Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 11, November -2017

# Short Survey on Naive Bayes Algorithm

Pouria Kaviani[1], Mrs. Sunita Dhotre[2]

[1]*M.Tech student, Department of Computer Engineering, Bharati Vidyapeeth University, College of Engineering, Pune*
[2]*Associate Professor,Department of Computer Engineering, Bharati Vidyapeeth University, College of Engineering, Pune*

**Abstract**—*Naive Bayes is a classification algorithm which is based on Bayes theorem with strong and naïve independence assumptions. It simplifies learning by assuming that features are independent of given class.This paper surveys about naïve Bayes algorithm, which describes its concept, hidden naïve Bayes, textclassification, traditional naïve Bayes and machine learning. Also represents augmented naïve Bayes by examples. And at the end some applications of naïve Bayes and its advantages and disadvantages has discussed for a better understanding of the algorithm.*

***Keywords***-*Naive Bayes; Classifier; Big Data; Algorithm; Data Mining; Classification; Machine Learning*

## I. INTRODUCTION

Naïve Bayes is a subset of Bayesian decision theory. It's called naive because the formulation makes some naïve assumptions. Python's text-processing abilities which split up a document into a vector are used. This can be used to classify text. Classifies may put into human-readable form. It is a popular classification method in addition to conditional independence, overfitting, and Bayesian methods.

In the face of the simplicity of Naive Bayes, it can classify documents surprisingly well. Instinctively a potential justification for the conditional independence assumption is that if the document is about politics, this is a good evidence of the kinds of other words found in the document. Naive Bayes is a reasonable classifier in this sense and has minimal storage and fast training, it is applied to time-storage critical applications, such as automatically classifying web pages into types and spam filtering.

Considering a set of objects, each of which belongs to a known class, and each of which has a known vector of variables, the aim is to create a rule which enables to allocate future objects to a class, given just the vectors of variables marking out the future objects. These problems are known as "supervised classification problem", are worldwide, and most of the methods for constructing such rules have been developed. It is very easy to establish, and no need any complicated repetitive parameter estimation schemes. This means it should be applied to huge data sets. It is easy to interpret, so unskilled users in classifier technology can make out the reason for it is making the classification it makes. Finally, it often does surprisingly well: it should not be the best possible classifier in any particular application, but it can usually be relied on to be robust and to do well.

## II. LITERATURE SURVEY

In the paper [1], the existing improved algorithms are summarized and a novel Bayes model is proposed: hidden naive Bayes (HNB). In HNB, a hidden parent is created for each attribute which combines the influences from all other attributes. A systematic experimental study on the classification, class probability estimation and ranking performance of HNB is done. The experimental results show that HNB has a better overall performance compared to the other state-of-the-art models for augmenting naive Bayes.

The structure of HNB contains one hidden layer. It can be expected to extend to more complex structures, such as double layers. This will be another topic for future research.

Naive Bayes is among the simplest probabilistic classifiers [2]. It often shows amazingly well in many real-world applications, in the face of the strong assumption that all features are provisionally independent given the class. The values of these variables are found by solving the equating optimization problems. The obtained results specify that the proposed models can importantly improve the display of the naive Bayes classifier, yet at the same time conserve its simple structure.

This paper [2] has presented three different optimization models for the naive Bayes classifier. Then applied optimization techniques to find the optimal values for these variables. Also, compared the proposed models with NB, TAN, the SVM, C4.5, and 1-NN on 14 real-world binary classification data sets. The values of features in data sets are individualized by using a median-based individualization method and applying two different individualization algorithms, the Fayyad and Irani method and the algorithm SOAC. Authors have presented results of numerical experiments. The results demonstrate that the proposed models show better than NB, TAN, the SVM, C4.5, and 1-NN in terms of accuracy, yet at the same time they maintain the simple structure of NB. Especially Model 3 expanded the test set accuracy of each dataset and

 

this is predictable due to considering more variables replacing class probabilities and conditional probabilities which enables to build a more accurate model. At the end, the main focus is on binary classification datasets since the simplest among the main classification categories. The applications of the proposed models for other types of data sets, and also normalizing Model 3 to any discrete features, persist in being significant questions for future work.

Traditional machine learning algorithms suppose that data are faithful. Although, this assumption [3] may not hold in some cases by reason of data uncertainty appearing from measurement errors, data staleness, and replicated measurements, etc. With uncertainty, the value of each data item is represented by a probability distribution function (pdf). In this paper, a novel naive Bayes classification algorithm for uncertain data with a pdf is proposed. The Key solution is to extend the class conditional probability estimation in the Bayes model to handle pdf.

In the paper [3], the author has addressed the problem of extending traditional naïve Bayes model to the classification of uncertain data. The key problem in naive Bayes model is a class conditional probability estimation, and kernel density estimation is a common way for that. The kernel density estimation method has extended to handle uncertain data. This decreases the problem to the consideration of double-integrals. For specific kernel functions and possibility distributions, the double integral can be analytically assessed to give a closed-form formula, authorizing an efficient formula-based algorithm. In general, however, the double integral cannot be simplified in closed forms. In this case, a sample-based approach is proposed. Extensive experiments on several UCI datasets show that the uncertain naive Bayes model considering the full pdf information of uncertain data can produce classifiers with higher accuracy than the traditional model that utilizing the mean as the representative value of uncertain data. Time complexity analysis and performance analysis based on experiments perform that the formula-based approach has special advantages over the sample-based approach.

The naive Bayes classifier especially simplifies learning by expecting that features are autonomic given class. However, independence is generally a poor expectation, in practice, naive Bayes sometimes approaches well with more sophisticated classifiers. [4]

Despite its unrealistic independence assumption, the naïve Bayes classifier is surprisingly effective in practice since its classification decision may often be correct even if its probability estimates are inaccurate. Although some optimality conditions of naive Bayes have been already identified in the past, a deeper understanding of data characteristics that affect the performance of naive Bayes is still required.

Instead, a better predictor of accuracy is the information loss that features contain the class when assuming naive Bayes model. However, further empirical and theoretical study is required to better understand the relationshipbetween those information-theoretic metrics and the behavior of naive Bayes. Further directions also include analysis of naive Bayes on the practical application that has almost-deterministic dependencies, characterizing other regions of naive Bayes optimality and studying the effect of various data parameters on the naive Bayes error. Better approximation techniques for learning efficient Bayesian net classifiers, and for probabilistic inference are devised.

Naïve Bayes classifiers which are vastly used for the text classification in machine learning are based on the conditional possibility of features attributed to a class, which the features are selected by feature selection methods. In the paper [5], an auxiliary feature method is proposed. It determines features by an existing feature selection method and selects an auxiliary feature which can reclassify the text space aimed at the chosen features. After that, the corresponding conditional probability is modified to refine classification accuracy. Illuminative examples perform that the proposed method indeed improves the performance of naive Bayes classifier.

After feature selection in text classification, naive Bayes classifier partition the text subspace composed of all document that presents based on each since naïve Bayes classifier gives the document to the class with the highest probability, naïve Bayes classifier is optimal in probability sense. The auxiliary feature method proposed here partition the text subspace again, so it outperforms the traditional way, and illustrative examples show that the proposed method indeed improves the performance of naive Bayes classifier.

Since the auxiliary feature method need choose features twice, how to give the auxiliary directly is meaningful and can reduce the computation complexity. Meanwhile, the relationship between existing features methods and our method is promising. In addition, due to the sparsity problem in text classification, whether to take the feature total of the document into account when adjusting the probability is worth to work other than substitution in this paper.

## III. NAIVE BAYES CLASSIFIER

A naïve Bayes classifier corresponds to a Bayesian network, as in Eq (1). Here, a single class variable $C$ and $m$ attribute variables $X_i$(for simplicity of exposition, attributesare discrete). Let $c$ denote a class label and $x_i$ denote a value of an attribute $X_i$. A naïve Bayes induces a distribution:

$$Pr(c, x_1, \ldots, x_m) = Pr(c) \cdot \prod_{i=1}^{m} Pr(x_i \mid c) \qquad \textbf{Eq (1)}$$

 

Where we have a class prior $Pr(C)$ and conditional distributions $Pr(X_i|C)$. We can estimate these parameters from(labeled) data, using maximum likelihood or MAP estimation.Once we have learned a naïve Bayes classifier from data, we can label new instances by selecting the class label $c^*$ that has maximum posterior probability given observation s$x_1,...,x_m$. select:

$$c^\star = \underset{c}{\operatorname{argmax}}\, Pr(c \mid x_1, \ldots, x_m). \qquad \textbf{Eq (2)}$$

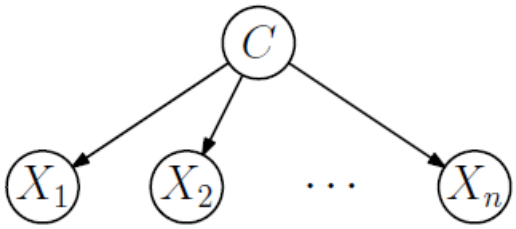


***Figure 1. A Naïve Bayes Classifier***

## IV. NAIVE BAYES AND AUGMENTED NAIVE BAYES

Classification is a fundamental issue in machine learning and data mining. In classification, the goal of a learning algorithm is to construct a classifier given a set of training examples with class labels. Regularlyexample $E$ is represented by attribute values by a tuple($x_1,x_2,\cdots,x_n$), where $x_i$is the value of attribute $X_i$. Let $C$ represent the classification variable, and let $c$ be the value of $C$. There are only two classes here: +(the positive class) or −(the negative class).

A classifier is a function that assigns a class label to an example. From the probability perspective, according to Bayes Rule, the probability of an example $E$=($x_1,x_2,\cdots,x_n$) being class $c$ is:

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)}. \qquad \textbf{Eq (3)}$$

E is classified as the class C =+ if and only if:

$$f_b(E) = \frac{p(C=+|E)}{p(C=-|E)} \geq 1, \qquad \textbf{Eq (4)}$$

Where $f_b$ ($E$) is called a Bayesian classifier.

Assume that all attributes are independent given the value of the class variable; that is,

$$f_{nb}(E) = \frac{p(C=+)}{p(C=-)} \prod_{i=1}^{n} \frac{p(x_i|C=+)}{p(x_i|C=-)}. \qquad \textbf{Eq (5)}$$

The function $f_{nb}(E)$ is called a naïve Bayesian classifier, or simply naïve Bayes(NB). In naïve Bayes, each attribute node has no parent except the class node.

Naïve Bayes is the simplest form of Bayesian network, in which all attributes are independent given the value of the class variable. This is called conditional independence. It is clear that the conditional independence assumption is infrequently correct in most of the real-world applications. A straightforward approach to control the limitation of naïve Bayes is to increase its structure to represent explicitly the dependencies among attributes. An augmented naive Bayesian

 

network, or simply augmented naive Bayes (ANB), is an extended naive Bayes, that the class node points to all attribute nodes directly, and there found links among attribute nodes. Figure 2 shows an example of ANB. From the view of possibility, an ANB G shows a joint probability distribution depicted below.

$$p_G(x_1, \cdots, x_n, c) = p(c) \prod_{i=1}^{n} p(x_i | pa(x_i), c), \qquad \textbf{Eq (6)}$$

Where $pa(x_i)$ denotes an assignment to values of the parents of $X_i$. the$pa(X_i)$ is used to denote the parents of $X_i$. ANB is a remarkable form of Bayesian networks that no node is identified as a class node. It has been shown that any Bayesian network can be depicted by an ANB. Thus, any joint probability distribution can be rendered by an ANB.

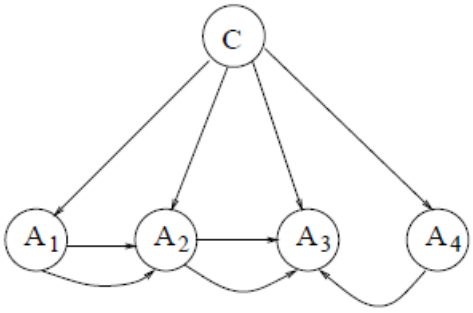


***Figure 2. An Example of ANB***

When a logarithm applies to $f_b(E)$ in Equation1, the resulting classifier log $f_b(E)$ is the same as $f_b(E)$, in the sense that example$E$ belongs to the positive class, if and only if log $f_b(E) \geq 0$. $F_{nb}$in Equation2 is similar. Given a classifier $f$,an example $E$ belongs to the positive class, if and only if $f(E) \geq 0$.

## V. NAIVE BAYES APPLICATIONS

The Naive Bayes algorithm is used in multiple real-life scenarios such as:

### 5.1. Text classification:

It is used as a probabilistic learning method for text classification. The Naive Bayes classifier is one of the most successful known algorithms when it comes to the classification of text documents, i.e., whether a text document belongs to one or more categories (classes).

### 5.2. Spam filtration:

It is an example of text classification. This has become a popular mechanism to distinguish spam email from legitimate email. Several modern email services implement Bayesian spam filtering. Many server-side email filters, such as DSPAM, Spam Bayes, Spam Assassin, Bogofilter, and ASSP, use this technique.

### 5.3. Sentiment Analysis:

It can be used to analyze the tone of tweets, comments, and reviews—whether they are negative, positive or neutral.

### 5.4. Recommendation System:

The Naive Bayes algorithm in combination with collaborative filtering is used to build hybrid recommendation systems which help in predicting if a user would like a given resource or not.

 

## VI. ADVANTAGES AND DISADVANTAGES OF NAIVE BAYES ALGORITHM

### 6.1. Advantages

- It is a relatively simple algorithm to understand and build.
- It is faster to predict classes using this algorithm than many other classification algorithms.
- It can be easily trained using a small dataset.

### 6.2. Disadvantages

- One of the problems of Naïve Bayes is known as the "Zero Conditional Probability Problem." This problem wipes out all the information in other probabilities too. There are several sample correction techniques to fix this problem such as "Laplacian Correction."
- Another disadvantage is the very strong assumption of independence class features that it makes. It is near to impossible to find such data sets in real life.

## VII. CONCLUSION

This paper presents the survey of Naive Bayes Algorithm whichdiscussed augmented Naïve Bayes text classification, Spam filtration, Sentiment analysis, and Recommendation System are some of the important applications of this algorithm. It has also some problems such as Zero Conditional Probability Problem and how to solve it. The key problem in naive Bayes model is a class conditional probability estimation, and kernel density estimation is a common way for that. This reduces the problem to the evaluation of double-integrals. For particular kernel functions and probability distributions, the double integral can be analytically evaluated to give a closed-form formula, allowing an efficient formula-based algorithm. In general, however, the double integral cannot be simplified in closed forms.

## REFERENCES

[1] Liangxiao Jiang, Harry Zhang, and ZhihuaCai, “A Novel Bayes Model, Hidden Naive Bayes”, IEEE Transaction on Knowledge and Data Engineering, Vol 21, No. 10, pp. 1361 – 1371
[2] S. O. N. A. Taheri, "Learning the naive Bayes classifier with optimization models,” vol. 23, no. 4, pp. 787–795, 2013.
[3] J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng, and D. Cheung, “Naive Bayes Classification of Uncertain Data,” no. 60703110.
[4] I. Rish, "An Empirical Study of the Naïve Bayes Classifier," no. January 2014.
[5] W. Zhang and F. Gao, “Procedia Engineering An Improvement to Naive Bayes for Text Classification,” vol. 15, pp. 2160–2164, 2011.
[6] SiddharthBanga, SakshamMongia, Vaibhav Tiwari, Mrs. SunitaDhotre, “Regression and Augmentation Analytics on Earth’s Surface Temperature”, IJCST, 2017.
[7] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey, J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, “Top 10 algorithms in data mining”, Springer-Verlag London, 2007
[8] Arthur Chol, NazgolTavabi, Adnan Darwiche, “Structured Features in Naive Bayes Classification”, Association for the Advancement of Artificial Intelligence, 2016.
[9] Harry Zhang, “The Optimality of Naive Bayes”, American Association for Artificial Intelligence, 2004.
[10] Toon Calders, SiccoVerwer, “Three naive Bayes approaches for discrimination-free classification”, Data Min Knowl Disk, 2010.

 