ISSN:0975-9646

An Empirical Study of Geographic Popularity Patterns of Motion Picture Content across Europe Using Unsupervised Machine Learning

Pawan Dwivedi, Pentyala Srinivasa Rao

Department of Applied Mathematics
Indian Institute of Technology (Indian School of Mines)
Dhanbad, India

***Abstract*—Unsupervised Machine Learning is used to study the covered structure inside the data in which training labels are not available or are not relevant to the problem. Motion picture content across Europe is expected to see some geographic popularity pattern. We are proposing a sequence of methods to find these patterns. First we define a measure of popularity using the available data. The key idea is to use the best dimensionality reduction and clustering algorithms. We have taken grossing of different movies across different countries in Europe for the year 2016.[1] With the advent of dimensionality reduction algorithms it is possible to analyze very high dimensional data efficiently without losing much information. Using proper scaling and principal component analysis, we are able to reduce the number of variables that can be used to get the entire information. Applying clustering algorithms like K-Means, Gaussian Mixture Model and Hierarchical Clustering on this data would give clusters of contents having similar popularity patterns. The next step is to use these clustering algorithms to find geographic regions which show similar popularity patterns. All statistical and programming implementations are done in R.**

***Keywords*—Unsupervised Machine Learning, Clustering, Dimensionality Reduction, Principal Component Analysis, KMeans, Gaussian Mixture Model, Hierarchical Clustering, tdistributed stochastic neighbor embedding**

I. INTRODUCTION

Unsupervised Clustering techniques have applications in many practical scenarios like recommendation systems, search engines, health care, etc.[2][3][4] Collaborative filtering is a well-known technique in recommender systems that establishes relationship between users based on their preference. But it gives inaccurate results in high dimensional and sparse data. That makes the computation of similarity between users imprecise and reduces the accuracy of collaborative filtering algorithms.[5] The solution to this is clustering. Similar objects are grouped together in search engines using clustering. By grouping similar patient records in clusters, health care patterns are discovered. To make these things possible, many clustering algorithms are developed.

The real challenge with these clustering algorithms is when the data is very high dimensional and sparse. One phenomenon associated with high dimensional data is the curse of dimensionality, i.e. when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. In such high dimensional data, all objects appear to be sparse and dissimilar.[6] It is not reliable to implement these unsupervised clustering algorithms using this data as it is very unlikely to show any patterns of similarity. This paper describes a solution to this problem by describing a sequence of steps that can take a very high dimensional data to form clusters. Our data has observations for 22 countries and over 600 movies.[1] We take grossing value of each movie-country pair, divide it by the total grossing in the country. We call the obtained value as grossing score. Next we apply proper filtering and scaling. Then using Principal Component Analysis, we are able to reduce the number of variables under consideration significantly by compromising 10% of the cumulative variance. The number of dimensions got reduced significantly remembering 90% of the original information. Then we have applied various clustering algorithms like KMeans, Gaussian Mixture Model and Hierarchical Clustering to get actual clusters. Further we discuss the future implications of this analysis.

II. METHOD

A. Variation in popularity

The analysis starts by taking grossing value for each movie across 22 European countries. Next this value is scaled by total grossing value of all movies in that country. Further this value is used as the regional popularity score. The first thing we did was finding out movies that show maximum variation in popularity score by taking mean and standard deviation of score values of each movie across different countries. Next step is to apply certain filters in the data. We have taken only those movies into consideration which contribute more than

0.1% in the total grossing of the country from movies.

B. Principal component Analysis (PCA)

PCA is applied to find the aspects that are most important in the analysis. The main idea is to reduce the number of variables under consideration. It does this by creating new variables that can summarize the original dimension. PCA doesn't discard any original variable. It just transforms them. It uses Orthogonal Transformation to form Principal Components. The first principal component would capture most of the data followed by second, third and so on. All Principal Components are linearly uncorrelated and orthogonal to each other.[7]

C. k-means Clustering

Given n observations (x_1, x_2,..., x_n). k-means Clustering partitions these *n* observations into *k* ($\leq$ *n*) clusters. The algorithm decides *k* centers and assigns each observation to the cluster whose mean yields the least within-cluster sum of squares of distance of each observation from its center.

$$arg\ min_S \sum_{i=1}^{k} \sum_{x \epsilon S_i} \parallel x - \mu_i \parallel$$

k-means algorithm uses Euclidean distance as the metric.[8] One of the drawbacks of this algorithm is that the number of clusters *k* is an input parameter. The clusters formed by k-means algorithm are always spherical. This results in some inaccuracy in the analysis. To overcome these drawbacks mixture model is used.

D. Gaussian Mixture Model

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians. The Gaussian Mixture object implements the expectation-maximization (EM) algorithm for fitting mixture-of-Gaussian models.[9] It can also draw confidence ellipsoids for multivariate models, and compute the Bayesian Information Criterion to assess the number of clusters in the data. Just like k-means, Gaussian Mixture Model also maps *n* observations to *k* clusters. The difference between k-means algorithm and Gaussian Mixture Model is that GMM is able to form non-spherical clusters also which makes it more accurate than k-means.[10] Because of this, we have used Gaussian Mixture Model instead of k-means in this analysis.[11]

E. Hierarchical Clustering

Hierarchical clustering is a statistical method that builds a hierarchy of clusters. The first step is to create a matrix of observations and variables. Then this matrix is centered and scaled by subtracting each element with mean of observation and then dividing by standard deviation. The next step is to create a distance matrix. The choice of appropriate metric decides the shape of the clusters

$$Euclidean\ distance\ \|a-b\| = \left(\sum(a_i - b_i)^2\right)^{1/2}$$

In this analysis, we have used agglomerative strategy, i.e. each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. The output of the hierarchical clustering is visualized using a dendogram.

F. t-distributed stochastic neighbor embedding (t-SNE)

It is a machine learning algorithm for dimensionality reduction. It maps down high dimensional data into two or three dimensions.[12] The main purpose of using t-SNE in this analysis is to visualize different clusters formed by the clustering algorithms. We have used t-SNE to map high dimensional output to 2 dimensional vectors and drawing a scatter plots of the same. This way we are able to visualize all clusters.

III. RESULTS

A. Variation in popularity

It can be seen in Figure 1, movies moving away from the origin along the y-axis show maximum variation in popularity. While those moving away along x-axis are having maximum contribution to the total grossing in that country. For example, Rogue One: A Star Wars Story and Finding Dory are the two movies showing maximum geographic variation in popularity, whereas The Conjuring 2 and X-Men: Apocalypse are the ones showing minimum geographic variation in popularity.

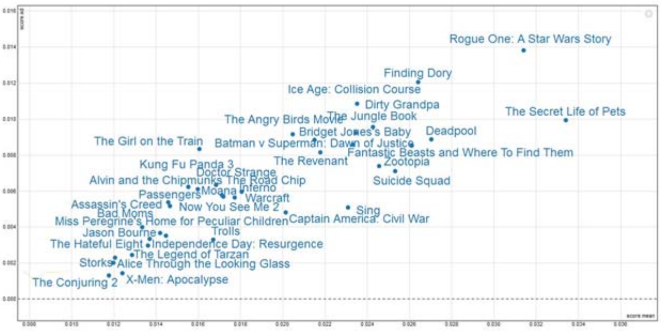

Fig. 1. Variation in popularity

B. Principal Component analysis

By maintaining 90% cumulative variance, we are able to reduce the number of variables under consideration to less than 50%.

C. Gaussian Mixture Model

We filter out any content that is contributing less than 1% in the countries' grossing from movies. Then we have estimated the number of components in Gaussian Mixture Model using Bayesian Information Criterion. The Bayesian Information Criterion (BIC) is a criterion for model selection among a finite set of models. It is based on the likelihood function.[13] The formula for BIC is given by

$$BIC = \ln(n)k - 2\ln(L)$$

where *L* is the maximum value of the likelihood function of the model, *n* is the number of observations and *k* is the number of parameters to be estimated. The number of components in the Gaussian Mixture Model is the value which yields lowest BIC . We calculate the BIC value for different number of components for the Gaussian Mixture Model. The results can be seen in Figure 2 and 3.

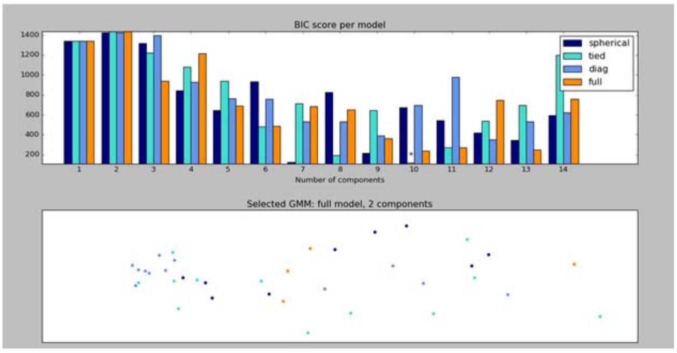

Fig. 2. BIC for movies

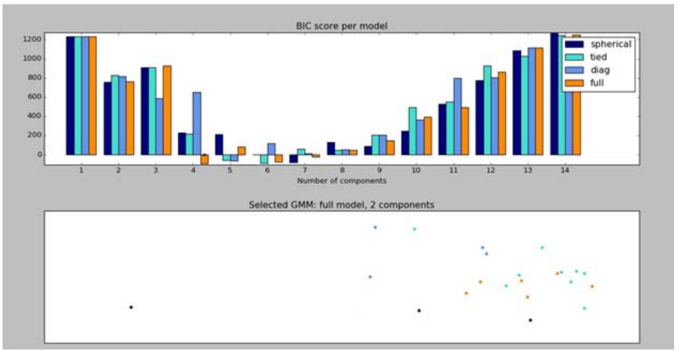

Fig. 3. BIC for countries

The optimal number of components comes out to be 10 for movies and 4 for countries. This means, we can effectively put movies into 10 clusters and countries into 4 clusters. To visualize the output, we pass this result into t-SNE. It maps the result into 2 dimensions. Then we show the partition of n movies to 10 clusters in a scatter plot in Figure 4. For instance, we can see from the scatter plot that Finding Dory, Deadpool and The Revenant are in one cluster whereas Kung Fu Panda 3, The Angry Birds Movie and Trolls are in another cluster. Repeating the same process for countries. We can see the output in Figure 5. It can be observed from the scatter plot in Figure 5 that Austria, Belgium and Luxembourg are in one cluster and Ukraine, Bulgaria and Romania are in another cluster.

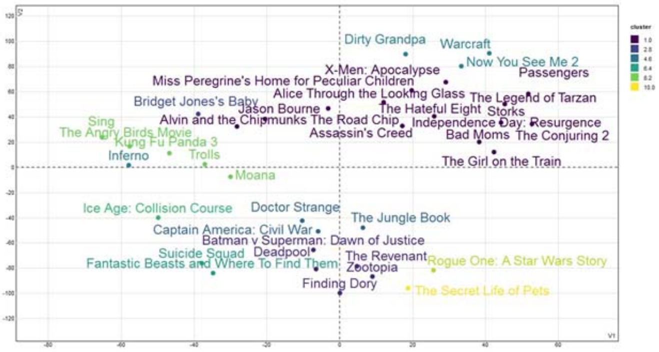

Fig. 4. Scatterplot of t-SNE output for GMM clusters of movies

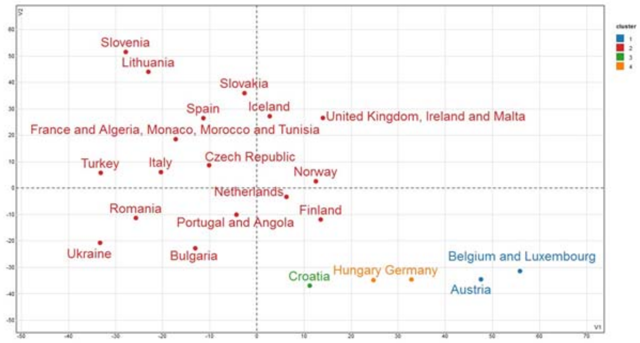

Fig. 5. Scatterplot of t-SNE output for GMM clusters of countries

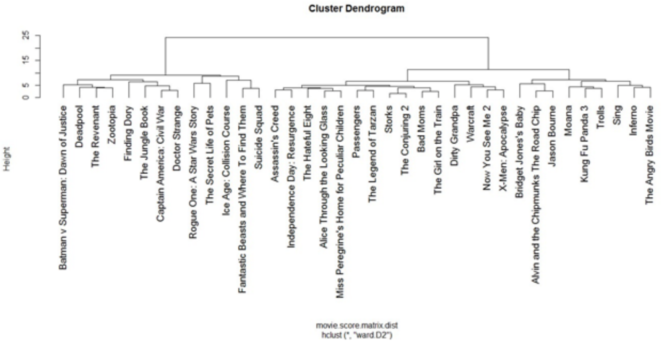

Fig. 6. Dendogram of hierarchical clusters of movies.

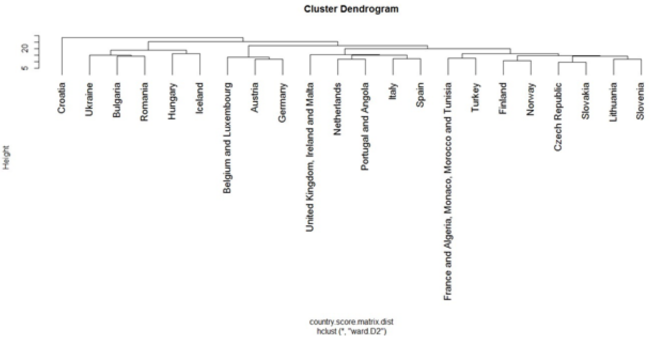

Fig. 7. Dendogram of hierarchical clusters of countries

D. Hierarchical Clustering

Hierarchical clustering is a very useful method of clustering. The output is best visualized using a dendogram. The dendogram is a tree that illustrates the hierarchy of clusters. The output for clusters of contents can be seen in Figure 6. The same output for clusters of countries can be seen in Figure 7. By comparing Figure 5 and 7, it can be seen that there exists some geographic pattern in popularity of contents.

IV. Discussion and Future Work

Clustering is one of the initial challenges of any machine learning problem. With the advent of big data, performing unsupervised cluster analysis is becoming difficult. This analysis serves as an example of how to find geographic variation in popularity using unsupervised machine learning. The analysis could serve applications in recommendation systems, search engines and models for many other systems that require an insight on geographic variation in popularity of contents. Further statistical models can be built by making use of the hierarchical structure in the data, which will be able to predict popularity of contents in different geographic regions.

V. Conclusion

By applying unsupervised machine learning techniques and comparing the output, it is observed that there exists some variation in popularity of contents across Europe. The analysis clearly defines sequence of methods to find patterns of geographic variation in popularity of contents using unsupervised machine learning techniques. The clusters of similar objects are displayed is Figures 4, 5, 6 and 7.

VI. Acknowledgement

I am thankful for the guidance and support of my mentor Dr. P. S. Rao, Assistant Professor, Department of Applied Mathematics.

References

[1] (2017) Information courtesy of box office mojo. [Online]. Available: http://www.boxofficemojo.com

[2] Y. Sun, "Conversational recommendation system with unsupervised learning," *ACM Digital Library*, 2005.

[3] T. Liu, "Clustering billions of images with large scale nearest neighbor search," *IEEE Workshop on Applications of Computer Vision*, 2007.

[4] X. Y. Wang and J. M. Garibaldi, "A comparison of fuzzy and nonfuzzy clustering techniques in cancer diagnosis," *The University of Nottingham*, 2006.

[5] M. C. Pham, "A clustering approach for collaborative filtering recommendation using social network analysis," *Journal of Universal Computer Science, vol. 17, no. 4*, 2011.

[6] M. Dash, "Dimensionality reduction," *Nanyang Technological University, Singapore*, 2006.

[7] I. Jollife, *Principal Component Analysis*, 2nd ed. Springer, 2002.

[8] X. H. Chris Ding, "K-means clustering via principal component analysis," 2004.

[9] I. Naim and D. Gildea, "Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients," *Department of Computer Science, University of Rochester Rochester, NY 14627, USA*, 2012.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, 2006.

[11] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis and density estimation," *Journal of the American Statistical Association 97:611-631*, 2002.

[12] G. H. Laurens van der Maaten, "Visualizing data using t-sne," *Journal of Machine Learning Research 9 (2008) 2579-2605*, 2008.

[13] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics, Vol. 6, No. 2.*, 1978.