Data Mining Methodology for Web Users' Demographic Data Prediction

Vesna Gega Kumbaroska[1] and Ilija Jolevski[2]

Abstract –**In this paper we present a data mining methodology used for web users' demographic prediction. Nowadays, the usage of web is directed toward the user, so following that trend, using this methodology, demographic targeted web advertising and dynamic web content personalization will be improved. Demographic attributes of interest for our study are: gender and age. The gender attribute is qualitative or discrete variable which contains two values: male and female. The age attribute is quantitative or continuous variable and it contains user's age. This variable is discretized to these categories of ages: A – aged 20 and younger, B – aged 21-30, C – aged 31-40, D – aged 41-50 and E – aged 51 and older. Experiments are performed on a real data obtained from a web log file and a survey, including following classification algorithms: *K*-Nearest Neighbours, Naïve Bayes, Bayesian Network and Decision Trees as single classifiers, and, Bagging, Boosting and Random Forest as ensembles. Four evaluation measures are used to evaluate the performance of each classifier on the pre-processed version of the data set: Precision, Recall, F-Measure and ROC Area or Area Under Curve (AUC). Our comparative analysis is performed using F-Measure and AUC on Female and B aged class attribute value, regarding the survey, where 62% of the visitors are women and 73% of the visitors are aged between 21 and 30 years. The results show that with the both measures, for the Female class value prediction, *K*-Nearest Neighbours and Random Forest are most superior algorithms of all single classifiers and ensembles, respectively. For the B aged class value prediction, once again *K*-Nearest Neighbours is most superior algorithms of all single classifiers and Bagging in combination with Naïve Bayes is preferred over all ensembles.**

Keywords –**Classification, Prediction, Naïve Bayes, K-Nearest Neighbors, Decision Trees, Random Forest, ID3, Weka, Evaluation, Precision, Recall, F-Measure, AUC, ROC curve.**

I. INTRODUCTION

On one hand, if a site contains personal data, it is not exposed in public in order to protect the right for privacy. On the other hand, some sites don't contain functionality for storing and using personal data associated with a specific user. However, such information can play a main role in personalization/characterization of different web services, including targeted advertising [9], [12], in order to improve user experience, user engagement and user satisfaction [8], [14].

Obtaining demographic information is not easy procedure. Several proposed approaches exist[2], [4], [7], and [15]. One of the possible alternatives is to predict users' demographics, such as gender and age, using a data mining methodology. In this paper we develop a case study for an entertainment site. The basic idea is to discover (one of several) classification algorithms that leads to best scores in demographic data prediction. For this purpose, first, we created a data set. One part of the data set was collected as a real web log file on the server side, where each visit during one week was recorded. Because the entertainment site does not support user profiles, a survey was inevitable to be performed in order to collect demographic data. The survey was offered to a group of daily visitors of this site during one week, where they were asked for several data, including their gender and age. The rest of the paper provides thorough explanation of the tasks performed, as follows.

The second section refers to the methodology used. We start with our demographic prediction problem, and then we continue with the pre-processing task used for converting the data set into a cleaned version. At the end we give brief theoretical introduction of the classification algorithms chosen for experimenting. In the third section, we put special emphasis on the measures chosen for performance evaluation. Next, we present the results obtained and provide their comparative analysis. Finally, in the last section, we draw a conclusion and give some future research directions.

II. METHODOLOGY

Demographic attributes of interest for our study are: gender and age. In this section we introduce our methodology for demographic data prediction, including pre-processing task and classification algorithms used.

Pre-processing

Before the classification task has been performed, the pre-processing task was undertaken. Modification and removing some attributes that were not important for our analysis were done using MS Visual C# script, which means converting the data set into cleaned version. The URL visited attribute contains information about the category, the subcategory and the ID of the article, thus we parsed it to these three attributes of interest. The gender attribute is qualitative or discrete variable which contains two values: male and female. This attribute was not changed. The age attribute is quantitative or continuous variable and it contains visitor's age. This variable

[1]Vesna Gega Kumbaroska is with the University for Information Science and Technology, Partizanska bb, Ohrid 6000, Republic of Macedonia, E-mail: vesna.gega@uist.edu.mk.

[2]IlijaJolevskiis with the Department of Computer Science and Engineering, Faculty of Technical Sciences, St. Clement of Ohrid University, Ivo Lola Ribar bb, 7000 Bitola, Republic of Macedonia, E-mail: ilija.jolevski@uklo.edu.mk.

was discretized using Weka[1], the same software used for the classification task[1], [5].We ended up with these categories of ages: A – aged 20 and younger, B – aged 21-30, C – aged 31-40, D – aged 41-50 and E – aged 51 and older. The final version of the data set contains five attributes: URL, category, subcategory, ID, gender/age.

Classification

As we mentioned before, for the classification task we used Weka open source data mining software [5]. We randomly split the already cleaned and modified data set into a training set which contains 2/3 and a testing set which contains 1/3 of the whole data set. Both in the training and testing phase we performed 10-fold cross validation. The classification algorithms employed during classification task are: Naïve Bayes, Bayesian Network, Ensembles, *K*-Nearest Neighbors, and Decision Trees[1], [13]. Brief theoretical background of these algorithms is given below.

Naïve Bayes is simple probabilistic classifier based on Bayes Theorem, where it is assumed the attributes relationship is naïve, or the attributes are independent [13]. This is represented using the following equation, where *A* is related to attributes and *C*refers to a class in the appropriate problem domain:

$$classify(C = c|A_1, A_2, \dots, A_n) = \text{argmax}_c\, p(C = ci{=}1 np(Ai{=}ai|C{=}c)$$

Bayesian Network is used as a classifier using the inference algorithm in the following way[5], [12]:

$$classify\ (C = c|A_1, A_2, \dots, A_n) = \underset{c}{\text{argmax}} \prod_{u \in U} p(u|parents(u))$$

Here, *U* is a set of known attributes represented as a network structure which is a direct acyclic graph over *U*and contain probability tables. Bayesian network represents probability distribution $P(U) = \prod_{u \in U} p(u|parents(u))$. Both, the Naïve Bayes and the Bayesian Network were used with their default settings in Weka.

K-Nearest Neighbors is one of the most simple classification algorithms, which is an instance-based learning algorithm, or also known as a lazy learning algorithm [13]. The main functionality is to find *K* most similar samples (neighbors) to the test sample, where *K* is usually an odd integer, and the class is chosen using the majority rule. Different distance metrics can be used in order to measure the similarity, such as Euclidean distance metric. If the value for *K* is very small number it may be sensitive to noise. Contrary to this, if *K* is very large number it destroys locality. Thus, we were very careful with choosing the right value for *K* to be 3, following the rule: $K = \sqrt{number\ of\ attributes}$, where the *number of attributes* in our case is 5.

The idea behind the ensemble classifiers is to learn and combine the predictions of multiple classifiers in order to achieve better predictive performance. Producing a weighted vote with a collection of classifiers in an iterative way is implemented in the method called Boosting. The most common version of this method is Ada Boost [13].Averaging the prediction using the majority vote rule over a set of classifiers is implemented in the method called Bagging [10], [13]. We used the both methods with their default settings in Weka, in combination with Naïve Bayes algorithm.

Decision Trees classifier is also one simple and widely used algorithm which uses a decision tree as a predictive model [13].The tree is constructed in a recursive top-down approach. The root and the leaves contain attribute conditions to separate the data and the terminal nodes represent class labels. There are several algorithms of this type, which select attributes based on different statistical measures. We decided to use Decision Trees based on the Information Gain Ratio or ID3 measure, with its default settings in Weka. Also, there are techniques that combine more than one tree with a notation of ensemble, such as Random Forest [11], [13]. As we mentioned above, the result may either be an average or weighted average of all of the terminal nodes that are reached, or, a voting majority. Once again, we used Random Forest with its default settings in Weka.

III. EVALUATION

Evaluation Measures

Four evaluation measures are used to evaluate the performance of each classifier on the preprocessed version of the data set: Precision, Recall, F-Measure, ROC Area or Area Under Curve (AUC) [13]. Precision is defined as a ratio of the number of records that are classified as true to the total number of records classified in the class (both true and false). Recall is defined as a ratio of the number of records that are classified as true to the total number of records that are correctly classified or belong in the class. Usually, there is an inverse relationship between Precision and Recall, when one goes up, the other goes down. Mostly, they are expressed in percentage. F-Measure is very useful measure, defined as a combination of Precision and Recall, as follows:

$$F - Measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Finally, the ROC curve is created by plotting several pairs of True Positive Rate and False Positive Rate for different threshold values. There are several measures to express the ROC curve as a single number, such as the popular one AUC measure. It expresses the probability of a model correctly determining which out of two possible classes; is the one that is providing the most truthful match [13]. The values can range from 0.5 to 1 and higher values mean better model performance.

We will use F-Measure, AUC and ROC curves to compare the performance scores of the classifiers used and to interpret their meaning.

[1]http://www.cs.waikato.ac.nz/ml/weka/

Evaluation Results

We picked to use Female class attribute value for which we describe the performance of the classifiers for the gender prediction. Maybe, this is not solely an intuitive (instinctive)decision, because if we look at the survey data collection, 62% of the visitors are women.

We observe that, according to the F-Measure, *K*-Nearest Neighbors achieves best results (68.3%) compared to all single classifiers. Also, Bayesian Networkis slightly better (64.8%) than Naïve Bayes, which does not have low performance score (64.2%). Decision Trees demonstrate worst performance score (57.8%). Random Forest seems to show better performance (65.6%) versus the other ensembles. Also, Bagging (63.0%)) is preferred over Boosting (56.6%). Using the AUC measure, once again, *K*-Nearest Neighbors achieves best result (62.7%) compared to all single classifiers. We can see reversed situation here, Naïve Bayes (59.0%) is slightly better than Bayesian Network (57.9%). Decision Trees demonstrate worst performance score (57.8%) compared to all single classifiers. The performance ratio of the ensembles using AUC is very same as using F-Measure. All these comparisons arise from Table 1. and they are visually plotted in Fig.1. Also, AUC performances are depicted in the ROC curve analysis in Fig.2.

TABLE I
GENDER PREDICTION EVALUATION RESULTS

Classification Algorithm	TP Rate	FP Rate	Precision	Recall	**F-Measure**	**ROC Area**	Class
3NN	0.872	0.780	0.562	0.872	**0.683**	**0.627**	f
AdaBoost	0.596	0.585	0.538	0.596	0.566	0.523	f
Bagging	0.723	0.659	0.557	0.723	0.630	**0.595**	f
Bayes Net	0.723	0.585	0.586	0.723	0.648	0.579	f
ID3	0.684	0.684	0.500	0.684	0.578	0.475	f
Naïve Bayes	0.723	0.610	0.576	0.723	0.642	0.590	f
Random Forest	0.851	0.854	0.533	0.851	**0.656**	0.509	f

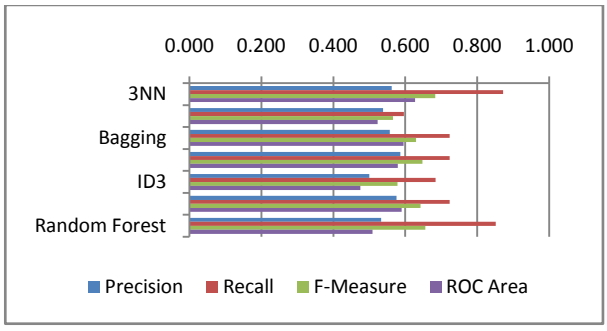

Fig.1. Comparison of the performance of different classifiers for gender prediction

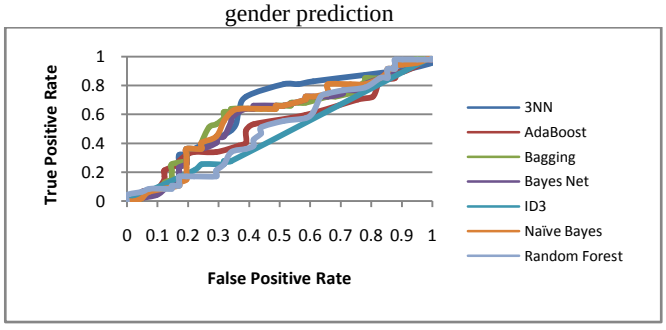

Fig.2 ROC analysis of the performance of different classifiers for Gender Prediction

According to the survey, 73% of the visitors are aged between 21 and 30 years. That is the reason we picked group B as a target class attribute value for which we describe the performance of the classifiers for the age prediction.

We observe that, according to the F-Measure, *K*-Nearest Neighbors and Naïve Bayes achieves best results (64.6%) compared to all single classifiers. But, a bit higher recall value (100%) for *K*-Nearest Neighbors makes it better than Naïve Bayes (97.6%). Also, Bayesian Network is slightly better (62.2%) than Decision Trees which demonstrate worst performance score (57.1%). In this case, Random Forest seems to show worse performance (61.8%) versus the other ensembles. Bagging (64.6%) is preferred over Boosting (61.7%). Using the AUC measure, we can observe that Naïve Bayes achieves best performance (57.1%).Bayesian Network shows slightly decreased performance (56.1%), but it is better than *K*-Nearest Neighbors (54.9%). Once again it is proofed that Decision Trees are the worst classifier (47.6%). Random Forest is not good choice this time (45.9%) over the other ensembles. Bagging (56.2%) is once again preferred over Boosting (50.5%). All these comparisons arise from Table 2. and they are visually plotted in Fig.3. Also, AUC performances are depicted in the ROC curve analysis in Fig.4.

TABLE II
AGE PREDICTION EVALUATION RESULTS

Classification Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
3NN	1.000	1.000	0.477	1.000	**0.646**	0.549	B
AdaBoost	0.881	0.891	0.474	0.881	0.617	0.505	B
Bagging	0.976	0.957	0.482	0.976	**0.646**	**0.562**	B
Bayes Net	0.881	0.870	0.481	0.881	0.622	**0.561**	B
ID3	0.778	0.850	0.452	0.778	0.571	0.476	B
Naïve Bayes	0.976	0.957	0.482	0.976	**0.646**	**0.571**	B
Random Forest	0.905	0.935	0.469	0.905	0.618	0.459	B

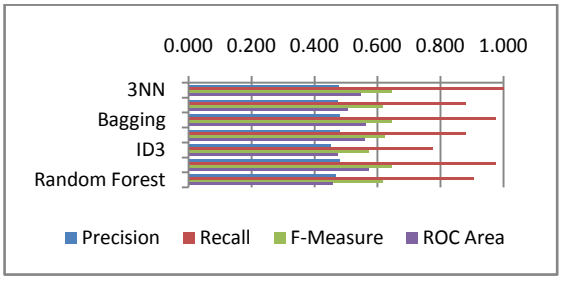

Fig.3. Comparison of the performance of different classifiers for age prediction

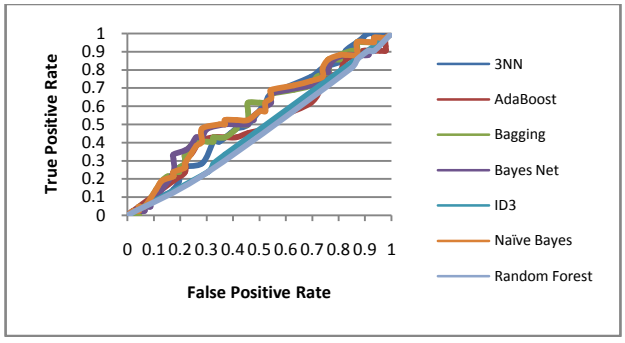

Fig.4. ROC analysis of the performance of different classifiers for Age Prediction

IV. CONCLUSIONS

This paper focuses on the data mining methodology used for web users' demographic prediction [2], [4], [7], and [15].Nowadays, the usage of web is directed toward the user, so following that trend, this methodology will improve user satisfaction, user experience, sales on web, which mean remaining the visitors on the website longer, downloading more, and purchasing more products. Especially this approach will be useful in demographic targeted web advertising [9], [12] and dynamic web content personalization [17] based on demographic criteria.

Experiments are performed on a real data including following classification algorithms: *K*-Nearest Neighbors, Naïve Bayes, Bayesian Network and Decision Trees as single classifiers [1], [13], and, Bagging [10], Boosting and Random Forest [11] as ensembles [13]. The evaluation results indicate that: i) Best results for Female class value prediction using both F-Measure and AUC measure are achieved with*K*-Nearest Neighbors (*K*=3), (68.3% and 62.7 respectively), compared to all single classifiers. Also, Random Forest seems to show better performance (65.6%) versus the other ensembles.ii) Best results for B aged (21-30 aged) as a targeted class value, using F-Measure are achieved with *K*-Nearest Neighbors (*K*=3)and Naïve Bayes (64.6%) compared to all single classifiers. Also, Naïve Bayes is preferred using the AUC measure. In this case, Bagging in combination with Naïve Bayes is preferred over the other ensembles (64.6% and 56.2%) using F-Measure and AUC measure respectively.

In future work we are interested in prediction additional demographic data, such as: marital status and education. Also, we plan to focus on demographic prediction inferred from users' behavior patterns discovered from web log files [7], [16].

REFERENCES

[1] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi and E. M. Al-Shawakfa, "A comparison study between data mining tools over some classification methods", International Journal of Advanced Computer Science and Applications, Special Issue, pp. 18-26, 2011.

[2] B. Bi, M. Shokouhi, M. Kosinski and T. Graepel, "Inferring the demographics of search users", Prof. of World Wide Web (WWW), 2013.

[3] B. Mobasher, R. Cooley, and J. Srivastava.."Automatic personalization based on Web usage mining". Commun. ACM 43, DOI=10.1145/345124.345169, pp. 142-151, 2000.

[4] D. Murray and K. Durrell, "Inferring demographic attributes of anonymous Internet users", Proc. International Workshop on Web Usage Analysis and User Profiling (B. Masand and M. Spiliopoulou, Ed.), LNCS 1836, 2000.

[5] E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, 1(1), 2009,

[6] F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer, 2001.

[7] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior". In Proc. 16th WWW, pp. 151–160, 2007.

[8] J. Lehmann, M. Lalmas., E. Yom-Tov and G. Dupret. *Models of user engagement*, In User Modeling, Adaptation, and Personalization. Springer Berlin Heidelberg, pp. 164-175, (2012).

[9] K. De Bock and D. Van den Poel, "Predicting Website Audience Demographics for Web Advertising Targeting Using Multi-Website Clickstream Data", Fundam. Inf. 98, 1, pp. 49-70, 2010.

[10] L. Breiman, "Bagging predictors", *Machine Learning, 24(2*), pp. 123-140, 1996.

[11] L. Breiman, "Random forests", *Machine Learning, 45(1*), pp. 5-32, 2001.

[12] P. Kazienko and M. Adamski, "AdROSA - Adaptive personalization of web advertising", *Information Sciences*, 177(11), pp. 2269-2295, 2007.

[13] P-N Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2006.

[14] S. Attfield, G. Kazai, M. Lalmas and B. Piwowarski, "Towards a science of user engagement" (Position Paper), In WSDM Workshop on User Modelling for Web Applications, 2011.

[15] S. Speltdoorn and D. Van den Poel, "Predicting demographic characteristics of web users using semi-supervised classification techniques".

[16] V. Gega and P. Mitrevski, "Using Generalized Stochastic Petri Nets to Model and Analyze Search Behaviour Patterns", In International Journal of Reasoning-based Intelligent Systems 6(2), pp. 26-33, 2014.