Sparse Poisson Latent Block Model for Document Clustering

G.Snigdha
M.Tech (CSIE) Dept. of CSE JNTUHCE

Abstract - Clustering has been successfully used in wide range of application domain such as pattern recognition and learning, marketing and economics, medicine and biology, text and document processing. Objects are clustered based on their properties and computing similarities and dissimilarities among them. Document clustering is automatic organization of documents into clusters so that documents with in cluster have high similarity in comparison to document in other clusters. In general, two types of algorithms are used like hierarchical based algorithms and partition based algorithm. Document clustering faces challenges like selection of appropriate feature of the document, similarity measure, clustering methods etc. Sparsity is the property of being scanty or scattered. In case of high dimensionality and sparsity it would be a challenging task for any clustering methods. So, co-clustering can better handle the challenging task. Co-clustering can be seen as a method of co-grouping two types of entities simultaneously based on similarities of their pair-wise interaction. Co-clustering can be used in various applications like neuro-science, computational biology, computer vision and so on. Sparse Poisson Latent Block Model is parsimonious and takes into account a diagonal structure of a data matrix to make a distinction between relevant block and block outside the diagonal. The Sports dataset is considered for the experimentation using Python.

I. INTRODUCTION

Clustering is a concept of grouping similar objects together. This is a very useful unsupervised learning technique to deal with large volumes of data. Most of the clustering measures focus either on objects or features but not both together. Even though clustering is used in various applications domains such as Marketing (Help marketers discover different groups in their customer bases, and then use this knowledge to develop targeted marketing programs), Land use (Identification of areas of similar land use in an earth observation database), Insurance (Identifying groups of motor insurance policy holders with a high average claim cost), City-planning (Identifying groups of houses according to their house type, value, and geographical location) and etc. Clustering approaches may sometimes be challenging task for high dimensional and sparse data illustrated from some datasets like document χ term matrices from text mining. Clustering has wide applications in economic science (market research), WWW (document classification and cluster weblog data to discover group of similar access patterns), pattern recognition, spatial data analysis (creating thematic maps in GIS by clustering features), image processing The advantages of clustering are automatic recovery from failure, ability to perform maintenance and upgrades with limited down time and disadvantages are shared storage failure, network service failure, operational errors and site disasters.

Document clustering is the method of collecting similar documents into bins, where similarity is some function on a document. It has been studied very thoroughly, as this appropriate in various areas like web mining, information retrieval and search engines. It is considered as similarity measure between documents and grouping similar documents together. It provides well-organized representation and visualization of the documents which help in easy retrieval of information

II. RELATED WORK

Clustering is divided into differrent algorithms, as mentioned in the fig: 1.

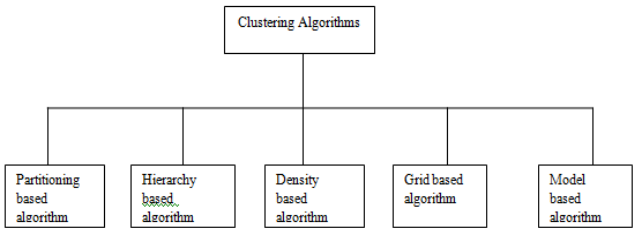

Fig 1: Clustering techniques

Co-clustering is method of grouping, objects and features concurrently. Co- grouping two types of entities simultaneously based on similarities of their pair-wise interaction. When we consider high dimensional and sparse datasets, co- clustering in compare to clustering, it groups objects and features simultaneously, which turns out to be more effective. Consider a data matrix x of size n X d where I is the set of n rows, and J is a set of d columns, the basic plan of co-clustering is making permutations of objects and attributes to draw a correspondence structure on I X J to present data easier to handle and understand. Various co-clustering application are Simultaneous clustering of documents and words in text mining, Microarray (i.e., genes and experimental conditions) in bioinformatics, Tokens and contexts in natural language processing, Content-based image retrieval, Auditory scene categorization, Video content recognition, Users and movies in recommender systems, Missing value prediction in recommender systems, Co-clustering categorical data matrices.

Co-clustering is preferred over clustering because clustering

can't deal with high dimensional and sparse data. Sparse data is a data composed of zeros. Here we are concerned with co-clustering of sparse high dimensional data. While dealing with such data, seeking homogeneous block may not always be enough to produce useful and ready-to-use results. In fact due to data sparsity, several co-clusters may form of sparse data (composed of zeros), such co- clusters, with homogenous data need to be filtered in post processing phase. In other words, the user can select the most useful co-clusters to determine which document clusters must go with which term clusters; however it is not easy task even with a reasonable number of document and term clusters.

A way to solve the sparsity problem is to use block diagonal algorithm. These algorithms try to find an optimal block diagonal clustering (that the objects and features have the same number of clusters) and after a proper permutation of the rows and columns the algorithm produces a block diagonal matrix as a result. These kind of approaches have the advantage of directly producing the most significant co-clusters, making the results much easier to analyse and understand. Another advantage of the diagonal assumption is that it puts particular focus on the most selective terms, which makes it possible to obtain high quality document and word clusters.

III. PROPOSED WORK

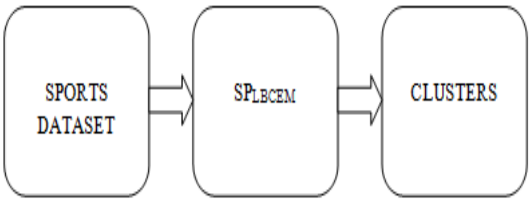

Fig 2: Block Diagram

In this work, Sparse Poisson Latent Block Model is implemented. This model is cost-conscious and takes into account a diagonal structure of a data matrix to make a difference between related blocks and the block outside the diagonal. The constraint forced on Poisson parameters of blocks is helpful and reflects the objective of document clustering. Placing the co- clustering problem within the Classification Maximum Likelihood approach, the novel diagonal co clustering algorithm (SPLBCEM) is been derived.

SPLBCEM ALGORITHM:

Step 1: Given n x d data matrix **X** defined on two sets I and J, g = no of rows.

Step 2: Initialize z, w (partitions on sets I and J), $л_k$, $ρ_l$, $γ_{kk}$, γ (parameters).

Maximization step:

Step 3: calculate complete-data log-likelihood (L_c) is given in Appendix A [1].

Expectation-classification step:

Step 4: computation of γkk for all k, and γ to show these are maximizing L_C (z,w,Θ) refer [1] Appendix A

Step 5: Step 3 and 4 are repeated for columns

Step 6: Steps 3, 4, 5 are repeated until convergence.

Step 7: return z, w, л, ρ, γ.

The experiment show that our algorithm is very effective on document x term matrices compared to traditional clustering and general partitioned co- clustering algorithms.

IV. IMPLEMENTATION

The implementation of this work is performed in python language. Our experiments will be performed on a pc (windows 7, 4 GB RAM, 1TB Hard Disk, Intel i3 processor with 1.70 GHz).

In this work, the sport dataset containing various text documents are considered. The data is pre-processed by eliminating special characters, and stop words. Then the word count and term frequency is calculated. Based on the term frequency the PDF (probability density function) is defined. Depending upon the probabilities the co-clustering is applied on data to form clusters.

The sports dataset is download from http://mlg.ucd.ie/datasets/bbc.html. This dataset contains various text documents based on 5 different sport categories.

V. REFERENCES

[1]. M. Aliem, F. Role, M. nadif, " Sparse Poisson Latent Block Model For Document Clustering", IEEE Trans. Knowledge and Data Engineering., Vol. 29, No. 7, pp. 1563-1576, July 1 2017.

[2]. N. Shah and S. Mahajan, "Document Clustering: a detailed review", International journal of computer application, vol.4, No.5, pp. 30-38, 5 Oct 2012.

[3]. Gerald J. Kowalski, Mark T. Maybury**. "**Information Storage and Retrieval Systems: Theory and Implementation", New York: Kluwer academy publisher, 2002.

[4]. M. Ailem, F. Role, and M. Nadif, "Co-clustering document-term matrices by direct maximization of graph modularity," in Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. 2015, pp. 1807–1810. Govaert and M. Nadif. Co-Clustering. Hoboken, NJ, USA: Wiley, 2013.

[5]. http://www.cs.cmu.edu/~lemur/3.1/cl uster.html

[6]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7100283

[7]. http://www.cs.put.poznan.pl/jstefanowski/sed/DM 7clusteringnew.pdf