Clustering Algorithms and Weighted Instance Based Learner for User Profiling

Ayse Cufoglu
Dept. of Computing and Technology
Faculty of Science and Technology
Anglia Ruskin University
Cambridge, UK
a.cufoglu@ieee.org

Mahi Lohi, Colin Everiss
Dept. of Electronic, Networks and Computer Engineering
Dept. of Business Information Systems
University of Westminster
London, UK
lohim@wmin.ac.uk, c.g.everiss@wmin.ac.uk

***Abstract*—User profiling has created opportunities for service providers to make available a channel for user awareness as well as to achieve high user satisfaction. Apart from traditional collaborative and content-based methods, a number of classification and clustering algorithms have been used for user profiling. In our previous work, a weighted classification method, namely Weighted Instance Based Learner (WIBL), was proposed and evaluated for user profiling. In this paper, we aim to compare the performance of a WIBL algorithm with well known clustering algorithms for user profiling. Simulations showed that a WIBL is capable of outperforming the clustering algorithms.**

Keywords-User Profiling; Weighted Instance Based Learner (WIBL); Clustering Algorithms.

I. INTRODUCTION

Personalization of services, is an opportunity to help improve the quality of service. The success of these services relies on how well the service provider knows the user requirements and how well this can be satisfied. The user profile is the representation of the user and holds information about the user such as personal profile data (demographic profile data), interest profile data and preference profile data. These profiles are the outcome of the user profiling. In user profiling applications a major challenge is to build and handle user profiles. In the literature, two fundamental user profiling methods have been proposed for this purpose. These are the collaborative and the content-based methods. It is also possible to use a hybrid of these two methods [1]-[3].

The collaborative method has been built on the assumption that similar users, with respect to the age, sex, and social class, behave similarly, and therefore have similar profiles [1][4]. The content-based method, on the other hand, has been built on the concept of content similarity and assumes that users behave similarly under the same circumstances [1][4]. Apart from the traditional profiling methods, well known data mining and machine learning algorithms have found applications within the user profiling process in personalization [5]-[7]. This paper is the first in the literature to compare the performance of Weighted Instance Based Learner (WIBL) [8], with selected algorithms for user profiling.

The paper is organized as follows: Section II, provides related works for this study. Section III, provides information about the algorithms, while Section IV, presents the simulation results. Finally, Section V, concludes this paper.

II. RELATED WORKS

Various research works have been carried out with user profiling methods [9]-[13]. For example, the moreTourism, mobile recommendations for tourism [9], uses a hybrid method. The proposed recommended system takes into account the tags, provided by the users, to provide tourist information profiled for users with similar likes depending on the user profile (user tag cloud), location in time and space, and the nearby context (e.g., nearby historical places and museums). In [10], Fernandez et al. proposed a tourism recommendation system that offers tourist packages (i.e., include tourist attractions and activities), that best matches with the user's social network profiles. Different from [9], the proposed hybrid system provides recommendations based on both the user's viewing histories (in this instance, Digital Television (DTV) viewing histories received from the user's set-top boxes via a 2.5/3G communication network) and the preferences in the social network (i.e., preferences of the user's friends). In [11], collaborative filtering was employed together with techniques from the Multi Criteria Decision Analysis (MCDA) for item recommendation. In this system, the MCDA was used to find the similar users while collaborative filtering was used to recommend items. In another work, a hybrid TV program recommendation system, gueveo.tv [12], has been proposed. According to Martinez et al. [12], the proposed system works well because both methods complement each other in a way, that the content-based method recommends usual programs while collaborative method provides the discovery of new shows.

The significance of user profiles for various personalization applications has triggered the use of classification and clustering algorithms in user profiling [5]-[7]. In [5], Irani et al. focused on the social spam profiles in MySpace. In their work, they compared well known machine learning

 ISBN: 978-1-61208-311-7

algorithms (AdaBoost algorithm, C4.5 Decision Tree (DT), Support Vector Machine (SVM), Neural Networks (NNs), and Naive Bayesian(NB)) with respect to their abilities to distinguish spam profiles from legitimate profiles. According to the simulations carried out on over 1.9 million MySpace profiles, the C4.5 DT algorithm achieved the highest accuracy of 99.4% in finding the spam profiles, while NB achieved 92.6% accuracy. Simulations were performed on the Weikato Environment for Knowledge Analysis (WEKA) platform where classifiers' default settings were used with 10 fold-cross validation. Paireekreng and Wong [6] investigated the use of clustering and classification of user profile at the client-side mobile. Here, the authors focused on content personalization to help mobile users retrieve information and services efficiently. In their proposed two phase framework, clustering was used to construct a user profile, while classification was classifying user profile based on the class information from clustering. In this work, K-means, TwoStep, Anomaly and Kohenen clustering algorithms were compared for clustering. Moreover, Locally Weighted Learning (LWL), RepTree, Decision Table and SVMReg classifiers were compared for classification. According to simulations, authors state that, for this framework, K-means and RepTree were the best options for classification and clustering respectively.

In our previous work [8], we have proposed a weighted classification method, WIBL. In this paper, however, we aim to compare the performance of WIBL with well known clustering algorithms on user profile.

III. ALGORITHMS

A. Weighted Instance Based Learner (WIBL)

Instance Based Learner (IBL), is a comprehensive form of the Nearest Neighbour (NN) algorithm which normalizes the range of its attributes, processes instances incrementally and has a simple policy for tolerating missing values [14]. In contrast to IBL, the WIBL [8] assigns weights to the attributes and considers the weighted distance of the instances for classification. Here, relevant attributes are aimed to have more influence on classification than irrelevant attributes. In WIBL the function that calculates the distance between test instance (new user) X_i and the training instance (existing user) Y_j is $dist(X_i, Y_j) = \sqrt{\sum_{k=1}^{A} w_{k,l}(C_m)\, g(x_i(k), y_j(k))}$, where $w_{k,l}(C_m) = P(C_m|f_k(l))$ [8]. Here, l is equal to the value of the $x_i(k)$. Therefore, the selection of which weight is to be used for a particular attribute value is based on k and $x_i(k)$. Note that $g(x_i(k), y_j(k))$ is evaluated as it is in IBL [8].

B. Clustering Algorithms

Clustering, also called unsupervised classification, is the process of segmenting heterogeneous data objects into a number of homogenous clusters [15]. Each cluster is a collection of data objects that are similar to one another and dissimilar to the data objects in other cluster/s [16]. A successful clustering algorithm has clusters with high intra-class similarity and low inter-class similarity [16] (see Figure 1 [17]).

Each clustering algorithm uses a different method to cluster the information. In the literature, the most popular clustering methods can be categorized into three subsections. These are Hierarchical, Partitional and Density-Based Clustering (DBC).

1) Hierarchical Clustering: Hierarchical clustering, is the process to create a hierarchical decomposition (dendogram) of the set of data objects [16]. The well known hierarchical clustering algorithms are Single-Linkage, Complete Linkage and Average-Linkage.

In Single-Linkage Clustering (SLC), the resulted distance between two clusters is equal to the shortest distance from any member of one cluster, to any member of the other cluster [18]. Here, the shortest distance reflects the maximum similarity between any two data objects in two different clusters.

The Complete-Linkage Clustering (CLC), is the opposite form of the single-linkage clustering since, in complete-linkage, the link between two different clusters is expected to be the maximum distance from any data object of one cluster to any data object of the other cluster [18]. The maximum distance reflects the minimum similarity between two data objects in two different clusters.

The Average-Linkage Clustering (ALC), can be considered as a combination of single and complete-linkage algorithms. The link between two clusters is equal to the average greatest distance of all paired data objects of these clusters.

2) Partitional Clustering: Partitional clustering is a non-hierarchical clustering method. This method creates disjoint clusters in one step, by decomposing the dataset. Therefore, there is no relationship among the clusters [19].

K-means, is the most representative algorithm of partitional clustering [17]. In this algorithm, the number of clusters, Q, is defined by the user. Then, randomly selected Q data objects become the center (cluster centroid) of the Q clusters. The rest of the data objects are assigned to the closest clusters. The cluster center is represented by the mean values of the data objects within the cluster. Therefore, every time that the cluster centroid is being updated, a new data object becomes a member of a cluster. This process is repeated until no change can occur. Figure 2 [4], summarizes the convergence of the K-means clustering algorithm.

3) Density-Based Clustering: Clusters have various sizes and shapes. Clustering based on the similarity distance between the data objects, results in only spherical shaped objects. To find clusters with complex shapes, requires a more comprehensive method than partitional clustering methods. DBC methods have been developed to find the clusters with arbitrary shapes. Such methods use connectivity and density

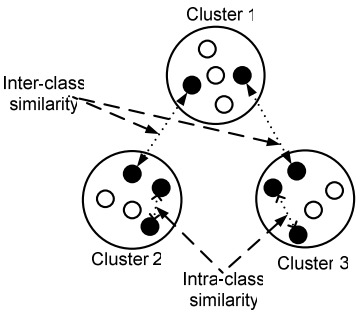

Figure 1. Intra and inter cluster similarity

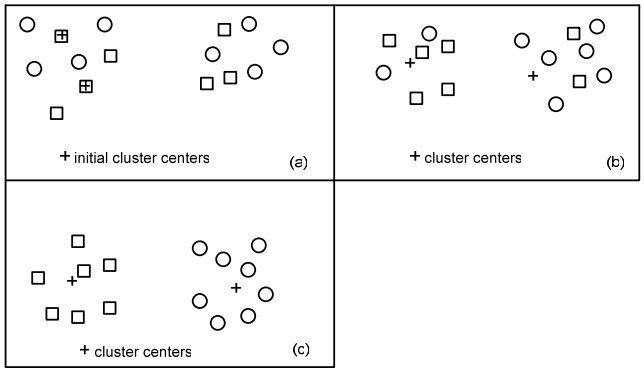

Figure 2. Convergence of K-means partitional clustering: (a) first iteration; (b) second iteration; (c) third iteration

functions to find arbitrary shape clusters [16]. In the data space, these methods consider clusters as dense regions of data objects which are separated by low density regions [20].

IV. SIMULATIONS

In this paper, we are comparing the accuracy performance of SLC, CLC, ALC, and K-means clustering algorithms with WIBL for user profiling. The following two sections provide detailed information about the dataset for the simulations and the results of these simulations.

A. Dataset

For the simulations, the dataset used was provided in [21], named 'Adult Data Set'. This dataset was created by Becker via extracting information from the 1994 census dataset and denoted to UCI (University of California, Irvine) Machine Learning Repository [21] by Kohavi and Becker for data mining applications. In this dataset, the demographic information of 303894 users is listed. 2000 selected users were adopted from this dataset. Some of the demographic information has been discarded and new information has been added to create a complete dataset of user profiles for the simulations.

In this study, each user is represented with three sets of profile information; demographic, interest and preference data. These profiles include information such as Age, Annual

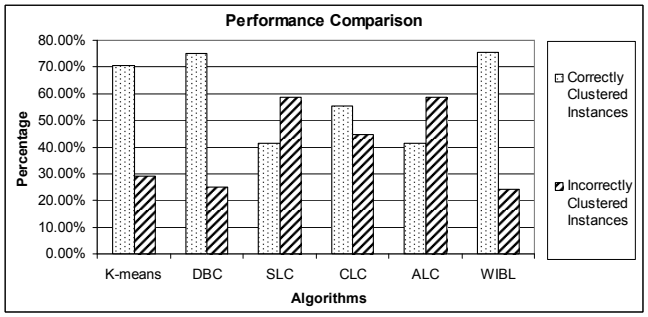

Figure 3. Performance comparison of the algorithms

Table I
PERFORMANCE COMPARISON OF THE ALGORITHMS ON USER PROFILE DATASET WHERE N=1000 AND M=1000

Algorithms	Correctly Clustered	Incorrectly Clustered
K-means	707(70.7%)	293(29.3%)
DBC	751(75.1%)	249(24.9%)
SLC	413(41.3%)	587(58.7%)
CLC	552(55.2%)	448(44.8%)
ALC	414(41.4%)	586(58.6%)
WIBL	756(75.6%)	244(24.4%)

Income, Sex, Sport Interest, Music Interest, Book Interest, Marital Status, Employment, Education and Profession. Simulations were carried out with two sets of datasets, which were training and test datasets. Both datasets have been selected from the complete user profile dataset and both has 1000 instances and 15 (A=15) attributes respectively. It is also worth mentioning that the content of both datasets are different from the ones which were used in [8].

Clustering algorithms were tested on the WEKA machine learning platform providing a benchmarking, consisting of a collection of popular existing learning schemes that can be used for practical data mining and machine learning applications [22].

B. Simulation Results

This subsection discusses the simulation results of SLC, CLC, ALC, K-means and WIBL, conducted on the above defined user profile dataset. The simulation parameters were set to be A = 15, Q = 5, N =1000 and M= 1000. Other simulation parameters (i.e., distance algorithm (Euclidean Distance)) were set as the default by WEKA, except the 'number of iterations' value for K-means, being taken as 7. Here, dissimilar training and test datasets have been used that includes information of different users. All algorithms were evaluated on the same training dataset and tested on the same test dataset, to obtain the classification/clustering accuracy results.

Table 1 and Figure 3 show the results of each simulation. From Figure 3, we can clearly see the performance comparison of the algorithms. Here, it can be seen that the lowest incorrectly clustered instance percentage, was archived by

 ISBN: 978-1-61208-311-7

WIBL and DBC. On the other hand, highest incorrectly clustered percentage is performed by the SLC and ALC hierarchical clustering algorithms. From the table, it can be seen that the best result is achieved by the WIBL algorithm, with 756 correctly clustered instances out of 1000 instances. The DBC follows the WIBL algorithm, with 751 correctly and 249 incorrectly clustered instances. The third best result is achieved by the K-means algorithm, with 707 correctly clustered instances. From Table 1, it can also be observed that the lowest performance was achieved by the SLC algorithm. The SLC clustered 413 instances correctly. With 414 correctly clustered instances, the ALC performs the second lowest result. Here, simulations revealed that the SLC and ALC algorithms perform similar, with user profile dataset, by clustering more than half of the instances incorrectly. The CLC algorithm performs better than the SLC and ALC algorithms by clustering 552 instances correctly.

In general, simulations showed that hierarchical clustering algorithms do not perform very well with user profiles. On the other hand, the DBC, with arbitrary clusters, gives one of the best results with user profiles. Moreover, using feature weighting to emphasize the relevancy of features during user profiling, has resulted in the WIBL achieving the highest performance among all the used algorithms.

V. Conclusion and Future Works

This paper aimed to evaluate the performance of the Weighted Instance Based Learner (WIBL) together with the well known clustering algorithms on a user profile dataset. The simulations were conducted on user profile dataset that reflects the user's demographic, interest and preferences information. Two sets of user profile dataset, training and test datasets, were used for the simulations. Here, all algorithms were trained on the same training dataset and tested on the same test dataset. According to the simulation results, Single-Linkage Clustering (SLC) has the lowest performance. The best performance, on the other hand, is achieved by the WIBL. The WIBL algorithm outperformed all the algorithms by classifying 75.6% of the instances correctly. Hence, it can be conclude that, compared to the well known clustering algorithms, with WIBL we can achieve the highest accuracy in user profiling.

This work is the first in the literature to present the comparison of classification accuracy of the WIBL and well known clustering algorithms with user profiles. Future studies could compare the performance of WIBL with well known classifiers. It would also be interesting to test and evaluate the performance of these algorithms with different real word user profile dataset.

References

[1] G. Araniti, P. D. Meo, A. Iera, and D. Ursino, "Adaptive controlling the QoS of multimedia wireless applications through user profiling techniques", IEEE Journal on Selected Areas in Communication, 21(10), December 2003, pp. 1546-1556.

[2] S. Henczel, "Creating user profiles to improve information quality", Factiva, 23(3), May 2004, p. 30.

[3] D. Poo, B. Chng, and J. M. Coh, "A hybrid approach for user profiling", Annual Hawaii International Conference on System Sciences, 4(4), January 2003, pp. 1-9.

[4] A.K. Jain and R.C. Dubes, "Algorithms for clustering data" , 1st Edition, Prentice-Hall Advanced Reference Series, Prentice Hall, Inc., Upper Saddle River, NJ, 1998, pp. 1-304.

[5] D. Irani, S. Webb, and C. Pu, "Study of static classification of social spam profiles in MySpace", International Conference on Weblogs and Social Media, May 2010, pp. 82-89.

[6] W. Paireekreng and K. W. Wong, "Client-side mobile user profile for content management using data mining techniques", International Symposium on Natural Language Processing, October 2009, pp. 96-100.

[7] W. Paireekreng, K. W. Wong, and C. C. Fung, "A model for mobile content filtering on non-interactive recommendation systems", IEEE International Conference on Systems, Man and Cybernetics, October 2011, pp. 2822-2827.

[8] A. Cufoglu, M. Lohi, and C. Everiss, "Weighted Instance Based Learner (WIBL) for user profiling", IEEE International Symposium on Applied Machine Intelligence and Informatics, January 2012, pp. 201-205.

[9] M. R. Lopez, A. B. B. Martinez, A. Peleteiro, F. A. M. Fonte, and J. C. Burguillo, "MoreTourism:mobile recommendations for tourism", IEEE International Conference on Consumer Electronics, January 2011, pp. 347-348.

[10] Y. B. Fernandez, M. L. Nores, J. J. P. Arias, J. G. Duque, and M.I.M. Vicente, "TripFromTV+:Exploiting social networks to arrange cut-price touristic packages", IEEE International Conference on Costumer Electronics, January 2011, pp. 223-224.

[11] K. Lakiotaki, N. F. Matsatsinis, and A. Tsoukias, "Multicriteria user modelling in recommender systems", IEEE Intelligence Systems, 26 (2), April 2011, pp. 64-76.

[12] A. B. B. Martinez et. al., "A hybrid content-based and item-based collaborative filtering to recommend TV programs enhanced with singular value decomposition", Elsevier Information Sciences, 180(22), November 2010, pp. 4290-4311.

[13] D. Xu, Z. Wang, Y. Zhang, and P. Zong, "A Collaborative tag recommendation based on user profile", International Conference on Intelligent Human-Machine Systems and Cybernetics, August 2012, pp. 331-334.

[14] D. W. Aha, D. Kibler, and M. K. Albert, Instance-based learning algorithms, Machine Learning journal, 1(6), 1991, pp. 37-66.

[15] M. J. A. Berry and G. S. Linoff, "Data mining techniques: for marketing, sales and customer relationship management", 2nd Edition, Electron. Book, John Wiley and Sons, Inc. (US), 2004, pp. 11, 165-167, [Retrieved: June, 2013].

[16] M. Sushmita and A. Tinku, "Data Mining: Multimedia, soft computing and bioinformatics", Electron. Book, Hoboken, John Wiley and Sons, Inc.(US), 2003, pp. 18-19, [Retrieved: July, 2013].

[17] A. L. Symeanidis and P. A. Mitkas, "Agent intelligence through data mining multiagent systems, artificial societies and simulated organizations; 14", Electron. Book, New York: Springer Science and Business Media, 2005, pp. 21-23, 27-28, [Retrieved: August, 2013].

 ISBN: 978-1-61208-311-7

[18] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, “Cluster Analysis”, 5th Edition, John Wiley and Sons, Ltd.(London), 2011, pp. 73-78.

[19] M. Khosrowpour, “Encyclopaedia of information science and technology”, Electron. Book, Hershey, PA Idea Group Reference, 2005, pp. 2063- 2067, [Retrieved: August, 2013].

[20] J. Wang, “Encyclopaedia of data warehousing and mining”, Electron. Book, Hershey, PA Information Science Reference, 2006, p. 144, [Retrieved: June, 2013].

[21] A. Asuncian and D. J. Newman, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, 2007, [Retrieved: July, 2013].

[22] H. Mark et al., “The WEKA data mining software: an update”, SIGKDD Explorations, 11(1), June 2009, pp. 10-18.

 ISBN: 978-1-61208-311-7