UNIVERSITY OF WESTMINSTER

WestminsterResearch
http://www.wmin.ac.uk/westminsterresearch

A comparative study of selected classification accuracy in user profiling

Ayse Cufoglu
Mahi Lohi
Kambiz Madani

School of Informatics, University of Westminster

2008 Seventh International Conference on Machine Learning and Applications

A Comparative Study of Selected Classification Accuracy in User Profiling

Ayse Cufoglu, Mahi Lohi and Kambiz Madani
Department of Electronics, Communication and Software Engineering
University of Westminster
London, UK
{A.Cufoglu1, lohim, madanik}@westminster.ac.uk

***Abstract*— In recent years the used of personalization in service provisioning applications has been very popular. However, effective personalization cannot be achieved without accurate user profiles. A number of classification algorithms have been used to classify user related information to create accurate user profiles. In this study four different classification algorithms which are; Naïve Bayesian (NB), Bayesian Networks (BN), Lazy Learning of Bayesian Rules (LBR) and Instance-Based Learner (IB1) are compared using a set of user profile data. According to our simulation results NB and IB1 classifiers have the highest classification accuracy with the lowest error rate.**

I. Introduction

In literature there are various definitions for user profile [1]-[3]. However, we can define it as the description of the user interests, characteristics, behaviors and preferences. User profiling is the practice of gathering, organizing and interpreting the user profile information [4]-[6].

As previously mentioned, user profiles include various information about each user. For instance, if we assume that user profiles are three dimensional matrices, each dimension of the matrix will represent a particular user related information such as; personal profile data (demographic profile data), interests profile data and preference profile data.

There are few works that compare some of the classification algorithms. In [8] Huang *et al.* compared AUC – known as the area under the ROC (Receiver Operating Characteristics) Curve- and accuracy of Naïve Bayes, Decision Trees and Support Vector Machine (SVM). Authors claimed that AUC is a better measure of accuracy with respect to the degree of discriminancy and consistency. According to their experimental results Naïve Bayesian, Decision Trees (C4.5, C4.4) and SVM are very similar with respect to the average predictive accuracy. In addition, Naïve Bayesian, C4.4 [19] and SVM have a similar average predictive AUC which is significantly higher than C4.5.

In another work Wang *et al.* [9] compared and constructed the relative performance of LBR and TAN (Tree Augmented Naïve Bayesian). In this work TAN algorithm approximates interactions between attributes by using a tree structures imposed on Naïve Bayesian structure [10]. LBR is desirable when small numbers of objects to be classified while TAN is desirable when large numbers of objects to be classified [14].

In [15] authors proposed Lazy Naïve Bayesian (LNB) algorithm and compare it with SNNB (Selective Neighborhood based Naïve Bayesian), LWNB (Locally Weighted Naïve Bayesian) and Lazy of Bayesian Rules (LBR). According to the authors, SNNB and LWNB improve classification accuracy of Naïve Bayesian (NB) while LNB improve ranking accuracy of NB. LNB spends no effort during training time and delays all computation until classification time. LNB learning algorithm deals with Naïve Bayes' unrealistic attribute conditional independence assumption by cloning each training instance to produce an expanded training instance. Based on the AUC measurement SNNB and LWNB can not significantly improve the NB, and LBR performs worse than NB. According to authors' experiments, LNB is slightly better than NB and C4.4 Decision Tree, with respect to the accuracy, robustness and stability.

In another work Zhang *et al.* [18] compared the ranking performance of NB and DT (C4.4) classifiers. The experiments conducted with using 15 dataset from UCI data repository [16] .According to the experimental results NB algorithm outperforms the C4.4 algorithm in 8 datasets, ties in 3 datasets and loses in 4 dataset. The average AUC of NB is 90.36% which is substantially higher than the average 85.25% of C4.4. Considering these results, authors argue that NB performs well in ranking, just as it does in classification.

This study is aimed to find the best classification algorithm for user profiling process.

In this paper Naïve Bayesian networks (NB), Bayesian Networks (BN), Instance-Based Learner (IB1) and Lazy Learning of Bayesian Rules (LBR) classification algorithms are compared in terms of classification accuracy of the user profile data. These four algorithms have been chosen since BN and NB algorithms are two of the most successful algorithms in Machine Learning (ML) and Data Mining (DM) fields; IB1 has never been considered for such a research work with BN, LBR and NB; and LBR is one of the best NB algorithms.

978-0-7695-3495-4/08 $25.00 © 2008 IEEE
DOI 10.1109/ICMLA.2008.139

II. NB, BN, LBR AND IB1 ALGORITHMS

The following section describes the NB, BN, LBR and IB1 classification algorithms.

Bayesian networks are probability based and are used for the reasoning and the decision making in uncertainty, and heavily rely on bayes' rule [7]. Bayes' rule can be defined as follows [7],

- Assume A_i attributes where i= 1,2,3,…,n, and which take values a_i where i= 1,2,3,…,n.
- Assume C as class label and $E = (a_1, a_2, ..., a_n)$ as unclassified test instance.
- E will be classified into class C with the maximum posterior class probability $P(C \mid E)$,

$$P(C \mid E) = \arg\max_{C} P(C)P(E \mid C) \tag{1}$$

Bayesian Networks can represent uncertain attribute dependencies, however it has been proven that learning optimal Bayesian network is NP (Non-deterministic Polynomial) hard [15].

Naïve Bayesian Classifier is one of the Bayesian Classifier techniques which also known as the state-of-the-art of the Bayesian Classifiers. In many works it has been proven that Naïve Bayesian classifiers are one of the most computationally efficient and simple algorithms for ML and DM applications [9] - [12]. Naïve Bayesian classifiers assume that all attributes within the same class are independent given the class label. Based on this assumption, the Bayesian rule has been modified as follows to define the Naïve Bayesian rule [7],

$$P(C \mid E) = \arg\max_{C} P(C)\prod_{i=1}^{n} P(A_i \mid C) \tag{2}$$

Naïve Bayesian classifiers are used within many interactive applications because of its efficiency and effectiveness. However, because of its naïve conditional independence assumption, optimal accuracy can not be achieved. LBR is one of the lazy learning algorithms that have been proposed to improve the accuracy performance of Naïve Bayesian classifier. LBR algorithm can be thought of as applying Lazy Learning techniques to Naïve Bayesian rule [9]. At the classification time of each test instance, LBR algorithm builds the most appropriate Bayesian rule for the test instance. Following formula shows the LBR Bayes rule that used for classification [17],

$$P(C_i \mid V_1 \wedge V_2) = P(C_1 \mid V_2) P(V_1 \mid C_i \wedge V_2) / P(V_1 \mid V_2) \tag{3}$$

Here we assume that V_1 and V_2 are any two conjunction of attribute values and $V = (a_1, a_2, ..., a_i)$ is an attribute vector. At each instance classification time each attribute values a_i from V are allocated to exactly V_1 or V_2 such that $V_1 = (A_1, A_2, ..., A_n)$ and $V_2 = (A_{n+1}, A_{n+2}, ..., A_t)$ where $A_i = a_i$.

IB1 or IBL (Instance-Based Learning) is one of the other classifiers and it is a comprehensive form of the Nearest Neighbor algorithm [13] [14]. IB1 generates classification predictions using only specific instances. Unlike Nearest Neighbor algorithm, IB1 normalizes its attributes' ranges, processes instances incrementally and has a simple policy for tolerating missing values [14]. IB1 uses simple normalized Euclidean distance (similarity) function to yield graded matches between training instance and given test instance [13]. Following function is the similarity that is used within IB1 algorithm [14],

$$\mathit{Similarity}(x, y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)} \tag{4}$$

Here, instances are represented by n attributes where $f(x_i, y_i) = (x_i - y_i)^2$ represents numeric valued attributes and $f(x_i, y_i) = (x_i \neq y_i)$ represents Boolean and symbolic attributes.

III. CLASSIFICATION ACCURACY

In this section we compare the results of four classifiers (NB, BN, LBR and IB1). The simulations conducted twice with using two different datasets. The first dataset reflects the users' personal information (demographic data) while the second dataset incorporates the user's personal information with the user's interests and preferences information. As a demographic profile data, UCI's adult dataset [16] has been modified and used. All simulations were performed in the Weka machine learning platform that provide a workbench which consist of collection of implemented popular learning schemes that can be used for practical data mining and machine learning works [13].

Below we highlighted the procedure for the simulations;

- Datasets have been converted into Weka readable ".cvs" format (see Table I). First 20 instances of the UCI's adult dataset have been chosen for the simulations.
- The first dataset, demographic user profile, includes 20 instances and 10 attributes (see Table I). These attributes are; Age, Work-class, Education, Education-num, Marital-status, Occupation, Relationship, Race, Sex and Native-country. In table I, missing values indicated with "?" symbol.
- The second dataset, extended user profile, consists of 20 instances and 18 attributes. These attributes are; Age, Work-class, Final-weight, Education, Education-num, Marital-status, Occupation, Relationship, Race, Sex, Native-country, capital-gain, capital-loss, Hours-per-week, Interest-music, interest-book, interest-sport and Preference-sound.

Table I. Personal User Profile Data in ".cvs" Format

Age	Work-class	Education	Education-num	Marital status	Occupation	Relationship	Race	Sex	Native country
25	Private	11^{th}	7	Never-married	Machine-op-inspct	Own-child	Black	Male	United-states
38	Private	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	United-states
28	Local-gov	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	United-states
44	Private	Some-collage	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	United-states
18	?	Some-collage	10	Never-married	?	Own-child	White	Female	United-states
34	Private	10^{th}	6	Never-married	Other-service	Not-in-family	White	Male	United-states
29	?	Hs-grad	9	Never-married	?	Unmarried	Black	Male	United-states
63	Self-emp-not-inc	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-states
24	Private	Some-collage	10	Never-married	Other-service	Unmarried	White	Female	United-states
55	Private	7^{th}-8^{th}	4	Married-civ-spouse	Craft-repair	Husband	White	Male	United-states
65	Private	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	United-states
36	Federal-gov	Bachelors	13	Married-civ-spouse	Adm-clerical	Husband	White	Male	United-states
26	Private	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	United-states
58	?	HS-grad	9	Married-civ-spouse	?	Husband	White	Male	United-states
48	Private	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	United-states
43	Private	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-states
20	State-gov	Some-collage	10	Never-married	Other-service	Own-child	White	Male	United-states
43	Private	HS-grad	9	Married-civ-spouse	Adm-clerical	Wife	White	Female	United-states
37	Private	HS-grad	9	Widowed	Machine-op-inspct	Unmarried	White	Female	United-states
40	Private	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac.	Male	?

- We chose 10 fold cross-validation as a test mode where 10 pairs of training sets and testing sets are created. All previously mentioned classification algorithms run on the same training sets and have been tested on the same testing sets to obtain the classification accuracy.
- Unlike other aforementioned three algorithms, LBR cannot handle numeric attributes. Therefore, before we do simulations with LBR, we normalized and binarised the attribute values of both datasets using unsupervised attribute filters "Normalized" and "Numeric-To-Binary".

A. Comparison of the Results

We conducted the first simulations on demographic user profile dataset to compare NB, BN, LBR and IB1 classifiers using classification accuracy as evaluation criterion. Table II demonstrates the classification accuracy results of these four classifiers. As we can see from table II, NB and IB1 classifiers have the result of 95% where 19 dataset instances have been classified correctly and 1 instance has been classified incorrectly. Moreover, with the second highest result that is 90%, LBR classifier followed the outcome of NB and IB1 algorithms. Bayesian classifier result is the lowest which is 85% (17 correctly classified and 3 incorrectly classified instances). Here, both NB and IB1 outperform the LBR and BN classifiers in terms of classification accuracy.

Table III shows that precision of the four classification algorithms are very similar.

Table II. Classification Accuracy Test Results (Simulation1)

Classifier	Correctly classified instances	Incorrectly classified instances
NB	19 (95%)	1 (5%)
IB1	19 (95%)	1 (5%)
LBR	18 (90%)	2 (10%)
BN	17 (85%)	3 (15%)

Table III. Classifiers vs. Precision

Classifier	Precision
NB	0.95
IB1	0.95
LBR	0.947
BN	0.944

Fig. 1 shows the error rate results. Here four different parameters are used to represent the error rate of the four classification algorithms. These are; Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE). It shows that NB and IB1 classifiers have the lowest error rate. Furthermore, BN classifier has the highest error rate and the difference is more in RRSE and RAE measurements, knowing that low error rate cause high accuracy or vice versa. Based on the above classification accuracy results (see Table II), the BN classifier demonstrates the highest error rate (see Fig 1).

In order to compare the classification accuracy performance of the NB, BN, LBR and IB1 classifiers with complete user profile data, a second simulation was performed on the extended user profile dataset. During the second simulation we have observed the following;

- The classification accuracy performance of the BN classifier was 80%. Therefore, when this result is compared with the first simulation we can see that BN classifiers performance degreases 5% from 85% to 80%. On the other hand, for NB, IB1 and LBR classifiers, first simulation results have remained the same during the second simulations (see Table IV). Therefore, NB and IB1 classification algorithms keep performing well with bigger user profile dataset.

According to our simulation results NB outperforms BN classifier. This is due to the fact that NB classifier assumes that class attributes within the same class are conditionally independent given the class label. Furthermore, we know that LBR classifier proposed to improve the performance of NB classifier by applying the lazy algorithm on the NB classifier. However, our results show that LBR classifier performs lower classification accuracy than NB.

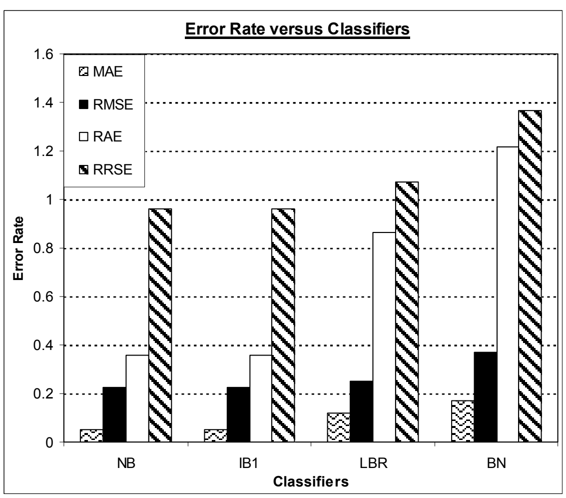

Figure1. Error rate measures of Classifiers (Simulation 1)

Table IV. Classification Accuracy Test Results (Simulation 2)

Classifier	Correctly classified instances	Incorrectly classified instances
NB	19 (95%)	1 (5%)
IB1	19 (95%)	1 (5%)
LBR	18 (90%)	2 (10%)
BN	16 (80%)	4 (20%)

- Fig. 2 shows the error rate results of the four classifiers respectively. According to these results, in the second simulations RAE of LBR and BN classifiers have increased significantly. This increment is much more in BN classifier where RAE increases from 121% to 162%.

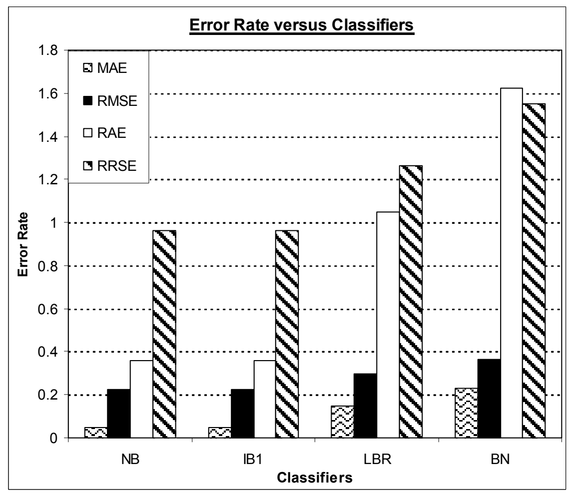

Figure2. Error rate measures of Classifiers (Simulation 2)

IV. CONCLUSION

In this paper we evaluated classification accuracy of four classification algorithms (BN, NB LBR and IB1). All simulations were performed in Weka [13] machine learning platform. Moreover, UCI adult dataset [16] has been modified and used as a demographic user profile data. The aim of these simulations was to find the best classification algorithm that has a high classification accuracy performance on the user profile data. According to the simulation results NB and IB1 classifiers perform the best classification on user related information. Furthermore, LBR shows similar results to NB and IB1 that are slightly different from BN. This indicates that NB and IB1 classification algorithms should be favored over LBR and BN classifiers in the personalization applications especially when the classification accuracy performance is important. In our future work, we will compare the well known DT and SVM classifiers with IB1 and NB classifiers with respect to classification accuracy performance on relatively larger user profile dataset.

REFERENCES

[1] Araniti G., Meo P. D., Iera A. and Ursino D. (2003) “Adaptive controlling the QoS of multimedia wireless applications through user profiling techniques”, IEEE Journal on selected areas in communication, Vol. 21, No. 10, Page(s) 1546-1556.

[2] ETSI (2005) Human Factors (HF). “User profile management” [Online] Available from: http://webapp.etsi.org

[3] Kuflik T. and Shoval P. (2000) “Generation of user profiles for information filtering- research agenda” [Online] Available from: http://delivery.acm.org

[4] Henczel S. (2004) “Creating user profiles to improve information quality**”** Factiva**,** Vol. 28, No 3, Page(s) 30, [Online] Available From: http://global.factiva.com/ha/default.aspx

[5] Usability by Design (2004) User Profiling [Online] Available from: www.usability.uk.com/glossary/user-profiling.htm

[6] Open Interface LTD (-) User Profiling [Online] Available from: http://www.openinterface.ie/development/design_user.html

[7] Jensen F.V. (1993) *Introduction to Bayesian Networks*. Denmark, Hugin Expert A/S.

[8] Huang J., Lu J. and X. Ling C. (2003) “Comparing naïve bayes, decision trees, and SVM with AUC and Accuracy” . 3rd conferences on Data Mining, IEEE , ICDM 2003, Page(s) 553 - 556

[9] Wang Z. and I. Webb G. (2002) “Comparison of lazy bayesian rule and tree-augmented bayesian learning”.IEEE conference on Data Mining, ICDM 2002, Page(s): 490 – 497.

[10] Shi Z., Huang Y. and Zhang S. (2005) “Fisher score based naive bayesian classifier”. International conference on Neural Networks and Brain, IEEE , ICNN&B 2005, Page(s) 1616- 1621.

[11] Xie Z. and Zhang Q. (2004) “A study of selective neigbourhood-based naïve bayes for efficient lazy learning”. 16th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2004.

[12] Santafe G., Loranzo J.A. and Larranaga P. (2006) “Bayesian model averaging of naive bayes for clustering” .IEEE Transactions on Systems, Man, and Cybernetics, Vol 36, No 5, Page(s) 1149 – 1161.

[13] Ian H. Witten and Frank E. (2005) *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[14] David W. Aha, Kibler D. and K. Albert M. (1991) “ Instance-based learning algorithms” . Machine Learning journal, Vol 1, No 6, ISDN 1573-0565, Page(s):37-66. [Online] Available from: http://www.springerlink.com/content/kn127378pg361187/fulltext.pdf

[15] Jiang L. and Guo Y. (2005) “Learning lazy Naïve Bayesian classifier for ranking”. 17th IEEE conference on tools with artificial intelligence, IEEE 2005, Page(s) 5pp.

[16] Asuncion A. and Newman D.J. (2007) UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science. [Online] Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html

[17] Zheng Z. and Webb G. I. (2000) “Lazy learning of bayesian rules”. Machine Learning , Vol 41 , No 1, Page(s): 53-87. [Online] Available from: http://www.csse.monash.edu.au/~webb/Files/ZhengWebb00.pdf

[18] Zhang H. and Su J. (2004) “Naive bayesian classifiers for ranking” . 15th European Conference on Machine Learning, ECML 2004. [Online] Available from: http://www.cs.unb.ca/profs/hzhang/publications/NBRanking.pdf

[19] Provost F. and Domingos P. (2003) “Tree induction for probability based ranking” . Kluwer Academic Publishers, Vol 52 , No 3 [Online] Available from: http://portal.acm.org