An Experimental Comparative Study of Web Mining Methods for Recommender Systems

SADDYS SEGRERA AND MARÍA N. MORENO
Department of Computing and Automatic
University of Salamanca
Plaza de la Merced s/n, 37008, Salamanca
SPAIN
saddys@usal.es, mmg@usal.es http://web.usal.es/~mmg

Abstract: - An essential goal of the present web engineering is the development of efficient and competitive applications. This objective can be achieved by building recommender systems endowed with suitable web mining algorithms. Multiclassifiers are reliable data mining models that have been hardly used in the web system area. The paper presents a comparative study among different simple classifiers and multiclassifiers using a dataset from MovieLens recommender system. The aim of the work is to identify when the use of multiclassifiers in this type of systems is efficient

Key-Words: - multiclassifiers, web mining, web engineering, recommender system

1 Introduction

The amount of available information in the Web, as a result of the increasing of the electronic business activities, is greater than a consumer can manage. In order to be competitive, E-commerce systems need to provide users with mechanisms for selective retrieval of web information. A way to obtain new customers and retain existing ones is the personalized product recommendation. E-commerce applications that incorporate recommender systems provide users with intelligent mechanisms to search products to purchase.

Data mining provides a number of algorithms to obtain profiles of users based on historical data, which are used to predict the preferences of new users. The process of applying data mining techniques on web data in order to obtain customer usage patterns is known as web mining. The predictive models induced by these algorithms are named classifiers. However, data from web environments used for building the models are heterogeneous, and the behaviour of classifiers in an individual way sometimes fails with some training sets, when a wide variety of data exist. Therefore, the use of multiclassifiers frequently is more feasible.

The multiclassifiers are the result of combining several individual classifiers. The methods for building multiclassifiers are divided in two groups: ensemble and hybrid methods. The first methods, such as Bagging [3] and Boosting [8], induce models that merge classifiers with the same learning algorithm, but introducing modifications in the training data set. The second type methods, such as Stacking [35], create new hybrid learning techniques from different base learning algorithms. The architectures and main methods of multiclassifiers were described in a previous study [31]. It was demonstrated by a case study their capacity to increase the precision in relation to the individual classifiers that compose them. This goal will be reached only if the individual classifiers are precise and diverse [11]. That is, a classifier is considered precise when its error is lower than 0.5 or its error is lower than the one obtained choosing a class arbitrarily. Two or more classifiers are diverse when their output errors are not correlated.

Most of the researches in web mining about the use of multiclassifiers are guided to the text mining and framed in the web content mining. At the present time text categorization has been topic of many researches, where the use of different algorithms is described. Many studies concerning the state of the art about classification algorithms have been developed, such as: support vector machine, nearest neighbour and neural networks [36], [32]. Also, papers that propose the multi-strategic learning or combination of classifiers as [29] have been studied; with the purpose of combining the advantages of different classification focuses and then, to increase the general precision.

Multiclassifiers also, are used in web usage mining [18], [2], [23]. The recommender systems apply personalization to the website and they use classification tasks. We have used data from MovieLens Recommender System [17] to make some experiments and analyze the behaviour of some classifiers with different learning techniques, individual classifiers and combination of them.

This paper is organized as follows: a brief description of multiclassifier application in text mining and web usage mining is given in section 2. Section 3 includes general aspects of the recommender systems. The evaluation of different classifiers with a dataset from MovieLens is presented in section 4. Finally, the conclusions are summarized in section 5.

2 Multiclassifiers in Web Mining

There are many researches in web mining where the multiclassifiers are applied. Most of them are related with text mining. Text mining is about looking for patterns in natural language text, and may be defined as the process of analyzing text to extract information from it for particular purposes. Text mining recognizes that complete understanding of natural language text, a long-standing goal of computer science, is not immediately attainable and focuses on extracting a small amount of information from text with high reliability [1].

Also, Web Usage Mining (WUM) has researches in this area. While content mining and structure mining utilize the real or primary data on the web, usage mining mines secondary data generated by the user's interaction with the web. Web usage data includes data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse-clicks and scrolls, and any other data generated by the interaction of users and the web. WUM works on user profiles, user access patterns, and mining navigation paths which are being heavily used by e-commerce companies for tracking customer behaviour on their sites [24].

2.1 Text Mining

Best Overall Results Generator System was proposed in [36]. It combines linear classification methods using the same weight for each individual classifier in the topic discovery. Some of the used methods were: Rocchio [28], nearest neighbour and language modeling. The increment yielded in the context of a text categorization problem was demonstrated in [15] by the use of a new query formulation and weighting methods combining three independent classifiers (nearest neighbour, relevance feedback and Bayesian classifier). Several researches as [12] examined different combination strategies in the context of documents filtering with learning algorithms as Rocchio, nearest neighbour, linear discriminant analysis and neuronal network. The evaluation of vote and meta-learning in partitioned data by inductive learning was presented in [4]. The effectiveness of Stacking generalization method to combine different types of learning algorithms was verified in [33]. Recently, a new method for web page classification that uses unlabeled data was presented in [25]. The learning method proposed first, trains a classifier with a small labeled training dataset. Later, a series of classifiers is built sequentially with unlabeled data.

The authors in [38] proposed an algorithm named tri-training, which approaches the problem to determine how to label the unlabeled examples and how to produce the final hypothesis. On the other hand, a better capacity of generalization is reached when three classifiers are combined. Let L denote the labeled example set and U denote the unlabeled example. In order to determine which example in U should be labeled and which classifier should be biased in prediction, the confidence of the labeling of each classifier must be explicitly measured. Assume that besides these two classifiers, i.e., h_1 and h_2, a classifier h_3 is initially trained from L. Then, for any classifier, an unlabeled example can be labeled for it as long as the other two classifiers agree on the labeling of this example, while the confidence of the labeling of the classifiers are not needed to be explicitly measure. For instance, if h_2 and h_3 agree on the labeling of an example x in U, then x can be labeled for h_1. In detail, the initial classifiers are trained from data sets generated via bootstrap sampling from the original labeled example set. These classifiers are then refined in the tri-training process, and the final hypothesis is produced via majority voting. The generation of the initial classifiers is similar to train an ensemble algorithm as Bagging from the labeled example set.

The taxonomies are used in the semantic web. A taxonomy, or directory or catalog, is a division of a set of objects (documents, images, products, goods, services, etc.) into a set of categories. There are a tremendous number of taxonomies on the web, and often it is necessary to integrate objects from various taxonomies into a master taxonomy. In [37] the Co-Bootstrapping technique is described, where a Boosting algorithm and the support vector machine are improved for the taxonomy integration.

2.2 Web Usage Mining

An empirical evaluation of classifier combination schemes for predicting user navigational behaviour was presented in [20]. The first one is built using decision trees for the whole data set, with the aim of studying user profiles and variable importance, while the second one combines simple classifiers based on small decision trees using a combination of the voting [14] and Cascading [9] paradigms, in order to make predictions which evolve during the period of time the website is collecting data. Results show that it is possible to extract useful information for studying user profiles and for predicting user behaviour using small decision trees.

An approach for classifying students was presented in [18] in order to predict their final grade based on features extracted from logged data in an education web-based system. A combination of multiple classifiers leads to a significant improvement in classification performance. A genetic algorithm was used to optimize the prediction accuracy. The classifiers included learning techniques such as: Quadratic Bayesian classifier, nearest neighbour, Parzen-window [26], multilayer perceptron and decision tree.

News Dude system [2], that reads new stories to the user, presents an automated induction to the user preferences and interests. The induction of user models consists of separate models for long-term and short-term. The nearest neighbour algorithm is used to assist to the short-term interests and a Bayesian classifier for the long-term interests. The approach that combines the predictions of multiple user models in [23] consists of leaning a set of referees, one for each prediction model, which characterizes the situations in which each of models is able to make correct predictions.

3 Recommender Systems

Recommender systems are directly related to the personalization on the website and the development of electronic commerce. Personalization includes a series of fundamental and interdependent processes [10]:

- User data acquisition: It is necessary to extend and use the contained information in the log files of the websites in order to enrich data about the user interaction.
- Model building: this refers to extend the information and the techniques to model building that support the anticipated adaptation tasks for the systems.
- Identification of adaptation tasks: It is related with the built models and the definition of adaptive tasks, it identifies the type of help that can be of utility for its accomplishment, for each cooperative learning task.

A way to build a recommender system using a classifier would be by the use of the information about a product and a consumer, also input data and to make that the output categories represent in what degree can be recommend the product to the client. The classifiers are implemented by different learning techniques. In spite of the advantages that the use of multiclassifiers offers, researches where multiclassifiers are applied to recommender system does not proliferate.

There are two main approaches, *memory-based* (*user-based)* and *model-based* (*item-based)* algorithms. ***Memory-based*** algorithms, also known as *nearest-neighbor* methods, were the earliest used [27]. They treat all user items by means of statistical techniques in order to find users with similar preferences (*neighbors*). The advantage of these algorithms is the quick incorporation of the most recent information, although the search for neighbors in large databases is slow [30].

Data mining technologies have also been applied to recommender systems. ***Model-based*** algorithms use these methods in the development of a model of user ratings. Some examples of these methods are the Bayesian network analysis [27], the latent class model [5], rule-based approaches, [16], association analysis [21] [22], decision tree induction combined with association rules [6], horting [34]. Web mining methods build models based mainly on users' behaviour more than in subjective valuations (ratings). The models are induced off-line, which allows a low user response time. This is the main advantage of this approach that avoids problems associated with traditional memory-based techniques [19].

Multiclasifiers belong to the last group of methods, however, in spite of the advantages that the use of multiclassifiers offers, researches where multiclassifiers are applied to recommender system does not proliferate.

In most cases, recommendation problems in e-commerce can be classified according to (1) whether customers for whom make recommendations, (2) whether the objective of recommendations is to predict how much a particular customer will like a particular product, or to identify a list of products that will be of interest to a given customer, and (3) whether the recommendation is accomplished at a specific time or persistently [13].

4 Case Study

MovieLens is a movie recommender system available in Internet based on GroupLens technology. It is an experimental data source. Actually, two datasets are in the GroupLens official website. One of them has been selected for this study in order to predict the ratings of movies. Finally, in this research only 1240 records were processed, due to many of them were eliminated in the pre-processing phase. Data appear in different files and some operations were necessary to merge all the content in only one file for introducing it, in the computer tools Mineset of Silicon Graphics, Inc. and WEKA from University of Waikato.

The ratings are the five values of an attribute, from 1 to 5, it is the opinion that users have about movies, where 1 means the lower rating or preference and 5 represents, the maximum. Each user has registered its gender, age, occupation and zip code. The attributes about movies are: title, release date, video release date and other 19 dedicated to each possible movie gender or category (unknown, action, adventure, animation, children, comedy, crime, documentary, drama, fantasy, film-noir, horror, musical, mystery, romance, science-fiction, thriller, war and western). These last features get value 1, if the movie belongs to a specific gender and 0 otherwise. This means a movie can belong to different film genders. The variable related to the video release date has been excluded directly because there is not registered information. There is also, the attribute timestamp that refers to the moment, the user made the rating.

Before making the attributes analysis, which are used to build the model, noise in data could be inferred. It does not understand that exist ratings over 2 when their title and gender movie are unknown. Also, there are other records where the user zip code does not appear or these data have less than five digits.

As for the variable to predict in the study "rating", necessary transformations were made to convert it in the variable "recom" with only two values: "Not recommended" for the values 1 and 2 of "rating", and "Yes recommended" for the values 3, 4 and 5. These changes are produced to simplify the problem, because the important is to determine if the movie has high ratings or not. Also, the number of the variables is high (22).

The behaviour of Bagging, Boosting and Stacking with different learning techniques was analyzed in order to determine if multiclassifiers could be used in the movie recommendation.

In WEKA for this case study, the individual classifiers by different algorithms showed high precisions (Fig. 1). Hence, could not be justified the use of multiclassifiers, that increase the model building and evaluation time only to overcome in hundredth to most of the individual classifiers.

Building and evaluation times of the individual algorithms are short in relation to Bagging and Boosting showed (Table 1). The last one increased its execution time significantly using nearest neighbour learning. Following the analysis of Fig. 1, we can discard the nearest neighbour individual classifier, which presented the lower precision value in comparison with the Bayesian learning method and decision tree. Moreover, the ensemble methods Bagging and Boosting that used nearest neighbour can be excluded to recommend movies.

However, multiclassifiers with decision trees improved the precision considerably. As much Bagging as Boosting increased the precision in relation to the individual method. Contrarily to the case study about land evaluation described recently in [31], the current data present noise, because as it was explained previously, there are irregularities due probably to deviations respect the real data, when the suppliers of the website make the first data cleaning and pre-processing. In this study, Bagging with decision trees demonstrated to have a better performance than Boosting (AdaBoost), which ratifies the studies made by [7].

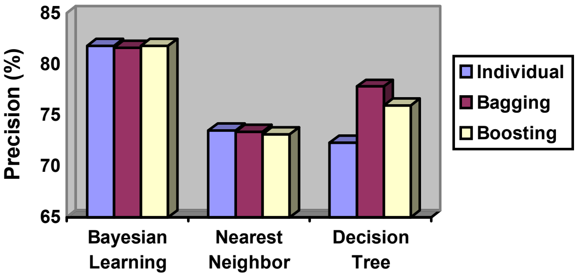

Fig. 1. Precision values for simple classifiers, Bagging and Boosting.

On the other hand, the hybrid method Stacking increased in 8% (80.08%) the precision achieved by the individual classifier (72.34%) with decision tree. Two decision trees (RandomTree and J48) were used as base classifiers to build Stacking and the meta-classifier applied was a nearest neighbor algorithm (IB1).

Table 1. Classifier building and evaluation times.

	Simple Classifier	Bagging	Boosting
Bayesian learning (BayesNet)	3	7	10
Nearest neighbor (IB1)	8	24	1354
Decision tree (RandomTree)	8	8	24

In spite of the precision increment by Stacking, the obtained precision by the individual Bayesian classifier was not surpassed. In addition, the execution time of this model was the lowest, only 3 seconds, reason why for this case study the use of multiclassifiers based on Bayesian learning is not recommended.

Multiclassifiers are sensitive to the data quality from the web. Its application in recommender systems must be considered if the employed time in the model building is not prolonged, since for this type of system the immediacy is one of the main factors to consider as indispensable requirement.

5 Conclusions

The use of multiclassifiers in web mining is more limited than in the traditional data mining. The combination of classifiers on the web is more frequent in the web content mining area, although it has been also applied in the user behaviour prediction and to study the evolution of such predictions.

The case study corroborated if data present noise, Bagging shows its superiority with respect to Boosting, when they are built with decision trees. Also, it was confirmed it is possible to build a hybrid method such as Stacking that, is even better than the two methods mentioned before.

The use of multiclassifiers in recommender systems must be justified with a very significant precision increment in comparison with the individual classifiers, but, in general, the multiclassifiers take more time in their execution. The priority of systems on the web is to give the precise answer, but also, to do it quickly. This fact does not limit the applicability of the classifiers in not very changing environments, when the recommendation models are built of line due to they are valid for long time.

References:

[1] Bainbridge, D., Frank E., Mahoui, M., Pfahringer, B., Wen, Y., Witten, I.H., Yeates, S., Text Mining http://www.cs.waikato.ac.nz/~nzdl/textmining, 2002.

[2] Billsus, D., Pazzani, M., A Hybrid User Model for News Story Classification, *Proceedings of the Seventh International Conference on User Modeling*, Banff, Canada, June 20-24, 1999, pp. 99-108.

[3] Breiman, L., Bagging predictors, *Machine Learning*, Vol. 24, No.2, 1996, pp. 123-140.

[4] Chan, P., Stolfo, J., Comparative evaluation of voting and metalearning on partitioned data. *Proceedings of the International Conference on Machine Learning (ICML '95)*, 1995, pp. 90-98.

[5] Cheung, K.W., Kwok, J.T., Law, M.H., Tsui, K.C., Mining customer product ratings for personalized marketing, *Decision Support Systems*, Vol. 35, 2003, pp. 231-243.

[6] Cho, H.C., Kim, J.K., Kim, S.H., A personalized recommender system based on web usage mining and decision tree induction, *Expert Systems with App.*, Vol. 23, 2002, pp. 329-342.

[7] Dietterich, T.G., An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, *Machine Learning*, Vol. 40, No.2, 2000, pp. 139-157.

[8] Freund, Y., Schapire, R.E., Experiments with a new boosting algorithm. *Proceedings of the 13th International Conference on Machine Learning*, 1996, pp. 148-156.

[9] Gama, J., Brazdil, P., Cascade Generalization, *Machine Learning*, Vol. 41, No.3, 2000, pp. 315-343.

[10] Gaudioso, E., Contribuciones al Modelado del Usuario en Entornos Adaptativos de Aprendizaje y Colaboración a través de Internet mediante técnicas de Aprendizaje Automático. Tesis Doctoral. Dpto. de Inteligencia Artificial, Facultad de Ciencias, Universidad Nacional de Educación a Distancia, Madrid, 2002.

[11] Hansen, L.K., Salamon, P., Neural network ensembles, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No.10, 1990, pp. 993-1001.

[12] Hull, D.A., Pedersen, J.O., Hinrich, S., Method combination for document filtering. *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, ACM Press, New York, US, 1996, pp. 279-288.

[13] Kim, J.K., Cho, Y.H., Kim, W.J., Kim, J.R., Suh, J.H., A personalized recommendation procedure for Internet shopping support, *Electronic Commerce Research and Applications*, Vol. 1, 2002, pp. 301-313.

[14] Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., On combining classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 3, No.3 March 1998, pp. 226-239.

[15] Larkey, L., Croft, W., Combining classifiers in text categorization, *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, ACM Press, New York, US, 1996, pp. 289-297.

[16] Lee, CH., Kim, Y.H., Rhee, P.K., Web personalization expert with combining collaborative filtering and association rule mining technique, *Expert Systems with Applications*, Vol. 21, 2001, pp. 131-137.

[17] Miller, B.N., Albert, I., Lam, SK., Konstan, J.A., Riedl, J., MovieLens Unplugged: Experiences with an Occasionally Connected Recommender System. *Proceedings of ACM 2003 International Conference on Intelligent User Interfaces*, January 2003.

[18] Minaei-Bidgoli, B., Punch, W.F., Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System. *GECCO'2003 Genetic and Evolutionary Computation Conference*, Springer-Verlag, Chicago, IL, July 2003, pp. 2252-2263.

[19] Mobasher, B., Cooley, R. and Srivastava, J., Automatic personalization based on web usage mining, *Communications of the ACM*, Vol. 43, No.8, 2000, pp. 142-151.

[20] Mor, E., Minguillón, J., An empirical evaluation of classifier combination schemes for predicting user navigational behavior. *Proceedings of the International Conference on Information Technology: Computers and Communications, ITCC'03*, 2003, pp. 467- 471.

[21] Moreno, M.N., García, F.J., Polo, M.J., López, V., Using Association Analysis of Web Data in Recommender Systems, *Lectures Notes in Computer Science*, LNCS 3182, 2004a, pp. 11-20.

[22] Moreno, M.N., García, F.J., Polo, M.J., An Architecture for Personalized Systems Based on Web Mining Agents, *Lectures Notes in Computer Science*, LNCS 3140, 2004b, pp. 563-567.

[23] Ortega, J., Exploiting multiple existing models and learning algorithms, *Working Notes of the AAAI Workshop on Integrating Multiple Learned Models*, 1996, pp. 101-106.

[24] Pal, S.K., Talwar, V., Mitra, P., Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions, *IEEE Transactions on Neural Networks*, Vol. 13, No.5, September 2002, pp. 1163-1177.

[25] Park, S.B., Zhang, B.T., Automatic Webpage Classification Enhanced by Unlabeled Data, *Proceedings of the 4th International Conference on Intelligent Data Engineering and Automated Learning*, Lecture Notes in Computer Science, Vol. 2690, 2003, pp. 821-825.

[26] Parzen, E., On the estimation of a probability density function and mode, *Annals of Mathematical Statistics*, Vol. 33, 1962, pp. 1065-1076.

[27] Resnick, P., Iacovou, N., Suchack, M., Bergstrom, P., Riedl, J., Grouplens: An open architecture for collaborative filtering of netnews. *Proceedings of ACM CSW'94 Conference on Computer. Supported Cooperative Work*, 1994, pp. 175-186.

[28] Rocchio, J.J., Relevance feedback in information retrieval. The SMART Retrieval System Experiments in Automatic Document Processing, Prentice Hall, 1971, pp. 313-323.

[29] Saleeb, H., Information Retrieval: A Framework for Recommending Text-based Classification Algorithms. Doctor of Professional Studies, Pace University, June 2002.

[30] Schafer, J.B., Konstant, J.A., Riedl, J., E-commerce recommendation applications, *Data Mining and Knowledge Discovery*, Vol. 5, 2001, pp. 115-153.

[31] Segrera, S., Moreno, M.N., Multiclasificadores: Métodos y Arquitecturas. Informe Técnico-Technical Report, DPTOIA-IT-2006-001, Departamento de Informática y Automática, Universidad de Salamanca, Mayo 2006.

[32] Thorsten, J., Text categorization with Support Vector Machines: Learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, 1398, Springer Verlag, Heidelberg, DE, 1998, pp. 137-142.

[33] Ting, K.M., Witten, I.H., Stacked Generalizations: When Does It work? *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI*, 1997, pp. 866-873.

[34] Wolf, J., Aggarwal, C. Wu, K.L., Yu, P., Horting hatches an egg. A new graph-theoretic approach to collaborative filtering. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, C.A., 1999.

[35] Wolpert, D.H., Stacked Generalization, *Neural Networks*, Vol. 5, 1992, pp. 241-259.

[36] Yang, Y., An evaluation of statistical approaches to text categorization, *Journal of Information Retrieval*, Vol. 1, 1999, pp. 69-90.

[37] Zhang, D., Lee, W.S., Learning to Integrate Web Taxonomies, *Journal of Web Semantics*, Vol. 2, No.2, 2004, pp. 131-151.

[38] Zhou, Z.H., Li, M., Tri-training: exploiting unlabeled data using three classifiers, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No.11, 2005, pp. 1529-1541.