IRAQI Academic Scientific Journals

العـــراقيـة المـجـلات الاكـاديـمـيـة الـعـلـمـيـة

Al-Kut University College Journal
K. U. C. J.

ISSN (E) : 2616 - 7808 I ISSN (P) : 2414 - 7419

Vol. 6 Issue 1 2021

Performance Analysis of LogitBoost and Naïve bayes Classification Algorithm for Data Classification

Rasha Hani salman[1] **Nadia Adnan Shiltagh**[2] **Mahmood Zaki Abdullah**[3]

[1]Informatics Institute for Postgraduate Studies, Iraqi Commission for Computer & informatics, Baghdad - Iraq
[2]University of Baghdad, College of Engineering, Baghdad - Iraq
[3]Mustansiriyah University, College of Engineering, Baghdad - Iraq

Rashahany609@gmail.com nadia.alijamali@coeng.uobaghdad.edu.iq drmzaali@uomustansiriyah.edu.iq

Abstract

Classification is a significant technique for data mining with wide applications to identify the different categories of data used in virtually every area in our lives. Classification is used to label the individual in relation to the predefined groups according to the characteristics of the individual. This paper sheds light on performance appraisal based on (precision, Recall, F-measure) data classification analysis using a classification algorithm (LogitBoost and Naïve Bayes). The Naïve Bayes algorithm, is based on probability and the LogitBoost algorithm is based on the finding that Adaboost basically matches the training data using an additive logistic regression model. The paper sets out to render comparative analyses of Naive Bayes and LogitBoost classifiers in the context of job classification dataset Experimental results revealed that LogitBoost has highest result in (precision = 82.73 percent, recall = 83.33 percent, F-measure = 82.31 percent) compared to the Naive bayes algorithm for the data set mentioned above.

Key word: classification algorithm, Naïve Bayes, LogitBoost, precision, Recall

تحليل اداء خوارزميات LogitBoost وNaïve bayes لتصنيف البيانات

رشا هاني سلمان[1] **نادية عدنان شلتاغ**[2] **محمود زكي عبد الله**[3]

[1]الهيئة العراقية للحاسبات والمعلوماتية، معهد المعلوماتية للدراسات العليا، العراق – بغداد
[2]جامعة بغداد، كلية الهندسة، العراق – بغداد
[3]الجامعة المستنصرية، كلية الهندسة، العراق – بغداد

drmzaali@uomustansiriyah.edu.iq nadia.alijamali@coeng.uobaghdad.edu.iq Rashahany609@gmail.com

الخلاصة

التصنيف هو تقنية مهمة لاستخراج البيانات مع تطبيقات واسعة لتحديد الفئات المختلفة من البيانات المستخدمة في كل مجال تقريبًا في حياتنا. يستخدم التصنيف لتسمية الفرد فيما يتعلق بالمجموعات المحددة مسبقًا وفقًا لخصائص الفرد. يلقي هذا البحث الضوء على تقييم الأداء القائم على تحليل تصنيف البيانات (الدقة، الاسترجاع، تعتمد خوارزمية تعتمد القياس F) باستخدام خوارزميات التصنيف (Naïve Bayes و

LogitBoost) تعتمد خوارزمية (Naïve Bayes) على الاحتمالات وتستند خوارزمية (LogitBoost) الى اكتشاف ان Adaboost يطابق بيانات التدريب بشكل أساسي باستخدام نموذج الانحدار اللوجستي الإضافي. تحدد الورقة، لتقديم تحليلات في مقارنة لمصنفات (Naïve Bayes و LogitBoost) سياق مجموعة بيانات تصنيف الوظائف. كشفت النتائج التجريبية أن (LogitBoost) لديه أعلى نتيجة في (الدقة = 82.73 بالمائة، الاسترجاع = 83.33 بالمائة ، قياسF = 82.31 بالمائة) مقارنة بخوارزمية (Naive bayes) لمجموعة البيانات المذكورة أعلاه.

الكلمات المفتاحية : خوارزميات التصنيف، LogitBoost ، Naïve Baye، الدقة، الاسترجاع

1. Introduction

Data mining is a tool used to characterize an exploration of information and to look for essential relationships such as patterns, correlations and variables within databases [1,2]. It is growing in numerous applications, such as organic compound analysis, diagnostic of drugs, targeted marketing, product design identification of credit card fraud, financial forecasting, automated classification, estimation of television audience shares, etc. Data mining refers to the review of large quantities of knowledge stored on servers [3]. It is not unique to one sort of media, or any kind of knowledge archive should be subject to data mining. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object oriented databases [4], data warehouses, transactional databases, unstructured and semi-structured repositories such as the World Wide Web, Advanced databases including relational databases, databases of multimedia, databases of time series and textual databases, and also flat files [5,6].There are many methods of data mining that can be used to derive important information from big data. Data mining has several roles including classification and predication, clustering and association rule mining [7]. In addition, classification is one of the most useful data mining strategies for constructing the classification models from an input data set [8]. The classification techniques used typically construct models which are used to forecast future patterns in results. There are several data classification algorithms which include the decision tree, Classifiers for Naïve Bayes and classifiers for LogitBoost [9]. The objective of this study therefore focuses on the data classification and performance measurement of the classifier algorithms based on precision, Recall, F-measure produced by the algorithms when applied to the data set. This paper is regular as follows: Section 2 presents the classification algorithm (LogitBoost, Naïve bayes). Data set Collection and Description is presented in section3. Measuring performance is presented in section4. The Experimental Work and Result are presented in section5. Section6 presents the conclusion.

2. Classification Algorithm

A-LogitBoost

This is a boosting algorithm (Jerome Friedman, Trevor Hastie, and Robert Tibshirani) which implements it. Newton 's adaptive measures were used by the Logit Boost algorithm to match the most identical logistic model with greater probability. [10]. The Logit Boost is intended to achieve a better version of Adaboost on the grounds that the binomial variance is constrained by this algorithm and that the binomial variance gives less weight to wrongly classified objects [11]. A Logit Boost thus defines higher weights for the inaccurate grouping of objects that are normally made up of more relevant details. A significant benefit in collecting data for multilevel categorical forecasts is the regression model for each of the fusion measures [12]. The algorithm begins with an introduction to the following parameters:

$wn_1 = 1/\, N_1$ (w: Each class's weight)

$F(x_1) = 0$ (The classifier of output)

$P(xn_1) = ½$ (Weighted probabilities in each class)

Then for $n_1 = 1, 2, \ldots . N_1$. The working, weighted, and weighted probability response is determined in the equations below [13]:

$$Y_1{}' = \frac{yn1 - p(xn1)}{p(xn1)(1 - p(xn1))}$$

$$wn_1 = p(xn_1)(1 - p(xn_1\ ))$$

$$p(x_1) = \frac{e\ F(X1)}{e\ F(X1) + e - F(X1)}$$

where $yn_1 * \in \{0,+1\ \}$(label for class)

$p(xn_1)$: The chance that the object belongs to a class. The weights and performance classifier are modified for each reiteration after calculating the regression model fm(x), where m is the repetition index, Updating the weights and output classifier:

$$F(X_1) = F(X_1) + \frac{1}{2}\, fm_1(x_1)$$

Lastly, normalized weights and the output classifier are calculated in equation:

$$\text{Sign } [F\ (X_1 i)] = \text{sign } (\ \textstyle\sum fm(x_1\ \ )\ M\ m_1=1\ )$$

B- Naive Bayes

This Bayesian Classification is used as a probabilistic learning strategy and each characteristic of the Bayesian algorithm is autonomous in estimating any other characteristics [14]. This classification process is a supervised classification of probabilistic learning and statistics. The clear probabilities of the hypothesis are determined and the noise in the input data is high. The probability of event A can well depend on events B and A, which are said to precede or coincide with case B. Based on any proof (x) that could be found, the basic concept of Pace 's base is that result an incident should be predicted. The Naïve Bayes is a basic classifier of probability statistics focused on the interpretation of the Bayesian theory of probability [15]. The principle of Bayes will measure the posterior likelihood of $P(c_1 / x_1)$ from $P(c_1\ )$, $P(x_1)$, and $P(x_1 / c_1)$. As seen below in the equation [16]:

$$p(c_1/x_1) = \frac{p(x1/c1) * p\ (c1)}{p(x1)}$$

where: $P(c_1/x_1)$: Posterior's probability of class (predictor (x_1, attributes), (c1, target)

$P(c_1)$: class's prior probability.

$P(x_1/c_1)$: It is the likelihood that is the probability of the index of a specific

$P(x_1)$: Is the previously probability of predictor of class.

The main benefit of Naive bayes, it takes a short computational time to plan training for, enhances the order classification by distinguishing the irrelevant properties and has great efficiency [17,18].

3. Data set Collection and Description

This study is using data source from the website (www. kaggle.com). It is a data collection obtained on January 7, 2017, providing some information on the work category Creation of a data classification model. For experimental purposes, this study explores 66 instances with 8 task attributes, the attributes are grouped into eight groups after deletion of inconsistencies in the data set (Education Level, Pay Grade, Org Impact, Experience, Supervision, Problem Solving, Contact Level, Financial).

4. Measuring performance

The most popular execution efficiency assessment approach is implemented to test the appearance of different entity collection methods and classifiers [19], Recall and precision are used to assess two types of classification accuracy. Precision indicates the percentage of right (true positive) classifications equivalent to the overall number of (false positive and true positive) classifications. While Recall means the number of right cases (true positive), which corresponds to the utter number of right cases (false negative and true positive) [20]. Precision is a measure of accuracy that defines the portion of relevant elements recovered from all recovered elements. The Recall is the measure of the completeness and specifies the portion of related elements recovered from all related element [21,22]. The Precision and Recall can be calculated in equations below [23]:

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FP)}}$$

Where: -

TP: True positive rate.

FP: False positive rate.

One of the composite tests of precision and Recall is the F-measure. In General, F1 is a common measure of F-measure. The ranking for F1 holds the equilibrium between memory and accuracy. It is possible to determine [24], F1 can be calculation in equation below [25].

$$F1 = 2 * \frac{precision*recall}{precision+recal}$$

5. Experimental Work and Result

This research performed the classification using LogitBoost and Naïve Bayes algorithm in a job classification dataset. Below tables show the performance of the two algorithms based on different parameters.

Table 1: Comparative study of Naïve Byes and LogitBoost algorithm based on precision measure

Sequence	Classifier	Precision%
1	Naive Bayes	68.50
2	LogitBoost	82.73

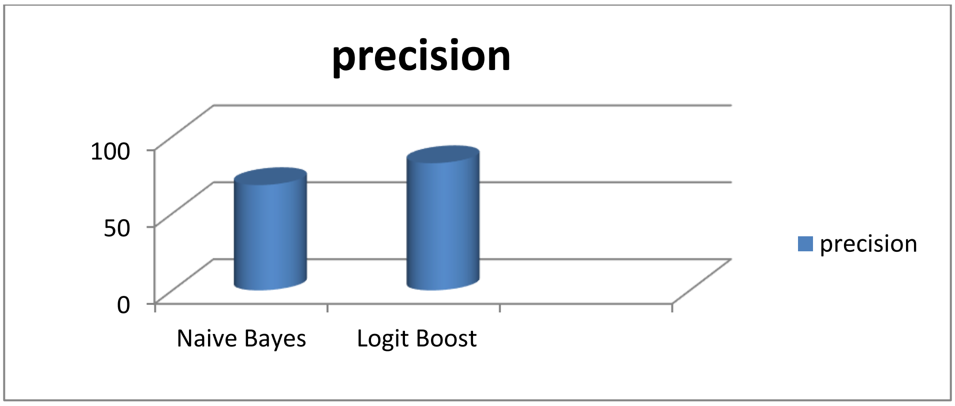

Figure 1: Algorithms vs. precision Graph

Observation from Table1 and Figure1: From the comparative study of the two algorithms, the LogitBoost algorithm shows the maximum rate of precision measure (82.73%).

Table 2: Comparative study of Naïve Bayes and LogitBoost algorithm based on Recall measure

Sequence	Classifier	Recall %
1	Naive Bayes	68.18
2	Logit Boost	83.33

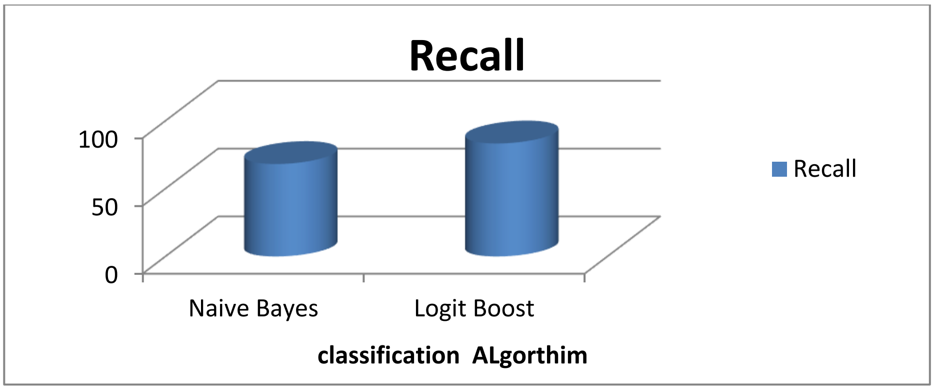

Figure 2: Algorithms vs. Recall Graph

Observation from Table 2 and Figure 2: In this comparison also LogitBoost algorithm is showing the recall value (83.33 %), which is greater than Naïve Bye algorithms. This means LogitBoost algorithm is highly significant based on recall measure.

Table 3: Comparative study of Naïve Bye and LogitBoost algorithm based on F- measure

Sequence	Classifier	F-Measure %
1	Naive Bayes	65.17
2	Logit Boost	82.31

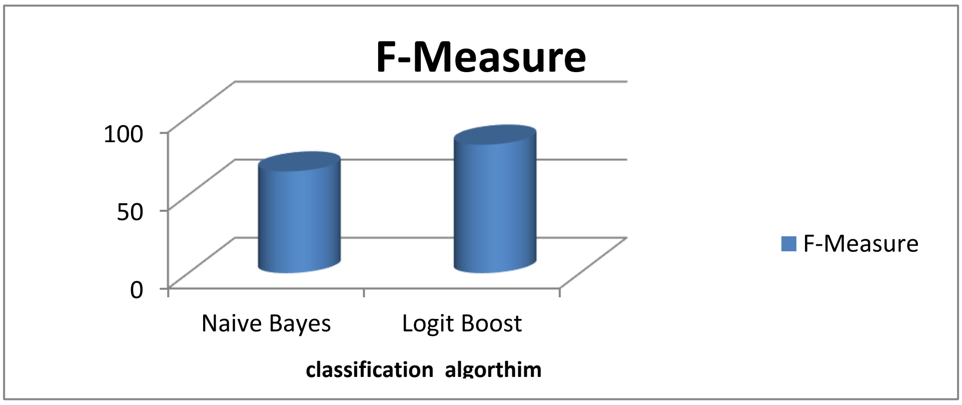

Figure 3: Algorithms vs. F-measure Graph

Observation from Table 3 and Figure 3: Here LogitBoost algorithm is showing high rate of m-measure. this algorithm is efficient than the other algorithm in terms of F-measure.

6 - conclusion

This study aims to explore and evaluate the performance of LogitBoost and Naïve Bayes algorithm. Effective result Taken from a job classification dataset. The results of the experiments are shown in study on accuracy measure (precision, recall, F-measure) Naive Bayes as well good results but the LogitBoost gives more rating in (precision, Recall and F-measure). It can be concluded that LogitBoost algorithm is highly efficient than algorithm mention above in the purpose of classification and analysis.

References

[1] Yang, Y., & Chen, W. (2016). Taiga: performance optimization of the C4.5 decision tree construction algorithm. Tsinghua Science and Technology, 21(4), 415-425.

[2] C.Rygielski , J.C. Wang, D.C. Yen, "Data mining techniques for customer relationship management," Technology in Society, vol. 24, pp. 483–502, 2002.

[3] Tesfaye, B., Atique, S., Elias, N., Dibaba, L., Shabbir, S. A., & Kebede, M. (2017). Determinants and development of a web-based child mortality prediction model in resource-limited settings: a data mining approach. Computer methods and programs in biomedicine, 140, 45-51.

[4] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[5] Margaret H. Danham, S. Sridhar, " Data mining, Introductory and Advanced Topics", Person education , 1st ed., 2006

[6] Wenke Lee, Salvatore J. Stolfo, Kui W. Mok, "A Data Mining Framework for Building Intrusion Detection Models".

[7] Ramani, R. G., & Sivagami, G. (2011). Parkinson disease classification using data mining algorithms. International journal of computer applications, 32(9), 17-22.

[8] Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36(4), 2431-2448.

[9] S.F. Shazmeen, M.M.A. Baig, and M.R. Pawar, "Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis," Journal of Computer Engineering, vol. 10(6), pp. 01-06, 2013.

[10] Kamarudin, M. H., Maple, C., Watson, T., & Safa, N. S. (2017). A logit boost-based algorithm for detecting known and unknown web attacks. IEEE Access, 5, 26190-26200.

[11] Song, J., Lu, X., Liu, M., & Wu, X. (2011, July). A new Logit Boost algorithm for multiclass unbalanced data classification. In 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) (Vol. 2, pp. 974-977). IEEE.

[12] WATSON, T., & SAFA, N. S. A LogitBoost-Based Algorithm for Detecting Known and Unknown Web Attacks.

[13] Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes

and decision tree classification techniques. International Journal of Science and Research (IJSR), 5(1), 1842-1845.

[14] Jantawan, B., & Tsai, C. F. (2013). The application of data mining to build classification model for predicting graduate employment. arXiv preprint arXiv:1312.7123.

[15] Gebrie, M. T., & Abie, H. (2017, September). Risk-based adaptive authentication for Internet of things in smart home eHealth. In Proceedings of the 11th European Conference on Software

[16] Bielza, C., Li, G., & Larranaga, P. (2011). Multi-dimensional classification with Bayesian networks. International Journal of Approximate Reasoning, 52(6), 705-727.Architecture: Companion Proceedings (pp. 102-108) .

[17] Qin, C., Zhu, H., Xu, T., Zhu, C., Jiang, L., Chen, E., & Xiong, H. (2018, June). Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 25-34).

[18] Rabcan, J., Vaclavkova, M., & Blasko, R. (2017, July). Selection of appropriate candidates for a type position using C4. 5 decision tree. In 2017 International Conference on Information and Digital Technologies (IDT) (pp. 332-338). IEEE.

[19] Zhu, Q., Wei, K., Ding, L., Lai, K. K., & Keung, K. (2017, May). Court Judgment Decision Support System Based on Medical Text Mining. In WHICEB (p. 2).

[20] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), e0118432.

[21] Gunawardana, A., & Shani, G. (2015). Evaluating recommender systems. In Recommender systems handbook (pp. 265-308). Springer, Boston, MA.

[22] Boutet, A., Frey, D., Guerraoui, R., Jégou, A., & Kermarrec, A. M. (2013, May). Whatsup: A decentralized instant news recommender. In 2013 IEEE 27th International Symposium on Parallel and Distributed Processing (pp. 741-752). IEEE.

[23] Flach, P., & Kull, M. (2015). Precision-recall-gain curves: PR analysis done right. In Advances in neural information processing systems (pp. 838-846).

[24] Çapraz, S. (2016). A content boosted hybrid recommendation system (Doctoral dissertation, Middle East Technical university).

[25] Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.