International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.6., 2015
(Nov.-Dec.,)

RESEARCH ARTICLE

ISSN: 2321-7758

A TWO TIER MARKOV MODEL BASED CLUSTER VALIDATION FOR CLUSTERING CATEGORICAL SEQUENCES

MEERA S NAIR[1], TALIT SARA GEORGE[2]

[1]4th Semester Mtech Student Caarmel Engineering College Perunad, India

[2]Assistant Professor, Department of Computer Science, Caarmel Engineering College
Perunad, India

International Journal of Engineering Research-online (IJOER)

ISSN:2321-7758
www.ijoer.in

ABSTRACT

Clustering categorical sequences is an important and difficult task in data mining which facilitate taxonomy formation. Due to the lack of an inherently meaningful measure of similarity between categorical data objects, clustering these data remains a more challenging effort than clustering numeric data. Here, proposes a framework which includes WCPD (Weighted Conditional Probability Distribution) model and WFI (Weighted Fuzzy Indicator) model. The model initialization problem is resolved by using WFI model built on fuzzy indicator vector representation of categorical sequences. For clustering categorical sequences a cascade optimization framework that combines WCPD and WFI models have been designed. In order to rectify the problem of how to choose the next cluster to split or when to terminate the splitting, cluster validation is proposed.

Keywords: Categorical sequences, clustering, similarity measure

1. INTRODUCTION

Data mining can be viewed as a result of natural evolution of information technology. The major reason for the great deal of attention in the information industry is due to the need for turning huge amounts of data into useful information and knowledge. With the increase in the amount of sequence data, sequence analysis has become increasingly vital, even though the problem of clustering categorical sequences remains an important theoretical and practical issue in data mining.

Clustering is a data mining technique which has been used to place data elements into related groups without advance knowledge of the group definitions. K-means clustering and expectation maximization (EM) clustering are the most popular clustering techniques. Cluster analysis itself cannot be considered as one specific algorithm but treated as the general task to be solved. It is an iterative process of knowledge discovery or interactive multi-objective optimization which involves trial and failure. Until the result achieves the desired properties, it will often be necessary to modify data preprocessing and model parameters.

The challenge of clustering sequences also arises from the fact that inherent characteristics of sequences are reflected by local patterns. These local patterns (also referred to as motifs in the case of biological sequences), often of different lengths, play a very important role in determining the natural properties of a sequence. For example, protein sequences may belong to a broad family due to the fact they have many similar motifs, even though they vary considerably in terms of global alignment. Also, significant local sequence patterns from web navigation histories are very indicative of web surfers' preferences and interests. However, neither pairwise comparison methods nor frequent-pattern-mining methods can discover these significant local sequence patterns that form the natural clusters. This is because pairwise comparison focuses on global alignment but ignores local composition structure; and significant patterns that are frequent in natural clusters are not necessarily frequent from the perspective of the entire sequence data set, especially when the distribution of the natural clusters is imbalanced.

In this paper, describes how to solve the clustering problem with splitting accuracy. To achieve the accuracy and reduce clustering error, a cluster validation is carried out for clustering the sequences efficiently.

The rest of the paper is organized as follows. Section 2 formulates the motivation for taking the subject. Next section describes the system model. Section 4 concludes the work.

2. Motivation and Overview

A straightforward way of clustering categorical sequences is by transforming the sequences into feature vectors, and then utilizing well-studied numeric data clustering techniques to cluster these transformed feature vectors. Existing methods use either predetermined features or frequent patterns to represent the vector space. It is difficult to choose informative features to compose the vector space even when some domain knowledge is available. Moreover, these predetermined features are often chosen in an ad hoc way. On the other hand, because frequent-pattern-mining techniques usually generate a large number of redundant patterns, using these patterns to compose the feature space results in very high-dimensional vectors, and clustering high-dimensional data is itself still an open problem. Another issue in using frequent patterns to compose the feature space is that some significant patterns may be missing, as they are frequent only in a natural cluster, but not necessarily frequent from the perspective of the whole data set.

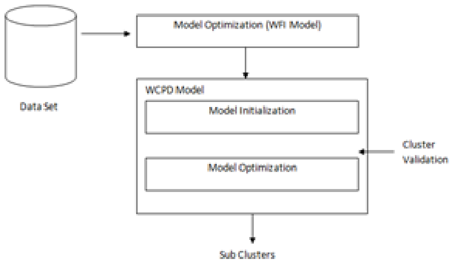
Pairwise measures of similarity between sequences have been proposed for comparing categorical sequences and agglomerative hierarchical clustering algorithms based on these measures are commonly used for sequence clustering. The q-grams distance and its variations are commonly used for sequence comparison, especially in text mining and sequence indexing. The edit (Levenshtein) distance is one of the preferred pairwise similarity measures for categorical sequence analysis. The major weakness of these similarity distance measures is that they have difficulty discovering local significant structures hiding in the sequences, and the essential nature of sequences in many domains lies in their local composition structure. Some statistical models are also used for pairwise comparison, although the intrinsic utility of these models is to measure how well a sequence fits a model: for example HMM (Hidden Markov Model). Because of the high complexity of HMM, a special Markov chain model with variable orders, the CPD model applied successfully to the correction of corrupted text and DNA sequence classification. This special Markov chain model is based on the conditional probability distribution of the next symbol, given a preceding subsequence. Supposing the sequences are composed from a finite alphabet, a CPD model is represented as P(s|), (s), which denotes the conditional probability of occurrence of the next symbol over given the preceding subsequence.

The CPD model is highly capable of modeling a set of categorical sequences in many domains. In current use of the CPD model, measures of similarity between a single sequence and a CPD model are calculated as a probability of generating the sequence by the model. However, existing similarity definitions suffer from one or more of the following drawbacks: Failure to satisfy the reflexivity condition: To be reasonable, a sequence should not be less similar to the model constructed from S itself than to the model constructed from other sequence. However, none of the existing definitions of a similarity measure satisfies this condition. Bias in the similarity calculation: By the existing definitions, a measure of similarity between S and the CPD model, i.e., sim(S), only takes the conditional probability distributions on the subsequences that appear in S into account, while completely ignoring the conditional probability distributions that are absent from S. For example, sequence S1 =ababab is quite dissimilar to sequence S2 =cdcdcdcdabab; however, by existing similarity definitions, sequence S1 may have the same similarity to the CPD model 1 constructed from S1 and the CPD model 2 constructed from S2. Sensitivity to noise symbols in the sequences. A noise symbol can cause the similarity between a single sequence and a CPD model (thus a sequence cluster) to be zero, no matter how similar the sequence would be to the CPD model if the noise symbols were filtered.

These drawbacks prompted to design a new CPD model and similarity measure. Based on the new CPD model, a similarity measure can be defined by taking all the discriminative patterns of each cluster into account to assign a sequence to the right cluster, so that reflexivity is satisfied and the similarity calculation is not biased towards partial specific conditional probability distributions of the statistical model. At the same time, the impact of noise symbols is reduced or eliminated. Although the CPD model has a desirable capacity to model a set of sequences, the problem of model initialization arises when it comes to solving clustering problems as the sequence set used to construct the model is initially unknown. At the present time, model/cluster initialization remains an unsolved problem for partition-based clustering algorithms. Existing partition-based algorithms implemented by the CPD model initialize the model from a single sequence via a heuristic method, in the same way as for numeric data. This results in a lack of statistical significance, especially when the sequences are short. It is clear that a statistical model should be constructed from a set of sequences showing great statistical similarity so that significant patterns are not diluted. How to effectively select a set of similar sequences to initialize a statistical model in the clustering process is still a challenging problem. This prompted to design a new Markov model which approximates new CPD model and is simple enough to consider the global information on variances among the sequences for model initialization.

Also, another factor that motivated is that the existing system doesn't pose an efficient method to select a cluster for splitting. If an appropriate cluster is not chosen then the entire clustering process fails. Therefore the issue needs to be overcome by applying an appropriate method for selecting a cluster for splitting.

3. System Architecture



The system architecture includes two different models such as WCPD model and WFI model. Both make use of the markov chain process. First tier is the WFI model which is initialized by using MCA (Multiple Correspondence Analysis) and is then used for initializing the second tier which is WCPD model. The WCPD model is constructed using PST (Probabilistic Suffix Tree). For the optimization of the two models, an algorithm named MCSC (Model-based Categorical Sequence Clustering) is used which starts with an all-inclusive cluster containing all the categorical sequences, and repeatedly chooses one cluster to split into two sub clusters. For cluster validation here proposes K-Means algorithm. Here the approach makes use of probability density rather than distance. K-Means tries to minimize the distance of objects belonging to a cluster from cluster center.

4. Algorithm Used

Algorithm 1: K-Means algorithm

Let $X = \{x_1,x_2,x_3,\ldots\ldots\ldots,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots\ldots,v_c\}$ be the set of centers.

Step 1: Select randomly '*c*' cluster centers.

Step 2: Calculate the distance between each data point and cluster centers.
Step 3: Assign data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
Step 4: Calculate the new cluster center.
Step 5: Again calculate the distance between each data point and new obtained cluster centers.
Step 6: In case if no data point was reassigned then stop, otherwise repeat from step 3.

5. Conclusion

Several similarity measures adopted for clustering possessed high computational complexity. Here, introduced a novel weighted CPD model (the WCPD model), which is significantly deferent from the conventional CPD model, to represent a set/cluster of categorical sequences, and a novel first-order Markov model (the WFI Markov model) to approximate the WCPD model. Also use Probabilistic Suffix Tree (PST) to select the memory subsequences of the WCPD model for clustering, and build these WCPD models. By combining the two statistical models, we propose a novel two-tier Markov model. The initialization and optimization of our model in the first tier provides a good initialization for the WCPD model in the second tier. The merits of the approach include the failure to satisfy the reflexivity condition is overcome, bias in the similarity calculation is maintained and sensitivity to noise symbols in the sequences also maintained. The drawback of the proposed system ie; how to choose a cluster for split- ting is overcome by incorporating cluster validation using K- Means algorithm.

Acknowledgment

We would like to thank reviewers for their helpful and constructive comments. We would also like to thank the faculties, the resource providers and the system administrator for the useful feedback and constant support.

References

[1]. Lidan Shou, He Bai, Ke Chen, and Gang Chen, “A Novel Variable-order Markov Model for Clustering Categorical Sequences,” vol. 26, no. 2, 2014.

[2]. M. Bicego, V. Murino, and M. A. Figueiredo, “Similarity-based classification of sequences using hidden Markov models,” Pattern Recognit., vol. 37, no. 12, pp. 2281–2291, 2004.

[3]. Q. Yang and X. Wu, “10 challenging problems in data mining research,” Int. J. Inform. Technol. Decis. Making, vol. 5, no. 4, pp. 597–604, 2006.

[4]. J. Wang, Y. Zhang, L. Zhou, G. Karypis, and C. C. Aggarwal, “Discriminating subsequence discovery for sequence clustering,” in Proc. SIAM SDM, 2007.

[5]. T. Xiong, S. Wang, A. Mayers, and E. Monga, “DHCC: Divisive hierarchical clustering of categorical data,” Data Min. Knowl. Discov., vol. 24, no. 1, pp. 103–135, 2012.

[6]. Kelil, S. Wang, and R. Brzezinski, “CLUSS2: An alignment independent Algorithm for clustering protein families with multiple biological functions,” Int. J. Comput. Biol. Drug Des., vol. 1, no. 2, pp. 122–140, 2008.