Model-based clustering with Hidden Markov Models and its application to financial time-series data

Bernhard Knab[1], Alexander Schliep[2], Barthel Steckemetz[3], and Bernd Wichern[4]

[1] Bayer AG, D-51368 Leverkusen, Germany
[2] Department Computational Molecular Biology, Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin, Germany
[3] Science Factory GmbH, D-50667 Köln, Germany
[4] ifb AG, D-50667 Köln, Germany

Abstract. We have developed a method to partition a set of data into clusters by use of Hidden Markov Models. Given a number of clusters, each of which is represented by one Hidden Markov Model, an iterative procedure finds the combination of cluster models and an assignment of data points to cluster models which maximizes the joint likelihood of the clustering. To reflect the partially non-Markovian nature of the data we also extend classical Hidden Markov Models to use a non-homogeneous Markov chain, where the non-homogeneity is dependent not on the time of the observation but rather on a quantity derived from previous observations.

We present the method and an evaluation on simulated time-series and large data sets of financial time-series from the Public Saving and Loan Banks in Germany.

1 Introduction

Grouping of data, or clustering, is a fundamental task in data analysis. Methods for clustering have been widely investigated (Everitt, 1993) and can be coarsely categorized into two classes: distance- and model-based approaches. The former base the decision whether to group two data points on their distance, the latter assign a data point to a cluster represented by a particular statistical model based on its likelihood under the model.

Model-based clustering is better suited for time-series data (MacDonald and Zucchini, 1997). Usually, there is no natural distance function between time-series. Several non-critical variances of signals — a delay, an overall slower rate, a premature cutoff — will be overly emphasized by, say, Euclidean distance. Hence, capturing the essential *qualitative* behavior of time-series is difficult.

Using stochastic models to represent clusters changes the question at hand from how close two given data points are to how likely one particular data point is under the model. One can expect a larger robustness with respect to noise virtue of the stochastic model. As it is straight-forward to generate

artificial data given a model-based clustering, an analysis of the clustering quality based on the *predictive* performance of the inferred set of models becomes feasible.

Our approach of using Hidden Markov Models (HMMs) as clusters is motivated by the well known k-means algorithm (Everitt, 1993). Already Bock (1974) describes that an analogon of k-means can be used for model-based clustering. McLachlan and Basford (1988) implemented that for multivariate Gaussians. In the k-means algorithm the median represents a cluster and a clustering is computed by an iterative application of the following steps.

1. Assign each data point to its closest median, and
2. Re-compute the median for each of the clusters.

When going over to HMMs as cluster representatives two modifications are necessary. The criterion for the re-assignment of data points to clusters is maximization of the likelihood of the data points. The re-computation of clusters is done by training the cluster models with the Baum-Welch re-estimation algorithm (Baum and Petrie, 1966; Baum et al., 1970).

The savings and loan bank application we considered implied contractual constraints which violated the Markovian assumption inherent in *classical* HMMs. We accounted for these constraints by a model extension, which can be thought of as a HMM based on a non-homogeneous Markov chain. The non-homogeneity is not conditioned on the time of the observation but on a function summarizing and, hence, dependent on previous observations. This extension required only minor modifications to the relevant HMM algorithms. The clustering with the extended model provided a powerful modeling and analysis framework which improved the quality of the modeling substantially when compared with the methods previously used.

This paper is organized as follows: After establishing notation and necessary concepts in Sec. 2 we introduce the algorithm, analyze its computational complexity and discuss implementational questions in Sec. 3. The setting of the application problem and the data used in the experimental validation is subsequently described. This motivates the following extension to non-homogeneous HMMs introduced in Sec. 5. Experimental results and a discussion conclude the paper.

2 Definitions and Notation

Hidden Markov Models (HMMs) can be viewed as probabilistic functions of a Markov chain (Burke and Rosenblatt, 1958; Petrie, 1969), where each state of the chain independently can produce emissions according to so-called emission probabilities or densities. We shall restrict ourselves to univariate emission probability densities. Extensions to multivariates or mixtures thereof as well as discrete emissions are routine.

Definition 1 (Hidden Markov Model). Let $O = (O_1, \ldots)$ be a sequence over an alphabet Σ. The following parameters fully determine a Hidden Markov Model λ: the states S_i, $1 \leq i \leq N$, the probability of starting in state S_i, π_i, the transition probability from state S_i to S_j, a_{ij}, and $b_i(\omega)$, the emission probability density of a symbol $\omega \in \Sigma$ in state S_i.

The obvious stochasticity constraints on the parameters apply. Rabiner (1989) gives a thorough introduction to HMMs. The problem we will address can be formally defined as follows.

Definition 2 (HMM Cluster Problem). Given n sequences O^i, not necessarily of equal length, with index set $\mathcal{I} = \{1, 2, \ldots, n\}$ and a fixed integer $K \ll n$. Compute a partition $\mathcal{C} = (C_1, C_2, \ldots, C_K)$ of $\mathcal{I}$ and HMMs $\lambda_1, \ldots, \lambda_K$ maximizing the objective function

$$f(\mathcal{C}) = \prod_{k=1}^{K} \prod_{i \in C_k} L(O^i|\lambda_k)\,. \tag{1}$$

Here, $L(O^i|\lambda_k)$ denotes the likelihood function, that is, the probability density for generating sequence O^i by model λ_k: $L(O^i|\lambda_k) := P(O^i|\lambda_k)$.

It has been implicitly discovered before, (e.g. Smyth, 2000), that the problem of computing a k-means clustering can be formulated as a joint likelihood maximization problem.

3 The clustering algorithm

Adapting the k-means algorithm, we propose the following maximum likelihood approach to solve a HMM Cluster Problem, given K initial HMMs $\lambda_1^0, \ldots, \lambda_K^0$.

1. **Iteration** ($t \in \{1, 2, \ldots\}$)**:**
   (a) Generate a new partitioning of the sequences by assigning each sequence O^i to the model k for which the likelihood $L(O^i|\lambda_k^{t-1})$ is maximal.
   (b) Calculate new parameters $\lambda_1^t, \ldots, \lambda_K^t$ using the re-estimation algorithm for each model with their start parameters $\lambda_1^{t-1}, \ldots, \lambda_K^{t-1}$ and their assigned sequences.
2. **Stop**, if the improvement of the objective function (1) is below a given threshold, ε, the grouping of the sequences does not change or a given iteration number is reached.

As there is a one-to-one correspondence between clusters and models we shall use the terms interchangeably in the following.

Convergence: The nested iteration scheme does indeed converge to a local maximum. This follows directly from the convergence of the Baum-Welch algorithm and the observation that re-assignment of sequences cannot decrease the likelihood. How to avoid the usual practical problems with local maximization is described later.

Implementation: The relevant data structures and algorithms are freely available in a portable C-library, the GHMM (Knab et al., 2002), licensed under the Library GNU General Public License (LGPL).

Initialization: A suitable model topology, i.e. the number of states and the allowed transitions (the nonzero transition probabilities), and the number of initial models should be motivated by the application. Note that the topology remains unchanged during the training process.[1]

Since the clustering algorithm will only converge to a local maximum the choice of the model's start parameters will affect the maximum computed. The simplest approach is to set all parameter to random values subject to stochasticity constraints. This can easily lead to an unbalanced assignment of sequences to models, as random models might have near zero probabilities of producing any sequences in the set at all. Alternatively, one can initially train one HMM with all sequences, and subsequently use K copies of that model as the input for the clustering, after adding small random perturbations to the parameters of the K copies individually. Training enforces divergence of clusters in this case. Generally, one has to pay attention that in the first iteration step each sequence can be generated from at least one model — i.e., the likelihood of the set of sequences may not be zero. If there is only limited amount of training data available, pseudo-counts or Dirichlet priors (Sjolander et al., 1996) can be used to dispatch with this over-fitting problem effectively.

4 Application to loan bank data

To evaluate the proposed clustering method, we use financial time-series data obtained from the public saving and loan banks in Germany for an ongoing co-operative research project (Knab et al., 1997). The fundamental concept behind saving and loan banks is to combine a period of saving money, usually until some threshold D has been reached — the prerequisite for taking out a loan — which then has to be repaid in fixed installments. Contractual details vary widely, but manual inspection suggested a number of prototypical contract histories.

[1] However, the trained model may contain transitions with low probability and therefore some states may hardly ever be reached. A pruning step can eliminate these states.

Each of the data points corresponds to an individual saving and loan contract. It consists of a time-series of feature vectors recorded in yearly intervals. Depending on the respective bank, there are as many as 3 million data points available.

There are about 40 individual quantities recorded in one feature vector. Out of those, we mainly consider the relative savings amount (RSA). The RSA quantifies the amount of money saved over the last period of twelve months relative, in percent, to the total volume of the loan. It is the most important feature of the time-series, since it is the dominant factor for the further development of the contract. Other recorded quantities, except demographical data etc., depend on it directly or indirectly. Modeling all 40 quantities can be easily accommodated in the HMM-Clustering framework.

In the RSA time-series data a number of typical patterns can be observed, which correspond to different types of behavior. This motivates a theoretical interest in classifying and clustering this data. From a practical point of view, the clustering process is highly relevant as it is the first step towards simulation of the whole collection of contracts. Simulation is used for liquidity forecasting and hence as the basis for executive decisions such as investment strategies or contract design.

The observed time-series exhibit global patterns that correspond to certain deterministic constraints imposed by the terms and regulations of loan banking (e.g. the threshold D which specifies the end of the saving period). A good model generates sequences also obeying those constraints. In the next section we demonstrate how this non-Markovian behavior can be accounted for in HMM modeling.

5 Model extensions

The basic idea of our Model extension is to allow transition probabilities to vary, similarly to time inhomogeneous Markov chains. However, in our case the transition probabilities do not depend on time but on the partial sequence observed so far. As an example we consider a sequence of savings which, when summed, exceed the threshold D. Usually the sequence will enter a state corresponding to amortizations instead of remaining in a saving phase state in the next time step.

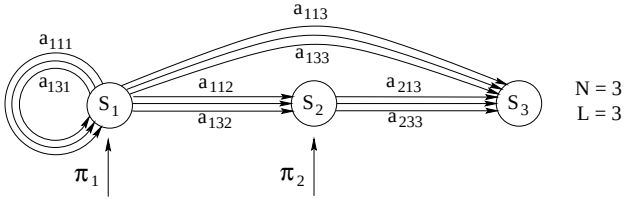

Fig. 1. Graph of an extended HMM with $L = 3$ conditional transition classes.

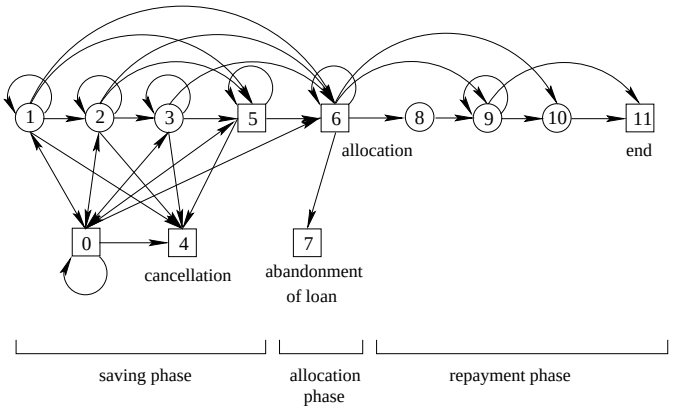

Fig. 2. Graph of an HMM for modeling the three phases of a loan banking contract.

To accomplish this for generated sequences we extend the transition Matrix $A = (a_{i,j})$ to a set of matrices, cf. Fig. 1, $A \to (A_1, \ldots, A_L)$. Suppose the model is in state i at time t and we already observed the partial sequence $(O_1, \ldots, O_t)$. The function $l = f(O_1, \ldots, O_t)$ determines the current transition matrix A_l. As a simple example we used the following step-function

$$f(O_1, \ldots, O_t) = \lceil L \sum_{\tau=1}^{t} O_\tau \rceil, \quad \text{where} \quad 0 \leq \sum_{\tau=1}^{t} O_\tau \leq 1.$$

Alternative choices for f have been developed by Wichern (2001). Knab (2000) shows how to modify the usual Baum-Welch reestimation formulas for this model extension.

6 Experimental Results

We tested our training algorithm with data sets containing up to 50,000 time series from savings and loan bank data. In a first step we restricted our model to the saving period. Several different model topologies with varying number of states were examined, as well as variations on the number of HMMs and transition matrices (not shown). We achieved optimal and stable results with a simple left-right model with $N = 13$ states and self-transitions ($a_{ij} = 0$ if $j < i$). The number of HMM clusters was $K = 9$ and the number of transition matrices was $L = 6$. This model parameters are used for the results in this section.

The first quantity to be examined was the sum of the relative savings amount (SRSA) per sequence. Fig. 3 displays the SRSA of the real data and the prediction of three different models: the currently used k-means model (Bachem et al., 1997), the naive HMM approach and our extended model of section 5. The SRSA of the real data is 0.0 until approximately

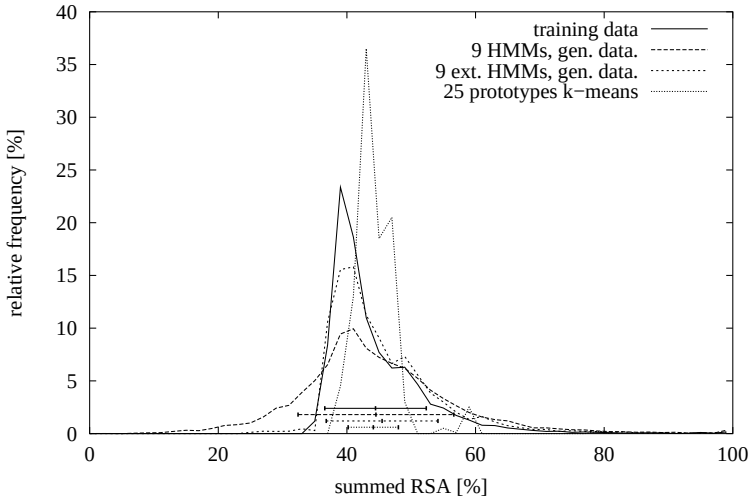

Fig. 3. Sum of relative savings amount (SRSA) for real data, generated data (HMMs and extended HMMs) and weighted k-means prototypes. The horizontal error bars show mean and standard deviation for the observed data.

37 %, reaches a sharp maximum at 39 % and has a long tail until 80 %. Note that the contracts require a savings amount of at least 40 % including interest. The k-means prediction has a much too sharp maximum and consequently a much too small variance due to the fixed time lengths of the k-means prototypes. The naive HMM approach achieves the maximum with high accuracy, but results in a much too broad length distribution. This can be avoided when using our extended model, where both the maximum and the variance are met. The trained models are distinguishable from each other and reflect certain well known and typical structures of saving behavior (Knab, 2000).

A more complex model: further events in loan banking

Fig. 2 shows one of the model topologies we investigated with regard to its capability in modeling the complete course of loan bank contracts. The three periods saving, allocation, and repayment correspond to three distinct groups of model states. The square states represent important discrete events such as canceling the contract; unique numerical values are chosen as almost-sure emissions. Emission probabilities of those states are excluded from training.

Furthermore, the emission parameters of these special states are never changed during training.

Another view is given in Fig. 4. Here the capability of extending truncated real sequences is displayed. The predicted data were generated by two different models (of same size and topology) which were trained on two different sets of sequences. The training sets are: *pred1* containing all contracts for the

year 1985 and *pred2* containing contracts regardless of the contract year. The truncated set consists of all contracts for the year 1986 and the sequences were truncated in 1992 and extended by the above mentioned models until a an end-state (e. g. a state with no outgoing transitions) was reached. Fig. 4 shows the yearly savings amount (YSA) and the yearly amortizations (YAM) summed over all sequences of the two generated sets (*pred1*, *pred2*) und of the real data (*real*, not truncated here). The YSA data is closely approximated by both predictions. For the YAM graph the prediction using the training set *pred1* is more accurate.

7 Conclusion and Outlook

We presented a new algorithm for clustering data, which performed well for the task of generating statistical models for prediction of loan bank customer collectives. The generated clusters represent groups of customers with similar behavior. The prediction quality exceeds the previously used k-means based approach.

HMMs lend themselves to various extensionsTherefore, we were able to incorporate many other relevant loan-bank parameters into our current model. These can then be estimated with *one* homogeneous statistical training algorithm, instead of using a collection of individual heuristics. We expect an even higher overall prediction accuracy and a further reduction of human intervention when applied to this and other application problems. Partial results are described elsewhere (Knab, 2000; Wichern, 2001). The clustering approach is general in its applicability: An analysis of gene expression time-series data from experimental genetics is forthcoming.

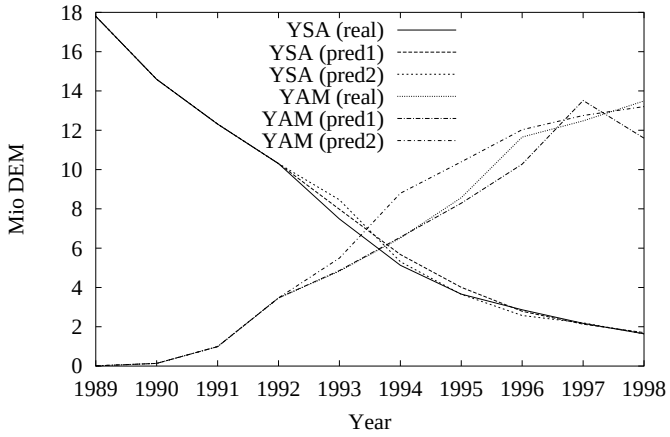

Fig. 4. Real and predicted yearly savings amount (YSA) and amortizations (YAM) for two different training scenarios. Prediction starts 1993.

Acknowledgments

The research was conducted at the Center for Applied Computer Science at the University of Cologne (ZAIK) and partially (BK, BW) funded by the German Landesbausparkassen. We would like to thank Prof. Dr. R. Schrader (ZAIK) for his support.

References

BACHEM, A. ET AL. (1997): Analyse großer Datenmengen und Clusteralgorithmen im Bausparwesen. In: C. Hipp, W. Eichhorn, W.-R., W.-R. Heilmann (eds.), *Beiträge zum 7. Symposium Geld, Finanzwirtschaft, Banken und Versicherungen, Dezember 1996*, no. 257, 955–961.

BAUM, L. E., PETRIE, T. (1966): Statistical inference for probabilistic functions of finite Markov chains. *Ann. Math. Statist.*, *37*, 1554–1563.

BAUM, L. E., PETRIE, T., SOULES, G., WEISS, N. (1970): A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, *41*, 164–171.

BOCK, H. H. (1974): Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten. Vandenhoeck & Ruprecht.

BURKE, C. J., ROSENBLATT, M. (1958): A Markovian function of a Markov chain. *Ann. math. stat.*, *29*, 1112–1120.

EVERITT, B. S. (1993): Cluster Analysis. Edward Arnold, London.

KNAB, B. (2000): Erweiterungen von Hidden-Markov-Modellen zur Analyse ökonomischer Zeitreihen. Ph.D. thesis.

KNAB, B., SCHLIEP, A., STECKEMETZ, B., WICHERN, B., GÄDKE, A., THORANSDOTTIR, D. (2002): The GNU Hidden Markov Model Library. Available from http://www.zpr.uni-koeln.de/ hmm.

KNAB, B., SCHRADER, R., WEBER, I., WEINBRECHT, K., WICHERN, B. (1997): Mesoskopisches Simulationsmodell zur Kollektivfortschreibung. Tech. Rep. ZPR97-295, Mathematisches Institut, Universität zu Köln.

MACDONALD, I. L., ZUCCHINI, W. (1997): Hidden Markov and other models for discrete-valued time series. Chapman & Hall, London.

MCLACHLAN, G., BASFORD, K. (1988): Mixture Models: Inference and Applications to Clustering. Marcel Dekker, Inc., New York, Basel.

PETRIE, T. (1969): Probabilistic functions of finite state Markov chains. *Ann. Math. Statist*, *40*, 97–115.

RABINER, L. R. (1989): A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*(2), 257–285.

SJOLANDER, K., KARPLUS, K., BROWN, M., HUGHEY, R., KROGH, A., MIAN, I. S., HAUSSLER, D. (1996): Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci*, *12*(4), 327–45.

SMYTH, P. (2000): A general probabilistic framework for clustering individuals. Tech. Rep. TR-00-09, University of California, Irvine.

WICHERN, B. (2001): Hidden-Markov-Modelle zur Analyse und Simulation von Finanzzeitreihen. Ph.D. thesis.