus detection, classification, privacy-preserving and etc. Dash [2] in his paper discusses and experiments with the Naïve Bayes method for masquerade detection. The majority of techniques employed in the field of masquerade detection are conforming the suspicious behavior with the learnt pattern from past behaviors. However, Dash identifies the problem in current techniques, where some users momentarily diverge from their past behavior that cause potential false alarm. To address the problem, Dash applies naïve Bayes to improve the masquerade detection method by comparing the momentary deviation of behavior of the user to the consistent deviation of the attacker. With datum from Schonlau dataset, Lane and Brodley dataset, and Greenberg dataset, Dash concluded that the implication of naïve Bayes model substantially improves the accuracy of masquerade detection.

Another application of the Naïve Bayes classification in the field of computer science is Zhang's [3] research in the field of multi-label is aiming to address the label sets of unobserved instances with a technique called MLNB. The method modifies the conventional Naïve Bayes classification method to deal with multi-label situations. Zhang employs artificial data sets to assess the execution of the multi-label training procedures and real-world data sets to evaluate the performance of the corresponded algorithms. The result suggests that the MLNB method that incorporated feature selection mechanisms achieves similar performance to other well-known multi-label learning methods.

Benferhat[4] applies both the Naïve Bayes and decision tree classification methods for intrusion detection, he performs an experiment base on the speculation of better performance detection system–decision tree or Naïve Bayes–perform better. He compares the accuracy and the time decision tree, and the Naïve Bayes model takes to construct, learn, and classify with KDD'99 data set for intrusion detection. The lab report suggests, even though in some cases decision tree outperform naïve Bayes, but from the computation perspective, the building of naïve Bayes is mostly faster than decision trees.

Similarly, Bouckaert[5] proposes that the current three methods, parametric approach, nonparametric approach, and discretization, for handling continuous variables show no particular method perform better without exception. Alternatively, he offers a process for deciding the best methods to approach continuous variables that could enhance the overall performance of the naïve Bayes classifier. For the selection of the most effective method, Bouckaert executed v-fold cross-validation for all three methods. Every data Bouckaert employed, comes from two experiments, the first experiment compared each of the three methods on their own, and the second experiment exhibits results on a 10*10 cross-validation test that selected the best-performing methods out of three for managing continuous variables. He, then, presumed using 10-fold cross-validation can primarily and undoubtedly improve the overall performance of naïve Bayes classifiers. His approach to continuous variables continually exceeds any of the three conventional methods on their own.

The Naïve Bayes classification method is widely used in the computer science field. Last but not least application of the method is used for privacy-preserving, Huai[6] proposes an unconventional private protocol with Naive Bayes classification over distributed data, which believed to be stronger than existing works. Not only are both the miner and parties can be randomly malicious and can collide with each other to disrupt the remaining honest parties' privacy, but also all interaction channels between them can be expected to be insecure. Huai employs both theoretical analysis and simulations with vertically and horizontally distributed data to ensure the more effective and low-cost way to preserve privacy and

security. Base on the experiment Huai believes Naïve Bayes performs as well as the existing model with less computation and communication costs.

2.3. Biology

The simplicity of the Naïve Bayes algorithm is also applicable in the field of biology. For example, Rytkonen[7] uses the algorithm to analyze sleeping records; states of sleeping are required. The sleep recordings are classified into three different states–fully awake, NREM (non-rapid-eye-movement) sleep and REM (rapid-eye-movement) sleep. However, manual scorings of the states are not only time-consuming and expensive but also prone to human errors. Rytkonen suggests that instead of using the labor-intensive manual scoring method, apply the naïve Bayes classifier that adapts to every recording by establishing the probability model from manual scores to generate an undemanding MATLAB-based automated scoring system. Rytkonen pulls the raw data from an experiment of 30 recordings from mice and rats, where he extrapolated the conclusion that the automated scoring system has high accuracy regardless of interdependencies in the data and have an overall auto-rater agreement of 93%.

Youself[8], on the other hand, points out that the majority of existing computational procedures for micro RNA gene estimation employ techniques build on sequence conservation and structural affinity. However, he innovatively discovers a new technique that is based on machine learning, which is a generalization of multiple species for miRNA gene prediction. Yousef employs the Naïve Bayes classifier that spontaneously extrapolates a model from training sets that are found of miRNAs from known species. The resulting procedure with the Naïve Bayes method displays more substantial specificity and similar sensitivity compared to existing procedures that depend on conserved genomic regions to decline the rate of FPs. Yousef concludes that the Naïve Bayes classifier is a valuable and inclusive method for miRNA Bayes classification.

2.4. Data Analysis

A powerful and straightforward classification method, such as the Naïve Bayes algorithm, is undeniable popular in the field of data analysis. Ren[9] suggests that conventional data analysis methods presume that data are accurate or precise. Nevertheless, the presumption does not last in many circumstances due to data uncertainty. Instead of using the conventional learning methods, Ren experimented with the UCI database by extending the class conditional probability evaluation in the Baes model to manage probability distribution functions. Base on the experimental results, he concluded that the precision of the naïve Bayes model could be enhanced by considering the uncertainty information.

Another implementation of the algorithm in the data analysis field is suggested by Frank[10] for regression purposes. According to Frank, Naïve Bayes performs well for categorical data, but how it operates in domains where the anticipated value is numeric, which is more sensitive to inaccurate probability estimates. To solve the enigma, Frank applies naïve Bayes methodology, with data from Kilpatrick and Cameron-Jones, StatLib repository, Simonoff, to numeric prediction tasks and examines them to linear regress, instance-based learning, and "model trees." The outcome suggests the Naive Bayes perform comparably to linear regression, and it outperforms the other methods concerning both error measures by a large margin on the standard datasets used during the experiment.

Furthermore, Lowd[11], in his paper, suggests Naïve Bayes models have been used occasionally for generic probabilistic learning and inference, yet it has the precision and learning time tantamount to the

Bayesian network with a broad range of standard datasets. Lowd use datasets, 47 datasets from the UCI repository, ranging in number of 5 to 618 variables, and size from 57 to 57000 samples, to perform ten-fold cross-validation for datasets less than 200, and single train-test split for more massive datasets to compare the training time of the WinMine, software for learning Bayesian networks, with the NBE model. Lowd concludes from the experiment, both naïve Bayes models and Bayesian networks take a similar time with comparable accurate, but naïve Bayes inference is orders of magnitude faster.

2.5. Library Science

Due to the nature of the naïve Bayes algorithm, it has been prevailingly used for classifying documents, texts, and other branches of library science. Information nowadays is exceedingly accessible and blooming faster than ever, yet the lack of sound organization of the documents from the Web makes retrieving documents from the loaded Web a strenuous and time-consuming task. Wang[12], the author of web documents using a Naïve Bayes method, realized that even though the many method was considered, most of them are still laborious and costly. Instead, Wang employs Naïve Bayes methods that focus on two different probability smoothing methods–additive and good-turning smoothing method. Wang adopts 722 documents from the Web and uses them to generate and experiment, where documents are firstly assigned by expert librarians with accurate LCSHs, and also divided by five-fold-cross-validation method into training and test set for testing and comparison using Naïve Bayes method. The experiment suggests that the Naïve Bayes method of classifying documents is faster than that of expert librarians, yet it fails to improve the performance due to the absence of a substantial number of training documents.

Other presentation of the algorithm in the library science can be found in Kim's[13] research. Kim indicates in his paper, even though the Naive Bayes classification method has been straightforward and powerful in many data analysis tasks, it demonstrated unsatisfactory results in the automatic text classification for the field of librarian science. To address the problem, he practices per-document text normalization and feature weighting methods to improve the performance of Naïve Bayes for the experiment, which used commonly used text categorization: Reuters21578 and 20 Newsgroup. Kim concludes based on the experiment that the Naive Bayes classification performs well enough to compete with far more complex learning method including SVM, which solves the difficulty in the parameter assessment process that cause unsatisfactory results in the text classification field.

Last but not least, Dai[14] wanders if training and test data distribution have to be identical for the test classification. Dai proposes a learning algorithm based on an EM-based Naive Bayes classification method in the domain of text classification. He employs a wide range of datasets from 20 Newsgroups, SRAA and Reuters-21578, which yields results showing the algorithm is highly efficient in various pairs of domains. The algorithm exceeds the performance of the conventional supervised algorithms despite the fact that the distributions of the training and test sets are remarkably different, which solves the problem of expensive data of the domain of interest, and makes transfer learning is affordable and effective.

2.6. Medical

The naïve Bayes classification method is likewise prevalent in the medical field for predicting the injury or disease. The exposure to the radiation is harmful, particularly for children. In Klement's[15] paper, the

researcher wants to promote the adoption of a comprehensive approach, which minimized the use of computed tomography imaging for patients with head injuries. Klement predicted the computed imaging decision based collection of multiple Naïve Bayes classifiers. The data he uses in the experiment is from the CATCH dataset, which was created by a group of pediatric doctors to support the decision of indicating children with secondary head injury computed tomography imaging. According to Klement, the prediction model demonstrates the best presentation in terms of AUC< G-mean and sensitivity measure, and the ensemble model achieved a more harmonious predictive execution than the CATCH rule.

Also, in the medical field, Wei[16] holds that clinical care can be dramatically improved if doctors can predict patient outcomes from a genome-wide measurement. Therefore, he constructed a method of predicting patients outcomes from genome-wide data by productively model the average over an exponential number of the naïve Bayes models. During the experiment, Wei applied the model-averaged naïve Bayes method to predict late-onset Alzheimer's disease in 1411 individuals whom each had 312318 SNP measures available as genome-wide predictive features. The result suggests that MANB(model-averaged naïve Bayes) performed well in predicting a clinical outcome from a high-dimensional genome-wide dataset.

2.7. Educational

Unsurprisingly the model is also implemented in the field of education to predict the performance of students and help them to succeed in their study. Razaque[17] conceptualizes a model of predicting the possible failure performance of students, which aid institutions in providing useful, timely assistant and necessary steps to improve the performance of those students. He utilizes enormous data stored in academic datasets that include relevant information for assessing the performance of students, which is a classifier depend on the Naive Bayes algorithm and used for Academic data mining. His model is proven to help schools and students who required special attention to improving performance and taking proper action for forthcoming exam to succeed academically.

The traditional testing indeed evaluates the knowledge of the test taker, but Agarwal[18] argues that the static test is not effective to compare the to the adaptive testing, which adapts to the user's knowledge and apace evaluate the true ability of the test taker. Agarwal utilizes the naïve Bayesian classification method to assign future questions that are appropriate for the user according to the previous set of questions. Base on this assumption, Agarwal designs an experiment where he induces an adaptive test of forty-five questions, which adapt to test taker's knowledge after every fifteen questions. Based on data from the experiment, Agarwal believes that the adaptive test not only helps to identify test taker's areas of weakness quickly but also helped to improve those weaknesses.

2.8. Energy

Finally, the Naïve Bayes classification method is used by Ng[19] to discover the robustness of lithium-ion batteries, where she argues that the management of battery health must generate a model for estimation of the remaining functional life of lithium battery. To explore the problem, Ng applies the Naïve Bayes model on the NASA Ames Li-ion battery cycle life test data, which consists of 18,650 sized lithium batteries that completed adequate cycles of complete charges and complete discharges to precipitate the aging process. Ng concluded that the naïve Bayes method produces more accurate and

stable predictions than other more complicated methods.

3. Conclusion

Out of the nineteen different circumstances, six of the lab reported suggested the Naive Bayes not only successfully replaced more complicated and time/human labor-consuming methods, but it also outperforms them. Five of the results suggest equivalent performances to more complex models, which significantly reduced the time and resources required for researches. Three suggest some areas of outperformance and other regions of equivalence. Two of the result fail to improve one due to lack of training set, and the other fails to have equivalent performance than more complicated models. The remaining three researchers discovered new models based on Naive Bayes models to predict the clinical outcomes, to transfer datasets form different fields to reduce the cost of the study, and to predict the performance of students.

Reference

[1] LENG Cuiping1, WANG Shuangcheng1,2 et al. Method of dynamic Bayesian classifier for analysis of macroeconomic risk[J]. CEA, 2016, 52(3): 224-229.

[2] Dash, Subrat & Reddy, Krupa & Pujari, Arun K. (2011). Adaptive Naive Bayes method for masquerade detection. Security and Communication Networks. 4. 410-417. 10.1002/sec.168.

[3] Zhang, Min-Ling & Peña, José & Robles, Víctor. (2009). Feature selection for multi-label naive Bayes classification. Information Sciences. 179. 3218-3229. 10.1016/j.ins.2009.06.010.

[4] Amor, Nahla & Benferhat, Salem & Elouedi, Zied. (2004). Naive Bayes vs decision trees in intrusion detection systems. Proceedings of the ACM Symposium on Applied Computing. 1. 420-424. 10.1145/967900.967989.

[5] Bouckaert, Remco. (2004). Naive Bayes Classifiers That Perform Well with Continuous Variables. 3339. 1089-1094. 10.1007/978-3-540-30549-1_106.

[6] Huai M., Huang L., Yang W., Li L., Qi M. (2015) Privacy-Preserving Naive Bayes Classification. In: Zhang S., Wirsing M., Zhang Z. (eds) Knowledge Science, Engineering and Management. KSEM 2015. Lecture Notes in Computer Science, vol 9403. Springer, Cham

[7] Rytkönen, Kirsi-Marja & Zitting, Jukka & Porkka-Heiskanen, Tarja. (2011). Automated sleep scoring in rats and mice using the naive Bayes classifier. Journal of neuroscience methods. 202. 60-4. 10.1016/j.jneumeth.2011.08.023.

[8] Yousef, Malik & Nebozhyn, Michael & Shatkay, Hagit & Kanterakis, Stathis & Showe, Louise & Showe, M.K.. (2006). Combining multi-species genomic data for MicroRNA identification using a Na??ve Bayes classifier. Bioinformatics (Oxford, England). 22. 1325-34. 10.1093/bioinformatics/btl094.

[9] Ren, Jiangtao & Lee, Sau & Chen, Xianlu & Kao, Ben & Cheng, Reynold & Cheung, David. (2009). Naive Bayes Classification of Uncertain Data. 944-949. 10.1109/ICDM.2009.90.

[10] Frank, Eibe & Trigg, Len & Holmes, Geoffrey & Witten, Ian & Aha, W.. (2001). Naive Bayes for Regression. Machine Learning.

[11] Lowd, Daniel & Domingos, Pedro. (2005). Naive Bayes models for probability estimation. ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning. 529-536. 10.1145/1102351.1102418.

[12] Wang, Yong & Hodges, Julia & Tang, Bo. (2003). Classification of Web Documents Using a Naive Bayes Method. IEEE Transactions on Applications and Industry. 560- 564. 10.1109/TAI.2003.1250241.

[13] Kim, Sang-Bum & Han, Kyoung-Soo & Rim, Hae-Chang & Myaeng, Sung-Hyon. (2006). Some Effective Techniques for Naive Bayes Text Classification. Knowledge and Data Engineering, IEEE Transactions on. 18. 1457-1466. 10.1109/TKDE.2006.180.

[14] Dai, Wenyuan & Xue, Gui-Rong & Yang, Qiang & Yu, Yong. (2007). Transferring Naive Bayes Classifiers for Text Classification.. 540-545.

[15] Klement, William & Wilk, Szymon & Michalowski, Wojtek & Farion, Ken & Osmond, Martin & Verter, Vedat. (2011). Predicting the need for CT imaging in children with minor head injury using an ensemble of Naive Bayes classifiers. Artificial intelligence in medicine. 54. 163-70. 10.1016/j.artmed.2011.11.005.

[16] Wei, Wei & Visweswaran, Shyam & Cooper, Gregory. (2011). The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data. Journal of the American Medical Informatics Association : JAMIA. 18. 370-5. 10.1136/amiajnl-2011-000101.

[17] Razaque Mughal, Fahad & Soomro, Nareena & Shaikh, Shoaib & Soomro, Safeeullah & Samo, Javed & Kumar, Natesh & Dharejo, Huma. (2017). Using naïve bayes algorithm to students' bachelor academic performances analysis. 1-5. 10.1109/ICETAS.2017.8277884.

[18] Agarwal, Sanjana & Jain, Nirav & Dholay, Surekha. (2015). Adaptive Testing and Performance Analysis Using Naive Bayes Classifier. Procedia Computer Science. 45. 10.1016/j.procs.2015.03.088.

[19] Ng, Selina & Xing, Yinjiao & Tsui, Kwok-Leung. (2014). A naive Bayes model for robust remaining useful life prediction of lithium-ion battery. Applied Energy. 118. 114–123. 10.1016/j.apenergy.2013.12.020.