Optimality of Naïve Bayes Classifier in Comparison to Other Complex Classifiers

Sandeep Katypally
NC State University
Centenial Campus
Raleigh
Telephone number, incl. country code
1st author's E-mail address

Michael Herzog
NC State University
Centenial Campus
Raleigh
Telephone number, incl. country code
2nd E-mail

Tim Menzies
NC State University
Centenial Campus
Raleigh
Telephone number, incl. country code
3rd E-mail

ABSTRACT

The simple Bayes classifier is known to be optimal when the attributes are independent in the data given the classes. However, Pazani et al [1] suggest that Naïve Bayes performs well in many domains which contain attribute dependencies and they also suggest that this classifier often outperforms more powerful classifiers. We reproduce these suggestions and verify if it is consistent with software datasets. We also check how Naïve Bayes and other classifiers perform when the SMOTE method is performed on the skewed dataset.

In this paper, we compare Naïve Bayes classifier with 9 other classifiers under different conditions and give an analysis of how Naïve Bayes compares against the other classifiers.

CCS Concepts

Computing methodologies → Supervised learning by classification • Computing methodologies → Cross-validation;

Keywords

SMOTE, sklearn, …

1. INTRODUCTION

In the field of machine learning, supervised learning problems are those where classification models learn the model from a set of training examples and their corresponding class labels, then ouputs a classifier. Now, the classifier takes unlabeled data and assigns it to the class label.

The proceedings are the records of the conference. ACM hopes to give these conference by-products a single, high-quality appearance. To do this, we ask that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download a template from [2], and replace the content with your own material.

2. PAGE SIZE

All material on each page should fit within a rectangle of 18 × 23.5 cm (7" × 9.25"), centered on the page, beginning 1.9 cm (0.75") from the top of the page and ending with 2.54 cm (1") from the bottom. The right and left margins should be 1.9 cm (.75"). The text should be in two 8.45 cm (3.33") columns with a .83 cm (.33") gutter.

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

3. TYPESET TEXT

3.1 Normal or Body Text

Please use a 9-point Times Roman font, or other Roman font with serifs, as close as possible in appearance to Times Roman in which these guidelines have been set. The goal is to have a 9-point text, as you see here. Please use sans-serif or non-proportional fonts only for special purposes, such as distinguishing source code text. If Times Roman is not available, try the font named Computer Modern Roman. On a Macintosh, use the font named Times. Right margins should be justified, not ragged.

3.2 Title and Authors

The title (Helvetica 18-point bold), authors' names (Helvetica 12-point) and affiliations (Helvetica 10-point) run across the full width of the page – one column wide. We also recommend phone number (Helvetica 10-point) and e-mail address (Helvetica 12-point). See the top of this page for three addresses. If only one address is needed, center all address text. For two addresses, use two centered tabs, and so on. For more than three authors, you may have to improvise.[1]

3.3 First Page Copyright Notice

Please leave 3.81 cm (1.5") of blank text box at the bottom of the left column of the first page for the copyright notice.

3.4 Subsequent Pages

For pages other than the first page, start at the top of the page, and continue in double-column format. The two columns on the last page should be as close to equal length as possible.

Table 1. Table captions should be placed above the table

Graphics	Top	In-between	Bottom
Tables	End	Last	First
Figures	Good	Similar	Very well

[1] If necessary, you may place some address information in a footnote, or in a named section at the end of your paper.

3.5 References and Citations

Footnotes should be Times New Roman 9-point, and justified to the full width of the column.

Use the "ACM Reference format" for references – that is, a numbered list at the end of the article, ordered alphabetically and formatted accordingly. See examples of some typical reference types, in the new "ACM Reference format", at the end of this document. Within this template, use the style named *references* for the text. Acceptable abbreviations, for journal names, can be found here: http://library.caltech.edu/reference/abbreviations/. Word may try to automatically 'underline' hotlinks in your references, the correct style is NO underlining.

The references are also in 9 pt., but that section (see Section 7) is ragged right. References should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible (i.e. you can give the address to obtain the report within your citation) and may be obtained by any reader. Proprietary information may not be cited. Private communications should be acknowledged, not referenced (e.g., "[Robertson, personal communication]").

3.6 Page Numbering, Headers and Footers

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

4. FIGURES/CAPTIONS

Place Tables/Figures/Images in text as close to the reference as possible (see Figure 1). It may extend across both columns to a maximum width of 17.78 cm (7").

Captions should be Times New Roman 9-point bold. They should be numbered (e.g., "Table 1" or "Figure 2"), please note that the word for Table and Figure are spelled out. Figure's captions should be centered beneath the image or picture, and Table captions should be centered above the table body.

5. Experiment

The heading of a section should be in Times New Roman 12-point bold in all-capitals flush left with an additional 6-points of white space above the section head. Sections and subsequent sub-sections should be numbered and flush left. For a section head and a subsection head together (such as Section 3 and subsection 3.1), use no additional space above the subsection head.

5.1 Setup

Python with Numpy and Sci-Kit Learn libraries been used to perform the experiment on a Macintosh operating system.

10 Classifiers as mentioned in the previous sections including Naïve Bayes has been used on 10 software datasets. The datasets are taken from the Promise repository[2].

5.2 Preprocessing

Discretization of the datasets has been performed before any further processing. 2 kinds of discretization have been used on the datasets 1) Equal Frequency Discretization 2) Equal Width Discretization.

SMOTE technique has been performed on the training dataset since the datasets are skewed (i.e positive to negative ratio of the labels was much greater or less than 1).

5.3 Results

Naïve Bayes classifier is compared to other classifiers for every dataset. A plot per performance metric i.e. for accuracy, precision, recall, F beta score and run time have been plotted as below.

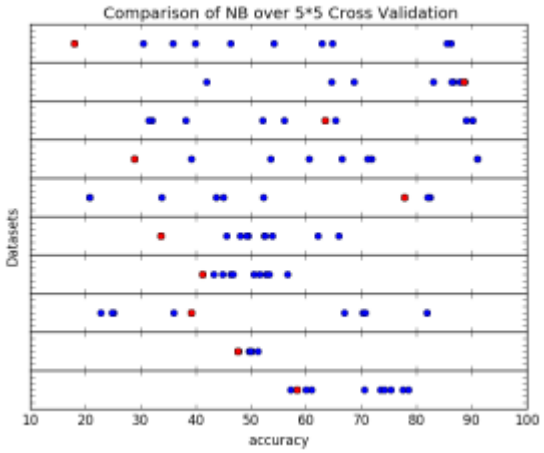

Figure 1. Accuracy comparison of 10 datasets on 10 classifiers 5×5 cross-validation. Red dot represents the Naïve Bayes classifier; blue dots represents other 9 classifiers. SMOTE has been performed on the training dataset only.

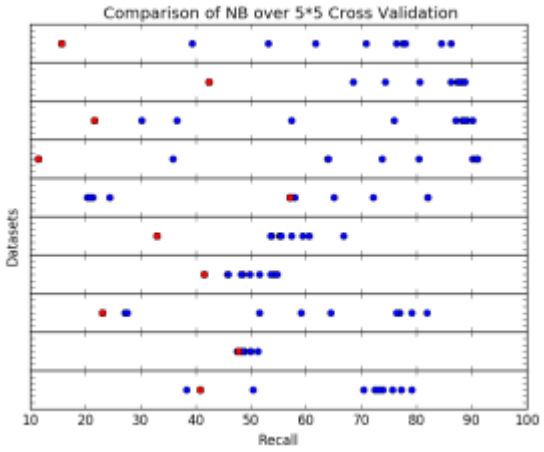

Figure 2. Recall comparison of 10 datasets on 10 classifiers 5×5 cross-validation. Red dot represents the Naïve Bayes classifier; blue dots represents other 9 classifiers. SMOTE has been performed on the training dataset only.

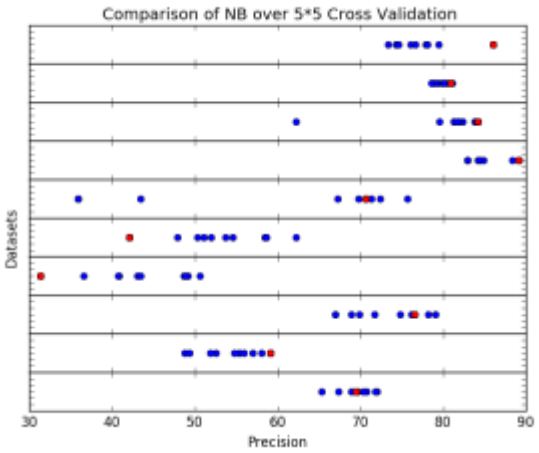

Figure 3. Precision comparison of 10 datasets on 10 classifiers 5×5 cross-validation. Red dot represents the Naïve Bayes classifier; blue dots represent other 9 classifiers.

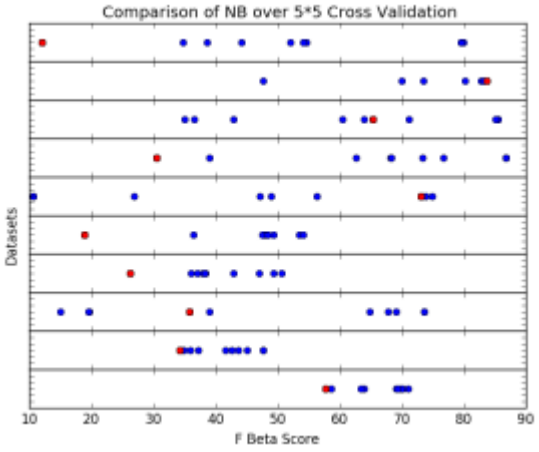

Figure 4 F Beta Score comparison of 10 datasets on 10 classifiers 5×5 cross-validation. Red dot represents the Naïve Bayes classifier; blue dots represent other 9 classifiers.

From the above plots, in most of the datasets accuracy of the NB classifier is worse than most classifiers. Likewise, recall of NB is also worse compared to most other classifiers. However, precision of NB is better than most of the classifiers. F Beta score(Beta=1) is worse for NB classifier. In some cases, F Beta score of NB is the worst compared to other classifiers. The performance of the NB classifier is inconsistent throughout 10 datasets but mostly performing worse than other classifiers.

Subsubsections

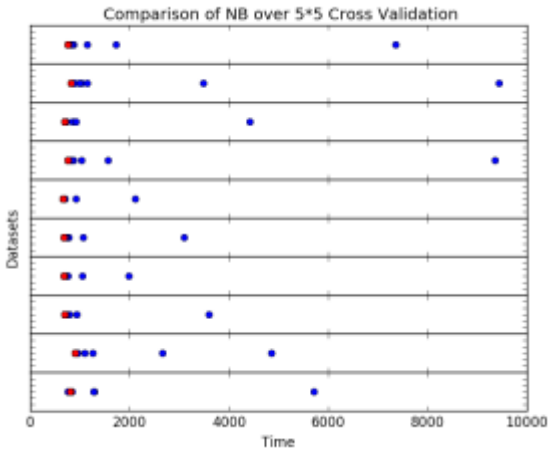

Figure 5. Run-time comparison of 10 datasets on 10 classifiers 5×5 cross-validation. Red dot represents the Naïve Bayes classifier; blue dots represent other 9 classifiers.

From above run-time graph, NB classifier runs faster than always when compared to other classifiers. This is the only consistency we have seen for NB so far. We also plotted the performance metrics of the classifiers on datasets without applying SMOTE on them so that we could compared the performance of NB and other classifiers due to SMOTE.

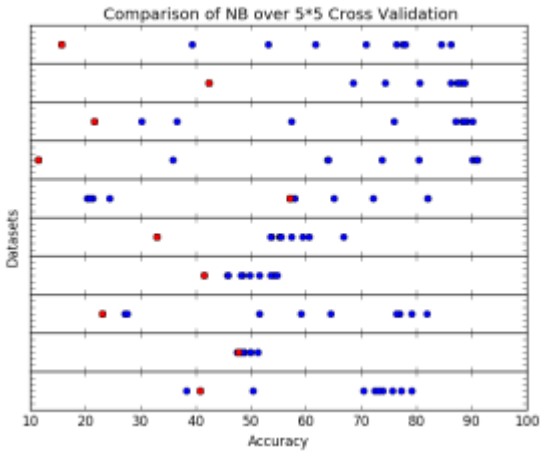

Figure 6 Accuracy comparison of 10 datasets on 10 classifiers 5×5 cross-validation with NO Smote.

Accuracy is much worse for NB compared to when we did SMOTE.

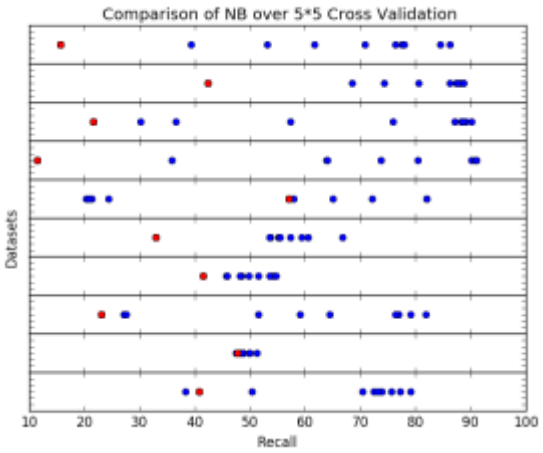

Figure 6 Recall comparison of 10 datasets on 10 classifiers 5×5 cross-validation with NO Smote.

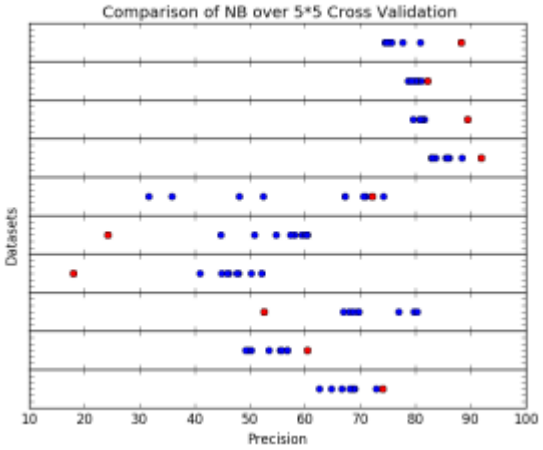

Figure 6 Precision comparison of 10 datasets on 10 classifiers 5×5 cross-validation with NO Smote.

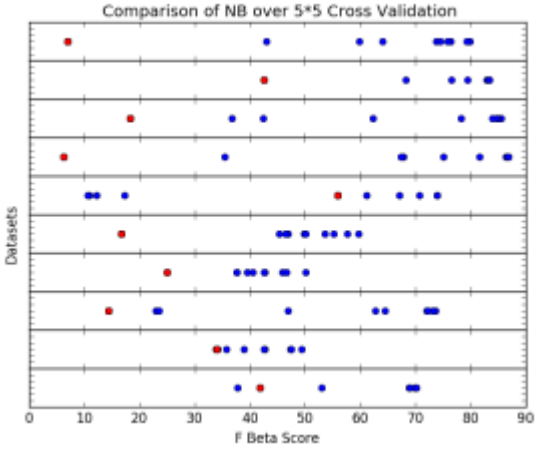

Figure 6 F beta score comparison of 10 datasets on 10 classifiers 5×5 cross-validation with NO Smote.

To understand the effect of Smote better, following plots showing the difference in the performance metrics of the classifiers for same datasets with and without Smote have been plotted.

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

5.3.1.1 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

The heading of subsections should be in Times New Roman 12-point bold with only the initial letters capitalized. (Note: For subsections and subsubsections, a word like *the* or *a* is not capitalized unless it is the first word of the header.)

5.3.2 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized and 6-points of white space above the subsubsection head.

5.3.2.1 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

5.3.2.2 Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

6. ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

7. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.16147.

[2] Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.

[3] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands, April 01 - 06, 2000). CHI '00. ACM, New York, NY, 526-531. DOI= http://doi.acm.org/10.1145/332040.332491.

[4] Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.

[5] Sannella, M. J. 1994. *Constraint Satisfaction and Debugging for Interactive User Interfaces*. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM, New York, NY, 1-10. DOI= http://doi.acm.org/10.1145/964696.964697.

[8] Yu, Y. T. and Lau, M. F. 2006. A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions. *J. Syst. Softw.* 79, 5 (May. 2006), 577-590. DOI= http://dx.doi.org/10.1016/j.jss.2005.05.030.

[9] Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738.

Columns on Last Page Should Be Made As Close As Possible to Equal Length