Understanding Defect Prediction Models

Shrenuj Gandhi
Computer Science
North Carolina State University
sgandhi4@ncsu.edu

Neng Jiang
Computer Science
North Carolina State University
njiang@ncsu.edu

Abstract— **This electronic document is a live template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document.**

Index Terms— **This electronic document is a live template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document.**

I. INTRODUCTION

This template, modified in MS Word 2003 and saved as Word 97-2003 & 6.0/95 RTF for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

II. BASELINE STUDY

In the paper "Data Mining Static Code Attributes to Learn Defect Predictors", authors Tim Menzies, Jeremy Greenwald and Art Frank demonstrate that *naive Bayes* data miner with a *log filtering* preprocessor on the numeric data outperforms rule-based and decision-based learning methods.

A. Method

The authors have used the following algorithm

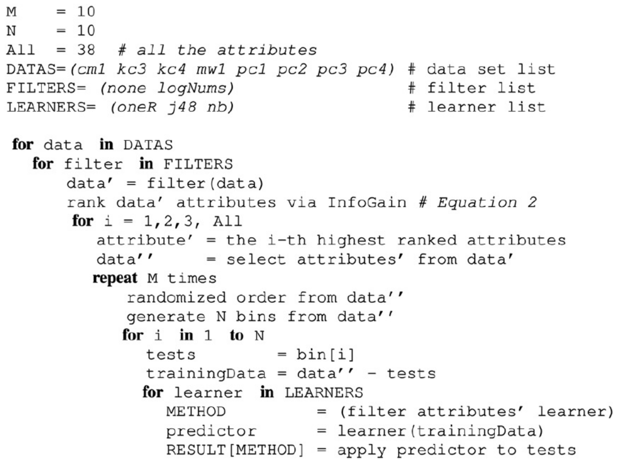
B. Results

III. METHODOLOGY

As part of this project, we perform 4 experiments in order to understand the behavior of defect prediction models.

A. Experiment 1 - Reproducing Baseline Results

The main reason that we see a different in performance is that the . Moreover the data was cleaned after the paper was published



```
M     = 10
N     = 10
All   = 38   # all the attributes
DATAS=(cm1 kc3 kc4 mw1 pc1 pc2 pc3 pc4) # data set list
FILTERS= (none logNums)                 # filter list
LEARNERS= (oneR j48 nb)                 # learner list

 for data  in DATAS
   for filter  in FILTERS
      data' = filter(data)
      rank data' attributes via InfoGain # Equation 2
       for i = 1,2,3, All
         attribute' = the i-th highest ranked attributes
         data''     = select attributes' from data'
         repeat M times
            randomized order from data''
            generate N bins from data''
           for i  in 1  to N
             tests        = bin[i]
             trainingData = data'' - tests
             for learner  in LEARNERS
               METHOD         = (filter attributes' learner)
               predictor      = learner(trainingData)
               RESULT[METHOD] = apply predictor to tests
```

Fig. 1. Data is filtered and the attributes are ranked using InfoGain. The data is then shuffled into a random order and divided into 10 bins. A learner is then applied to a training set built from nine of the bins. The learned predictor is tested on the remaining bin.

B. Experiment 2 - Performing SMOTE

C. Experiment 3 - Tuning Feature Selection

D. Experiment 4 - Comparing Learners

Each dataset goes through the following procedure

E. Data Cleaning

Any string column is ignored. Only numeric columns are processed. The idea is to keep it simple. String columns will

F. Pre-processing

The numeric data varies a lot. The following figure shows the min and max values of all the columns in the dataset -. Due to high variance, we are applying a filtering layer. We applying log filtering and then normalize the data.

G. Processing Method

The dataset is broken into testing and training sets. We are using 10×10 cross evaluation i.e. 10 times we train on 90% of data and test on the remaining 10%. Then SMOTE is applied to training set (keeping the size constant). Figure x shows the defect percentage in the 7 datasets. Clearly there is a minority of the defective class. Due to this we apply the Synthetic Minority Over-Sampling Technique (SMOTE). (source

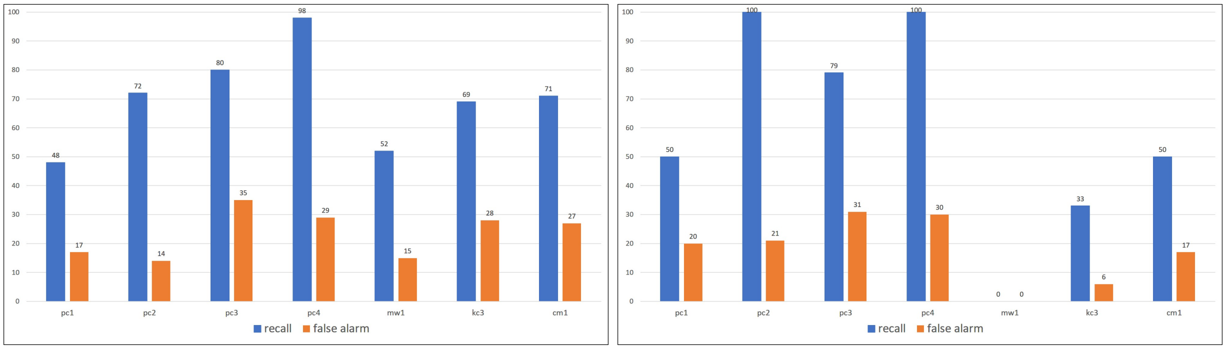


Fig. 2. The left hand side graph are the baseline results. It shows the average recall and average false alarm rate of Naive Bayes classifier with log filtering when top 3 features (based on InfoGain) are selected. The right hand side graph depicts the results we produced by following the steps in stated in Figure 1.

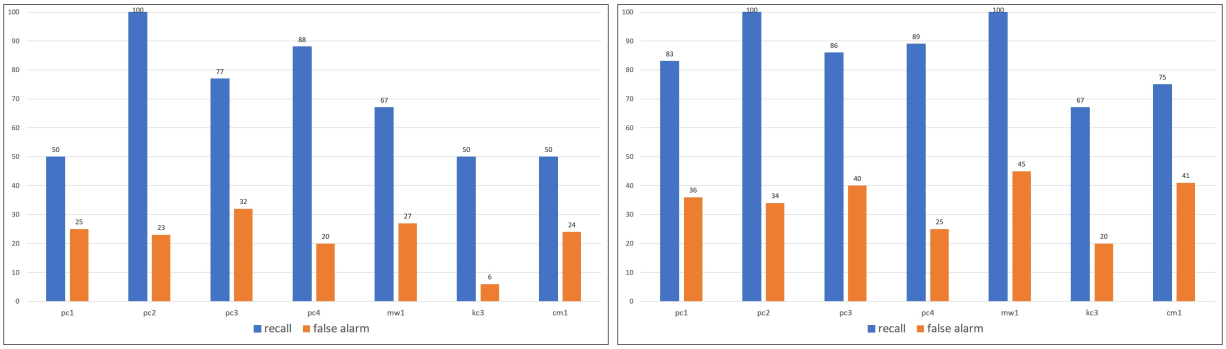


Fig. 3. The left hand side graph shows average recall and average precision values calculated by Naive Bayes after log filtering when top 5 features are selected based on InfoGain. The right hand side graph shows the average recall and average false alarm values when SMOTE is applied to the training set, while keeping the other procedure same.

- http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/) This is done by replacement (resulting in inflated training set), or by SMOTE and then reconstructing original bin size by random selection// How SMOTING works - It works by creating synthetic samples from the minor class instead of creating copies. The algorithm selects two or more similar instances (using a distance measure) and perturbing an instance one attribute at a time by a random amount within the difference to the neighboring instances.(source -http://www.jair.org/papers/paper953.html)

OR examine results of SMOTE and try re-sampling i.e. duplicating data Side-effect of SMOTE-ing You will have more and different data, but the non-linear relationships between the attributes may not be preserved.

The training and testing datasets are fed to the learners after SMOTE. For evaluating the effect of SMOTE with the baseline result, we have used Naive Bayes classifier. Naive Bayes uses Bayes Theorem to model the conditional relationship of each attribute to the class variable.

Prior works(mention references) have also used different learning methods. To get a good grasp of the performance of different learners, we have also used, in this experiment, Support Vector Machine (SVM), Random Forest, Classification and Regression Trees(CART), and Logistic Regression(LR).

SVM method uses points in a transformed problem space that best separate classes into two groups. Classification for multiple classes is supported by a one-vs-all method. SVM also supports regression by modeling the function with a minimum amount of allowable error[MLM]. Random Forest is an extension of Bootstrap Aggregation or bagging that in addition to building trees based on multiple samples of your training data, it also constrains the features that can be used to build the trees, forcing trees to be different. CART are constructed from a dataset by making splits that best separate the data for the classes or predictions being made. Logistic regression fits a logistic model to data and makes predictions about the probability of an event (between 0 and 1). Different learners (source-http://machinelearningmastery.com/get-your-hands-dirty-with-scikit-learn-now/)

Decision Trees perform well on imbalanced datasets.The splitting rules that look at the class variable used in the creation of the trees, can force both classes to be addressed.try a few popular decision tree algorithms like C4.5, C5.0, CART, and Random Forest.

The above logic is wrapped around a feature selection process. This spits out the top K attributes (based on info gain). This process is tuned to select top k attributes that provide the most recall value. (provide reason do doing this)

H. Reporting Results

sdfsdf

IV. Results

V. Procedure for Paper Submission

A. Selecting a Template (Heading 2)

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the US-letter paper size. Please do not use it for A4 paper since the margin requirements for A4 papers may be different from Letter paper size.

B. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations

VI. Math

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as 3.5-inch disk drive.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: Wb/m2 or webers per square meter, not webers/m2. Spell out units when they appear in text: . . . a few henries, not . . . a few H.
- Use a zero before decimal points: 0.25, not .25. Use cm3, not cc. (bullet list)

C. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

$$\alpha + \beta = \chi \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use (1), not Eq. (1) or equation (1), except at the beginning of a sentence: Equation (1) is . . .

D. Some Common Mistakes

- The word data is plural, not singular.
- The subscript for the permeability of vacuum ?0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter o.
- In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an inset, not an insert. The word alternatively is preferred to the word alternately (unless you really mean something that alternates).
- Do not use the word essentially to mean approximately or effectively.
- In your paper title, if the words that uses can accurately replace the word using, capitalize the u; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones affect and effect, complement and compliment, discreet and discrete, principal and principle.

- Do not confuse imply and infer.
- The prefix non is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the et in the Latin abbreviation et al..
- The abbreviation i.e. means that is, and the abbreviation e.g. means for example.

VII. USING THE TEMPLATE

Use this sample document as your LaTeX source file to create your document. Save this file as **root.tex**. You have to make sure to use the cls file that came with this distribution. If you use a different style file, you cannot expect to get required margins. Note also that when you are creating your out PDF file, the source file is only part of the equation. *Your TEX → PDF filter determines the output file size. Even if you make all the specifications to output a letter file in the source - if you filter is set to produce A4, you will only get A4 output.*

It is impossible to account for all possible situation, one would encounter using TEX. If you are using multiple TEX files you must make sure that the “MAIN“ source file is called root.tex - this is particularly important if your conference is using PaperPlaza’s built in TEX to PDF conversion tool.

A. Headings, etc

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named Heading 1, Heading 2, Heading 3, and Heading 4 are prescribed.

B. Figures and Tables

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation Fig. 1, even at the beginning of a sentence.

TABLE I

AN EXAMPLE OF A TABLE

One	Two
Three	Four

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity Magnetization, or Magnetization, M, not just M. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write Magnetization (A/m) or Magnetization A[m(1)], not just A/m. Do not label axes with a ratio of quantities and units. For example, write Temperature (K), not Temperature/K.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an document, this method is somewhat more stable than directly inserting a picture.

Fig. 4. Inductance of oscillation winding on amorphous magnetic core versus DC bias magnetic field

VIII. CONCLUSIONS

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

APPENDIX

Appendixes should appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word acknowledgment in America is without an e after the g. Avoid the stilted expression, One of us (R. B. G.) thanks . . . Instead, try R. B. G. thanks. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References are important to the reader; therefore, each citation must be complete and correct. If at all possible, references should be commonly available publications.

REFERENCES

[1] G. O. Young, Synthetic structure of industrial plastics (Book style with paper title and editor), in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 1564.

[2] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123135.

[3] H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.

[4] B. Smith, An approach to graphs of linear forms (Unpublished work style), unpublished.

[5] E. H. Miller, A note on reflector arrays (Periodical styleAccepted for publication), IEEE Trans. Antennas Propagat., to be publised.

[6] J. Wang, Fundamentals of erbium-doped fiber amplifiers arrays (Periodical styleSubmitted for publication), IEEE J. Quantum Electron., submitted for publication.

[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style), IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[9] M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[10] J. U. Duncombe, Infrared navigationPart I: An assessment of feasibility (Periodical style), IEEE Trans. Electron Devices, vol. ED-11, pp. 3439, Jan. 1959.

[11] S. Chen, B. Mulgrew, and P. M. Grant, A clustering technique for digital communications channel equalization using radial basis function networks, IEEE Trans. Neural Networks, vol. 4, pp. 570578, July 1993.

[12] R. W. Lucky, Automatic equalization for digital communication, Bell Syst. Tech. J., vol. 44, no. 4, pp. 547588, Apr. 1965.

[13] S. P. Bingulac, On the compatibility of adaptive controllers (Published Conference Proceedings style), in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 816.

[14] G. R. Faulhaber, Design of service systems with priority reservation, in Conf. Rec. 1995 IEEE Int. Conf. Communications, pp. 38.

[15] W. D. Doyle, Magnetization reversal in films with biaxial anisotropy, in 1987 Proc. INTERMAG Conf., pp. 2.2-12.2-6.

[16] G. W. Juette and L. E. Zeffanella, Radio noise currents n short sections on bundle conductors (Presented Conference Paper style), presented at the IEEE Summer power Meeting, Dallas, TX, June 2227, 1990, Paper 90 SM 690-0 PWRS.

[17] J. G. Kreifeldt, An analysis of surface-detected EMG as an amplitude-modulated noise, presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.

[18] J. Williams, Narrow-band analyzer (Thesis or Dissertation style), Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

[19] N. Kawasaki, Parametric study of thermal and chemical nonequilibrium nozzle flow, M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

[20] J. P. Wilkinson, Nonlinear resonant circuit devices (Patent style), U.S. Patent 3 624 12, July 16, 1990.