In Anthony Jameson, Cécile Paris, and Carlo Tasso (Eds.), *User Modeling: Proceedings of the Sixth International Conference, UM97*. Vienna, New York: Springer Wien New York. Available on−line from http://um.org.

Symbolic Data Analysis With the K-Means Algorithm for User Profiling

Anne-Claude Doux[1]*, Jean-Philippe Laurent[1], and Jean-Pierre Nadal[2]

[1] Laboratoires d'Electronique Philips S.A.S., France
[2] Laboratoire de Physique Statistique, Ecole Normale Superieure, Paris, France

Abstract. We propose to simplify human-machine interaction by automating device settings that are normally made manually. We present here a classification scheme of user behaviours based on an adaptation of the K-means algorithm to symbolic data representing user behaviours. This classification enables a system to derive prototypical behaviours and to control device settings automatically.

1 Introduction

The general framework is the following: Some users, or agents, are involved in a particular task or activity, such as watching TV, where they are to set the parameters of a system according to their preferences. The chosen *actions* usually depend on external conditions, to be referred to as the *environment* (e.g., room lighting for TV picture contrast and brightness setting). Our goal is to characterize each user *behaviour* (set of pairs {environment, action}) in order to automatically generate the action which best matches the one the user would have chosen in a given environment, even if the user is not yet known.

We may presume that users will most often accept settings sufficiently close to those they would have chosen. Therefore, it should be possible to classify user behaviours, so that each user could be satisfied with the actions defined by one of the corresponding prototypes. We introduce a new method for performing behaviour classification, by means of a new generalization of the standard K-means algorithm to complex data. Our classification scheme rests on applications to real and simulated data for which we developed prototypes.

2 The Data: Symbolic Profiles of Users

Data are collected from dedicated experiments during which users are asked to give their preferences in various environmental conditions.

1. Real data set: Characteristics of 9 environments were specified. A panel of 120 users participated in the experiments, which gave rise to 38,000 observations, out of which 11 typical actions were extracted.
2. Simulated data set: Our set of simulated data, with a known underlying structure, includes 300 users. As for real data, it deals with 9 environments and 11 actions.

* We thank E. Diday and his colleagues from the LISE-CEREMADE group (University Paris-Dauphine) for fruitful discussions on data analysis.

We have to cope with different kinds of problems inherent to real data. Data are incomplete (many users are observed in only some of the predefined environments) and inconsistent (returning to the same environment, many users gave different preferences). From problems encountered with real users in a real application and thanks to symbolic data analysis (Diday, 1993), we build symbolic objects representing user behaviours (Polaillon et al., 1996). A user u is characterized by the set of probability rules (one per environment he had access to), which give the empirical probability that he chooses one action in a given environment.

3 Methodology of Classification With K-Means on Symbolic Objects

In order to define prototypical behaviours, so that any user is close to one of these, we propose classifying the user behaviours into K classes and computing K typical behaviours. We need an unsupervised algorithm (Duda and Hart, 1973) which can handle symbolic data, generate a representative for each class, provide the best intrinsic K, and have the ability to generalize to unknown users too (Jain and Dubes, 1988). We thus choose to adapt the K-means algorithm for these symbolic objects.

3.1 Two Main Adaptations of the K-Means Algorithm

We propose a training criterion $\mathcal{E}$ which favours homogeneous and compact classes to improve the assignment of the users and therefore to improve the quality of the partition. We perform its optimization with the K-means algorithm (Doux et al., 1996).

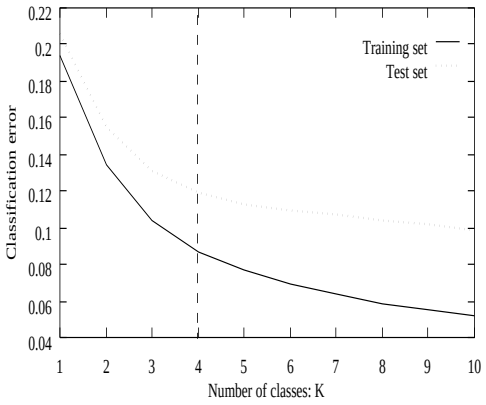
We then define and compute three kinds of representatives u_k^* for each K: the *center of gravity*, which is theorically the best choice; the *paragon* (i.e., the real recorded user who is nearest to the middle of the considered class), since in some applications this realistic behaviour is required; the *horde representative* (i.e., a hybrid user with real parts of behaviours belonging to different users of the class), which represents an intermediate solution between the two previous ones. In our real application, the horde is much better than the paragon representative (see Figure 1). In fact, because of missing data, some users cannot be compared with the paragon. This problem, however, does not occur with the simulated data, and all representatives then remain acceptable. It is therefore better to choose the paragon in that case, because it is a strong and realistic behaviour.

3.2 Methodology to Validate the Partition

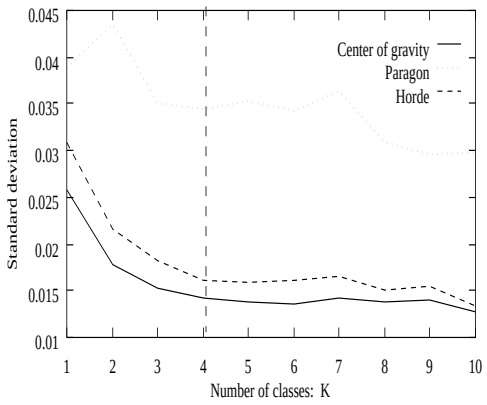
To check the partition's ability to generalize—that is, to match unknown user behaviours—we split the population into training and test sets. The training set is classified and provides a final partition with a number K of classes. Each user u of the corresponding test set is then assigned to a class. We finally define a generalization criterion $\mathcal{E}_{\mathcal{G}}$ (Doux et al., 1996) to measure the quality of an unknown user assignment.

3.3 Intrinsic Data Structure: The Best Number of Classes

Without knowledge of the intrinsic data structure, we have to find the best K, which is linked to the ability to generalize. The F-maximum criterion (Milligan and Cooper, 1985) provides the best



Classification error using the center of gravity



Standard deviation

Figure 1. Results on the real data set obtained by averaging over 100 random samples of the population (test, training), for each number of classes.

K on the well-structured data but fails on our real data. In our method, the standard deviation of the generalization criterion $\mathcal{E}_{\mathcal{G}}$ decreases when the number of classes increases. But in fact, when reaching the optimal K, $\mathcal{E}_g$ significantly decreases and then stabilizes with the addition of more classes. With the simulated data, we find the same K as computed by the F-maximum criterion. With the real data, our method still leads to a good partition (4 or 5 classes, see Figure 1). Since our application calls for the use of as few classes as possible, the 4-class partition is chosen.

4 Conclusion

The next step of this study is to investigate further the assignment scheme for new users when very little is known about them. Assessment of user dissatisfaction also remains a question. Furthermore, we can make profiles dynamically evolve. Some work is underway to improve the assignment of users through the symbolic interpretation of classes.

References

Diday, E. (1993). An introduction to symbolic data analysis. In *Proceedings of the 4th International Conference of the Federation of Classification Societies*. Paris. Springer Verlag.

Doux, A. C., Laurent, J. P., Nadal, J. P., and Diday, E. (1996). User profiling: Dynamic clustering on symbolic objects. Manuscript submitted for publication.

Duda, R. O., and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. NJ : Wiley.

Jain, A. K., and Dubes, R. C. (1988). *Algorithms for Clustering Data*. NJ : Prentice Hall.

Milligan, G. W., and Cooper, H. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50:159-179.

Polaillon, G., Getter-Summa, M., Pardoux, C., and Laurent, J. P. (1996). Approche numérique symbolique pour le codage et la classification des comportements d'utilisateurs. In *Proceedings of the 28th Days of Statistics*, 608-611.