**Relevancy Index: Gaussian Mixture Model Feature Selection Embedded with EM Algorithm**

**Abstract**

Gaussian Mixture Model (GMM) is a popular clustering algorithm due to its neat statistical properties, which enable the “soft” clustering and the determination of the number of clusters. To estimate the GMM parameters, Expectation Maximization (EM) is applied. While promising, inclusion of features not contributing to clustering may confuse the model and increase computational cost. Recognizing the issue, this paper proposes a new algorithm, termed ESM by adding a feature selection step (S). Specifically, we introduce a relevancy index (*RI*), a metric indicating the probability of assigning a data point to a certain clustering group. The *RI* index reveals the contribution of the feature in the clustering process thus can assist the feature selection. We conduct theoretical analysis to justify the use of *RI* for feature selection. In addition, to demonstrate the efficacy of the proposed ESM, two synthetic datasets, two benchmark datasets and an Alzheimer’s Disease dataset are studied.

Key Words: Gaussian Mixture Model (GMM); Expectation Maximization (EM); feature selection

**1. Introduction**

Clustering is an unsupervised data mining technique to group the data into different segments by discovering the patterns within the dataset [1]. Data points from the same group are more like each other than the data from other groups. Due to the nature of the problem, clustering has been broadly applied to many fields including manufacturing [2], biology [3], finance [4] and astronomy [5], just to name a few.

Generally, clustering algorithms are categorized as hierarchical and partitional methods [1]. The hierarchical clustering starts by assigning each data point as its own cluster and obtaining the proximity distance matrix (e.g., single, complete link, group average) between each pair of data points (clusters). The algorithm then finds the closest pair of clusters to merge into a single cluster, and the new distances between the clusters are reassessed accordingly. Iteratively, the process ends until all data points are formed into one cluster as the root of the cluster tree.

The cluster tree, a.k.a. dendrogram is generated illustrating the cluster structures at different levels. Other than simplicity and intuitive visual presentation of the data, one known advantage of hierarchical clustering is its flexibility in deriving the clusters. That is, the user can visualize or use domain knowledge to decide the number of clusters as well as the level of the cluster structures to be investigated. However, as pointed out by [6], hierarchical clustering suffers from quadratic time complexity limiting its application to large dataset. In addition, for the data with noise, the performance of hierarchical clustering is unsatisfactory.

In traditional partitional methods such as k-means and its extensions k-prototype, the number of clusters usually needs to be pre-specified. Centroid, the center of each cluster is used to guide the assignment of the data point to the cluster. Centroid is iteratively updated during the clustering process. Due to its computational effectiveness, these partitional methods have great advantage over hierarchical algorithms for large scale datasets [1], [7]. However, it is a non-trivial task to determine the number of clusters. Often, it requires either domain knowledge as prior or empirical experiments to identify the appropriate number of clusters. In addition, k-means methods are known as “hard assignment” in which the data point is assigned to one specific group with certainty. However, during the clustering process, it is very likely that a data point is assigned to one cluster initially, and is reverted in the subsequent steps. The robustness of the model, especially to a noisy dataset, is thus questionable. Lately, mixture based partitional methods have attracted more attentions. Example are Gaussian mixture models [8], Dirichlet process mixture models[9]. These mixture models are built upon a well-studied statistical inference framework that provides guidelines for determining the optimal number of clusters. In these models, some statistical metrics derived from the data distribution instead of the distance between the data points are used in the modelling. The model uses the probability to describe the degree of the data point belonging to the cluster and the probability is updated iteratively during the clustering. Since most real world problems are uncertain by nature, the use of this “soft assignment” approach may be a better alternative comparing to the “hard assignment” (e.g., k-means). Indeed, Gaussian mixture model (GMM) is a generalized modelling approach and k-means is a special case of GMM [10].

While GMM has demonstrated its superior performance in handling noisy data in the field of image classification and segmentation [11], automatic speaker recognition [12], there exists some challenges in GMM research for high dimensional dataset [13]. It is noted that among the large number of features, some may not truly contribute to delineate the cluster profiles. Inclusion of these “noisy features” requires more parameter estimations for GMM, which are computationally costly. Additionally, the noisy features will confuse the model to identify the true structure of the clusters [14]. Recognizing this issue, researchers consider feature selections for solution. One approach is to conduct the feature selection independent from the model development. For example, Ghosh & Chinnaiyan [15] and Bernardo et al. [16] employ dimension reduction techniques such as principal component analysis. However, the leading components of PCA are linear combination of original variables, the use of PCA for subsequent clustering may destroy the clustering structure on the original data [17]. Research has also explored the univariate analysis to identify a subset of the features with statistical significances [18]. The study fails to consider the joint effects of multiple features in deriving the clusters. Often, some individually insignificant features may jointly play an important role in forming the clusters.

Here we contend separating feature selection from the clustering procedure is problematic as the statistical significance is not equivalent to clustering contributions. There is another important concept: relevancy vs. irrelevancy. In the clustering procedure, there are features directly contributing to the clustering, termed relevant features, and otherwise, irrelevant [19]. This concept leads to a second approach, embedding the feature selection into the clustering algorithm, e.g., expectation maximization (EM) algorithm. Law et al. [20] introduce a latent variable for each feature indicating whether the feature is relevant or not and the probability of the feature being relevant is defined as feature saliency. The posterior probability estimates of the latent variables are updated in the EM procedure. Others introduce a penalty term in the forms of log-likelihood function to regularize parameter estimation in the EM [21]–[23]. To obtain analytical solutions, both approaches require the diagonal covariance matrix for the EM implementation. In other words, both are under the strict assumption that the features are independent with each other. To relax assumption, Raftery & Dean [19] take a forward selection approach. The univariate analysis is launched to identify the first most important feature for the clustering as the initial relevant feature set. Remaining features, one by one, is evaluated via Bayes factor, the likelihood of adding vs. excluding the feature from the relevant feature set. The features excluded from the clustering form the irrelevant feature set. In calculating the Bayes factor, Raftery & Dean [19] assume the irrelevant features are dependent on the features from the relevant feature set. Maugis et al. [14] argue such assumption may not hold. Therefore, Maugis et al. [14] propose a backward stepwise strategy starting with all variables so the model takes block interactions between features into account. In both forward selection (Raftery & Dean [19]) and backward selection (Maugis et al. [14]), the comparison between two models: the model with and without the feature, is conducted to make the feature selection. For a dataset with D features, both approaches need to build at least number of models which raises the computational concern [24].

Delving into the detailed EM implementation, one interesting observation is responsibility metric calculated as an intermediate step. Responsibility is to measure the probability assigning the data point to a specific cluster. Given a feature, if the inclusion and the exclusion of the feature show no significant differences in responsibility measure, we conclude this feature is not truly contributing to the clustering. We define this responsibility difference as *RI*, the relevancy index for the feature. Based on *RI*, here we propose ESM algorithm, the assessment of the *RI* is embedded naturally in between the E and M steps in the EM implementation for feature selection. One advantage of our proposed ESM is that it follows GMM principle on data dependencies, it is generalized thus can handle the dataset with dependent and independent variables. The second advantage is that the proposed ESM is embedded in the EM procedure which guarantees the convergence.

The remainder of this study is organized as follows. Section 2 reviews the basics of GMM and EM algorithm. Section 3 presents the proposed ESM algorithm in detail with theoretical analysis on the *RI*. Next, three sets of experiments on two synthetic datasets, two benchmark datasets and one Alzheimer’s Disease dataset are illustrated in Section 4. In Section 5, the conclusion and future direction are presented.

**2. Review of GMM and EM**

A finite Gaussian mixture model, is the weighted sum of *K* Gaussian components (clusters) and can be written as

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where is the mean vector of component, is the covariance matrix of component, is the mixing coefficient representing the proportion of component, and is the probability distribution of component shown in Equation (2).

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

As in [25], we use a *K*-dimensional binary random variable having a one of *K* representation in which an element is equal to 1 and all other elements are equal to 0. The marginal distribution over is specified in terms of the mixing coefficient ,

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

where and . Suppose we have a data matrix with *N* data points and *D* features in which the row is . If the data points are drawn independently from the distribution, the log likelihood function is given by

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

In a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters including means , covariances and mixing coefficients , *k*=1, 2…, *K*. EM is a commonly used four-step algorithm to estimate these parameters. The algorithm starts from initializing , and , and evaluates the initial value of log likelihood function. In the second step (known as E step), the EM evaluates the responsibilities under the current parameter settings. The responsibility is defined as the probability of assigning a data point to a specific clustering group:

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

In the third step, the EM re-estimates the parameters given the responsibilities. The estimation method is maximum likelihood, and therefore the third step is called “M step”. The final step is to check the convergence of log-likelihood. If the log likelihood difference between two iterations is small enough (e.g., less than a small number), it is converged. Otherwise, the algorithm goes back to the E step initiating the next iteration. For details of the EM algorithm, interested readers are to refer to [25]. We present our proposed ESM algorithm based on responsibility metric in the next section.

**3. Proposed Method: ESM**

**3.1. Relevancy Index and ESM**

The proposed ESM algorithm takes advantage of the responsibility measures in the E step. Let us consider the responsibilities , the probability of assigning the data point to cluster *k*, if we remove one specific feature, responsibilities shall change. Specifically, let the full feature space with *D* features be , the feature space excluding feature be . Here, we denote the responsibility on the full feature space as and the responsibility on the reduced feature space (excluding feature) as which is related to the data point and the cluster. Relevancy index (*RI*) is defined as the difference between two responsibilities averaged over *N* data points and *K* clustering groups to evaluate the importance of feature to the clustering. It is written as:

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

The assumption behind our proposed method is that if is smaller than a pre-defined threshold, feature is neglectable in assigning data points to clusters. Thus, feature can be removed in feature selection process under the condition the converges over the iterations. Concerning the convergence criteria for , we evaluate the changes of between the current and previous iteration, let say, if it is less than a small number, e.g., 0.0005, we conclude converges. Regarding the pre-defined threshold for feature selection, it can be set based on the approximate number of features to be selected (see experiments in Section 4 for details).

Table 1 summarizes the ESM algorithm with the proposed S step highlighted.

Table 1 ESM Algorithm Pseudo Code

1. Initialize the means , covariances and mixing coefficients , and evaluate the initial value of the log likelihood.
2. **E step.**  Evaluate the responsibilities using the current parameter values

|  |  |  |
| --- | --- | --- |
|  |  |  |

and responsibilities after excluding each feature

where , and are the corresponding vector of , and after excluding variable.

1. **S step.** Calculate the difference between responsibilities before and after excluding feature at iteration t.

If (converged) and is small enough, then discard the feature with smallest and update the full feature space *F*.

1. **M step.** For reduced data with feature space *F*, re-estimate the parameters using the current responsibilities
2. Evaluate the log likelihood

|  |  |  |
| --- | --- | --- |
|  |  |  |

If the parameters or the log likelihood are not converged, go back to step 2.

**3.2. Theoretical Analysis on Relevancy Index**

As stated in [26], for a specific feature, when the variances are the same among clusters, if the mean of a cluster on the feature is equal to the global mean, this feature is uninformative or irrelevant to clustering. Intuitively, the bigger difference between the means, the more relevant the feature is. The following theorem gives the relationship between *RI* and difference of means which justifies the use of *RI* can identify the irrelevant features to be removed.

**Theorem 1.** Given a dataset with *D* features and *K* clusters, let the conditional mean of cluster on the feature given all the other features be , the conditional mean difference between two clusters and be , the lower bound of is an increasing function of for some . Specifically, if , .

**Proof**:

Let be a data point with *D* features and the variable is denoted as . Let be the corresponding vector of after excluding the feature. Let be the mean and covariance of cluster on full *D* feature dataset and and be the corresponding vector of and after excluding variable.

The joint distribution of *D* features can be decomposed into the joint distribution of all the other features except feature and the conditional distribution of feature. That is,

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

where is a linear function of .

Hence, we can rewrite as

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  | (8) |
|  |  |  |

Given any data point , where , there exists a component , which has the largest likelihood:

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

Let the ratio of likelihood between any two components, component *l* and component *k* be

|  |  |  |
| --- | --- | --- |
|  |  | (10) |

where . By Equation (8), we have

Expanding the Ratio in Equation (9), we obtain

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

Since we have

, which can be rewritten as =.

Therefore,

Hence the lower bound of is an increasing function of conditional mean difference of two clusters . If for any and conditional variances are the same for all clusters, the Ratio between any two components equals to 1. By equation (8), .

**End of Proof.**

We prove that an irrelevant feature having the same mean and variance among all clusters given all the other features has an *RI* being zero. In addition, the bigger mean difference between clusters can lead to a bigger lower bound value of the *RI*. If the lower bound of is small, the mean differences are small indicating the feature does not contribute to clustering. One advantage of *RI* over mean differences used in [26] is *RI* considers both mean difference as well as variance differences. As shown in Equation (8), the mean and the variances are both incorporated in the formula. Often, there are cases that the means from two clusters are the same, but the variances differ. Simply using the mean difference will not capture the features contributing. Here we use an illustration example to explain this idea.

[**Illustration example**] For a two-dimensional dataset X, there are 100 samples. The first feature is generated from normal distribution . The second feature is generated from two normal distributions with 50 samples from and another 50 samples from . In the setting, only the second feature contributes to clustering. If we use the mean differences, both clusters have [3, 1.5] as the mean, the conclusion is both features are not contributing to the clustering. Now, if we study the EM implementation, given some initialized mean and variances, the *RI* values are calculated by Equation 8, we get:

We conclude the second feature is contributing while the first feature should be treated as an irrelevant feature based on the *RI*. This simple example illustrates the advantages of *RI* over the mean differences. Some experiments to validate the efficacy of *RI* are discussed in the following sections.

**4. Experiments**

In the section, two synthetic datasets, two benchmark dataset and one medical application (Alzheimer’s Disease) dataset are studied to demonstrate the performance of proposed ESM algorithm. On the benchmark dataset, we compare the proposed algorithm with other existing algorithms from R software packages that do variable selection for model-based clustering. Please note our proposed ESM develops one model within which we decide the features to be kept or removed. In comparison, both the forward feature selection algorithm [19] and backward feature selection algorithm [14] need to develop a large number of models with each requiring the parameter estimations. For the dataset with large number of features, we expect the computational advantages of the proposed ESM will show. In addition, our experiments on the synthetic datasets indicate all three algorithms including ours are able to identify the relevant features, here we choose to report the experimental results from our proposed ESM comparing to EM in the following sections.

Since the ground truth of all the datasets are known, we use the following two metrics for the performance evaluation: (1) RFS: the percentage of relevant features being selected; (2) Accuracy: the percent of instances correctly clustered.

**4.1. Experiments on Synthetic Datasets**

We design both synthetic datasets with two relevant features, and two clusters. Additional irrelevant features are generated as “noise features”. In the first dataset, we have 10 features, all independent. As we argue dependencies among the feature prevail in most real-world data, we have 15 features in the second dataset with correlations added.

**4.1.1 Experiment I**

In this experiment, we get 10 features with and being the relevant features for clustering and other 8 features are irrelevant features. The relevant features are simulated from two-component mixture of Gaussian distributions with equal number of data points for each component. The irrelevant features are randomly generated from normal distributions. Since the data size may affect the performance of clustering performance, we create the 10 datasets from 100 to 1000 data points with 100 increment (see Table 2 for the experiment settings). For each setting, 10 experiment runs are conducted.

Table 2. Experiment setting for synthetic dataset I

|  |  |
| --- | --- |
| # of features | 2 relevant +8 irrelevant |
| # of groups/components | a mixture of 2 components |
| # of data points | 100, 200,300, 400,500,600,700,800,900,1000 |
| # of repeated runs | 10 |
| Distribution of relevant features for each component |  |
| Distribution of irrelevant features |  |
| # of repeated runs | 10 |

Figure 1 illustrates that the clustering performance varies with different data sizes. We observe that as the data size increases from 100 to 300, the clustering accuracy improves. The performance remains the same afterward. This concurs with the challenges commonly identified in the clustering research. That is, when the ratio between the number of data points vs. the number of features is small, it is difficult to derive the stable cluster structures. The second observation is that the standard deviation of the accuracy has similar pattern. However, the performance is relatively unstable (large standard deviation) for the experiments on 100 data points and 200 data points depending on the ESM initializations. In the 10 repeated runs, about three to five with good initializations are able to successfully identify the relevant features. Here we argue this should not be a concern since a run with a bad initialization tends to converge slower. It is our intention to design an initialization mechanism to improve the performance as a future task.

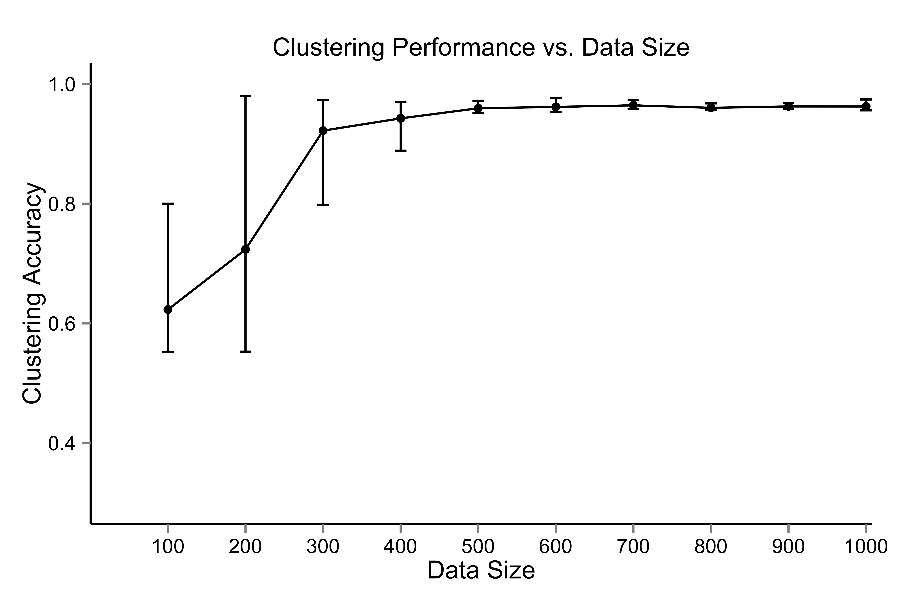


Figure 1. Synthetic dataset I: the clustering performance vs. data size

As shown in Figure 1, the ESM shows good performance when the ratio reaches 30 (300 data points, 10 features). We will use this experiment setting to illustrate the value of the *RI* in feature selection.

Figure 2 shows the evolving of the *RI* for each feature over the iterations. One distinct observation is both and have significantly higher values than the other 8 irrelevant features. In addition, the accuracy of ESM is 97.0% while the accuracy of EM (on the full feature set) is only 71.5%. We conclude ESM can identify the relevant features (independent features) and exclude the irrelevant “noisy features” resulting much improved clustering performance.



Figure 2. for each feature over iterations on 300 data points. The large values of for and in the final iteration show that and are critical for clustering.

**4.1.2 Experiment II**

In this experiment, we include 15 features with and being the relevant features for clustering and remaining 13 features being irrelevant features. The relevant features are generated from two-component mixture of Gaussian distributions with 225 data points for each component. In addition, we purposely add correlations between the features including ~ , ~ and ~ correlations. The experiment setting is summarized in Table 3. Similarly, 10 runs are conducted for each experiment.

Table 3. Experiment setting for synthetic dataset II

|  |  |
| --- | --- |
| # of features | 2 relevant + 13 irrelevant |
| # of groups/components | a mixture of 2 components |
| # of data points | 450 |
| # of repeated runs | 10 |
| Distribution of relevant features for each component |  |
| Distribution of irrelevant features |  |
| # of repeated runs | 10 |

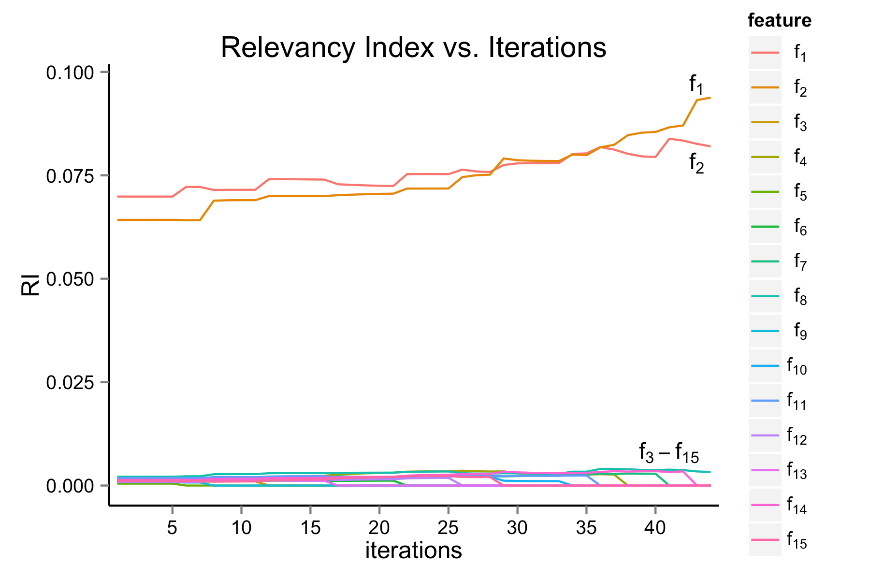


Figure 3. for each feature over iterations on 450 data points. The large values of for and in the final iteration show that and are critical for clustering.

Figure 3 shows both and have significantly higher values than the other 13 irrelevant features. In addition, the accuracy of ESM is 97.5% while the accuracy of EM on the full feature set is only 84.7%. We conclude ESM is able to identify the relevant features with dependencies and exclude the irrelevant “noisy features” resulting much improved clustering performance.

**4.2. Benchmark Dataset**

In this part, we compare the proposed algorithm with other existing algorithms. To get a fair comparison, we focus on the feature selection algorithms for model-based clustering. The survey paper [27] summarized six GMMs variable selection R packages named as sparcl[28], clustvarsel [29], VarSelLCM [30], vscc [31], SelvarMix [32], bclust [33]. Among the six methods, ‘sparcl’ mainly performs sparse hierarchical and sparse K-means clustering; the ‘bclust’ package requires a deliciated initial transformed model. Thus, we exclude these two methods and only compare the proposed method with the other four methods.

We will test the performance including accuracy and running time on two benchmark datasets called ‘Crab’ and ‘Wine’. The Crab dataset has 200 rows and 8 columns, among which five columns describe the morphological measurements. The remaining three columns are the color (orange and blue), sex (Female and Male) and index that divide the 200 crabs evenly into four groups by the combination of color and sex. The Wine dataset is a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The wine data contains three types of wines with 178 rows and 27 columns.

****

Figure 4. Clustering Accuracy and Running time on Crab Data. The proposed method has 2nd best accuracy but much lower running time than the algorithm with the best accuracy.

****

Figure 5. Clustering Accuracy and Running time on Wine Data. The proposed method has comparable accuracy and running time as ‘VarSelLCM’ and ‘vscc’.

For the Crab dataset, the proposed algorithm has the second-best accuracy but only uses less than half of the time spent by the algorithm with best accuracy. For the Wine dataset, the proposed algorithm is not the best but still has comparable accuracy and running time with the algorithm VarSelLCM. It is notable that the second algorithm ‘VarSelLCM’ gains the best performance on the Wine dataset (comparable accuracy and much lower running time) but worst performance (much lower accuracy compared with other algorithms) on the Crab dataset. This observation reveals the fact that the performance of the algorithm may depends on dataset. Fortunately, the proposed algorithm has a relative robust performance on both datasets.

**4.3. Real World Application: Alzheimer’s Disease**

Alzheimer’s Disease (AD) is a progressively neurodegenerative disease which is the most frequent type among elderly dementia patients. In the U.S., approximately 5.2 million people over 60 are afflicted by AD [34]. This drives a great amount of research investigating ways to slow down the AD progression and detect AD at early stage for better treatment or even prevent the disease. Mild cognitive impairment (MCI) is a syndrome defined as cognitive decline greater than expected for individuals during the course of aging but that does not interfere notably with activities of daily life [35]. It is an intermediate stage between normal aging with mild cognitive decline and dementia where cognitive impairment is more severe even impacting daily function. Though it is distinct from dementia, MCI patients with memory complaints and deficits (amnestic mild cognitive impairment) have high risks of progression to AD [35], [36]. The early diagnosis of MCI stage is becoming essential when the interventional strategies may be more effective.

Extensive research has investigated predictive model for AD in hoping to predict the risk of each individual patient converting to AD and this is still on-going effort. The focus of this research is to identify the underlying patient cohort structures which may discover patient subtypes for interventional treatment. In this study, we have collected 317 patients’ data from Alzheimer’s disease neuroimaging initiative [37], a large scale online repository designed to identify more sensitive and accurate methods to detect Alzheimer’s disease at earlier stage and mark its progress via biomarkers. Specifically, the baseline data is collected to evaluate the efficacy of our proposed ESM method in AD early detection. Among all these patients, 22 are AD, 172 are MCI and 123 are Normal Controls (NCs). For each patient, we obtain PET, and MR images and cognitive tests (see Table 4).

Table 4. Summary of features for Alzheimer’s Disease data.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Notation | Mean ± SE | Category |
| Age |  | 72.9 ± 7.3 | Demographic |
| Mini-Mental State Examination (MMSE) |  | 28.2 ± 2.3 | Cognitive Test |
| Clinical Dementia Rating (CDR) Score |  | 1.1 ± 1.6 | Cognitive Test |
| Volume of hippocampus |  | 7157 ± 1160 | MRI biomarkers |
| Volume of ventricles |  | 35086 ± 19333 | MRI biomarkers |
| Whole Brain |  | 1048732 ± 112296 | MRI biomarkers |
| Entorhinal |  | 3677 ± 741 | MRI biomarkers |
| Volume of Intracranial |  | 1514196 ± 156568 | MRI biomarkers |
| Hypometabolic Convergence Index (HCI) |  | 10.9 ± 5.6 | FDG-PET biomarkers |
| Statistical region of interest (sROI) |  | 1.2 ± 0.07 | FDG-PET biomarkers |
| mean cortical Standard Uptake Value Ratio with cerebellum as reference region (mcSUVRcere) |  | 1.1 ± 0.2 | F-AV45-PET biomarkers |
| mean cortical Standard Uptake Value Ratio with corpus callosum and centrum semiovale combined as reference region (mcSUVRwm) |  | 0.77 ± 0.17 | F-AV45-PET biomarkers |

Table 5 presents the correlation matrix of the features. Clearly, some features are highly dependent to each other, for example, the clinical test scores MMSE (*f*2) vs. CDR (*f*3), MRI biomarkers volume of hippocampus (*f*4) vs. whole brain (*f*6), FDG-PET biomarkers HCI (*f*9) vs. sROI (*f*10).

Table 5. Correlation Matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *f1* | *f2* | *f3* | *f4* | *f5* | *f6* | *f7* | *f8* | *f9* | *f10* | *f11* | *f12* |
| *f1* | 1.00 | -0.09 | -0.02 | -0.37 | 0.41 | -0.34 | -0.15 | 0.02 | 0.03 | -0.35 | 0.12 | 0.19 |
| *f2* |  | 1.00 | **-0.72** | 0.44 | -0.19 | 0.19 | 0.41 | 0.03 | -0.48 | 0.41 | -0.38 | -0.47 |
| *f3* |  |  | 1.00 | -0.41 | 0.22 | -0.14 | -0.42 | 0.02 | **0.55** | -0.45 | 0.38 | 0.48 |
| *f4* |  |  |  | 1.00 | -0.33 | **0.56** | **0.64** | 0.29 | -0.39 | **0.54** | -0.27 | -0.39 |
| *f5* |  |  |  |  | 1.00 | 0.06 | -0.11 | 0.45 | 0.28 | -0.43 | 0.11 | 0.44 |
| *f6* |  |  |  |  |  | 1.00 | 0.49 | **0.79** | 0.00 | 0.35 | -0.07 | -0.05 |
| *f7* |  |  |  |  |  |  | 1.00 | 0.30 | -0.27 | 0.38 | -0.19 | -0.25 |
| *f8* |  |  |  |  |  |  |  | 1.00 | 0.15 | 0.05 | 0.04 | 0.14 |
| *f9* |  |  |  |  |  |  |  |  | 1.00 | **-0.65** | 0.32 | 0.49 |
| *f10* |  |  |  |  |  |  |  |  |  | 1.00 | -0.30 | -0.41 |
| *f11* |  |  |  |  |  |  |  |  |  |  | 1.00 | **0.84** |
| *f12* |  |  |  |  |  |  |  |  |  |  |  | 1.00 |

One challenge in applying traditional clustering approach to the data is that some noise features degrade the performance of GMM. Table 6 shows the clustering results from original EM using all 12 features. The overall accuracy of correctly identifying patients to disease types is only 58.68%. For the AD cohort, EM clusters 20 out of 22 AD patients correctly (90.91%). For the NC cohort, EM though identifies 87 out of 123 correctly, 2 NCs are put into the AD group, and another 34 NCs are labeled as MCI. The accuracy of NC cluster is 70.73%. The results on MCI cohort is even worse with 79 out of 172 are correctly labeled, and 58 MCIs are put into the NC group, remaining 35 are grouped to AD cohorts resulting in 45.93% accuracy.

Table 6. Confusion matrix of EM using all features

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | GroundTruth |  |  |  |
|  |  | NC | AD | MCI | Overall accuracy |
| Clustering | NC | 87 | 0 | 58 |  |
| AD | 2 | 20 | 35 |  |
| MCI | 34 | 2 | 79 |  |
|  | Total | 123 | 22 | 172 |  |
|  | Accuracy | 70.73% | 90.91% | 45.93% | **58.68%** |

Applying ESM algorithm, three out of 12 features are selected: CDR(), HCI () and mcSUVRcere (). As shown in Figure 6, the three features have significantly higher *RIs* than other features. Using the three features, the overall clustering accuracy is 84.86% (Table 7). In comparing the results from Table 6 and Table 7, the accuracy of AD improves from 90.91% to 100%. The accuracy of NC improves from 70.73% to 92.68% where one NC is mislabeled as AD and 8 out of 123 NCs are labeled as MCIs. The accuracy of MCI improves from 45.93% to 77.33% with majority of the MCIs (133 out of 172) are correctly labeled, one MCI is put into the NC group and 38 MCIs are put to the AD group.

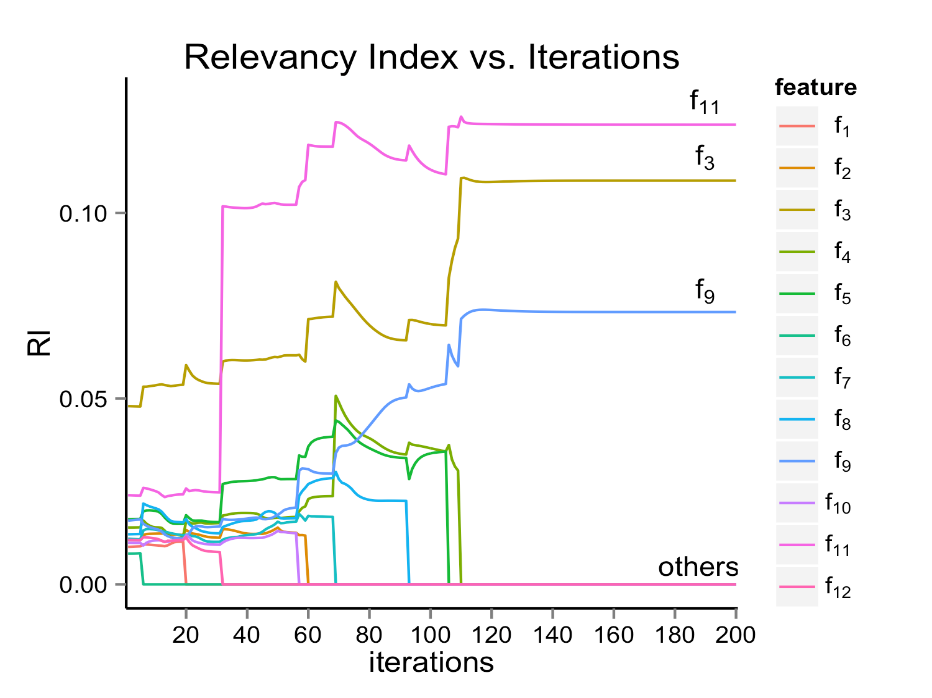


Figure 6. for each feature over iterations on AD data. The large values of for , and in the final iteration show that , and are critical for clustering.

Table 7. Confusion matrix for ESM clustering

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | GroundTruth |  |  |  |
|  |  | NC | AD | MCI | Overall accuracy |
| Clustering | NC | 114 | 0 | 1 |  |
| AD | 1 | 22 | 38 |  |
| MCI | 8 | 0 | 133 |  |
|  | Total | 123 | 22 | 172 |  |
|  | Accuracy | 92.68% | 100% | 77.33% | **84.86%** |

We want to emphasize that the results on MCI are not surprising. Clinically, MCI cohort has subtypes: MCI converter and MCI non-converter. MCI converter refers to the patient positively diagnosed as AD in the follow-up exams. Fortunately, ANDI is a rich data repository with longitudinal data available. We collect the updated patient staging information from the follow-up visit to explore the composition of the MCI group. For illustration purpose, we show a 2D plot of the two most relevant features CDR (*f*3) and mcSUVRcere (*f*11).

C:\Users\yinlinfu\Dropbox\ESM paper\figures\figure5.tif

Figure 7. Clustering shown in two feature space: CDR (*f*3) and mcSUVRcere (*f*11)

The points are colored by the real diagnosis results: AD (red), MCI (green) and NC (blue). The shapes signify the converters: triangle for NC converting to MCI, square for NC converting to AD and cross for MCI converting to AD. In Figure 7, 16 out of 26 green crosses are on the boundary between MCI and AD clusters, but close to AD. That is, among the 172 MCIs, 26 are staged as AD in the follow-up visit. Using baseline data, the 17 out of 38 MCIs mislabeled as AD (Table 7) are indeed the converted. The blue triangle represents the patient converting from NC to MCI. In the 7 blue triangles, there is one special point, which is diagnosed as NC in the first visit. However, in clustering, the point is closer to MCI than NC in Figure 7 and mislabeled as MCI in Table 7, which is verified by the diagnosis of MCI in the second visit. Indeed, the clustering technique can help to capture the convert between disease types.

**5. Conclusion and Future Work**

Gaussian Mixture Model (GMM), as a soft clustering methodology, has attracted great attention due to the distinct advantages from its statistical foundation. However, its performance deteriorates notably if the dataset has many noisy features irrelevant to the clustering process. This research proposes ESM algorithm based on a new metric: relevancy index (*RI*). Traditional EM algorithm for GMM modeling parameter estimation is extended with a S step using *RI* for feature selection. ESM preserves the good properties of EM algorithm such as guaranteed convergence and optimum determination of the clustering number. To evaluate the performance of ESM algorithm, experiments on two synthetic datasets (with independent features, with dependent features), two benchmark datasets and one Alzheimer’s Disease (AD) dataset are conducted. The experiments on synthetic dataset show ESM is able to identify the relevant features and improved clustering accuracy comparing to EM. The experiments on two benchmark datasets show ESM has competitive performance on accuracy and running time compared with existing algorithms. Other than improved clustering result, experiment on AD indicates ESM may potentially identify the patient subtypes which is crucial for patient treatment planning. While promising, the algorithm is limited for applications on more complex data such as mixed data with both continuous and categorical features. In the future, we may tackle the issue for more general data types. Also, for data with limited number of samples like 100 and 200 data size in simulation experiments, the performance of ESM algorithm depends on the initialization as in original EM algorithm. We may improve the robustness of ESM for small datasets.

**Conflict of Interest**

None declared.

**Acknowledgments**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Reference**

[1] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data clustering: a review,” *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[2] J. Yu, “Fault detection using principal components-based gaussian mixture model for semiconductor manufacturing processes,” *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 3, pp. 432–444, 2011.

[3] B. Mazoyer *et al.*, “Gaussian Mixture Modeling of Hemispheric Lateralization for Language in a Large Sample of Healthy Individuals Balanced for Handedness,” *PLoS One*, vol. 9, no. 6, p. e101165, Jun. 2014.

[4] A. Lindemann, C. L. Dunis, and P. Lisboa, “Probability distributions, trading strategies and leverage: An application of Gaussian mixture models,” *J. Forecast.*, vol. 23, no. 8, pp. 559–585, 2004.

[5] K. J. Lee, L. Guillemot, Y. L. Yue, M. Kramer, and D. J. Champion, “Application of the Gaussian mixture model in pulsar astronomy - pulsar classification and candidates ranking for the Fermi 2FGL catalogue,” *Mon. Not. R. Astron. Soc.*, vol. 424, no. 4, pp. 2832–2840, 2012.

[6] M. Steinbach, G. Karypis, and V. Kumar, “A Comparison of Document Clustering Techniques,” *KDD Work. text Min.*, vol. 400, pp. 1–2, 2000.

[7] S. Ben-David, “A framework for statistical clustering with constant time approximation algorithms for K-median and K-means clustering,” *Mach. Learn.*, vol. 66, no. 2–3, pp. 243–257, Mar. 2007.

[8] C. Fraley and a E. Raftery, “How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis,” *Comput. J.*, vol. 41, no. 8, pp. 578–588, 1998.

[9] D. B. Dahl, “Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model,” *Bayesian inference gene Expr. proteomics*, pp. 201–218, 2006.

[10] G. C. & G. Govaert, “A classification EM algorithm for clustering and two stochastic versions,” vol. 14, 1992.

[11] H. Permuter, J. Francos, and I. Jermyn, “A study of Gaussian mixture models of color and texture features for image classification and segmentation,” *Pattern Recognit.*, vol. 39, no. 4, pp. 695–706, 2006.

[12] D. a. Reynolds, “Automatic speaker recognition using Gaussian mixture speaker models,” *Lincoln Lab. J.*, vol. 8, no. 2, pp. 173–192, 1995.

[13] B. Thiesson, C. Meek, and D. Heckerman, “Accelerating EM for large databases,” *Mach. Learn.*, vol. 45, no. 3, pp. 279–299, 2001.

[14] C. Maugis, G. Celeux, and M. L. Martin-Magniette, “Variable Selection for Clustering with Gaussian Mixture Models,” *Biometrics*, vol. 65, no. 3, pp. 701–709, 2009.

[15] D. Ghosh and A. M. Chinnaiyan, “Mixture modelling of gene expression data from microarray experiments.,” *Bioinformatics*, vol. 18, no. 2, pp. 275–86, Feb. 2002.

[16] J. M. Bernardo *et al.*, “Bayesian Clustering with Variable and Transformation Selections,” *BAYESIAN Stat.*, vol. 7, pp. 249–275, 2003.

[17] C. Bouveyron, S. Girard, and C. Schmid, “High-dimensional data clustering,” *Comput. Stat. Data Anal.*, vol. 52, no. 1, pp. 502–519, 2007.

[18] G. J. McLachlan, R. W. Bean, and D. Peel, “A mixture model-based approach to the clustering of microarray expression data,” *Bioinformatics*, vol. 18, no. 3, pp. 413–422, Mar. 2002.

[19] A. E. Raftery and N. Dean, “Variable Selection for Model-Based Clustering,” *J. Am. Stat. Assoc.*, vol. 101, no. 473, pp. 168–178, 2006.

[20] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, “Simultaneous feature selection and clustering using mixture models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, 2004.

[21] W. Pan and X. Shen, “Penalized Model-Based Clustering with Application to Variable Selection,” *J. Mach. Learn. Res.*, vol. 8, pp. 1145–1164, 2007.

[22] S. Wang and J. Zhu, “Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data,” *Biometrics*, vol. 64, no. 2, pp. 440–448, Jun. 2008.

[23] B. Xie, W. Pan, and X. Shen, “Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables,” *Electron. J. Stat.*, vol. 2, pp. 168–212, 2008.

[24] C. Bouveyron and C. Brunet-Saumard, “Model-based clustering of high-dimensional data: A review,” *Comput. Stat. Data Anal.*, vol. 71, pp. 52–78, 2014.

[25] C. M. Bishop and N. Nasrabadi, “Pattern Recognition and Machine Learning,” *Pattern Recognit.*, vol. 4, no. 4, p. 738, 2006.

[26] P. D. Hoff, “Model-based subspace clustering,” *Bayesian Anal.*, vol. 1, no. 2, pp. 321–344, 2006.

[27] M. Fop and T. B. Murphy, “Variable Selection Methods for Model-based Clustering,” *Stat. Surv.*, vol. 12, no. 0, pp. 18–65, 2017.

[28] A. M. Daniela Witten and R. Tibshirani Maintainer Daniela Witten, “Package ‘sparcl’ Type Package Title Perform Sparse Hierarchical Clustering and Sparse K-Means Clustering,” 2018.

[29] L. Scrucca and A. E. Raftery, “clustvarsel: A Package Implementing Variable Selection for Model-based Clustering in R,” Nov. 2014.

[30] M. Marbac, M. S. Maintainer, and M. Sedki, “Package ‘VarSelLCM’ Title Variable Selection for Model-Based Clustering of Mixed-Type Data Set with Missing Values,” 2018.

[31] J. L. Andrews and P. D. Mcnicholas, “Package ‘vscc’ Title Variable selection for clustering and classification,” 2015.

[32] M. Sedki, G. Celeux, C. Maugis, and R. Maintainer, “Package ‘SelvarMix’ Type Package Title Regularization for Variable Selection in Model-Based Clustering and Discriminant Analysis,” 2017.

[33] P. Nia and A. C. Davison, “Package ‘bclust’ Type Package Title Bayesian Hierarchical Clustering Using Spike and Slab Models,” 2015.

[34] “2008 Alzheimer’s disease facts and figures,” *Alzheimer’s and Dementia*, Mar-2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1552526008000332. [Accessed: 18-Aug-2017].

[35] S. Gauthier *et al.*, “Mild cognitive impairment,” *Lancet*, vol. 367, no. 9518. pp. 1262–1270, 2006.

[36] M. Castro and G. E. Smith, “Mild cognitive impairment and Alzheimer’s disease.,” in *APA handbook of clinical geropsychology, Vol. 2: Assessment, treatment, and issues of later life.*, Washington: American Psychological Association, 2015, pp. 173–207.

[37] Michael W. Weiner MD, “ADNI Alzheimer’s Disease Neuroimaging Initiative,” 2013. [Online]. Available: http://www.adni-info.org/. [Accessed: 18-Aug-2017].