HIERARCHICAL BAYESIAN MODELING FOR CLUSTERING SPARSE SEQUENCES IN THE CONTEXT OF GROUP PROFILING

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes a hierarchical Bayesian model for clustering sparse sequences.This is a mixture model and does not need the data to be represented by a Gaussian mixture and that gives significant modelling freedom.It also generates a very interpretable profile for the discovered latent groups.The data that was used for the work have been contributed by a restaurant loyalty program company. The data is a collection of sparse sequences where each entry of each sequence is the number of user visits of one week to some restaurant. This algorithm successfully clustered the data and calculated the expected user affiliation in each cluster.

1 INTRODUCTION

This paper describes a Sequence Clustering algorithm which combines sequence modeling with Clustering to find latent groups.Additionally, the algorithm generates group/cluster profiles of the latent groups.This algorithm can be used to explore data which contains events that can be linked as a sequence. Some examples of the business data/scenario that could be modelled by this algorithm are:

Sequence of weekly visit-counts of users to a restaurant or shop collected for several weeks. Latent user groups can be identified and group profiles can be created by this algorithm.

Click stream data of user navigation . Latent user group identification and profiling is possible by this algorithm.

The order in which items are added in a shopping cart. User behavior modeling and group identification can be done. Also group user-profile creation is possible.

Gene(DNA-sequence) Clustering. Gene-Group profile discovery can be done via the algorithm.

Topic Modeling.Documents are sequence of words.

This paper describes a soft clustering technique for variable length sparse sequence of small positive numbers.A mixture model is built to describe the assignment of the sequences to the clusters.The modeling assumption is that the sequences are generated from mixture of distributions; the distributions being associated to different clusters and each sequence has a mixing proportion associated with each cluster-distribution. This technique is a static clustering technique,is assumed that this mixing proportion is constant for the whole sequence. This technique can be enhanced by assuming that each sequence entry has a time-stamp and at each time-stamp the mixing proportion might be correlated to its last value.

The existing Sequence Clustering Techniques follow variations of three basic ideas. In several bioinformatics papers on gene-sequence clustering is based on non sequence attributes, such as the length of the sequence and similarity of the entries are considered to be the main features based on which the clustering is done. This technique is not a model based technique. Also as variable length sequences are not feature vectors, the 'distance' between sequences is an ambiguous term, which should be defined clearly in context.Based on the idea of distance the 'similarity' between sequences is measured, thus this measurement needs to be defined too. The second technique is based on modeling sequences as Markov chains where the order of the chain determines number of past states affecting current state.As the order increases the cost of modeling also scales up in this

model and also this model does not support the idea of an unobserved variable which might control the generation of the sequence, which is often the case in practical scenario. The third technique is based on hidden Markov model where the observed outcome is a resultant of hidden latent state.This papers presents a hierarchical Bayesian model for soft clustering of the sequences. The proposed technique is ideologically closest to an HMM, since HMM can be considered a generalization of a mixture model where the hidden variables (or latent states), which control the mixture component to be selected for each observation, are related through a Markov process. This model however is more interpretable in terms of its parameters associated to each sequence and the clusters/groups of sequences.Thus the parameters and some quantities derived from the parameters can be mapped to real physical quantities and can be used to create group profiles.

1.1 Business Scenario

The data used in this paper is real industrial data from a restaurant loyalty program where by visiting the restaurant people can earn points or rewards.In a restaurant loyalty program users enroll, visit the restaurants using the program and after a period of time stop using the system and thus churn out of the system.It is important to study user behavior to predict his next visit, predict probability of churning out and study his buying behavior. Not only in case of individual users, but identifying the latent groups and studying user-group behavior for next-visit, churn-probability and buying behavior is also important for user segment targeting in a campaign.In this paper we have modelled user visit behavior.We have found latent groups among users through hierarchical Bayesian clustering.Also, we have generated user group profile as function of user parameters. The user visit-behavior is recorded as a sequence of number of visits to a restaurant every week. Hence all the users are represented by variable length sequences where every entry of a sequence is the number of visits of the corresponding user in that particular week. By hierarchical modeling we find every user's mean visits, every user-group's mean visits and average user participation in every group.The parameters associated to cluster, for example, mean visit frequency of a cluster, mean cluster membership proportion, is used for building a group profile.

2 Literature Review

In this paper we describe sequence modeling, latent group identification from user-visit sequence and clustering as a result of sequence modeling. Hence the Literature review section has been divided into three sub parts: Discovering groups among users, Clustering techniques in general and Sequence clustering.

2.1 Discovering Groups among Users

The existing literature related to user group profiling does not necessarily look for latent groups that naturally existed among the users, but somehow manages to put some users together as a group for eliciting a preference of the group of users for certain product or item given the data about individual user's preference for that product or item (Senot, 2010). For example in a rating based system, the objective of the existing literature is to generate a group rating for an item from individual ratings for that item. In the existing literature, the problem has been approached in two different ways. One way is aggregating the ratings of the users' group and generating a single rating for the group. (Masthoff, 2002) Another way is to create a representative user for the user group and predicting the rating by that representative user for the given item. (McCarthy, 2006) (OConnor, 2001)

The group of users thus selected by the existing literature lack certain characteristics. First, these people may not really be connected in any intrinsic manner, they are apparently just random people thrown together. For example, MUSICFX (McCarthy, 1998) selects music channels for the music to be played in a fitness center. Based on the preferences that have been previously specified by the members who are currently working out, the system chooses one of 91 possible music channels, including some randomness in the process.

INTRIGUE (Ardissono, 2003) recommends tourist attractions for heterogeneous groups of tourists that include relatively homogeneous subgroups (e.g. children).

However, the people working out on a particular day and time has no connection between their tastes in music! Similarly, random tourists making a tour plan have no inherent connection. Hence, they share no commonality that a group should share. In this paper, a group is a latent group, sharing similar behavior pattern. Latent groups are discovered by hierarchical Bayesian modeling of the user-visit data. Hence, they are more connected as a group and thus a more fit candidate for group profiling.

In case of a social network based user base, the existence of social groups is investigated (Richter Y, 2010) for estimating the probability of churning out of closely connected users from a loyalty program. The group among the users has been determined based on their social connected-ness. This process does use an intrinsic connection to define a group, but it is doubtful whether the churn behavior is at all impacted by social connection. In this paper,based on user behavior data we aim to find latent groups and to create a group user profile, which is a novel idea compared to these papers.

2.2 Clustering techniques

Clustering techniques can be divided into two groups based on their style of associating an item to a cluster: hard assignment and soft assignment.The algorithm proposed in this paper uses soft assignment technique. On the other hand K-means makes a hard assignment of the data points to the cluster depending on the euclidean distance. EM Clustering assumes that the data points are generated by a mixture of Gaussians and determines the probability of a data point being generated by a particular Gaussian (assigned to a cluster). Application of EM clustering is not possible if the data is not generated by a mixture of the Gaussian distributions. So, EM clustering cannot be applied to a data-set in which every data point is a sequence of zeros and positive numbers and cannot be represented as Gaussian. Our algorithm is conceptually similar to EM in the sense that effectively it creates a mixture model.But it gives the freedom that the distributions need not be Gaussian. Hierarchical Bayesian Clustering (Heller & Ghahramani, 2005), a probabilistic clustering technique, where we do not have to define the number of clusters. This technique probabilistically allocates a data point to a cluster and merges two clusters when they are very close. But as our data is very skewed, all the cluster centers it has are close, so merging becomes very arbitrary. Another good approach is the information-theoretic approach to Clustering (Slonim N, 2005).

2.3 Sequence Clustering

Sequence Clustering by similarity measures has been mostly applied to biological sequences, the similarity measure being edit distance, hamming distance,longest common sub-sequence.The features are extracted and then converted into a vector and similarity between these vectors are calculated. Another way of representing and clustering sequences is done by suffix tree. (Xing, 2010)(Agarwal, 2014). Probabilistic models has also been used for generating and clustering sequences. Most common among them is using HMM, of which one of the best papers is (Smyth, 1997).Microsoft paper uses Markov chains to model and cluster click user navigation pattern (Cadez, 2000). The following link discusses Microsoft's model for sequence clustering.

`https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/microsoft-sequence-clustering-algorithm-technical-reference?view=sql-server-2017/`

There are many variations of sequence modeling, but the basic theme centres on the discussed techniques.

3 Hierarchical Bayesian Clustering Model

3.1 Description

Consider a data set D consisting of N sequences, $D = \{Y_1, Y_2, \ldots, Y_N\}$ and $Y_i = \{y_{i1}, y_{i2}, \ldots, y_{iL}\}$ is a sparse sequence of length L_i of small positive numbers.The problem addressed here is to find K latent groups in the data. The data used is described in detail in the business scenario section.N sequences are sequences of restaurant visits of N users. Thus y_{iw} is the number of times the ith user

visited the restaurant in the wth week. Every user-visit sequence i is generated from a Poisson distribution with mean-visits λ_i.There are K clusters in the data with cluster-mean-visits η_k. So every cluster k has a distribution with mean η_k. Now these distributions are sort of mixed in proportion α_{ik} to contribute to the distribution of every user i . Thus mixture proportion α_{ik} is a measure of proportional impact of cluster mean η_k on mean user-visits λ_i of every user-sequence i.

3.2 MODEL

$shape \sim$ uniform (2,5)
$rate \sim$ uniform (0.1,1.0)
$\epsilon_k \sim$ uniform (0.1,1.0)
$\delta_k \sim$ uniform (0.1,1-ϵ_k)
$\eta_k \sim$ gamma($shape$, $rate$)
$\alpha_{ik} \sim$ Beta(ϵ_k,δ_k)
(Normalized over k)
$\lambda_{ik} \sim$ gamma(η_k , α_{ik})
$\lambda_i = \sum \lambda_{ik}$
$y_{iw} \sim$ Poisson(λ_i)

As we see from the model λ_{ik} is the contribution of the distribution of cluster K to the user-visit-mean λ_i.The mean of the distribution of λ_{ik} is η_k*α_{ik} i.e., cluster-mean multiplied by mixture proportion.

3.3 ALGORITHM

This algorithm assumes that number of clusters is known.The number of clusters K can also be estimated by calculating log-likelihood of the model for different values of K and by maximizing log-likelihood we can find a most possible K.

For every cluster k, we choose hyper-parameters shape, rate, ϵ_k and δ_k from uniform distribution. For each cluster k parameter η_k is the representative value for number of restaurant visits for cluster k. η_k is drawn from a gamma distribution with hyper-parameters $shape$ and $rate$. So the mean of the distribution of η_ks is $shape/rate$ and variance is $shape/rate^2$. Every cluster k has a parameter α_{ik}, the mixture proportion which has a beta distribution with hyper parameter ϵ_k and δ_k. So, ϵ_k/(ϵ_k + δ_k) is the mixture proportion averaged over the users for cluster k.This is an estimate of mean proportional user affiliation to cluster K. The algorithm derives the user group profile as a two dimensional vector where the entries are η_k, visit frequency of the user group and ϵ_k/(ϵ_k + δ_k), mean mixture proportion (user affiliation) to group K.

After α_{ik} is drawn it is normalized over k values for each User i. λ_{ik} is drawn from a gamma distribution with parameter η_k and α_{ik}. λ_{ik} is the contribution of the cluster k to λ_i, the expected number of weekly visits of user i. The distribution of λ_i is sum of distributions of λ_{ik}s. As λ_{ik} is drawn from K gamma distributions , λ_i will also have a gamma distribution with ϕ and θ where ϕ = $\sum(\alpha_{ikinv} * \eta_k)^2 / \sum(\alpha^2_{ikinv} * \eta_k)$
and
θ = $\sum(\alpha_{ikinv} * \eta_k)/\phi$

by approximation proposed by WelchSatterthwaite equation (Satterthwaite, 1946). α_{ikinv} is $1/\alpha_{ik}$. So we draw λ_i from the approximate distribution stated above and use it as the parameter of the Poisson distribution of user visit frequency per week.

Since the above calculation is complicated and it does not give any computational advantage, we apply a coarser approximation for calculating the distribution of λ_i.We assume that the rate parameter α_{ik} of all the gamma distributions of λ_{ik}s are equal. As all the α_{ik}s are within the range (0,1) and sum to 1, we approximate θ = $\sum \alpha_{ik}/k$
The sum of gamma distributed RVs with same rate parameter has a shape parameter which is sum of the shape parameters of the components ϕ =$\sum \eta_k$.We use this approximation in the implementation.

Table 1: Group profile for real data

cluster	η	$\epsilon_k/(\epsilon_k + \delta_k)$
1	0.317743	0.470588
2	0.640771	0.895
3	0.739492	0.46980
4	0.853064	0.47456
5	1.07128	0.34046

Table 2: Group Profile for synthetic data

cluster	η	$\epsilon_k/(\epsilon_k + \delta_k)$
1	0.326886	0.45398
2	0.420915	0.4483566
3	0.62439	0.51660901

4 Experiments and Result Discussions

4.1 Data Description

The data is collected from a marketing company which promotes loyalty programs for its client company. Each record contains user id, the total number of restaurant visits by the user, number of visits every week for 113 weeks and the interval between visits . In this paper, 'number of visits every week' is the user behavior that has been modeled. Almost 20000 in 25000 total number of users visited at most 7 times in 113 weeks. However,there are at least 100 users who visited almost every week.From the business perspective these 100 people are very important and cannot be treated as outliers. This skewness is the reason why a clustering technique which tries to get equal size clusters, such as K-means is ineffective because all the cluster centers merge to one. We have taken a less sparse subsection of the data for the experiment, but still our algorithm also is impacted by the skewness. Thus it is important to effectively initialize the algorithm and also it is important to prune it at certain step when all the cluster centers are not too close.

From the viewpoint of sequence modeling every User's data is a sequence made of zeros and positive numbers, each entry representing the number of user visits every week. The data is very sparse and skewed because very few users visited very frequently and most of the users visited below seven times in the whole observation period. There are huge number of zeroes in the user visit Sequences since there is no restaurant visits most of the weeks.For the experiment we have selected users who meet at least a visit threshold, empirically determined. Still for all users, zero is the value for maximum entries.

Other than using the industrial data we used synthetic data for the experiment. Three sets of data generated from three distinct Poisson distributions with parameters 0.2, 0.7 and 1.5. The size of these three sets were 100(lambda=0.2), 70 (lambda = 0.7) and 30 (lambda = 1.5). The difference in size of data is maintained for imitating the skewed nature of the real data.The parameters of Poisson distributions were chosen to be small numbers so that the synthetic data would have a huge number of zeroes in it and thus replicate the sparse nature of the actual sequence data.

The number of clusters K is chosen to be three in the experiment to match the number of distributions used to generate the data. In this experimental setting, the goal was to see the performance of the algorithm in finding three distinct latent groups, when there actually existed three latent groups in the data.We can see that each of the group parameters η_ks have three distinctly separated values, which shows the model could clearly cluster the synthetic sequences generated from three different distributions into three well separated clusters.Thus the model could bring out the latent structure in the data.

Table 3: mixture proportion of example sequence

cluster	η	λ_k
1	0.326886	0.0206129
2	0.420915	0.000552983
3	0.62439	0.0208292

4.2 EXPERIMENTAL PROCEDURE

The model was implemented in pystan. Pystan is a Bayesian inference platform where if a model is compiled and the data is input,the software generates posterior and samples it using No-U-Turn Sampler or NUTS, which adaptively sets path length in Hamiltonian Monte Carlo.

The objective of the experiment was to learn the parameters of the distribution of the clusters$\epsilon_k/(\epsilon_k + \delta_k)$ and $\epsilon_k/(\epsilon_k + \delta_k)$. We started using 1000 iterations, but the model was converging too quickly to form K very close clusters whose η_k values are very close. This is because of the inherent structure of the data, where the latent clusters are widely varying in size.To handle this skewness we recorded the parameters of the model at 500 iterations and that is reported in the table 1.However, In case of the synthetic data we did not have to prune the process, since the data were not that highly skewed and also the data had a clear underlying group structure as the data were generated by three different distributions.

4.3 RESULTS

The values of the cluster level parameters are listed in Table 1 for industrial data and Table 2 for synthetic data. The clusters are distinguishable, but as the sequences are sparse, the effect of high number of visits in a few weeks does not change the expected number of visits in a sequence in a big way.Thus the mean visits of every sequence has a low value and thus all the cluster means has low values.

5 CONCLUSIONS

This paper presents a hierarchical Bayesian model for clustering sparse sequences.The model is a mixture model, but does not need to be a Gaussian distribution, which gives much freedom in modeling. The algorithm creates latent group profiles of the discovered clusters in the data.The uncertainty in the parameter values can be estimated by the hyperparameters.In case of the skewed and sparse data that have been used in this experiment the algorithm has performed reasonably well.

5.1 FUTURE WORK

The future work will be to build a dynamic model in which the mixture proportion of the model will vary with time and depend on past values for mixture proportions. This model can be used to model changes of user behavior with time.

REFERENCES

CC Agarwal. *Data Classification Algorithms and applications*. Chapman and Hall, 2014.

Goy A. Petrone G. Segnan M. Torasso Ardissono, L. Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices. *Applied Artificial Intelligence17(8-9)*, pp. 687–714, 2003.

Meek Smyth Cadez, Heckerman. Visualization of navigation patterns on a website using model based clustering. *KDD '00 Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 280–284, 2000.

K. Heller and Z. Ghahramani. Bayesian hierarchical clustering. *In Proceedings of the 22nd International Conference on Machine Learning*, pp. 297–304, 2005.

Luckin R Masthoff, J. Future tv: Adaptive instruction in your living room. *Proceedings of the workshop associated with the Intelligent Tutoring Systems conference ITS02*, 2002.

Anagnost T.D McCarthy, J.F. Musicfx: An arbiter of group preferences for computer supported collaborative workouts. *In: Proc. of the ACM 1998 Conf. on Comp. Support. Coop. Work. CSCW 98*, pp. 363–372, 1998.

Salam M. McGinty L. Smyth B. McCarthy, K. Cats: A synchronous approach to collaborative group recommendation. *In: Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference*, pp. 8691, 2006.

Cosley D. Konstan J.A. Riedl J. OConnor, M. Polylens: A recommender system for groups of users. *In: Proceedings of ECSCW 2001, Bonn, Germany*, pp. 199–218, 2001.

Slonim N Richter Y, Yom-Tov E. Predicting customer churn in mobile networks through analysis of social groups. *Proceedings of the 2010 SIAM international conference on data mining (SDM 2010*, pp. 732–741, 2010.

Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin 2*, pp. 110114, 1946.

Bouzid-Picault Aghasaryan Bernier Senot, Kostadinov. Analysis of strategies for building group profiles. *LNCS*, 6075:40–51, 2010.

Tkacik G Bialek W Slonim N, Atwal GS. Information-based clustering. *Proc Natl Acad Sci U S A*, pp. 40–51, 2005.

Smyth. Analysis of strategies for building group profiles. *Advances in neural information processing Systems*, 1997.

et.al. Xing, Pei. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12 Issue 1:40–48, 2010.