UNSTRUCTURED CANCER DOCUMENT CLASSIFICATION USING ASSOCIATION RULE BASED BINOMIAL NAÏVE BAYES

K. Aadinath

Department of Computer Science and Engineering, Mar Baselios College of Engineering and Technology, India

Abstract

The readability and accuracy of any excellent classifier are two essential characteristics. Associative classifiers have lately been utilised for many classification problems, for reasons such as acceptable accuracy, fast training, and good interpretability. While features may be extremely helpful for categorization of texts, owing to the great dimensionality of text documents, both training time and the number of rules generated will substantially rise. In this article we present an algorithm for classifying texts, which comprises a selection phase for features to pick essential characteristics and a classifying phase to address this shortcoming. The experimental findings from the application of the suggested algorithm show that our method surpasses others in efficiency and performance compared with the results of a chosen well-known classification algorithm.

Keywords:

Classification, Text dataset, Naïve Bayes, Feature Selection

1. INTRODUCTION

The need to acquire and handle information has grown quickly with the exponential development of information. In organised as well as unstructured formats, information may be shown. This article is named class labels to automatically categorise a collection of papers into specified categories. The categorization of texts provides a better understanding of information as a text in terms of the language most often employed. Several kinds of applications, including genre classification [1], opinion mining [2], text-based webpage classification [3], and spam filtering [4], were reported in the domain.

They may be classified into two major categories: traditional such as K Neighbors [6]-[12], Decision Trees [5] and C4.5 [5], and Association Classification Algorithms. Classification algorithms are available in two primary groups. Experimental findings in [6] demonstrate the greater precision of association classification algorithms than conventional methods.

The whole process of association classification starts with one of the association rules mining algorithms such as Apriori [8] and Fp-Growth [9]. A series of association rules are generated and a limited number of high-quality rules are then chosen, which are ultimately used to forecast. As a classifier based on the rules, man may comprehend the functioning of the Associative Classifier, and the results of the prediction give a clear, straightforward interpretation [10]. The fact that Associative Classification frequently leads to a large number of rules in association rule mining, and also strives to choose high-quality rules [11], causes difficulties in efficiency.

In [16], Agrawal presented the issue of mining association regulations with basket data. An illustration of this rule might be that 98 percent of consumers who buy tyres and car accessories also need car service. For cross-marketing and connected mailing applications, it is important to find all these rules. Other uses include catalogue design, sales add-ons, shop layout and purchasing pattern-based consumer segmentation. Due to their extremely huge databases, it is thus essential for quick algorithms to be executed throughout the work of these applications [21].

Most text data has a large space in which many association rules are generated. Two popular methods for high-quality regulations exist. First, it produces all the regulations of the organisation. Therefore, in the induction process for the associative classifier, many high-order association rules are produced. High-order rules are often more informative; thus, the categorization using these rules is superior [22]. The complexity of these techniques is very high [10]. Second, a few high-quality regulations were applied irrespective of the ruling decision (or preferring high-order rules). These techniques eliminate superfluous words and choose high-quality characteristics alone to decrease calculation time.

High frequency words are initially detected and then grouped into class labels in our suggested approach. High-quality characteristics are therefore identified for each class label. The association rules for every grade label are created using these characteristics, and thus, all words for each grade label must not be taken into account.

2. LITERATURE SURVEY

Several text classifiers, such as machine learning methods and probabilistic models, have been suggested in the literature, e.g. decision trees, nave Bayes [5], closest neighbours. A set of classifiers known as association-regulatory classifiers, which are successful and similar to most well-known text classifiers [13], has just been suggested. The CBA [7], CMAR [6], CPAR [11], Harmony [14], MCAR [15] and CACA [12] are well-known in association regulatory base classifiers, and use diverse techniques in rules discovery, rule ranking, rules trimming, rules prevision, and rules assessment methods. The following is a short description of each of them:

One of the earliest algorithms for classification using the association rule mining method was suggested in [7] and is called CBA. It uses the apriori algorithm [16] to find common objects in transaction data. Once the frequent things are discovered, the CBA generates rules for any frequent items with minimal certainty. These criteria are assessed and a subset of these rules is chosen and utilised for the final classification. In two distinct stages, CBA is applied for the identification of frequent itemsets and rules.

The frequent pattern mining technique has been enhanced by CMAR; FP-growth; and the distribution-related class FP-tree is constructed and huge datasets are quickly broken down. The CR-tree structure utilises association rules for the storage and retrieval of and prints them for trust, correlation and database coverage. Harmony directly mines one of the top classification rules in each

training instance, which supports and builds a classification model from the union of these rules across the whole collection of examples.

Harmony CPAR aims to integrate both associative and conventional rule-based categorization benefits. It inherits the fundamental concept [17] is a rule generator and incorporates the associative classification characteristics into the predictive rule analysis using a 'great algorithm.' Harmony uses a framework for the creation of instance-centric rules and ensures that the optimal rule for each training instance will be found and included [14]. This approach, however, may function negatively when a database with unequal class distribution labels is manufactured, because a class with numerous training instances may be highly predictive of the label, which leads to an incorrect prediction [18]. A method for the discovery of frequent item sets is used in the MCAR algorithm. Two major stages comprise MCAR: rule generation and the function Object () [native code] of a classification. The first phase scans the training data set to find the common '1 item' in a group of items, then creates combined criteria for producing candidate items with more. The training data set will then be scanned. A "candidate rule" is produced as any rule supporting more than minimal support. In phase two, rules are established to construct a classifier in view of its efficiency in the training data set. In the classifier [15], only rules are maintained that cover a specific number of training instances.

A method named CACA was also proposed [12] for the classification of the association. First, CACA searches the data set before vertically storing it like the MCAR method. The next frequency counts the frequency of each value and is ordered according to the frequencies in decreasing order. TIDs are intersected to limit the search range of common patterns with all frequent 'disjoint attribute value' values. A TID with a frequent attribute value contains the row numbers where the items in the training data set appear. Finally, in each class group attribute passing the minimal level of confidence, the root node support, confidence and class of a class attribute are put into the Ordered Rule Tree (OR-Tree). CACA categorises data that is invisible, such as the CBA. Experimental findings indicate that CACA is superior to other associative algorithms in UCI data sets with regards to accuracy and calculation times [19] [20].

The filtering of the frequent search space, while obtaining the high quality features differently, is inspired by CACA.

The aforementioned classifiers are not used for text categorization and have broad applicability. Certain algorithms, such as [9] [13] [18] were specifically designed for texts and are classified in the Association Family.

3. PROPOSED METHODOLOGY AND RESULTS

The whole text classification method association structure is shown in Figure 1. As shown in Figure 1, these processes have two stages: a training phase and a test phase. Raw data on training papers is created throughout the training phase. In this section, words and unnecessary characters are deleted and words are stemmed and indexed according to necessity (step 1). We get the categorization rules by extracting data from this pre-processed database (step 2). The extremely high number of categorization rules filters out some superfluous rules and selects a limited number of key rules. This is called tailing (step 3). We now have a classification rules database. In the test phase, unknown materials should be preprocessed before any activity and transformed into a word pattern and matched to the categorization criteria (step 4). In step 5, the class labels of unknown documents are predicted in the final phase according to the matching criteria.

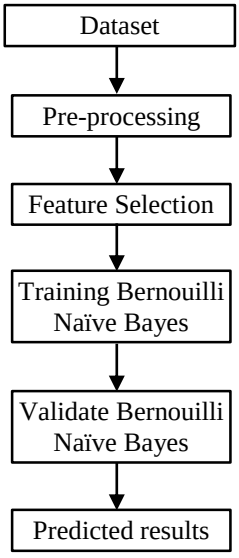

Fig.1. Text classification structure

The Bernouilli Naïve Bayes procedure looks at word occurrences; all that matters is whether a word is absent or present in a document. The Binomial Naïve Bayes procedure discussed here uses word frequencies. The prior probabilities are estimated the same way as explained previously, but the conditional joint probabilities $P(x_1, x_2, ..., x_m \mid y = 1)$ and $P(x_1, x_2, ..., x_m \mid y = 2)$ in Eq.(1) are estimated differently. For multinomial Naïve Bayes the marginal conditional probabilities of the occurrence of the i^{th} word,

$$P(Occurrence_i \mid y = 1) \text{ and } P(Occurrence_i \mid y = 2),$$

are obtained by dividing the number of times a given word occurs in documents of a given group (either group 1 or group 2) by the number of all word occurrences in that group; Laplace smoothing is implemented as well to avoid problems when a given word does not occur in one of the groups. Note this is different from the Bernouilli version which ignores how often a word is included in a document and works with occurrence indicators. Furthermore, the calculation of the posterior probability of the test case $P(y \mid x_1, x_2, ..., x_m)$, where n is the total number of words of the test case and $x_1, x_2, ..., x_m$ are its number of occurrences of the m words, incorporates the word frequencies x_i of the new case; the terms on the right-hand side of Eq.(1) are calculated from the multinomial distribution as

$$P(x_1, x_2, ..., x_m \mid y) = \frac{n!}{x_1! x_2! ... x_m!} \prod_{i=1}^{m} [P(Occurrence_i \mid y)]^{x_i} \quad (1)$$

Note that the Eq.(1) cancels out the factorial conditions, thus no calculation is necessary. Fig.2 shows the common kernel function using Bernouilli Nave Bayes.

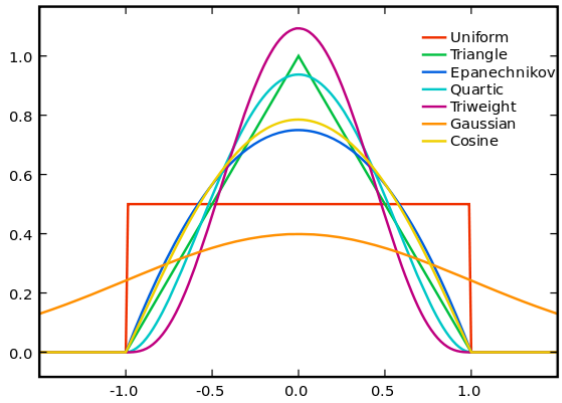

Fig.2. Performance of Bernouilli Naïve Bayes on Various Fitness Function

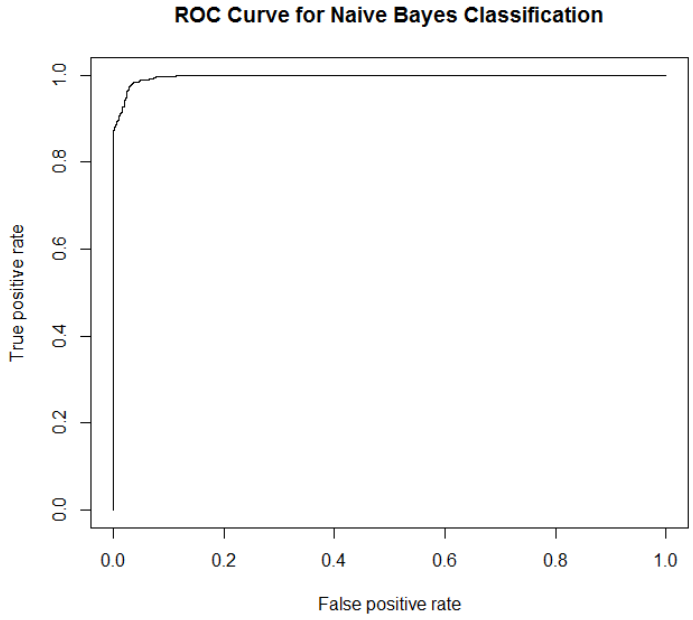

Fig.3. RoC Curve

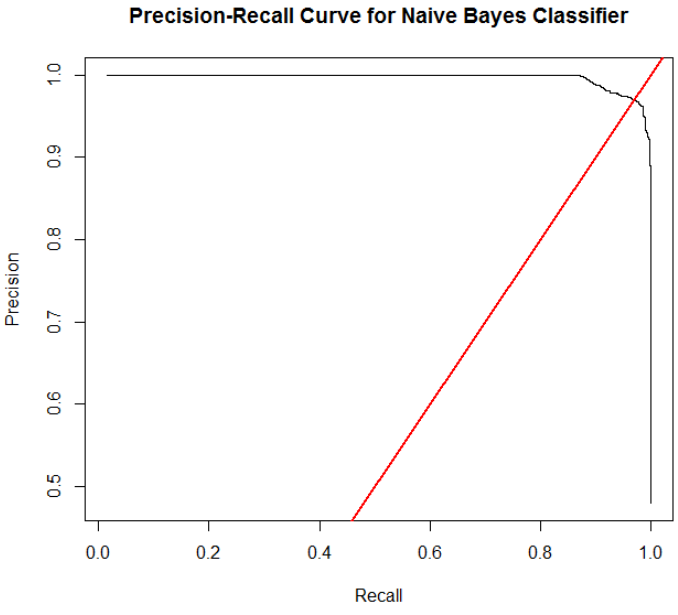

Fig.4. Precision-Recall Curve

The true positive rate and the false positive rate of the ROC curve in Fig.3 are verified using Bernouilli Nave Bayes. In addition, the application of the system to an unstructured data set is evaluated in Fig.4 in terms of the precision recall curve.

4. CONCLUSION

A novel method for the categorization of texts has been developed, called Bernouilli Nave Bayes. The technique presented has many characteristics enhanced compared to conventional and association classification methods used to classify documents. It produces understandable rules for man-made interpretations, contains a selection part to decrease text dimensions, clusters of important label-based features, needs a single data scan, saves features as a binary stream, and uses binary operations for all processes to reduce the necessary memory space.

REFERENCES

[1] Igor Mekterovic, Ljiljana Brkic and Mirta Baranovic, "A Systematic Review of Data Mining Approaches to Credit Card Fraud Detection", *WSEAS Transactions on Business and Economics*, Vol. 15, No. 2, pp. 437-444, 2018.

[2] S. Vimala and K.C. Sharmili, "Survey Paper for Credit Card Fraud Detection using Data Mining Techniques", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 6, No. 11, pp. 357-364, 2017.

[3] Ong Shu Yee, Saravanan Sagadevan and Nurul Hashimah Ahamed Hassain Malim, "Credit Card Fraud Detection using Machine Learning as Data Mining Techniques", *Journal of Telecommunication, Electronic and Computer Engineering*, Vol. 10, No. 1, pp. 23-27, 2018.

[4] X. Luo, J. Deng, J. Liu and W. Wang, "A Quantized Kernel Least Mean Square Scheme with Entropy-Guided Learning for Intelligent Data Analysis", *China Communications*, Vol. 14, No. 7, pp. 1-10, 2017.

[5] T.N. Lal, O. Chapelle and J. Weston, "*Embedded Methods*", Springer, 2006.

[6] S. Zhou, Q. Chen and X. Wang, "Active Semi-Supervised Learning Method with Hybrid Deep Belief Networks", *PloS One*, Vol. 9, No. 9, pp. 1-9, 2014.

[7] C. Szegedy, W. Liu, Y. Jia and P. Sermanet, "Going Deeper with Convolutions", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[8] C.C. Aggarwal and C. Zhai, "*A Survey of Text Classification Algorithms*", Springer, 2012.

[9] W. Aziguli, Y. Zhang, Y. Xie and D. Zhang, "A Robust Text Classifier based on Denoising Deep Neural Network in the Analysis of Big Data", *Scientific Programming*, Vol. 2017, pp. 1-20, 2017.

[10] Vinay Kumar and Shishir Kumar. "*Predictive Analysis of Emotions for Improving Customer Services*", IGI Global Publisher, 2017.

[11] Shadi Shaheen, Wassim El-Hajj, Hazem Hajj and Shady Elbassuoni, "Emotion Recognition from Text based on Automatically Generated Rules", *Proceedings of IEEE International Conference on Data Mining Workshop*, pp. 383-392, 2014.

[12] K. Jain and B. Yu, "Automatic Text Location in Images and Video Frames", *Pattern Recognition*, Vol. 31, No. 12, pp. 2055-2076, 1998.

[13] O.D. Trier and A.K. Jain, "Goal Directed Evaluation of Binarization Methods", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17, No. 12, pp. 1191-1202, 1995.

[14] V. Wu, R. Manmatha and E.M. Riseman, "Text finder an Automatic system to Detect and Recognize Text in Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 11, pp. 1224-1229, 1999.

[15] Sunil Kumar, Rajat Gupta, Nitin Khanna, Santanu Chaudhury and Shiv Dutt Joshi, "Text Extraction and Document Image Segmentation using Matched Wavelets and MRF Model", *IEEE Transactions on Image Processing*, Vol. 16, No. 8, pp. 2117-2128, 2007.

[16] Y.I. Chucai and Yingli Tian, "Localizing Text in Scene Images by Boundary Clustering Stroke Segmentation and String Fragment Classification", *IEEE Transactions on Image Processing*, Vol. 21, No. 9, pp. 4256-4268, 2012.

[17] Boris Epshtein, Eyal Ofek and Yonatan Wexler, "Detecting Text in Nature Scenes with Stroke Width Transform", *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2963-2970, 2010.

[18] M. Liu, G. Haffari, W. Buntine and M. Ananda-Rajah, "Leveraging Linguistic Resources for Improving Neural Text Classification", *Proceedings of the Australasian Language Technology Association Workshop*, pp. 34-42, 2017.

[19] A. Dasgupta, P. Drineas, B. Harb and V. Josifovski, "Feature Selection Methods for Text Classification", *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 230-239, 2007.

[20] X. Chen, J. Yang, J. Zhang and A. Waibel, "Automatic Detection and Recognition of Signs from Natural Scenes", *IEEE Transactions on Image Processing*, Vol. 13, No. 1, pp. 87-99, 2004.

[21] D.T. Chen, J.M. Odobez and H. Bourlard, "Text Detection and Recognition in Images and Videos Frames", *Pattern Recognition*, Vol. 37, No. 3, pp. 595-608, 2004.

[22] C. Mancas-Thillou and B. Gosselin, "Spatial and Color Spaces Combination for Natural Scene Text Extraction", *Proceedings of IEEE Conference on Image Processing*, pp. 985-988, 2006.