Proportional Data Modeling using Unsupervised Learning and Applications

Jai Puneet Singh

A Thesis

in

The Department

of

Concordia Institute of Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Information Systems Security) at

Concordia University

Montréal, Québec, Canada

May 2017

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Jai Puneet Singh**

Entitled: **Proportional Data Modeling using Unsupervised Learning and Applications**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Information Systems Security)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________________ Graduate Program Director
Dr. Dr. W. Lucia

______________________________ External Examiner
Dr. A. Hammou-Lhadj

______________________________ Examiner
Dr. A. Mohammadi

______________________________ Supervisor
Dr. N. Bouguila

Approved by ______________________________
C. Assi, Graduate Program Director
Department of Concordia Institute of Information Systems Engineering

______________ 2017

____
Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Proportional Data Modeling using Unsupervised Learning and Applications

Jai Puneet Singh

In this thesis, we propose the consideration of Aitchison's distance in K-means clustering algorithm. It has been used for initialization of Dirichlet and generalized Dirichlet mixture models. This activity is then followed by that of estimating model parameters using Expectation-Maximization algorithm. This method has been further exploited by using it for intrusion detection where we statistically analyze entire NSL-KDD data-set.

In addition, we present an unsupervised learning algorithm for finite mixture models with the integration of spatial information using Markov random field (MRF). The mixture model is based on Dirichlet and generalized Dirichlet distributions. This method uses Markov random field to incorporate spatial information between neighboring pixels into a mixture model. This segmentation model is also learned by Expectation-Maximization algorithm using Newton-Raphson approach. The obtained results using real images data-sets are more encouraging than those obtained using similar approaches.

“Time is the longest distance between two places.”

Tennessee Williams

Acknowledgments

I wish to thank everyone who helped me to complete this thesis. I would like to express my sincere gratitude to my supervisor Dr. Nizar Bouguila. I will be forever grateful to him for giving me the opportunity to work under his guidance and improve my academic knowledge. The completion of this thesis would have not been possible without his endless support, effort, encouragement and most importantly patience.

I would like to thank Mitacs Inc. and Quebec government for giving me an opportunity to study in Canada through Globalink fellowship. I would like to thank all my friendly labmates. Especially I would like to thank Dr.Yewang Chen, Walid Masoudimansour, Elise Eppaillard, Eddie and Muhammad Azam.

My gratitude extends to my friends Marie Epain, Zaida Escila Martinez and Asama Alglawe for their continuous support and encouragement.

Last but not the least, my gratitude goes to my family for their endless love and support throughout my life, without their support, simply I could not do this. I also extend my warmest thanks to my little sister...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and Related Work

Data clustering is one of the main problems in data mining. It is an unsupervised classification of patterns (observations, data items or feature vectors) into groups (clusters) (34). Data clustering is used in computer vision, decision making, pattern analysis and machine learning applications such as image segmentation, information retrieval, anomaly detection and character recognition. Due to the advancement of technology, a huge amount of data is generated every day. Proportional data, in particular, are naturally generated by different fields such as geology, Bio-informatics, computer vision, etc. Proportional data can be defined as any vector $x = (x_1, x_2..., x_D)$ subject to unit sum constraint where $(x_1 + x_2... + x_D) = 1$ and $x_d \geq 0$, $d = 1, 2, ..., D$. The clustering of this type of data requires efficient algorithms (39). Cluster analysis is prevalent in any discipline that involves analysis of multivariate data (34). The problem of unsupervised clustering using mixture models requires good initialization which is generally done with the help of K-means algorithm. However, K-means uses Euclidean distance which finds the shortest distance between two samples. Unfortunately, the Euclidean distance is not appropriate for proportional data. Despite the fact that Aitchison and other distance metrics are appropriate for proportional data, they have not received much attention compared to Euclidean distance. However, some related works do exist. For instance, Kashima et al. proposed a L1 distance (36) based K-means algorithm to address the problem of proportional vectors clustering. Hijazi et al. used Dirichlet regression models for modeling compositional data and came to the conclusion that Dirichlet regression model is just an alternative to log-ratio analysis which can be done by Aitchison Log Ratio Analysis (30). In this thesis, I have compared different distance metrics when clustering proportional data with K-means algorithm.

A direct application of clustering is intrusion detection. Emerging growth of networks and rate of transfer of data through networks has increased the demand for network security. Machine learning approaches are one of the popular strategies in network security for finding attacks. The NSL-KDD data-set is widely used to validate network intrusion detectors, a predictive model capable of distinguishing between 'bad' connections, called intrusions or attacks and 'good' normal connections. There is a significant literature on Anomaly detection. Anomaly detection deviates from normal traffic and it is important to find an anomaly in an era of communication. Although, there are a lot of articles on intrusion detection, feature selection and unsupervised learning approaches are often underrepresented. There are very limited publicly available data sets for network-based anomaly detection. Earlier KDDCup99 was used heavily for all kind of intrusion detections through machine learning methodology. KDDCup99 has a huge number of redundant records (54). It was found that around 78% of records in KDDCup99 were duplicated. Mchugh (43) gave many critics on KDDCup dataset and DARPA data set of 1998 as it was not good for applying statistical approaches to learning. The new NSL KDD data set was proposed (5) to overcome the problems present in KDDCup99 and DARPA data sets (1). NSL KDD data set does not have redundant and duplicate records. There is a lot of work which has been done on NSL KDD data set to find intrusions. All existing learning approaches are supervised. The author in (29) had used principle component analysis for feature extraction followed by SVM for finding intrusions in NSL KDD data set. The author in (46) had used a combination of classifiers or clusters which are followed by supervised or unsupervised data filtering. The author in (64) had used feature selection with NSL KDD data set. In this thesis, we have used unsupervised learning using Dirichlet Mixture Model. The initialization of mixture model is done with K-means using different distance metrics. Aitchison distance metric showed better results than Euclidean distance. It is followed by feature selection on NSL KDD data which reduces features from 41 to 16 features. The comparative analysis has been drawn which shows how feature selection and proper initialization increases the detection rate in NSL-KDD data set.

Another direct application of clustering is image segmentation. It plays an important role in the field of computer vision, pattern recognition, and image processing. There has been a rapid growth in the domain of image segmentation. There are numerous methodologies being proposed and most of them are based on supervised approaches, with the availability of ground truth of image data sets like Berkeley data set, MIT image data sets, etc. In unsupervised approaches, the Gaussian distribution is heavily used for image segmentation purposes. However, it suffers from various limitations as discussed in (17). There are other models which have been proposed such as Dirichlet and generalized Dirichlet mixtures that provide better results for compactly supported data as discussed in (26) (42). The generalized Dirichlet mixture model has more flexible covariance structure and requirements of conditional independence are less restrictive as compared to the Dirichlet mixture model (61).

Previously, The Dirichlet and generalized Dirichlet mixture models have been deployed for image processing, clustering and various other pattern recognition applications (14) (13) (60). The Markov random field for image segmentation has been used in the past to find color boundaries (31). It has been used to eliminate noise and to fill in data without disrupting its discontinuities (8). Markov random field (MRF) uses image brightness of edges as its guide to find contours. Hence, this property of MRF has integrated with Gaussian mixture models using Expectation-Maximization technique in (44). In the segmentation approach proposed in this thesis, the integration of spatial information using MRF in Dirichlet and generalized Dirichlet mixtures is applied. To the best of our knowledge, this technique has not been used before for these mixtures. This combination aims to enhance image segmentation. The comparative results obtained are very encouraging.

1.2 Contributions

The contributions of the thesis are as follows:

(1) **Comparing different distance metrics for K-means clustering of proportional data**: In the proposed work, we present the K-means clustering approach using different distance metrics. In particular, we propose the consideration of the Aitchison's distance. Experimental

results are presented using silhouette plots for showing divergence from the center, and confusion matrices are used to validate our clustering of synthetic and real data sets. The algorithm with Aitchison's distance metric results in lower error rates.

(2) **Proposing k-means with feature selection to improve mixture model accuracy, sensitivity and precision**: Most of the approaches applied on a NSL KDD data set were supervised approaches. We had conducted statistical analysis on this data set using a Dirichlet Mixture model. We have performed initialization using Aitchison's distance metric. The feature selection highly affects both the performance and results leading to an improved evaluation of anomaly detection through an unsupervised approach.

(3) **Integration of spatial Information into Dirichlet and generalized Dirichlet finite mixture models**: Our approach suggests the integration of spatial information into two different finite mixture models (Dirichlet mixture model, generalized Dirichlet mixture model) to produce smooth and more meaningful regions in image segmentation while offering more flexibility and ease of use for data modeling in comparison to the well known Gaussian mixture model.

1.3 Thesis Overview

(1) **Chapter 1:** It introduces the thesis and gives some related works. This presents the predicament of clustering data sets in unsupervised learning.

(2) **Chapter 2:** It proposes new distance metrics that can be used in a K-means clustering algorithm which is capable of clustering proportional data sets. It proposes how proper initialization in mixture models using the above proposed method improves the results via an intrusion detection application. Hence, further improvement is depicted using feature selection and feature extraction methods.

(3) **Chapter 3:** This chapter shows the integration of spatial information using Markov Random Field into mixture models. The results have shown increased improvements when compared with similar previous approaches.

(4) **Chapter 4:** This gives the summary, conclusion and potential avenues for future work.

Chapter 2

Proportional Data Clustering using K-Means Algorithm: A comparison of different distances

In this work, we propose to use Aitchison distance metric for clustering of proportional data (6). Moreover, we compare it with several other distances in several applications. The rest of the chapter is organized as follows: In section 2.1, the proposed method is explained in details, and various distance metrics are introduced. In section 2.2, outlier detection method with the proposed algorithm is proposed. Section 2.3 gives different experimental results on many synthetic and real data sets. Finally, in section 2.6 concluding remarks are drawn.

2.1 The Proposed Method

K-means clustering uses distance metrics to find nearest neighbors and mostly Euclidean distance has been used. The objective function of K-means can be represented as follows:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left|\left| x_i^j - c_j \right|\right|^2 \tag{1}$$

In this equation x_i represents the data point and c_j represents the cluster center. This type of distance generates spherical shaped type of clusters as a result. Kashima et al. (36) proposed L_1 distance which is also known as Manhattan distance for optimization of K-means clustering on proportional data and successful improvement was seen. There are many different distances and divergence metrics. However, in our chapter we concentrated on Euclidean log transformed data, Aitchison's and Kullback-Leibler divergence. These distances have been known for a long time, but they have not been explored for proportional data. The only drawback of these distances is that they don't accept 0 values. Martin et al. (40) proposed an approach to deal with zeros in compositional data. We have applied this approach before clustering. In this chapter, we are proposing K-means clustering using Aitchison distance. As per our knowledge, this distance has not been considered in K-means in the past.

Algorithm 1 K-means Algorithm

```
1: Set the Initial number of centroids randomly or sequentially.
2: Calculate the distance between each data point and cluster centers.
3: repeat:
4:     Assign the minimum distance data points to cluster center whose distance is minimum to
       that point.
5:     Recalculate the cluster center using:
6: $c_i = \frac{1}{m_i} \sum_{j=1}^{m_i} x\left(i\right)$; $m_i$ represents total number of data points in $\left(i\right)$ cluster.
7:     Re-calculate the distance between each data point and newly obtained cluster center.
8: until : No data point is reassigned.
```

In the K-means algorithm, distance is used in steps 2 and 4 of Algorithm 1. The distance metrics that we have explored are given in Table 2.1.

2.2 Outlier Detection

Outlier detection is a deeply researched topic in both communities of statistics and data mining (21). Outlier detection methods are categorized as external and internal methods (63). In our case, we used internal outlier detection technique where after K-means clustering with various distance metrics, distance is compared with the other data in same group with centroids of particular group as shown in Algorithm 2.

Table 2.1: Different Distance Metrics used in K-means

S.No.	Distance Name	Distance Metrics
1	Euclidean Distance	$d_E^2\left(x,y\right)=\sum_i\left(x_i-y_i\right)^2$
2	EL transformed Data	$d_{EL}^2\left(x,y\right)=\sum_i\left(\log x_i-\log y_i\right)^2$
3	J-divergence	$d_{jd}^2\left(x,y\right)=\sum_i\left(\log x_i-\log y_i\right)\left(x_i-y_i\right)$
4	Jeffery's-Matusita Distance	$d_m^2\left(x,y\right)=\sum_i\left(\sqrt{x_i}-\sqrt{y_i}\right)^2$
5	Manhattan Distance (L1 Distance)	$d_{L1}^2\left(x,y\right)=\sum_i\lvert x_i-y_i\rvert$
6	Kullback-Leibler Divergence	$d_{KL}\left(x,y\right)=\sum_i\left(x_i\log\frac{x_i}{y_i}+y_i\log\frac{y_i}{x_i}\right)$
7	Aitchison's Distance	$d_{AD}\left(x,y\right)=\frac{1}{D}\sum_{i<j}\left(\log\frac{x_i}{x_j}-\log\frac{y_i}{y_j}\right)$ $d_{AD}^2\left(x,y\right)=\sum_{k=1}^D\left(\log\frac{x_i}{g(x_j)}-\log\frac{y_i}{g(y_j)}\right)$
8	Cosine Distance	$d_C\left(x,y\right)=1-\frac{\sum_i x_iy_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$
9	Mahalonbis Distance	$d_m^2\left(x,y\right)=\left(x-y\right)^T S^{-1}\left(x-y\right)$

Algorithm 2 Outlier Detection Algorithm

```
Perform K-means Algorithm (Algo. 1 )
INPUT: D-Dimensional Data $X_n, n = 1, ..., N$, No of Clusters and choose distance metrics (Aitchison Distance) for K-means.
Find distance between obtained centers and points using distance metrics.
Sort in descending order the distance obtained.
$d_{max} = max_i\left\{\|x_i - c_i\|\right\}, i = 1...N$
Highest distance between center and points is an outlier.
```

2.3 Experiments with Synthetic and Real Data

2.3.1 Synthetic Data sets:

We have taken synthetic data set to validate our proposed method. We have generated samples of data using 2 different mixtures of Dirichlet distributions by using different α parameters. This synthetic data set consists of 400 vectors with a dimension of 100. The Dirichlet distribution can be expressed as:

$$p\left(X|\alpha\right)=\frac{1}{\beta\left(\alpha\right)}\prod_{i=1}^{K}x_i^{\alpha_i-1} \tag{2}$$

$$\beta\left(\alpha\right)=\frac{\prod_{i=1}^{k}\Gamma\left(\alpha_i\right)}{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)} \tag{3}$$

Over here in above equation $\alpha=(\alpha_1,...,\alpha_K)$. When we performed K-means clustering with Euclidean and Aitchison distance matrices the results obtained are shown by confusion matrices in Table 2.2. it is clear that Aitchison distance is much better for proportional data clustering.

Table 2.2: Confusion matrices when clustering synthetic data set using Euclidean and Aitchisons distance

K-means Euclidean distance)

	Yes	No
Yes	155	45
No	200	0

K-means Aitchison distance

	Yes	No
Yes	90	110
No	105	95

2.3.2 Real Data sets:

In our experiments we have taken three real data sets for clustering. The data sets are:

- Data set 1 (Text Documents) (3)

  Instances: 3430

  Vocabulary words: 6906

  Number of words in collection: 467714.

  Number of Classes: 50

- Data set 2 (Spambase) (4)

  Instances: 4601

  Attributes: 57

  Labeled into into 2 groups (Spam and Non Spam Emails).

- Data set 3 : Human Face Identification (49)

  Number of Facial expression: 400 images

  Dictionary size: 1000

  classification is 40.

The first experiment is on document clustering using bag of words approach. We have taken the KOS blog entries (3) as our data set which contains 3430 documents and number of words in the vocabulary is 6906. The problem arises with value of zeros in the data-sets. So, before performing clustering we removed 0 values by using exponentially small value around $2^{(-52)}$. After processing, we normalize data using following equation:

$$x_i = \frac{x_i}{x_1 + x_2 .... + x_D} \tag{4}$$

The second experiment is on visual objects clustering using bag of visual words approach. The real image data set contains 400 images of 256 pixels each presenting facial expressions. In order to generate bag of visual words, SIFT descriptor (51) has been used. After this process, each image is represented as a proportional vector.

Figure 2.1: Data set 3 contains 400 facial images of 40 individuals, each showing 3 different but similar facial expressions. (49)

High dimensional clusters are very difficult to visualize. The best method to visualize high dimensional clustering is with the help of Silhouette. Peter J.Rousseew (48) explains how the cluster analysis can be done with the help of silhouette. Silhouette is a method which is used to find the consistency of clustered data. The values obtained for each vector can be combined and statistical values can be obtained to validate the clustering operation. The statistical values ranges between -1 to 1, and show how an object is well matched to its own cluster. Higher silhouette value means the clustering is appropriate. If a negative value is obtained it means more clustering can be done and the points does not lie within the same clusters. After finding the silhouette value it is necessary to find the group wise summary statistics which gives us a clear view of clustering. Table 2.5 shows the statistical values for clustering using different distance metrics of KOS blog entries. Our results for text clustering shows that the Aitchison distance metrics is most appropriate among all, as the higher silhouette value the more appropriate is the clustering. It is followed by Kullback distance metric and then Euclidean log distance metric. The Euclidean, Matisuita and the cosine show the poorest results for proportional data clustering, while, Cosine distance gives good results when data set is in form of frequency. For Spambase (4), we have obtained good results.

Table 2.3: Spambase data set clustering results using Euclidean and Aitchison distances

K-means Euclidean distance)

	Yes	No
Yes	345	1468
No	952	1836

K-means Aitchison distance

	Yes	No
Yes	219	1594
No	474	2314

2.3.3 Outlier detection results:

We have used Haberman's survival dataset for finding outliers. The data set consists of 306 instances and 3 attributes. Through the results obtained k-means with Aitchison distance was able to determine 5 outliers correctly whereas k-means with Euclidean distance determined 3 outliers correctly. Table 2.4 shows the confusion matrix after performing the experiment and figure 2.2 shows graphical output.

Table 2.4: The confusion matrices when clustering after outlier detection

K-means Euclidean distance)

	Yes	No
Yes	121	47
No	102	31

K-means Aitchison distance

	Yes	No
Yes	150	27
No	73	51

2.3.4 Error percentage:

To find error percentage Silhouette method is used as given by equation 5. In this equation, $a(i)$ is the average dissimilarity within the same cluster, $b(i)$ is lowest average dissimilarity of i to any other cluster where as i is a datum. It is used to find total sum of the silhouette values of each cluster on a data set. The error percentage is used to the dissimilarity of the clustered data as given by equation 6. It represents group wise statistics where mean of each cluster silhouette value is calculated and $\bar{x}$ gives the sum of the mean of silhouette value of each cluster. In equation 7, we find error percentage or dissimilarity percentage of K-means clustering based on distance metrics.

$$s\left(i,j\right) = \begin{cases} 1 - a\left(i\right) & : a\left(i\right) < b\left(i\right) \\ 0 & : a\left(i\right) = b\left(i\right) \\ b\left(i\right)/a\left(i\right) & : a\left(i\right) > b\left(i\right) \end{cases} \tag{5}$$

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} s\left(i,j\right) \tag{6}$$

$$e = \frac{N - \bar{x}}{N} \times 100 \tag{7}$$

The above process was performed on facial image data set (49) of 400 images which has 40 different classes of 10 images each. The error percentage or dissimilarity measure obtained by clustering into 40 different clusters was 24.75 % by k-means using Aitchison distance and 32.66 % by K-means using Euclidean distance. Table 2.5 provides the statistical values which are mean of each silhouette value that clearly determines Aitchison distance is much better distance metric when compared to other distance metrics.

S.No.	Cluster	Euc. Dist.	Ait. Dist.	KL. Div.	Cosine	EL. Dist	Mat. Dist.
1.	Cluster 1	-0.0795	1.000	0.2720	0.0047	0.4205	-0.1169
2.	Cluster 2	0.0477	-0.2244	-0.0805	-0.0994	0.2389	0.1666
3.	Cluster 3	0.1952	0	0.1569	0.0869	-0.2761	-0.0789
	Sum	0.1750	0.7756	0.3484	-0.0078	0.3832	-0.0293

Table 2.5: Statistical Values of different clusters by distance metrics for K Blog Entries Bag of words approach

2.4 Intrusion Detection System using Unsupervised Approach

In Section 2.4.1 of our chapter, we discuss same feature selection approaches and results obtained when applied to intrusion detection. In section 2.4.2, Dirichlet mixture model is discussed with Aitchison distance being applied on K-means. Section 2.4.3, gives the experimental results.

2.4.1 Feature Selection:

There are various feature selection methods example include: Stepwise Regression, Stability Selection, Significance Analysis for Micro arrays, Weight by Maximum relevance, Least Absolute Selection and Shrinkage Operator (LASSO), etc. While feature selection performs removal of relevant features, feature extraction transforms the attributes where transformed attributes are combination of the original ones. In this process linear dependency between the features is minimized and projection of original data is on new space. The common feature extraction methods are PCA (principal component analysis), ICA (independent component analysis), Multi factor dimensionality reduction, Latent semantic analysis, etc. A novel method for feature extraction in the case of proportional data was proposed in (41) using data separation by Dirichlet distribution. In our work, we have concentrated upon feature selection.

Weight by Maximum Relevance:

Weight by Maximum Relevance approach has been proposed in (10). It is a filter that measures the dependence between every feature x and the classification feature y (i.e., the label) using Pearson's linear correlation, F-test scores and mutual information (33) (10). The high score by mutual correlation reveals the features which are important. The NSL KDD Data set has 41 features and in order to reduce the complexity and finding an optimal solution we have reduced it to 16 features taking into an account that Weight by Maximum Relevance score of feature is $f \geq 0.05$. The output obtained is displayed in figure 2.5.

Weight by Maximum Relevance correlation vector can be defined by Pearson Correlation coefficient as:

$$R\left(i\right) = \frac{cov\left(X_i, Y\right)}{\sqrt{Var\left(X_i\right) Var\left(Y\right)}} \tag{8}$$

The equation can be written as:

$$R\left(i\right) = \frac{\sum_{k=1}^{M}\left(x_{k,i} - \bar{x}\right)\left(y_k - \bar{y}\right)}{\sqrt{\sum_{k=1}^{M}\left(x_{k,i} - \bar{x}\right)^2 \sum_{k=1}^{M}\left(y_k - \bar{y}\right)^2}} \tag{9}$$

This can only detect the linear dependency between variable and target (28).

Least Absolute Selection and Shrinkage Operator (LASSO)

Least Absolute Selection and Shrinkage Operator (LASSO) has been proposed in (55). It is a method which is used for estimation in Linear models. The method minimizes the residual sum of squares which is related to coefficient being less than a constant. It uses β as a checking vector which is a coefficient vector. It shrinks coefficients and set others to zero, therefore tries to retain the good features of both subset selection and ridge regression. It is given $(x_1, x_2, ..., x_D)$ and an outcome be y, the LASSO should fit linear model. The computation of LASSO is a quadratic problem and can be solved by standard numerical analysis algorithms. LASSO does shrinkage and variable selection where as in ridge regression only shrinkage takes place. The initial idea is to start working with large value of λ and slowly start decreasing it. The minimization for LASSO can be expressed as follow:

$$f = \sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{10}$$

In this equation y_i is the outcome variable, for cases $i = 1, 2, ..., n$ features x_{ij}, $j = 1, 2, ..., p$. Figure 2.6. represents feature selection by LASSO and reducing features to 16 features by taking into an account $f \geq 0.0053$.

2.4.2 Proposed Method

Let $\mathcal{X} = \left\{ \vec{X}_1, \vec{X}_2, ..., \vec{X}_N \right\}$ be a data set with N D-dimensional vectors modeled by a Dirichlet mixture model, then:

$$p\left(\vec{X}_i|\theta\right) = \sum_{j=1}^{M} p_j p\left(\vec{X}_i|\vec{\alpha}_j\right) \tag{11}$$

where $\vec{\alpha}_j$ is the parameter vector of component j, $\{p_j\}$ are the mixing proportions which should be positive and always sum to 1. $\theta = \{p_1, p_2, ..., p_M; \vec{\alpha}_1, \vec{\alpha}_2, ..., \vec{\alpha}_M\}$ is the complete set of parameters fully characterizing the mixture, and $M \geq 1$ is the number of components. Each Dirichlet distribution can be written in the form

$$p\left(\vec{X}_i|\vec{\alpha}_j\right) = \frac{1}{\beta\left(\alpha\right)} \prod_{d=1}^{D} X_{id}^{\alpha_{jd-1}} \tag{12}$$

$$\beta\left(\alpha\right) = \frac{\prod_{d=1}^{D} \Gamma\left(\alpha_{jd}\right)}{\Gamma\left(\sum_{d=1}^{D} \alpha_{jd}\right)} \tag{13}$$

where $X_{id} > 0$ $d = 1, 2, ..., D$, $X_{i1} + X_{i2}, ..., +X_{id} = 1$, and $\vec{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jD})$ represents parameter vector for j^{th} population. Let $\mathcal{X}$ be a data set with a common, but unknown, probability density function $p(\vec{X}_i|\theta)$ as given in above equation. We supposed that the number of mixture components is known. The ML estimation method consists of getting the mixture parameters that maximize log likelihood function. The below equation defines the posterior probability obtained after maximizing log likelihood function. This function is used in the E-step of Expectation Maximization (EM) algorithm.

$$p\left(j|\vec{X}_i, \vec{\alpha}_j\right) = \frac{p_j p\left(\vec{X}_i|\vec{\alpha}_j\right)}{\sum_{k=1}^{K} p_k p\left(\vec{X}_i|\vec{\alpha}_k\right)} \tag{14}$$

Now, using this expectation our goal is to maximize complete log likelihood. During the process we also have to ensure the constraints $p_j \geq 0$ as well as $\sum_{j=1}^{M} p_j = 1$. In maximization step of the algorithm, we have to update the parameters α until convergence to get the best result. As, it is to be noted that closed form solution for α does not exist. In the maximization step, iterative approach of newton raphson method has been used as explained in (16).

During the initialization we use K-means algorithm as given in Algorithm 1. We have used Euclidean and Aitchison distances. As, we know that Aitchison distance out performs Euclidean distance metric when proportional data are in question. In order to increase the performance of the algorithm, we have used feature selection methodology. In order to perform feature selection, the first step we have taken is to normalize the NSL KDD data set.

$$x_i = \frac{x_i}{x_1 + x_2 .... + x_D} \tag{15}$$

After obtaining proportional data, which act as an input for Weight by Maximum Relevance (WMR) proposed by Blum et al. (10) and Least Absolute Selection and Shrinkage Operator (LASSO) for selection of features from a data set.

Algorithm 3 EM Algorithm for Dirichlet Mixture Model

```
1: **Input:** Data set $\left(\vec{X}_1, \vec{X}_2, ..., \vec{X}_N\right)$ and specified number of components M.
2: Apply the k-means algorithm as given in Algorithm 1 on N $D$-dimensional vectors to obtain initial M clusters.
3: calculate $p_j = \frac{\text{Number of elements in class j}}{\text{N}}$
4: Apply moments method to obtain $\alpha$ parameters.
5: **Expectation-Maximization step** after Initialization
6: E-Step: Compute the posterior probability $p\left(j|\vec{X}_i, \vec{\alpha}_j\right)$
7: M-Step:
8: **repeat**:
9:     Update priors $p_j$ using equation 14 .
10:    Update the parameters $\alpha$ using Newton Raphson method.(16).
11: **until** : $p_j \leq \epsilon$, discard $j$ and go to E-Step.
12: if convergence test is passed then terminate, else go to E-Step.
```

2.4.3 Experiment with NSL KDD data set

We have taken NSL KDD 2009 data-set for performing Intrusion detection. The NSL KDD data set contains 2 classes which are normal and attack sets. The attacks can be divided into four parts which are: Denial of Service Attack (DoS), User to Root attacks (U2R), Remote to local attacks (R2L) and probing attacks. In our experiment we have taken only normal and attack sets into consideration without considering different types of attacks. In our methodology, we have used Dirichlet mixture model for clustering. While performing clustering using Dirichlet mixture model results into 51.12% of accuracy which was relatively increased to 53.44 % when clustering was performed with initialization of k-means using Aitchison distance. In our experiments, we have done feature selection using the methodology of Weight by maximum relevance where the number of features has been reduced to 16 instead of 41. The experiment on 16 features using Dirichlet mixture model with Euclidean distance in K-means during initialization results into 52.54 % accuracy and 56.37 % was obtained when initialization was done with K-means using Aitchison distance as shown by figure 2.7 and by table 2.8. To depict our results, we have used confusion matrix (see Table 2.6, 2.7). Accuracy is defined as percentage of correctly classified vectors:

$$Accuracy = 100 \times \frac{\text{Correctly identified vector}}{\text{total vectors}} \tag{16}$$

In our case we have used only test data without labels. Our results are better than SVM approach where accuracy determined is 51.90 % (33). The author in (25) obtained 47% which is comparably less than our approach.

Table 2.6: Confusion matrix of DMM, initialization with K-means using Euclidean and Aitchison distances

DMM (Euclidean distance)

	Yes	No
Yes	5386	1492
No	4300	672

DMM (Aitchison distance)

	Yes	No
Yes	4953	1564
No	3953	1380

Table 2.7: Confusion matrix of DMM after feature selection, initialization with K-means using Euclidean and Aitchison distance

Confusion Matrix FS WMR DMM (Euclidean distance)

	Yes	No
Yes	5587	1513
No	4111	639

Confusion Matrix DMM (Aitchison distance)

	Yes	No
Yes	5128	1285
No	3884	1553

S.No.	Method	Accuracy	Precision	Sensitivity
1.	DMM (Euclidean Distance)	51.12%	0.78	0.55
2.	DMM (Aitchison Distance)	53.44%	0.76	0.56
3.	FS WMR DMM (Euclidean Distance)	52.54%	0.78	0.58
4.	FS WMR DMM (Aitchison Distance)	56.37%	0.80	0.57

Table 2.8: Accuracy, precision and sensitivity obtained after applying different methods.

2.5 Bot detection using Mixture model for twitter tweets

In today's era, social media has taken its leap in every phase of life and become a part of everyday activities being elections, marketing, etc. It is an important tool for public policy as seen from examples such as Arab spring, japan earthquake, etc. Micro blogging websites like twitter, tumblr, etc are used to display smaller content which are useful in SEO (Search Engine Optimization) point of view. Twitter has become relatively popular as it gives real time information. Even major search engines like Google, Bing, etc uses twitter data for mining real time events happening around the world. As only 140 characters are allowed on twitter which are used by major companies to analyze these short messages being it media, marketing companies, etc. The trending topic on twitter is taken as an advantage by spammers or bot's to post pictures, tweets containing shortened URL which takes the user to unrelated websites. Hence, until now no proper mechanism has been found. In web services, phishing and malware attacks are the regular threats. The new methodologies in this domain by using unsupervised learning approach should be devised to counter such attacks. As, supervised learning consume lot of time and training cost. In the last decade there has been rapid shift from the static, editor-controlled Web 1.0 to the user-driven Web 2.0 paradigm (53). The web 2.0 allows us to post any type of content which becomes the target for spammers. The spam or bot has also been developed on mobile applications.

There is lot of research being done on twitter primarily on twitter sentiment analysis (37), (58) which is related to linguistics depicting emotions. There has been very little research conducted for detecting social spammers on twitter such as creation of duplicate accounts, automatic tweets at set amount of time which affects public opinion in making decision. The authors in (38) used machine learning approach to create classifier for identification of spam. Another research done by author in (57) used Bayesian classifier to detect spam and non spam tweets. The famous research done by author in (7) has collected billions of tweets from many different users and used SVM to classify spam or fake tweets.

All research done has focused on the use of supervised learning approaches. Supervised learning is usually not that much practical for online, ever changing, inconsistent data. Thus, we propose the use of unsupervised technique on such type of data to detect bots on twitter tweets of Colombian election. The results obtained are motivating and better than most of the supervised approaches proposed previously.

In this section, we use unsupervised approach using Dirichlet mixture model. Feature extraction method is used to reduce the features from the data set and they are reduced from 30 to 7.

2.5.1 Experimental Results

In our experiment we have compared our approach with Gaussian mixture model. We have taken Colombian election twitter tweets (19) and our main goal is to cluster the data in order to decide whether the tweet has been generated by the bot or it has been manually tweeted by the person. The bot tries to masquerade like humans and its quite difficult to identify them with present models. The tweets are converted from text to numeric form by following the bag of words (BoW) approach.

After we use feature extraction method to obtain the set of values which are linearly uncorrelated variables into an orthogonal plane. The method we use as feature extraction is Principal Component Analysis (35). The main goal to utilize feature extraction is to increase the efficiency of mixture model. Hence, feature extraction methodology is different from feature selection. In this case we don't lose any information and it is just projection of data into a plane. By using Principal Component Analysis we obtain negative scores which can be converted to positive without affecting the magnitude of the values by finding $val = max\left(x\right) - min\left(x\right)$ and dot product of the val with the data x. After which we use the normalizing equation to normalize the data again. The experiment shows that in unsupervised learning approach Dirichlet mixture model performs far better than Gaussian mixture model. The initial data we have consist of 1331 users with 30 attributes which consist of most recent tweets by the users. Over here we are taking 40 most recent tweets of each user and it can be depicted as follow:

The feature extraction step has reduced the features to 7. The various measures used to determine our results are as follows:

$$Accuracy = \frac{\text{TP+TN}}{TP + FP + TN + FN} \tag{17}$$

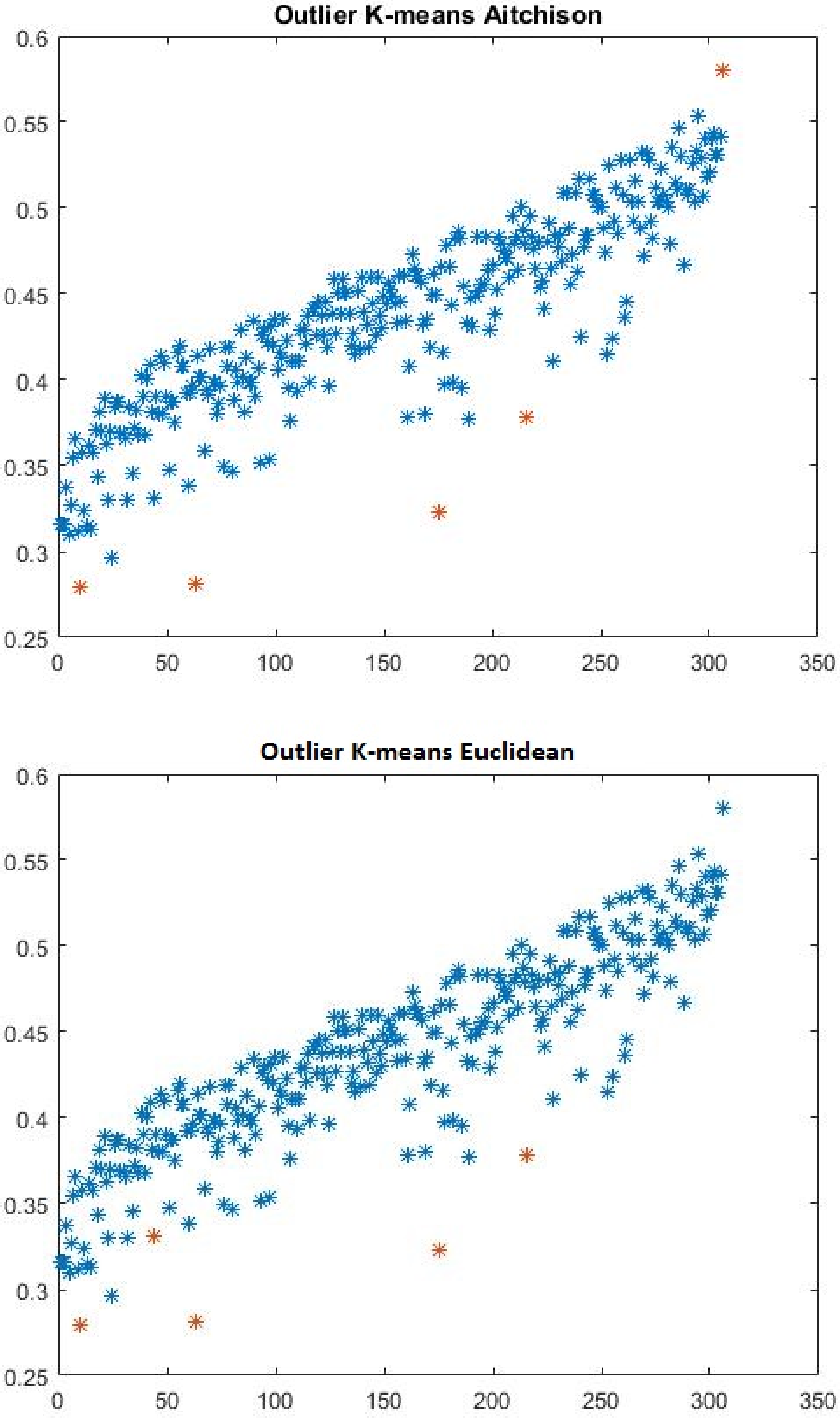
$$Precision = \frac{\text{TP}}{TP + FP} \tag{18}$$

$$Sensitivity = \frac{\text{TP}}{TP + FN} \tag{19}$$

The results obtained shows that accuracy has considerably increased by applying Principal component analysis. The Dirichlet mixture model outperforms Gaussian mixture model in terms of detecting bot tweets on collected twitter data. The reported result obtained by Dirichlet mixture model is 66.79% where as precision and sensitivity have also increased.

2.6 Conclusion

Different distance metrics have been investigated for the K-means clustering of proportional data. Aitchison distance has been shown to give the best performance. This distance can be used with different types of clustering methodologies. The proper initialization for mixture models is a crucial step in unsupervised learning. Using Aitchison distance as initialization step can give better mixture results applicable to proportional data. It has also been observed that Aitchison distance performs better for sparse data sets and for high dimension when compared with Euclidean distance for this particular type of data. By finding the K-means score it has been seen that Aitchison's distance is more viable solution as a distance metric for doing K-means clustering for proportional data. We have statistically analyzed the entire NSL KDD data set. In NSL KDD data set, 16 features had shown strong contribution for anomaly detection. Our basis is to state the baseline for unsupervised learning for future IDS solution


Outlier K-means Aitchison

Outlier K-means Euclidean

Figure 2.2: Outlier results of Haberman's survival Data set (2)

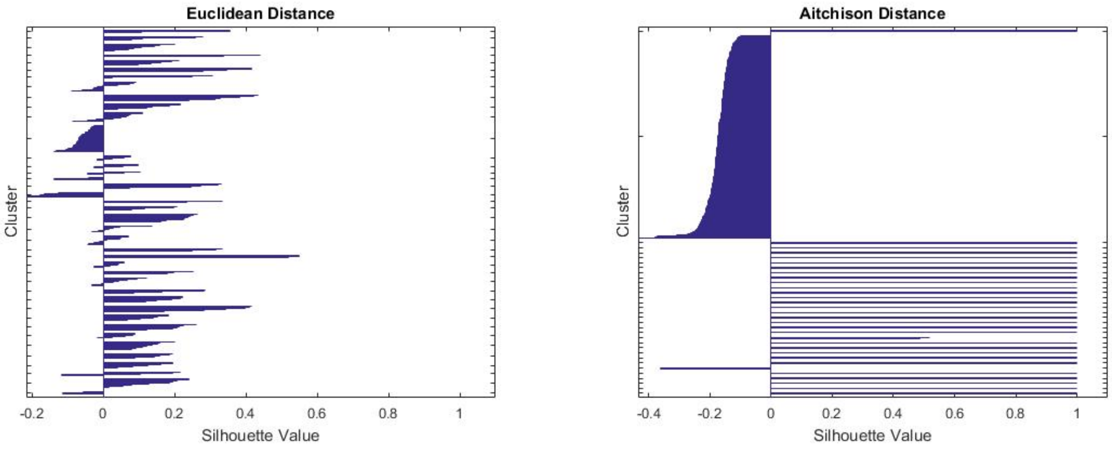


Figure 2.3: Silhouette value for each point with Euclidean and Aitchison distances when clustering 400 Image Dataset

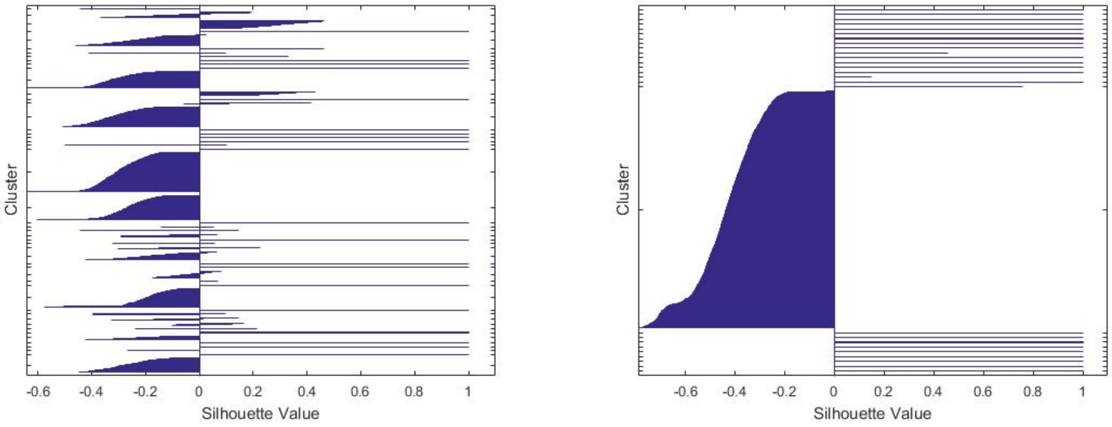


Figure 2.4: Silhouette value for each point with Euclidean and Aitchison distances when clustering text data. (N) is 50

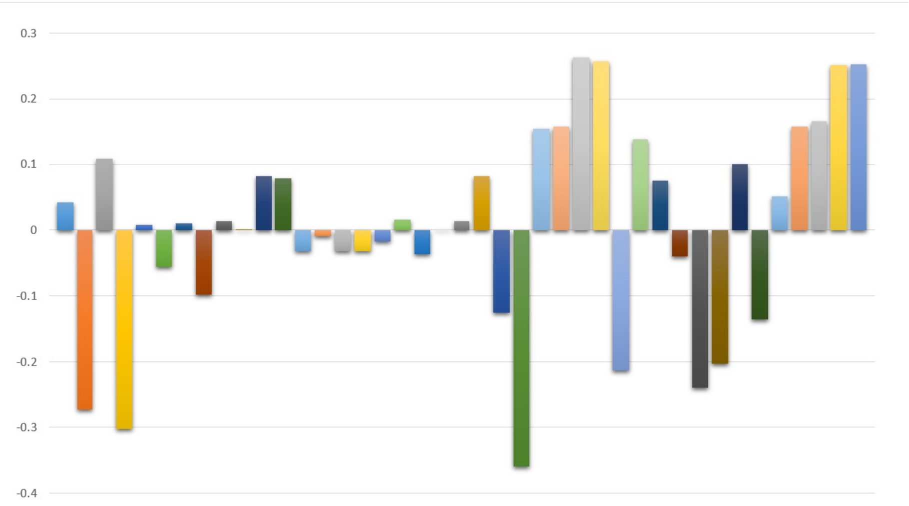


Figure 2.5: Score obtained after applying weight by maximum relevance feature selection technique.

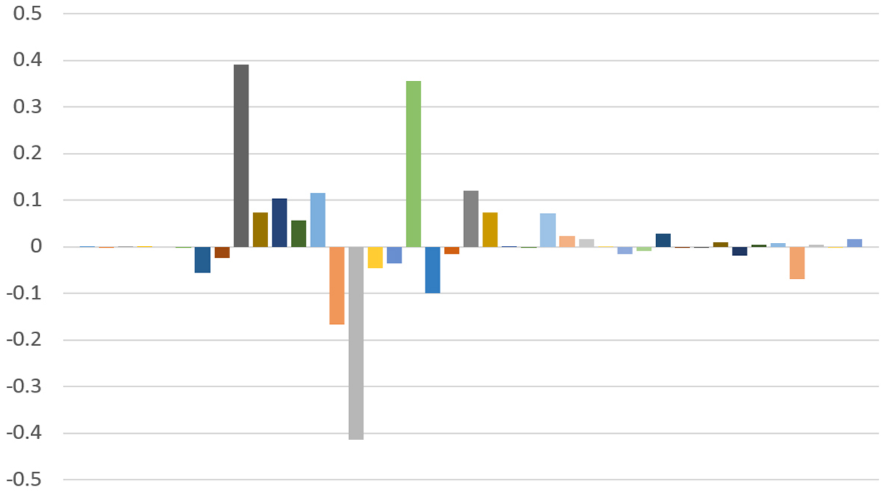


Figure 2.6: Score obtained after applying LASSO feature selection technique.

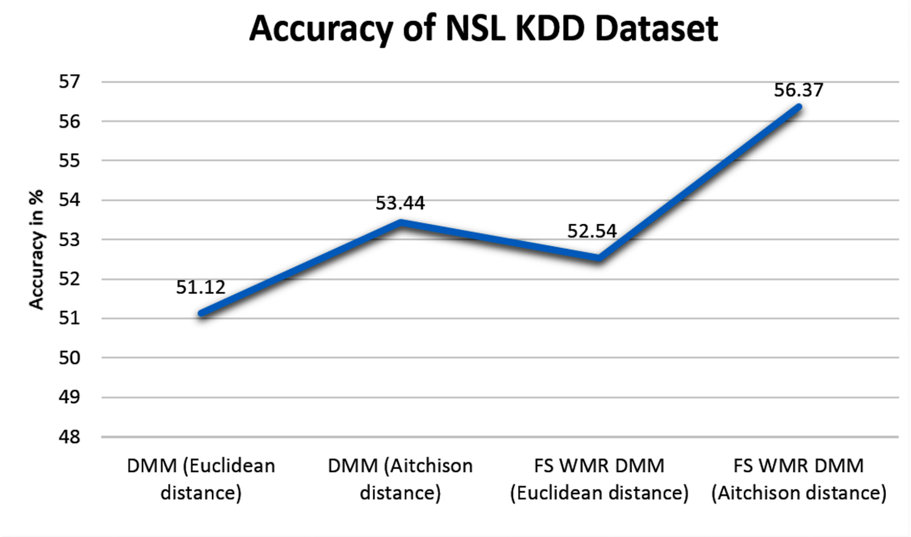


Figure 2.7: Accuracy of DMM model using different approaches.

S.No.	Features
1.	Profile description.
2.	Verified: when 1, ”indicates that the user has a verified account.”
3.	Age of the user account
4.	Number of followings.
5.	Number of followers
6.	reputation $= \frac{\text{(number of followers)}}{\text{(number of followings + number of followers)}}$
7.	User Mention
8.	Unique user mention (@) ratio.
9.	URL ratio.
10.	Hashtag () ratio.
11.	Average of tweet content similarity.
12.	Retweet rate.
13.	Reply rate.
14.	Number of tweets.
15.	Mean of inter-tweeting delay.
16.	Standard deviation of inter-tweeting delay.
17.	Average of tweets per day.
18.	Average of tweets per week.
19.	Number of tweets from manual devices.
20.	Number of tweets from automated devices.
21.	Fofo rate $= \frac{\text{number of followers}}{\text{number of followings}}$
22.	Following rate $= \frac{\text{number of followings}}{\text{age of the user account (in days)}}$
23.	Percentage of tweets posted during the 00:00:00 at 02:59:59 hours period.
24.	Percentage of tweets posted during the 03:00:00 at 05:59:59 hours period.
25.	Percentage of tweets posted during the 06:00:00 at 08:59:59 hours period.
26.	Percentage of tweets posted during the 09:00:00 at 11:59:59 hours period.
27.	Percentage of tweets posted during the 12:00:00 at 14:59:59 hours period.
28.	Percentage of tweets posted during the 15:00:00 at 17:59:59 hours period.
29.	Percentage of tweets posted during the 18:00:00 at 20:59:59 hours period.
30.	Percentage of tweets posted during the 21:00:00 at 23:59:59 hours period.

Table 2.9: Twitter Data set with different features

	A	B
A	216	382
B	416	317

Table 2.10: Confusion matrix obtained after applying Gaussian mixture model with 30 attributes

	Yes	No
Yes	337	221
No	303	430

Table 2.11: Confusion matrix obtained after applying Gaussian mixture model with 7 attributes obtained after PCA

	Yes	No
Yes	598	135
No	307	291

Table 2.12: Confusion matrix obtained after applying Dirichlet mixture model with 7 attributes obtained after PCA

S.No.	Process	Accuracy	Precision	Sensitivity
1.	GMM	40.00%	27.00%	34.17%
2.	GMM (7 Attributes)	57.62%	60.39%	52.65%
3.	DMM (7 Attributes)	66.79%	81.58%	66.07%

Table 2.13: Accuracy, Precision and Sensitivity obtained after applying different mixture models

Chapter 3

Spatially Constrained Mixture Models for Image Segmentation

The problem of image segmentation and grouping based on regions has remained as a great challenge in the field of computer vision. It is often the first step in variety of computer vision and image analysis tasks. There are various types of images we encounter in everyday life, for example, light intensity images, magnetic resonance image, etc (45). There has been a large number of approaches proposed in previous years for image segmentation. The problem of image segmentation of noisy images or corrupt images is still an open challenge. Image segmentation is widely used for anomaly detection (18) and medical image analysis (47). Various statistical models have been proposed in the past.

In this chapter, we develop an unsupervised approach for noisy image segmentation. Earlier similar unsupervised approaches have been used in research, for example, the author in (22) has used fuzzy c-means for medical image segmentation. In this chapter, we had the focus on a particular statistical model which is finite mixture. The main problems faced were 1) Integration of Markov Random Field (MRF) with Dirichlet based mixture models to achieve the segmentation of noisy images, 2) Estimation of parameters which is often a difficult task in mixture models, and 3) Initialization of parameters where we have used moments method with k-means. Markov Random Field has been heavily used for modeling spatial information for medical image segmentation

(27). Markov Random Field is heavily used for semantic segmentation of images in supervised learning approaches which is defined as multi-label classification problem. An interesting approach to integrate spatial information with Gaussian distribution has been proposed in (62). As we know, Gaussian mixture is a popular model in the field of computer vision. The Gaussian mixture model is restrictive as we have seen from previous works (52) (11) where Dirichlet and generalized Dirichlet distributions have generally shown better results.

Hence, In this chapter, we propose the integration of Markov Random Field in Dirichlet and generalized Dirichlet mixture models which are flexible (52) for data modeling. Experiments show that integrating spatial information into Dirichlet and generalized Dirichlet mixture models gives excellent results for image segmentation.

3.1 Dirichlet Mixture Segmentation approach

Let $\mathcal{X} = \left\{\vec{X}_1, \vec{X}_2, ..., \vec{X}_N\right\}$ representing a given image where N is the number of pixels and each pixel is denoted by random vector $\vec{X}_i = (X_{i1}, X_{i2}, ..., X_{iD})$. Now, the random vector $\vec{X}_i$ follows Dirichlet mixture model and is considered to be independent from the label. The density function can be presented as:

$$p\left(\vec{X}_i|\theta\right) = \sum_{j=1}^{M} p_j p\left(\vec{X}_i|\vec{\alpha}_j\right) \tag{20}$$

where $\vec{\alpha}_j$ is the parameter vector of component j which can be represented as $\vec{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jD})$. $\{p_j\}$ are the mixing proportions which should be positive and always sum to 1.
$\theta = \{p_1, p_2, ..., p_M; \vec{\alpha}_1, \vec{\alpha}_2, ..., \vec{\alpha}_M\}$ is the complete set of parameters fully characterizing the mixture, $M \geq 1$ is the number of components.

$$p\left(\vec{X}_i|\vec{\alpha}_j\right) = \frac{1}{\beta\left(\alpha\right)} \prod_{d=1}^{D} X_{id}^{\alpha_{jd-1}} \tag{21}$$

$$\beta\left(\alpha\right) = \frac{\prod_{d=1}^{D} \Gamma\left(\alpha_{jd}\right)}{\Gamma\left(\sum_{d=1}^{D} \alpha_{jd}\right)} \tag{22}$$

where $X_{id} > 0$, $d = 1, 2, ..., D$, $X_{i1} + X_{i2}, ... + X_{iD} = 1$, and $\vec{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jD})$ represents parameter vector for j^{th} component. The mean, variance and covariance of Dirichlet distribution are given as follows:

$$E\left(X_d\right) = \frac{\alpha_d}{|\vec{\alpha}|} \tag{23}$$

$$Var\left(X_d\right) = \frac{\alpha_d\left(|\vec{\alpha}| - \alpha_d\right)}{|\vec{\alpha}|^2 + 1} \tag{24}$$

$$Cov\left(X_i, X_j\right) = \frac{-\alpha_i\alpha_j}{|\vec{\alpha}|^2\left(|\vec{\alpha}| + 1\right)} \tag{25}$$

The image is composed of different regions. Thus it's appropriate to describe it by Dirichlet mixture model with M clusters as shown in equation 20. Moreover, for each pixel $\vec{X}_i \in \mathcal{X}$, there is a peer which has been arisen from the same cluster of $\vec{X}_i$. This spatial information can be used indirectly to estimate the number of clusters or regions in an image.

3.1.1 Segmentation approach

Let $\mathcal{X} = \left\{\vec{X}_1, \vec{X}_2, ..., \vec{X}_N\right\}$ be a data set of N D-dimensional positive vectors with a common, but unknown, probability density function $p(\vec{X}_i|\theta)$ as given in 20. The vectors are modeled as statistically independent, the joint conditional density of data can be presented as:

$$f\left(\mathcal{X}|\theta\right) = \prod_{i=1}^{N} p\left(\vec{X}_i|\theta\right) = \prod_{i=1}^{N}\sum_{j=1}^{M} p_j p\left(\vec{X}_i|\vec{\alpha}_j\right) \tag{26}$$

As we have taken each pixel to be independent of pixel label, the spatial correlation between nearby pixels is not taken into account. So, in this case, the image is relatively very sensitive to noise and illumination (50). To overcome the problem of noise and illumination the MRF (Markov Random Field) is used to create the spatial correlation between label values. The MRF distribution is given by:

$$f\left(\Pi\right) = Z^{-1}\exp\left\{-\frac{1}{T}U\left(\Pi\right)\right\} \tag{27}$$

In MRF, Z is considered as normalizing constant, T is a Temperature constant and $U\left(\Pi\right)$ is the smoothing prior where $\Pi = p_j$. The Bayes rule for posterior probability can be represented as:

$$f\left(\theta|\mathcal{X}\right) \propto f\left(\mathcal{X}|\theta\right) f\left(\Pi\right) \tag{28}$$

The log likelihood is given as follow:

$$L\left(f\left(\theta|\mathcal{X}\right)\right)=\log\left(f\left(\theta|\mathcal{X}\right)\right)=\sum_{i=1}^{N}\log\left\{\sum_{j=1}^{M}p_jp\left(\vec{X}_i|\vec{\alpha}_j\right)\right\}+\log f\left(\Pi\right) \tag{29}$$

$$L\left(f\left(\theta|\mathcal{X}\right)\right)=\sum_{i=1}^{N}\log\left\{\sum_{j=1}^{M}p_jp\left(\vec{X}_i|\vec{\alpha}_j\right)\right\}-\log Z-\frac{1}{T}U\left(\Pi\right) \tag{30}$$

There has been vast research which has already been conducted for determining smoothing prior of MRF distribution. The smoothing prior determined in most research is complex and requires lot of computation time when combined with mixture models. The example of such kind of smoothing prior is given by (9):

$$U\left(\Pi\right)=\kappa\sum_{i=1}^{N}\sum_{m\in\delta_i}\left[1+\left(\sum_{j=1}^{M}\left(p_{ij}-p_{mj}\right)^2\right)^{-1}\right]^{-1} \tag{31}$$

where κ represents a constant value. In the above equation, Z and T are set to 1 ($Z=1$ and $T=1$). Due to the complexity of this equation, the M-step of EM algorithm cannot be applied directly to prior distribution p_j. Various smoothing priors were proposed, but the major drawback of all of them has been that they are not robust to noise. In order to overcome this difficulty, prior distribution has been considered. A novel factor was proposed by (44) as follows:

$$G_{ij}^{t}=\exp\left[\frac{\kappa}{2N_i}\sum\left(z_{mj}^{(t)}+p_{mj}^{(t)}\right)\right] \tag{32}$$

In this equation κ is the temperature value and hence, changing the temperature value determines noise reduction of the image. It is used to determine neighborhood pixels around the pixel X_i. As proposed by authors (44) G_{ij} is only dependent on value of posteriors at previous step (t) and priors value.

The smoothing prior has been proposed in order to overcome the deficiencies which were seen previously in smoothing prior. Hence, smoothing prior is given by:

$$U\left(\Pi\right) = -\sum_{i=1}^{N}\sum_{j=1}^{M} G_{ij}^{(t)} \log p_{ij}^{(t+1)} \tag{33}$$

Maximizing equation 30 we get the expanded equation with the hidden variable z_{ij}

$$\begin{aligned} L\left(f\left(\theta|\mathcal{X}\right)\right) = \sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}^{t} \left\{\log p_{j}^{(t+1)} + \log p\left(\vec{X}_i|\vec{\alpha}_j\right)\right\} - \log Z \\ + \frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{M} G_{ij}^{t} \log p_{j}^{(t+1)} \end{aligned} \tag{34}$$

The hidden variable can be expanded as

$$z_{ij}^{(t)} = \frac{p_j^{(t)} p\left(\vec{X}_i|\vec{\alpha}_j\right)}{\sum_{k=1}^{K} p_k^{(t)} p\left(\vec{X}_i|\vec{\alpha}_k^{t}\right)} \tag{35}$$

Hence, putting the value of $U\left(\Pi\right)$ in equation 30 and expanding the equation with Dirichlet distribution as well as setting normalizing constant Z and Temperature Value T to be proportional over here (Z=1 and T=1), we get:

$$\begin{aligned} Q\left(f\left(\theta|\mathcal{X}\right)\right) = \sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}^{t} \Bigg\{ \log p_j^{(t+1)} + \log\Gamma\left(\sum_{d=1}^{D} \alpha_{jd}^{t+1}\right) - \sum_{d=1}^{D} \log\Gamma\left(\alpha_{jd}\right)^{t+1} + \\ \sum_{d=1}^{D}\left(\alpha_{jd}^{t+1} - 1\right)\log\left(X_{id}\right) + \sum_{i=1}^{N}\sum_{j=1}^{M} G_{ij}^{(t)} \log p_j^{(t+1)} \end{aligned} \tag{36}$$

Now, In M-Step of Expectation Maximization algorithm, $Q\left(\theta|\vec{X}\right)$ is maximized using Newton-Raphson approach as proposed in (16). Hence, for the $\vec{\alpha}$ parameters we have:

$$\alpha_{jd}^{(t+1)} = \alpha_{jd}^{(t)} - H^{-1}\left(\alpha_{jd}^{(t)}\right) \times \left(\frac{\partial Q\left(\theta|\vec{X}\right)}{\partial \alpha_{jd}}\right) \tag{37}$$

In the above equation H is the Hessian matrix which requires the calculation of second and mixed derivatives as presented in (16). To satisfy the condition of $\sum_{j=1}^{M} p_j = 1$, we use the Lagrangian multiplier Λ, which gives:

$$p_j = \frac{z_{ij}^{(t)} + G_{ij}^{(t)}}{\sum_{m=1}^{k} \left(z_{im}^{(t)} + G_{ik}^{(t)} \right)} \tag{38}$$

Now, we have done the integration of MRF into Dirichlet mixture model. Hence, we can see from the equation that MRF distribution is affecting the prior distribution which indirectly affects the estimation of the parameters.

3.1.2 Intialization and Segmentation Algorithm

Parameter initialization is important task for mixture models when parameter estimation is done through the EM algorithm. Our intialization algorithm is done through K-means using Aitchison distance metric and followed by method of moments (MM) algorithm (17). It can be summarized as follows:

$$\alpha_d = \frac{\left(x'_{11} - x'_{21}\right) x'_{1d}}{x_{21} - \left(x'_{11}\right)^2} \quad d = 1, ..., D \tag{39}$$

$$\alpha_{D+1} = \frac{\left(x'_{11} - x'_{21}\right)\left(1 - \sum_{d=1}^{D} x'_{1d}\right)}{x'_{21} - \left(x'_{11}\right)^2} \tag{40}$$

$$x'_{1d} = \frac{1}{N} \sum_{n=1}^{N} x_{nd} \quad d = 1, ..., D+1 \tag{41}$$

$$x'_{21} = \frac{1}{N} \sum_{n=1}^{N} x_{n1}^2 \tag{42}$$

Thus, the proposed learning approach is summarized in algorithm 4:

3.2 Generalized Dirichlet Mixture Segmentation approach

It is known that Dirichlet distribution has its own limitations given by (14) (23) as it has negative covariance matrix. Over here, we have tried to improve our results by using the Generalized form

Algorithm 4 EM Algorithm Dirichlet Mixture Model with MRF

```
Apply K-means on image data points to obtain initial $k$ clusters for segmentation.
Initialization using Method of Moments as proposed by the author in (11) to obtain $\alpha$ parameters.
Use the image data points to update the mixture parameters.
E-Step: Compute the posterior probability $z_{ij}^{(t)}$
M-Step:
repeat:
    Update priors $p_j$ using equation 38 .
    Update the parameters $\alpha$ using Newton Raphson method (16).
until : $p_j \leq \epsilon$, discard $j$ and go to E-Step.
if convergence test is passed then terminate, else go to E-Step.
```

of Dirichlet distribution. It can be given as follows:

$$p\left(\vec{X}|\vec{\alpha}\right) = \sum_{d=1}^{D} \frac{\gamma\left(\alpha_d+\beta_d\right)}{\gamma\left(\alpha_d\right)\gamma\left(\beta_d\right)} X_d^{\alpha_d-1}\left(1-\sum_{i=1}^{d} X_i\right)^{\gamma_d} \tag{43}$$

for $\sum_{d=1}^{D} X_d < 1$ and $0 < X_d < 1$ for $d = 1, 2, ..., D$ where $\gamma_d = \beta_d - \alpha_{d+1} - \beta_{d+1}$ for $d = 1, ..., D-1$ where $\gamma_d = \beta_d - \alpha_{d+1} - \beta_{d+1}$ for $d = 1, 2, ..., D-1$ and $\gamma_d = \beta_d - 1$. Note that generalized Dirichlet distribution is reduced to a Dirichlet distribution when $\beta_d = \alpha_{d+1} + \beta_{d+1}$. The mean, variance and covariance can be shown as below:

$$E\left(X_d\right) = \frac{\alpha_d}{\alpha_d+\beta_d}\prod_{i=1}^{d=1}\frac{\beta_i+1}{\alpha_i+\beta_i} \tag{44}$$

$$Var\left(X_d\right) = E\left(X_d\right)\left(\frac{\alpha_d+1}{\alpha_d+\beta_d+1}\prod_{i=1}^{d=1}\frac{\beta_i+1}{\alpha_i+\beta_i}+1-E\left(X_d\right)\right) \tag{45}$$

$$Cov\left(X_i,X_j\right) = E\left(X_d\right)\left(\frac{\alpha_i}{\alpha_i+\beta_i+1}\prod_{k=1} i=1\frac{\beta_k+1}{\alpha_k+\beta_k}+1-E\left(X_i\right)\right) \tag{46}$$

The interesting applications for generalized Dirichlet distribution can be found in (59). The approach can be explained as follows. For each pixel $\vec{X}_i \in \mathcal{X}$, there is a peer which has been arisen from the same cluster of $\vec{X}_i$. This spatial information can be used indirectly to estimate the number of clusters or regions in an image.

3.2.1 Segmentation approach

In the following, we adopt the segmentation approach, based on generalized Dirichlet mixture models with Markov Random Field (MRF) for the introduction of spatial information. The density function can be presented as:

$$p\left(\vec{X}_i|\theta\right) = \sum_{j=1}^{M} p_j p\left(\vec{X}_i|\vec{\alpha}_j, \vec{\beta}_j\right) \tag{47}$$

where $\vec{\alpha}_j$, $\vec{\beta}_j$ is the parameter vector of component j which can be represented as $\vec{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jD})$ and $\vec{\beta}_j = (\beta_{j1}, \beta_{j2}, ..., \beta_{jD})$. $\{p_j\}$ are the mixing proportions which should be positive and always sum to 1. $\theta = \left\{p_1, p_2, ..., p_M; \vec{\alpha}_1, \vec{\alpha}_2, ..., \vec{\alpha}_M, \vec{\beta}_1, \vec{\beta}_2, ..., \vec{\beta}_M\right\}$ is the complete set of parameters fully characterizing the mixture, $M \geq 1$ is the number of components.

$$p\left(\vec{X}_i|\vec{\alpha}_j, \vec{\beta}_j\right) = \prod_{d=1}^{D} \frac{\Gamma\left(\alpha_{jd} + \beta_{jd}\right)}{\Gamma\left(\alpha_{jd}\right)\Gamma\left(\beta_{jd}\right)} X_{id}^{\alpha_{jd-1}} \left(1 - \sum_{K=1}^{d} X_{iK}\right)^{\gamma_{jd}} \tag{48}$$

where $X_{id} > 0$, $d = 1, 2, ..., D$, $X_{i1} + X_{i2}, ... + X_{iD} = 1$, and $\vec{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jD})$, $\vec{\beta}_j = (\beta_{j1}, \beta_{j2}, ..., \beta_{jD})$ represents parameter vector for j^{th} component. In this case, where $\gamma_{jd} = \beta_{jd} - 1$, note that the generalized Dirichlet distribution is reduced to Dirichlet distribution (12) when $\beta_{jd} = \alpha_{jd+1} + \beta_{jd+1}$, $d = 1, ..., D$. Let $\mathcal{X} = \left\{\vec{X}_1, \vec{X}_2, ..., \vec{X}_N\right\}$ be a data set of N D-dimensional positive vectors with a common, but unknown, probability density function $p(\vec{X}_i|\theta)$ as given in above equation. The vectors are modeled as statistically independent, the joint conditional density of data can be presented as:

$$f\left(\mathcal{X}|\theta\right) = \prod_{i=1}^{N} p\left(\vec{X}_i|\theta\right) = \prod_{i=1}^{N} \sum_{j=1}^{M} p_j p\left(\vec{X}_i|\vec{\alpha}_j, \vec{\beta}_j\right) \tag{49}$$

As we have taken each pixel to be independent of pixel label, the spatial correlation between nearby pixels is not taken into an account. So, in this case, the image is relatively very sensitive to noise and illumination (50). To overcome the problem of noise and illumination the MRF (Markov Random Field) distribution is used as shown in 27.

The log likelihood for generalized Dirichlet distribution is given as follow:

$$L\left(f\left(\theta|\mathcal{X}\right)\right)=\log\left(f\left(\theta|\mathcal{X}\right)\right)=\sum_{i=1}^{N}\log\left\{\sum_{j=1}^{M}p_{j}p\left(\vec{X}_{i}|\vec{\alpha}_{j},\vec{\beta}_{j}\right)\right\}+\log f\left(\Pi\right) \tag{50}$$

$$L\left(f\left(\theta|\mathcal{X}\right)\right)=\sum_{i=1}^{N}\log\left\{\sum_{j=1}^{M}p_{j}p\left(\vec{X}_{i}|\vec{\alpha}_{j},\vec{\beta}_{j}\right)\right\}-\log Z-\frac{1}{T}U\left(\Pi\right) \tag{51}$$

There has been vast research which has already been conducted for determining smoothing prior of MRF distribution. The smoothing prior determined in most research is complex and requires lot of computation time when combined with mixture models. The example of such kind of smoothing prior is given by (9) which is shown in equation 31 and 33.

Maximizing equation 51 we get the expanded equation with the hidden variable z_{ij}

$$\begin{aligned}L\left(f\left(\theta|\mathcal{X}\right)\right)=\sum_{i=1}^{N}\sum_{j=1}^{M}z_{ij}^{t}\left\{\log p_{j}^{(t+1)}+\log p\left(\vec{X}_{i}|\vec{\alpha}_{j},\vec{\beta}_{j}\right)\right\}-\\ \log Z+\frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{M}G_{ij}^{t}\log p_{j}^{(t+1)}\end{aligned} \tag{52}$$

where

$$\begin{aligned}\log\left(p\left(\vec{X}_{i}|\vec{\alpha}_{j},\vec{\beta}_{j}\right)\right)=\sum_{d=1}^{D}\left[\log\Gamma\left(\alpha_{jd}+\beta_{jd}\right)\right]-\sum_{d=1}^{D}\left[\log\left(\Gamma\left(\alpha_{jd}\right)\Gamma\left(\beta_{jd}\right)\right)\right]\\ +\sum_{d=1}^{D}\left[\alpha_{jd-1}\log X_{id}-\gamma_{jd}\sum_{K=1}^{d}\log X_{iK}\right]\end{aligned} \tag{53}$$

The hidden variable can be expanded as

$$z_{ij}^{(t)}=\frac{p_{j}^{(t)}p\left(\vec{X}_{i}|\vec{\alpha}_{j},\vec{\beta}_{j}\right)}{\sum_{k=1}^{K}p_{k}^{(t)}p\left(\vec{X}_{i}|\vec{\alpha}_{k}^{(t)},\vec{\beta}_{k}^{(t)}\right)} \tag{54}$$

Hence, putting the value of $U\left(\Pi\right)$ in equation 51 and expanding the equation with generalized Dirichlet distribution as well as setting normalizing constant Z and Temperature Value T to proportional over here (Z=1 and T=1) we get:

$$Q\left(f\left(\theta|\mathcal{X}\right)\right) = \sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}^{t}\left\{\log p_j^{(t+1)}\right\} + \sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}^{t}\left\{\log\left(\Gamma\left(\alpha_{jd}+\beta_{jd}\right) - \log\left(\Gamma\left(\alpha_{jd}\right)\Gamma\left(\beta_{jd}\right)\right)\right)\right\} \\ + \sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}^{t}\left\{\alpha_{jd-1}\log X_{id} - \gamma_{jd}\sum_{K=1}^{d}\log X_{iK}\right\} + \sum_{i=1}^{N}\sum_{j=1}^{M} z_{ij}^{t}\left\{G_{ij}^{(t)}\log p_j^{t+1}\right\} \tag{55}$$

Now, In M-Step of Expectation Maximization algorithm, $Q\left(\theta|\vec{X}\right)$ is maximized using Newton-Raphson approach as proposed in (15). Hence, for the $\vec{\alpha}$, $\vec{\beta}$ parameters we have:

$$\begin{pmatrix} \alpha_{jd} \\ \beta_{jd} \end{pmatrix}^{(t+1)} = \begin{pmatrix} \alpha_{jd} \\ \beta_{jd} \end{pmatrix}^{(t)} - H^{-1} \times \begin{pmatrix} \frac{\partial Q\left(\theta|\vec{X}\right)}{\partial \alpha_{jd}} \\ \frac{\partial Q\left(\theta|\vec{X}\right)}{\partial \beta_{jd}} \end{pmatrix}^{(t)} \tag{56}$$

In the above equation H is the Hessian matrix which requires the calculation of second and mixed derivatives as presented in (15). To satisfy the condition of $\sum_{j=1}^{M} p_j = 1$, we use the Lagrangian multiplier Λ. Using the above methods of prior probability which gives:

$$p_j = \frac{z_{ij}^{(t)} + G_{ij}^{(t)}}{\sum_{m=1}^{k}\left(z_{im}^{(t)} + G_{ik}^{(t)}\right)} \tag{57}$$

Now, we have done the integration of MRF into generalized Dirichlet mixture model. Hence, we can see from the equation that MRF distribution is affecting the prior distribution which indirectly affects the estimation of the parameters.

3.2.2 Initialization and Segmentation Algorithm

Parameter initialization is an important issue in mixture models. The generalized Dirichlet mixture model initialization model has been proposed in (23). The integration of spatial information using MRF in generalized Dirichlet mixture model is summarized in algorithm 5:

Algorithm 5 EM Algorithm for generalized Dirichlet Mixture Model with MRF

```
Apply K-means on image data points to obtain initial $k$ clusters for segmentation as shown in Algorithm 1.
Initialization using Method of Moments as proposed by the author in (15) to obtain $\alpha$, $\beta$ parameters.
Use the image data points to update the mixture parameters.
E-Step: Compute the posterior probability $z_{ij}^{(t)}$
M-Step:
repeat:
    Update priors $p_j$ using equation 57 .
    Update the parameters $\alpha$, $\beta$ using Newton Raphson method (15).
until : $p_j \leq \epsilon$, discard $j$ and go to E-Step.
if convergence test is passed then terminate, else go to E-Step.
```

3.3 Experimental results

The main goal of this section is to investigate the performance of proposed method of Dirichlet and Generalized mixture models with Markov Random Field as compared with one developed by (44). The author in (44) has developed the model for Gaussian mixture with spatially constrained information. The work has not been carried out for Non-Gaussian Mixture models which give relatively better results of image segmentation of noisy images. As, we have considered Dirichlet and Generalized model where Dirichlet mixture model is a special case of the generalized Dirichlet mixture model. Evaluating segmentation results is an important problem and over here we are using NPR (Normalized Probabilistic Rand) (56) which can be given as follows:

$$\text{NPR Index} = \frac{\text{PR Index - Expected Index}}{\text{Maximum Index - Expected Index}} \tag{58}$$

The Expected value of PR Index can be given as follow:

$$E\left[PRI\left(S_{test}, \{S_k\}\right)\right] = \frac{1}{\binom{N}{2}} \sum_{\substack{i,j \\ i<j}} \left[p'_{ij} p_{ij} \left(1 - p'_{ij}\right)\left(1 - p_{ij}\right)\right] \tag{59}$$

This comparison model was proposed in (56) in order to provide comparison between image segmentation algorithms. In the experiment, the image is taken and converted to gray-scale after that we have induced three types of noise in an image which are: Gaussian noise, Poisson Noise and Salt and Pepper. The Gaussian noise image de-noising was earlier proposed by (32) who used soft threshold shrinkage method of sparse components. Poisson noise is also called as shot noise which is correlated with each pixel of an image. We add the Gaussian noise but Poisson is applied. Adaptive median filter with specialized regularization method has been used to reduce salt and pepper impulsive noises (20). We have observed that our method performs very well on all cases and image segmentation takes place without any difficulty and even the proposed method of Dirichlet and generalized Dirichlet distribution is better than median based filters.

In our experiment conducted, we have set the temperature value $(\kappa = 10)$. The other important factor in this equation is the determination of window size $(N_i = 25)$. The data set used is Berkeley Segmentation Data set 500 (BSDS500) which is an extension of Berkeley Segmentation data set 300 and BioId face database . These are publicly available data sets and heavily used in the field of computer vision. It is difficult to calculate NPR index of every image as this process is computationally expensive so we have calculated NPR index of a limited number of images followed by different segmentation approaches. The experimental results show that there is a large difference between the two different mixture models used. Tables 3.1, 3.2 and 3.3 show the NPR Index sample mean of images by different mixture models being performed on images. It can be seen that integrated model with Dirichlet and generalized Dirichlet performs way better than modified Gaussian mixture model. Fig 3.1 shows the comparison of image segmentation results obtained after applying modified Dirichlet mixture model with the modified Gaussian mixture model. Fig 3.2 shows the results of modified generalized Dirichlet mixture model with the modified Gaussian mixture model. Fig 3.3 shows the comparison of different images being segmented with different mixture models. All of these approaches are applied on noisy images.

	GMM	DMM	FRGMM	FRDMM
NPR Index Sample Mean	0.2833	0.4024	0.5462	0.6022

Table 3.1: NPR index sample for Gaussian Mixture model (GMM), Dirichlet Mixture model (DMM), Fast and Robust Gaussian mixture model (FRGMM) and Fast and Robust Dirichlet mixture model (FRDMM)

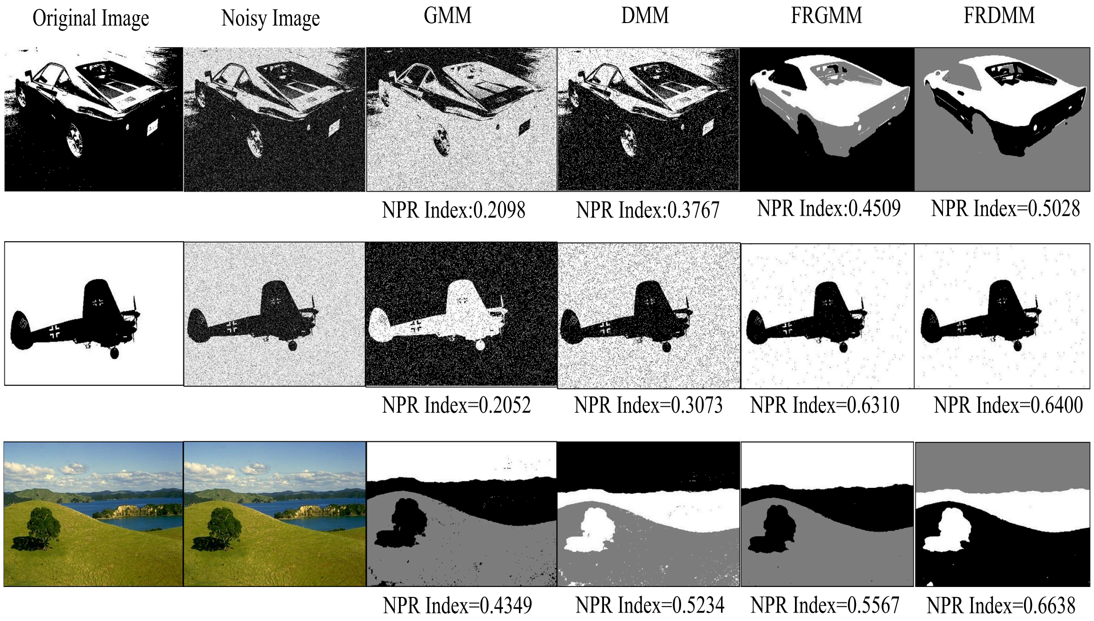


Figure 3.1: Segmentation of images from Berkeley 500 database. Column 1 gives the original image, column 2: Noisy Image, Column 3: Segmentation with Gaussian Mixture model, Column 4: Segmentation with Dirichlet Mixture model, Column 5: Segmentation with Markov Random field with Gaussian mixture model and Column 6: Proposed method with Dirichlet mixture model.

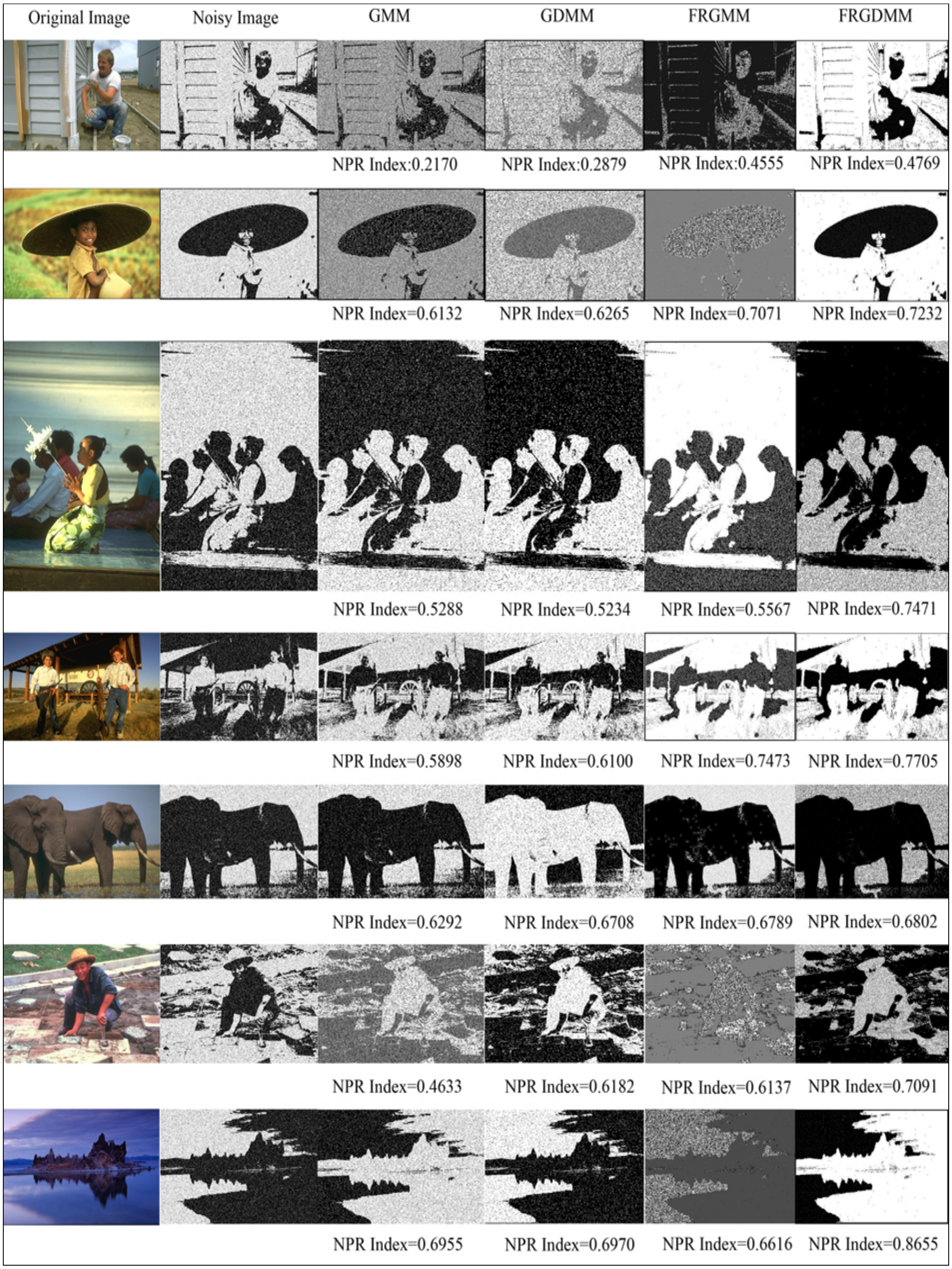


Figure 3.2: Segmentation of images from Berkeley 500 database. Column 1 gives the original image, column 2: Noisy Image, Column 3: Segmentation with Gaussian Mixture model, Column 4: Segmentation with generalized Dirichlet Mixture model, Column 5: Segmentation with Markov Random Field with Gaussian mixture model and Column 6: Proposed method with generalized Dirichlet mixture model.

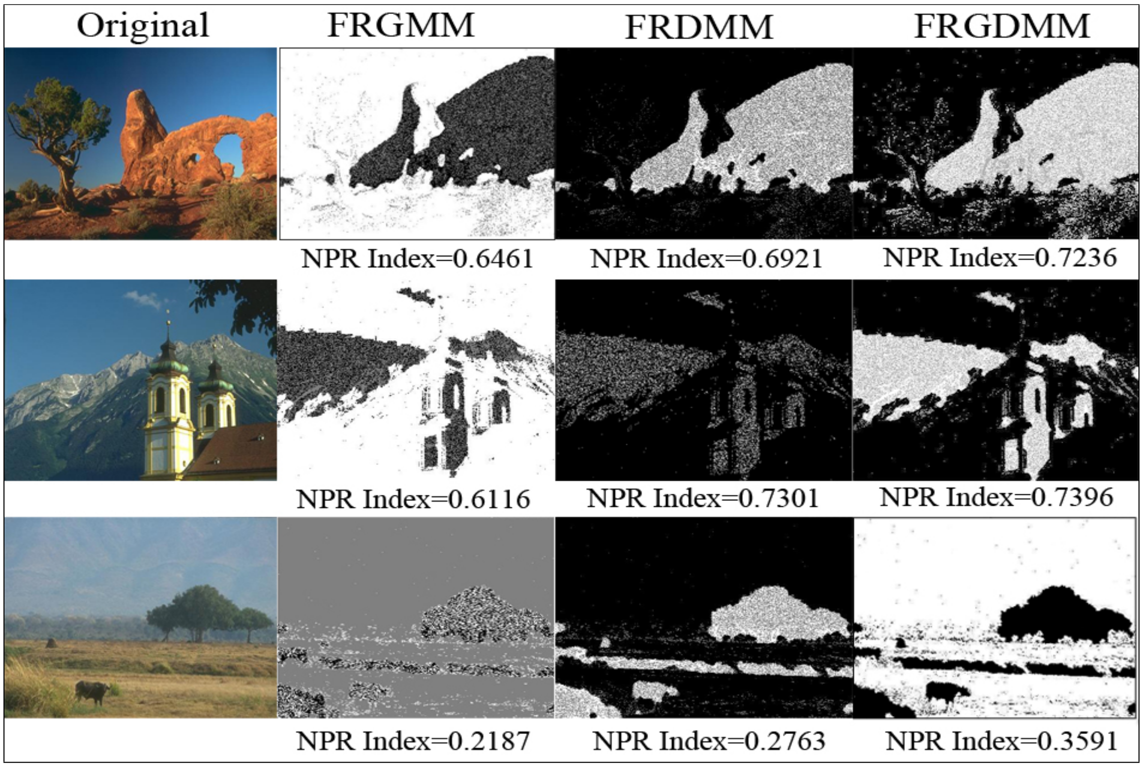


Figure 3.3: Segmentation of images from Berkeley 500 database. Column 1 gives the original image, Column 2: Segmentation with Markov Random Field with Gaussian mixture model, Column 3: Proposed method with Dirichlet mixture model and Column 4: Proposed method with generalized Dirichlet mixture model.

	GMM	GDMM	FRGMM	FRGDMM
NPR Index Sample Mean	0.4530	0.4864	0.6012	0.6499

Table 3.2: NPR index sample for Gaussian Mixture model (GMM), generalized Dirichlet Mixture model (GDMM), Fast and Robust Gaussian mixture model (FRGMM) and Fast and Robust generalized Dirichlet mixture model (FRGDMM)

	FRGMM	FRDMM	FRGDMM
NPR Index Sample Mean	0.4522	0.5771	0.6074

Table 3.3: NPR index sample for Fast and Robust Gaussian mixture model (FRGMM), Fast and Robust Dirichlet mixture model (FRDMM) and Fast and Robust generalized Dirichlet mixture model (FRGDMM)

3.4 Conclusion

In this chapter, we performed image segmentation based on Dirichlet and generalized Dirichlet mixture models with MRF (Markov Random field) to integrate the spatial information. It gave us good results when compared with modified Gaussian Mixture model. The generalized dirichlet mixture model is more flexible as it has two parameters when compared with Dirichlet mixture model. The selection of mixture model is motivated by its excellent results obtained when compared with other methodologies used in the past. The work can be extended for image segmentation with other mixture models and video segmentation being considered as another important application. Anomaly detection using video segmentation and learning approaches of the mixture can be done using for instance approaches previously proposed by (24). The drawback of a mixture model is the initialization of parameters as proper initialization becomes complex due to complexity of the mixture.

Chapter 4

Conclusion

In this thesis, we have presented different distance metrics which can be used with a K-means algorithm for the clustering of proportional data and yet, they have not been exploited. We have shown how proportional data can be clustered using Aitchison's distance which gives extremely good results. It was argued previously that Euclidean distance is not a universally defined distance to be used to measure distance between data points. We have shown how distance used with an initialization of Dirichlet distribution and generalized Dirichlet distribution gives better results. The above proposed method is exploited by using NSL-KDD data-set. This data-set is used for anomaly detection in network data. We have used feature selection methods such as LASSO and WMR after which mixture model is applied to get the results. The results had shown that Dirichlet mixture model works better when properly initialized.

In the second part of thesis, we have presented different algorithms for noisy image segmentation by integrating spatial information using Markov random field (MRF) into finite mixture models. The selection of mixture models is motivated by their flexibility for approximation of data points in different shapes where a well known Gaussian mixture model always keeps the symmetric bell shape. Firstly, we have taken a Dirichlet mixture model for its flexibility and lower number of parameters. We have initialized the parameters using a moments method by utilizing the Aitchison's distance metric in K-means. The main drawback of Dirichlet mixture model is its negative covariance matrix. This disadvantage is handled by generalized Dirichlet distribution where number of parameters is increased. Both mixture models are applied for image segmentation by integrating spatial information with MRF. To depict our results, we have used famous Berkeley Image data sets and our proposed algorithm performs better to deal with noise and illumination from an image. Future work can be devoted to an object detection and recognition. Video segmentation could be considered as another interesting application.

Abbreviations

GMM **G**aussian **M**ixture **M**odel

DMM **D**irichlet **M**ixture **M**odel

GDMM **G**eneralized **D**irichlet **M**ixture **M**odel

MRF **M**arkhov **R**andom **F**ield

PR **P**robabilistic **R**and

NPR **N**ormalized **P**robabilistic **R**and

KL **K**ullback **L**eibler

EL **E**uclidean **L**ogarithmic

LASSO **L**east **A**bsolute **S**election and **S**hrinkage **O**perator

WMR **W**eight by **M**aximum **R**elevance

FS **F**eature **S**election

Bibliography

[1] Darpa intrusion detection evaluation. http://www.ll.mit.edu/IST/ideval/data/dataindex.html. Last Accessed: 2016-11-05.

[2] Haberman's survival data set. https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival. Accessed: 2016-09-01.

[3] Kos blog enteries. http://dailykos.com. Accessed: 2016-08-10.

[4] Landsat satellite data set. https://archive.ics.uci.edu/ml/datasets/Spambase. Accessed: 2016-09-01.

[5] Nsl-kdd data set for network-based intrusion detection systems. http://nsl.cs.unb.ca/KDD/NSLKDD.html. Last Accessed: 2016-11-05.

[6] AITCHISON, J. The statistical analysis of compositional data.

[7] BENEVENUTO, F., MAGNO, G., RODRIGUES, T., AND ALMEIDA, V. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)* (2010), vol. 6, p. 12.

[8] BERTERO, M., POGGIO, T. A., AND TORRE, V. Ill-posed problems in early vision. *Proceedings of the IEEE 76*, 8 (1988), 869–889.

[9] BLEKAS, K., LIKAS, A., GALATSANOS, N. P., AND LAGARIS, I. E. A spatially constrained mixture model for image segmentation. *IEEE Transactions on Neural Networks 16*, 2 (2005), 494–498.

[10] Blum, A. L., and Langley, P. Selection of relevant features and examples in machine learning. *Artificial intelligence 97*, 1 (1997), 245–271.

[11] Bouguila, N., and Ziou, D. Dirichlet-based probability model applied to human skin detection [image skin detection]. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on* (2004), vol. 5, IEEE, pp. V–521.

[12] Bouguila, N., and Ziou, D. Unsupervised learning of a finite discrete mixture model based on the multinomial dirichlet distribution: Application to texture modeling. In *PRIS* (2004), Citeseer, pp. 118–127.

[13] Bouguila, N., and Ziou, D. A new approach for high-dimensional unsupervised learning: applications to image restoration. In *International Conference on Pattern Recognition and Machine Intelligence* (2005), Springer, pp. 200–205.

[14] Bouguila, N., and Ziou, D. A hybrid sem algorithm for high-dimensional unsupervised learning using a finite generalized dirichlet mixture. *IEEE Transactions on Image Processing 15*, 9 (2006), 2657–2668.

[15] Bouguila, N., and Ziou, D. A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized dirichlet mixture. *IEEE Trans. Image Processing 15*, 9 (2006), 2657–2668.

[16] Bouguila, N., Ziou, D., and Vaillancourt, J. Novel mixtures based on the dirichlet distribution: application to data and image classification. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition* (2003), Springer, pp. 172–181.

[17] Bouguila, N., Ziou, D., and Vaillancourt, J. Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application. *IEEE Transactions on Image Processing 13*, 11 (2004), 1533–1543.

[18] Brand, M., and Kettnaker, V. Discovery and segmentation of activities in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*, 8 (2000), 844–851.

[19] Cerón-Guzmán, J. A., and León, E. Detecting social spammers in colombia 2014 presidential election. In *Mexican International Conference on Artificial Intelligence* (2015), Springer, pp. 121–141.

[20] Chan, R. H., Ho, C.-W., and Nikolova, M. Salt-and-pepper noise removal by median-type noise detectors and detail-preserving regularization. *IEEE Transactions on image processing 14*, 10 (2005), 1479–1485.

[21] Chawla, S., and Gionis, A. k-means-: A unified approach to clustering and outlier detection. In *SDM* (2013), SIAM, pp. 189–197.

[22] Chuang, K.-S., Tzeng, H.-L., Chen, S., Wu, J., and Chen, T.-J. Fuzzy c-means clustering with spatial information for image segmentation. *computerized medical imaging and graphics 30*, 1 (2006), 9–15.

[23] Connor, R. J., and Mosimann, J. E. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association 64*, 325 (1969), 194–206.

[24] Epaillard, E., and Bouguila, N. Proportional data modeling with hidden markov models based on generalized dirichlet and beta-liouville mixtures applied to anomaly detection in public areas. *Pattern Recognition 55* (2016), 125–136.

[25] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*. Springer, 2002, pp. 77–101.

[26] Fan, W., and Bouguila, N. A variational component splitting approach for finite generalized dirichlet mixture models. In *Communications and Information Technology (ICCIT), 2012 International Conference on* (2012), IEEE, pp. 53–57.

[27] Greenspan, H., Ruf, A., and Goldberger, J. Constrained gaussian mixture model framework for automatic segmentation of mr brain images. *IEEE transactions on medical imaging 25*, 9 (2006), 1233–1245.

[28] GUYON, I., AND ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research 3*, Mar (2003), 1157–1182.

[29] HEBA, F. E., DARWISH, A., HASSANIEN, A. E., AND ABRAHAM, A. Principle components analysis and support vector machine based intrusion detection system. In *2010 10th International Conference on Intelligent Systems Design and Applications* (2010), IEEE, pp. 363–367.

[30] HIJAZI, R. An em-algorithm based method to deal with rounded zeros in compositional data under dirichlet models. In *Proceedings of the 4th International Workshop on Compositional Data Analysis* (2011), pp. 1–5.

[31] HURLBERT, A. C., AND POGGIO, T. A. A network for image segmentation using color. In *NIPS* (1988), pp. 297–304.

[32] HYVÄRINEN, A., HOYER, P. O., AND OJA, E. Sparse code shrinkage: Denoising by non-linear maximum likelihood estimation. *Advances in Neural Information Processing Systems* (1999), 473–479.

[33] IGLESIAS, F., AND ZSEBY, T. Analysis of network traffic features for anomaly detection. *Machine Learning 101*, 1-3 (2015), 59–84.

[34] JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters 31*, 8 (2010), 651–666.

[35] JOLLIFFE, I. *Principal component analysis*. Wiley Online Library, 2002.

[36] KASHIMA, H., HU, J., RAY, B., AND SINGH, M. K-means clustering of proportional data using l1 distance. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on* (2008), IEEE, pp. 1–4.

[37] KOULOUMPIS, E., WILSON, T., AND MOORE, J. D. Twitter sentiment analysis: The good the bad and the omg! *Icwsm 11*, 538-541 (2011), 164.

[38] LEE, K., CAVERLEE, J., AND WEBB, S. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010), ACM, pp. 435–442.

[39] Martín-Fernández, J., Barceló-Vidal, C., Pawlowsky-Glahn, V., Buccianti, A., Nardi, G., and Potenza, R. Measures of difference for compositional data and hierarchical clustering methods. In *Proceedings of IAMG* (1998), vol. 98, pp. 526–531.

[40] Martin-Fernandez, J. A., Palarea-Albaladejo, J., and Olea, R. A. Dealing with zeros. *Compositional data analysis: Theory and applications* (2011), 43–58.

[41] Masoudimansour, W., and Bouguila, N. Dimensionality reduction of proportional data through data separation using dirichlet distribution. In *International Conference Image Analysis and Recognition* (2015), Springer, pp. 141–149.

[42] Masoudimansour, W., and Bouguila, N. Generalized dirichlet mixture matching projection for supervised linear dimensionality reduction of proportional data. In *Multimedia Signal Processing (MMSP), 2016 IEEE 18th International Workshop on* (2016), IEEE, pp. 1–6.

[43] McHugh, J. Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Transactions on Information and System Security (TISSEC) 3*, 4 (2000), 262–294.

[44] Nguyen, T. M., and Wu, Q. J. Fast and robust spatially constrained gaussian mixture model for image segmentation. *IEEE transactions on circuits and systems for video technology 23*, 4 (2013), 621–635.

[45] Pal, N. R., and Pal, S. K. A review on image segmentation techniques. *Pattern recognition 26*, 9 (1993), 1277–1294.

[46] Panda, M., Abraham, A., and Patra, M. R. A hybrid intelligent approach for network intrusion detection. *Procedia Engineering 30* (2012), 1–9.

[47] Pham, D. L., Xu, C., and Prince, J. L. Current methods in medical image segmentation 1. *Annual review of biomedical engineering 2*, 1 (2000), 315–337.

[48] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20* (1987), 53–65.

[49] Samaria, F. S., and Harter, A. C. Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on* (1994), IEEE, pp. 138–142.

[50] Sanjay-Gopal, S., and Hebert, T. J. Bayesian pixel classification using spatially variant finite mixtures and the generalized em algorithm. *IEEE Transactions on Image Processing 7*, 7 (1998), 1014–1028.

[51] Scovanner, P., Ali, S., and Shah, M. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia* (2007), ACM, pp. 357–360.

[52] Sefidpour, A., and Bouguila, N. Spatial color image segmentation based on finite non-gaussian mixture models. *Expert Systems with Applications 39*, 10 (2012), 8993–9001.

[53] Tan, E., Guo, L., Chen, S., Zhang, X., and Zhao, Y. Unik: Unsupervised social network spam detection. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management* (2013), ACM, pp. 479–488.

[54] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A.-A. A detailed analysis of the kdd cup 99 data set. In *Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009* (2009).

[55] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[56] Unnikrishnan, R., Pantofaru, C., and Hebert, M. Toward objective evaluation of image segmentation algorithms. *IEEE transactions on pattern analysis and machine intelligence 29*, 6 (2007), 929–944.

[57] Wang, A. H. Detecting spam bots in online social networking sites: a machine learning approach. In *IFIP Annual Conference on Data and Applications Security and Privacy* (2010), Springer, pp. 335–342.

[58] WANG, H., CAN, D., KAZEMZADEH, A., BAR, F., AND NARAYANAN, S. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations* (2012), Association for Computational Linguistics, pp. 115–120.

[59] WONG, T.-T. Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation 97*, 2-3 (1998), 165–181.

[60] WONG, T.-T. Alternative prior assumptions for improving the performance of naive bayesian classifiers. *Data Mining and Knowledge Discovery 18*, 2 (2009), 183–213.

[61] WONG, T.-T. Parameter estimation for generalized dirichlet distributions from the sample estimates of the first and the second moments of random variables. *Computational Statistics & Data Analysis 54*, 7 (2010), 1756–1765.

[62] YANG, M.-S., AND TSAI, H.-S. A gaussian kernel-based fuzzy c-means algorithm with a spatial bias correction. *Pattern recognition letters 29*, 12 (2008), 1713–1725.

[63] YOON, K.-A., KWON, O.-S., AND BAE, D.-H. An approach to outlier detection of software measurement data using the k-means clustering method. In *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)* (2007), IEEE, pp. 443–445.

[64] ZARGARI, S., AND VOORHIS, D. Feature selection in the corrected kdd-dataset. In *Emerging Intelligent Data and Web Technologies (EIDWT), 2012 Third International Conference on* (2012), IEEE, pp. 174–180.