Feature Selection Based on the Rough Set Theory and Expectation-Maximization Clustering Algorithm

Farideh Fazayeli[1], Lipo Wang[1], and Jacek Mandziuk[2]

[1] School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore 639798
{fari0004,elpwang}@ntu.edu.sg
[2] Faculty of Mathematics and Information Science Warsaw University of Technology
Plac Politechniki 1,
00-661 Warsaw, Poland
mandziuk@mini.pw.edu.pl

Abstract. We study the Rough Set theory as a method of feature selection based on tolerant classes that extends the existing equivalent classes. The determination of initial tolerant classes is a challenging and important task for accurate feature selection and classification. In this paper the Expectation-Maximization clustering algorithm is applied to determine similar objects. This method generates fewer features with either a higher or the same accuracy compared with two existing methods, i.e., Fuzzy Rough Feature Selection and Tolerance-based Feature Selection, on a number of benchmarks from the UCI repository.

1 Introduction

The problem of reducing dimensionality has been investigated for a long time in a wide range of fields, e.g., statistics, pattern recognition, machine learning, and knowledge discovery. In order to reduce the input dimensionality, there exist two main approaches, i.e., feature extraction and feature selection (FS). Feature extraction maps the primitive feature space into a new space with a lower dimensionality. Two of the most popular feature extraction approaches include Principal Components Analysis [13], and Partial Least Squares [2]. There are numerous applications of feature extraction in the literature, such as image processing [9], visualization[29], and signal processing [21]. In contrast, the FS approach chooses the most informative features from the original features according to a selection method, e.g., t -statistic [17], f -statistic [15], correlation [34], separability correlation measure [7], or information gain [32]. The irrelevant and redundant features in the dataset lead to slow learning and low accuracy. Finding the subset of features that are enough informative is NP complete. Some heuristic algorithms are proposed to search through the feature space. The selected subset can be evaluated from some issues, such as the complexity of the learning algorithm and the accuracy.

C.-C. Chan et al. (Eds.): RSCTC 2008, LNAI 5306, pp. 272–282, 2008.

The Rough Set (RS) theory can be used as a tool to reduce the input dimensionality and to deal with vagueness and uncertainty in datasets. The reduction of attributes is based on data dependencies. The RS theory partitions a dataset into some equivalent (indiscernibility) classes, and approximates uncertain and vague concepts based on the partitions. The measure of dependency is calculated by a function of the approximations. The dependency measure is employed as a heuristic to guide the FS process. In order to obtain a significant measure, proper approximations of the concepts are required. Hence, the initial partitions play an important rule. Given a discrete dataset, it is possible to find the indiscernibility classes; however, in case of datasets with real-valued attributes, it is impossible to say whether two objects are the same, or to what extent they are the same, using the indiscernibility relation. A number of research groups [6, 20, 26, 27, 28, 30] extended the RS theory using the tolerant or similarity relation (termed tolerance-based Rough Set). The similarity measure between two objects is delineated by a distance function of all attributes. Two objects are considered to be similar when their similarity measure exceeds a similarity threshold value. Finding the best threshold boundary is both important and challenging. [14] used genetic algorithms to find the best similarity threshold. [8, 10, 22, 23, 25] used fuzzy similarity to cope with real-valued attributes. In this paper we use Expectation-Maximization (EM) [3, 5, 16, 24, 33, 35] clustering algorithm to determine the tolerance classes. The EM algorithm is a general statistical method for finding the maximum likelihood estimations of parameters in probabilistic models. In particular it can be applied in clustering problems. The EM algorithm allows for overlapping clusters and it is robust to noise and to highly skewed data.

The paper is organized as follows. Section 2 summarizes basics of the RS theory. A brief overview of the mixture model and EM algorithm is represented in section 3. In Section 4, the proposed method of feature selection using the RS theory and EM clustering algorithm is outlined. Section 5 shows the potential of the proposed method on some real datasets. We discuss our results and draw some conclusions in the final section.

2 Basics of the Rough Set Theory

Let $T(U, A, C, D)$ be a decision table, where U is a universe of objects, A is a set of primitive features, C is a set of conditional attribute, D is a decision attribute or class label, and $C, D \subseteq A$. For an arbitrary set $P \subseteq A$, an indiscernibility relation is defined as follows,

$$IND(P) = \{(x, y) \in U \times U : \forall a \in P, a(x) = a(y)\} \tag{1}$$

If $P \subseteq C$ and $X \subseteq U$ then the lower and upper approximations of X, with respect to P, are respectively defined as follow,

$$\underline{P}X = \{x \in U : [x]_{IND(P)} \subseteq X\} \tag{2}$$

$$\overline{P}X = \{x \in U : [x]_{IND(P)} \cap X \neq \phi\} \tag{3}$$

where

$$[x]_{IND(P)} = \{y \in U : a(y) = a(x), \forall a \in P\} \tag{4}$$

is the equivalence class of x in $U/IND(P)$.

A P-positive region of D is a set of all objects from the universe U which can be classified with certainty to one class of $U/IND(D)$ employing attributes from P,

$$POS_P(D) = \bigcup_{x \in U/IND(D)} \underline{P}X \tag{5}$$

A dependency of D on P is defined as,

$$\gamma_p(D) = \frac{|POS_P(D)|}{|U|}. \tag{6}$$

where $|A|$ is the cardinality of a set A.

A feature $a \in C$ is dispensable in P, if $\gamma_p(D) = \gamma_{p-a}(D)$; otherwise a is an indispensable attribute in P with respect to D. An arbitrary set $B \subseteq C$ is called independent if all its attributes are indispensable.

From these definitions a reduct set of features can be defined as follows, a set of features $R \subseteq C$ is called the reduct of C, if R is independent and $POS_R(D) = POS_C(D)$. In other words, the reduct is a set of attributes that conserves the partitions generated by C.

In [4] the QUICKREDUCT algorithm for determining the reduct set is proposed. It is a heuristic algorithm that avoids exhaustively generating all possible subsets. The greedy algorithm starts with an empty set and in each iteration adds the attribute that results in the greatest increase in the rough set dependency metric to the reduct set.

3 Mixture Model and EM Algorithm

The mixture model is an effective representation of the probability density function and consists of k component density functions. The objective of a mixture model is to fit the density functions to a given dataset to approximate the data distribution. The EM alogrithm can be used in solving the problem of the mixture models where Θ is the model parameters, and unknown-random variable $Y = \{y_i\}^N$ presents each object belongs to which model. That means $y_i = k$ if the $i - th$ object belongs to the component k. The EM algorithm allows for overlapping clusters hence each object can belong to more than one component. The EM algorithm is outlined in the Appendix.

Let D be a dataset with m objects and d attributes and $\mathbf{x} \in D$ be an object in the dataset. The mixture model probability density function, evaluated at $\mathbf{x}$, is defined as follows,

$$p(\mathbf{x}|\Theta) = \sum_{l=1}^{k} W_l.p(\mathbf{x}|\theta_l) \tag{7}$$

where

- W_l is the fraction of data points belonging to the cluster l, and $\sum_{l=1}^{k} W_l = 1$.
- $p(\mathbf{x}|\theta_l)$ is the cluster or component distribution models the records of the l-th cluster.
- θ_l is the model parameters of density function of cluster l. In case of Gaussian distribution, θ_l is the mean (μ_l) and covariance matrix (Σ_l).

The complete-data log-likelihood expression for this density from the data X and Y is given by:

$$\log(L(\Theta|X,Y)) = \log(P(X,Y|\Theta)) = \sum_{i=1}^{N} \log(P(x_i|y_i)P(y)) = \sum_{i=1}^{N} \log W_{y_i} p(x_i|\theta_{y_i}) \tag{8}$$

In this work a Gaussian distribution is used. The EM algorithm is used to determine the value of mean (μ_l), covariance matrix (Σ_l), and sampling probability (W_l) for each cluster. The attribute set will affect the distribution of data and lead to the different model parameters.

The algorithm is as follows,

1. **E Step.** For each object $\mathbf{x} \in D$, compute the membership probability of $\mathbf{x}$ in each cluster $l = 1 \cdots k$ at iteration j:

$$p(y_i|x_i, \boldsymbol{\mu}^j, \boldsymbol{\Sigma}^j) = \frac{W_{y_i}^j . p(x_i|\mu_{y_i}^j, \Sigma_{y_i}^j)}{p(x_i|\boldsymbol{\mu}^j, \boldsymbol{\Sigma}^j)} \tag{9}$$

2. **M Step.** Update mixture model parameters for each cluster $l = 1, 2, \cdots, k$ that maximize the value of $Q(\Theta, \Theta^{(j)})$:

$$W_l^{j+1} = \frac{1}{N} \sum_{\mathbf{x} \in D} pr(l|\mathbf{x}) \tag{10}$$

$$\mu^{j+1,l} = \frac{\sum_{\mathbf{x} \in D} \mathbf{x}.pr(l|\mathbf{x})}{\sum_{\mathbf{x} \in D} pr(l|\mathbf{x})} \tag{11}$$

$$\Sigma^{j+1,l} = \frac{\sum_{\mathbf{x} \in D} pr(l|\mathbf{x})(\mathbf{x} - \mu_{j+1,l})(\mathbf{x} - \mu_{j+1,l})^T}{\sum_{\mathbf{x} \in D} pr(l|\mathbf{x})} \tag{12}$$

3. If $|L^j - L^{j+1}| \leq \epsilon$, stop. Else set $j = j+1$ and go to 1. L^j is the log likelihood of the mixture model at iteration j,

$$L^j = \sum_{\mathbf{x} \in D} \log(pr^j(\mathbf{x})) = \sum_{\mathbf{x} \in D} \log(\sum_{l=1}^{k} W_l^j . pr^j(\mathbf{x}|\mu_l^j, \Sigma_l^j)) \tag{13}$$

4 Proposed Method

In the proposed method, each cluster represents a tolerance class. The tolerance classes that are generated by the EM clustering algorithm for an object x are defined as:

$$Clus_P(x) = \{Y \in U |\ x,\ and\ Y\ belongs\ to\ the\ same\ cluster\} \tag{14}$$

4.1 Approximations and Dependency

In a similar way to the original RS theory, the lower and upper approximations are then delineated as follow,

$$\underline{P}X = \{x \in U : Clus_P(x) \subseteq X\} \tag{15}$$

$$\overline{P}X = \{x \in U : Clus_P(x) \cap X \neq \phi\} \tag{16}$$

Based on this, the positive region and dependency functions can respectively be defined as follow,

$$POS_P(D) = \bigcup_{x \in U/IND(D)} \underline{P}X, \tag{17}$$

$$\acute{\gamma}_P(D) = \frac{|POS_P(D)|}{|U|} \tag{18}$$

Following the above definitions, a feature selection algorithm can be constructed that uses the tolerance-based degree of dependency, $\gamma_P(D)$, to evaluate the significance of feature subsets. The proposed FS algorithm are presented in Figure 1.

EM-CLUSTERING-REDUCT(C, D).
Inputs :
C, the set of all conditional attributes;
D, the set of decision attributes;
Output :
R, the Reduct Set

(1) $R = \phi$
(2) $\acute{\gamma}_{best} = 0$
(3) **do**
(4) $\quad \acute{\gamma}_{tmp} = \acute{\gamma}_{best}$
(5) $\quad T = R$
(6) $\quad$ **for** x **in** $(C - R)$
(7) $\quad\quad$ **if** $\acute{\gamma}_{R \cup \{x\}}(D) > \acute{\gamma}_T(D)$
(8) $\quad\quad\quad T = R \cup \{x\}$
(9) $\quad\quad\quad \acute{\gamma}_{best} = \acute{\gamma}_T(D)$
(10) $\quad R = T$
(11) **until** $\acute{\gamma}_{best} == \acute{\gamma}_{tmp}$
(12) **return** R

Fig. 1. EM Clustering QuickReduct

4.2 An Illustrative Example

In this section, a simple example is used to demonstrate the procedure of the proposed method (see Table 1). There are three continuous conditional attributes

Table 1. Example Table

Object	a	b	c	q
1	-0.4	-0.3	-0.5	0
2	-0.4	0.2	-0.1	1
3	-0.3	-0.4	-0.3	0
4	0.3	-0.3	0	1
5	0.3	-0.3	0	1
6	0.2	0	0	0

and a crisp-valued class attribute in the dataset. In this example, the number of clusters is set to 3.

The greedy algorithm starts with an empty reduct set. It checks each attribute separately and chooses the attribute that has the highest dependency degree. In this example the attribute c is chosen with the dependency degree of 0.33. Then the attribute c is added to the reduct set.

$$U/clust_{\{q\}} = \{\{1,3,6\},\{2,4,5\}\}$$

$$U/clus_{\{a\}} = \{\{3\},\{4,5,6\},\{1,2\}\}$$

$$\acute{\gamma}_a = \frac{|\{3\}|}{|\{1,2,3,4,5,6\}|} = \frac{1}{6} = 0.17$$

$$U/clust_{\{b\}} = \{\{1,3,4,5\},\{6\},\{2\}\}$$

$$\acute{\gamma}_b = \frac{|\{2,6\}|}{|\{1,2,3,4,5,6\}|} = \frac{2}{6} = 0.33$$

$$U/clust_{\{c\}} = \{\{3\},\{2,4,5,6\},\{1\}\}$$

$$\acute{\gamma}_c = \frac{|\{1,3\}|}{|\{1,2,3,4,5,6\}|} = \frac{2}{6} = 0.33$$

$$R \leftarrow \{c\}$$

The hill climbing forward selection algorithm chooses other attributes in the reduct set as follow,

$$U/clust_{\{a,c\}} = \{\{1,3\},\{4,5,6\},\{2\}\}$$

$$\acute{\gamma}_{a,c} = \frac{|\{1,2,3\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6} = 0.5$$

$$U/clust_{\{b,c\}} = \{\{1,3\},\{4,5\},\{2,6\}\}$$

$$\acute{\gamma}_{b,c} = \frac{|\{1,2,4,5\}|}{|\{1,2,3,4,5,6\}|} = \frac{4}{6} = 0.67$$

$$R \leftarrow \{b,c\}$$

$$U/clust_{\{a,b,c\}} : \{\{1,3\},\{4,5,6\},\{2\}\}$$

$$\acute{\gamma}_{a,b,c} = \frac{|\{1,2,3\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6} = 0.5$$

Finally, it returns $\{b,c\}$ as the reduct set which has the same size as the reduct set provided by the Fuzzy Rough Feature Selection (FRFS) and tolerance based FS methods in [11].

5 Simulation Result

In order to evaluate the proposed method, we applied it to a number of real datasets from the UCI repository [1] in Table 2. The EM clustering algorithm from the Weka software [31] was chosen where the number of clusters was selected empirically. The obtained reducts are evaluated via the accuracy of classification. J48, JRIP, and PART classifier in the Weka [31] are chosen as the classifier algorithms.

The obtained accuracies are compared with the accuracy of the FRFS and Tolerance-based FS in [11]. In [12] the FRFS method is compared with other FS methods (such as Relief-F, PCA, and entropy-based approaches) and has been shown that the FRFS method outperformed them. Hence in this paper, the proposed method is compared with only the FRFS and Tolerance-based FS. Table 3 shows the average classification accuracy of 10-fold cross validation as a percentage. The classification algorithms are performed on the original dataset and reduced datasets were obtained by the feature selection algorithms, i.e., the FRFS [11], the Tolerance-based FS [11], and the proposed method.

Table 2. Reduct Size For FRFS, Tolerance, and EM Clustering Methods

Dataset	Objects	Features	Reduct Size		
			FRFS[a]	Tol.[b]	EMRS[c]
Glass	214	10	9	7	5
Heart	270	14	11	10	3
Ionosphere	230	35	11	10	5
Iris	150	5	5	4	4
Water2	**390**	**39**	**11**	**8**	**3**
Wine	178	14	10	8	8

[a] FRFS : Fuzzy Rough Set Feature Selection [11].
[b] Tol. : Tolerance-based Feature Selection [11].
[c] EMRS : The proposed method, i.e., Feature Selection using the RS theory and EM algorithm

Table 3. Classification Accuracies(%) For Unreduced, FRFS, Tolerance, and Clustering Methods

CA[a] / FS[b] / Dataset	J48				JRIP				PART			
	Original[c]	FRFS[d]	Tol.[e]	EMRS[f]	Original[c]	FRFS[d]	Tol.[e]	EMRS[f]	Original[c]	FRFS[d]	Tol.[e]	EMRS[f]
Glass	67.29	69.63	69.16	69.16	69.16	67.76	67.76	69.16	67.76	68.22	69.62	69.16
Heart	76.67	78.89	80.37	79.59	79.63	81.85	82.59	79.59	73.33	78.52	80.37	79.59
Ionosphere	87.83	91.30	87.39	88.32	86.96	86.52	86.96	86.61	88.26	91.30	86.52	90.03
Iris	96.00	96.00	96.00	96.00	95.33	95.33	94.67	95.33	94.00	94.00	95.33	95.33
Water2	**83.33**	**80.26**	**81.79**	**81.77**	**81.03**	**80.51**	**82.31**	**81.57**	**85.64**	**82.56**	**81.28**	**82.34**
Wine	94.38	92.14	94.94	94.94	91.57	90.45	94.38	92.7	93.82	93.82	94.38	94.38

[a] CA : Classification Algorithm.
[b] FS : Feature Selection Algorithm used for each Classification Algorithm.
[c] Original : Original dataset.
[d] FRFS : Fuzzy Rough Set Feature Selection [11].
[e] Tol. : Tolerance-based Feature Selection [11].
[f] EMRS : The proposed method, i.e., Feature Selection using the RS theory and EM algorithm

It is evident from Table 2 that the proposed method generated fewer features compared with the two other FS methods. For the J48 classifier, the clustering based FS improved the average accuracy of the unreduced datasets except for the water2 dataset. The proposed method either unchanged or improved upon the performance of the reduced datasets with the other two FS algorithms in all but in the Ionosphere dataset. For the JRip classifier, the proposed method maintained the average accuracy of the unreduced datasets in all. It either improved or maintained the performance of the reduced dataset with the other two FS algorithms in all but two cases. For PART, the proposed method improved the average accuracy of unreduced datasets in all except the water2 dataset. It has the same behavior as the other two FS methods.

Overall, the proposed algorithm produced a smaller number of attributes compared to the other two FS algorithms and the average accuracy of classifiers is improved or in a few instances remains unchanged. For example, in the water2 dataset the proposed method chose 3 features among 39 features whereas the FRFS chose 10 and the Tolerance-based FS method chose 8 features. In addition, the proposed method has a similar average accuracy compared with the other two approaches.

6 Conclusion

In this work the EM clustering algorithm was applied to deal with the problem of determining initial tolerant classes to obtain a significant classification accuracy. Through some experiments, it was concluded that the proposed method generated a smaller size of feature sets in all datasets compared with the FRFS [11] and tolerance-based FS methods [11]. Beside that, the proposed method either improved or unchanged the average accuracy in all except a few datasets. For future work, an improvement of searching algorithm for finding the reduct set with the new definition of approximations is required. In Addition, an evaluation of the proposed method through experimental comparisons with the other methods in the literature is recommended.

References

[1] Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine (2007), http://www.ics.uci.edu/~mlearn/MLRepository.html

[2] Boulesteix, A.L.: PLS Dimension Reduction for Classification with Microarray Data. Statistical Applications in Genetics and Molecular Biology 3(1), Article 33 (2004)

[3] Bradley, P., Fayyad, U., Reina, C.: Scaling EM clustering to large databases. Technical report, Microsoft Research (1999)

[4] Chouchoulas, A., Shen, Q.: Rough Set-Aided Keyword Reduction for Text Categorisation. Applied Artificial Intelligence 15, 843–873 (2001)

[5] Dempster, A.P., Laird, N.M., Rubin, D.: Maximum likelihood estimation from incomplete data via the EM algorithm. Journal of The Royal Statistical Society 39, 1–38 (1977)

[6] Doherty, P., Szalas, A.: On the Correspondence between Approximations and Similarity. In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) RSCTC 2004. LNCS (LNAI), vol. 3066, pp. 143–152. Springer, Heidelberg (2004)

[7] Fu, X., Wang, L.: Data dimensionality reduction with application to simplifying rbf network structure and improving classification performance. IEEE Transactions on Systems, Man, and Cybernetics, Part B 33, 399–409 (2003)

[8] Greco, S., Inuiguchi, M., Slowinski, R.: Fuzzy rough sets and multiple-premise gradual decision rules. International Journal of Approximate Reasoning 41, 179–211 (2006)

[9] Hancock, P., Burton, A., Bruce, V.: Face processing: Human perception and principal components analysis (1996)

[10] Jensen, R., Shen, Q.: Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches. IEEE Transactions on Knowledge and Data Engineering 16, 1457–1471 (2004)

[11] Jensen, R., Shen, Q.: Tolerance-based and Fuzzy-Rough Feature Selection. In: Proceedings of the 16th International Conference on Fuzzy Systems (FUZZ-IEEE 2007), pp. 877–882 (2007)

[12] Jensen, R., Shen, Q.: Fuzzy-Rough Sets Assisted Attribute Selection. IEEE Transactions on Fuzzy Systems 15, 73–89 (2007)

[13] Kambhatla, N., Leen, T.K.: Dimension Reduction by Local Principal Component Analysis. Neural Comp. 9, 1493–1516 (1997)

[14] Kim, D.: Data classification based on tolerant rough set. Pattern Recognition 34, 1613–1624 (2001)

[15] Lai, Y., Wu, B., Chen, L., Zhao, H.: Statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics, 3146–3155 (2004)

[16] Ordonez, C., Cereghini, P.: SQLEM: Fast clustering in SQL using the EM algorithm. In: ACM SIGMOD Conference (2000)

[17] Pan, W.: A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18, 546–554 (2002)

[18] Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)

[19] Pawlak, Z.: Rough Sets, Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)

[20] Polkowski, L., Semeniuk-Polkowska, M.: On Rough Set Logics Based on Similarity Relations. Fundam. Inf. 64, 379–390 (2005)
[21] Porrill, J., Stone, J.: Independent components analysis for signal separation and dimension reduction (1997)
[22] Radzikowska, A., Kerre, E.E.: Fuzzy rough sets based on residuated lattices. In: Peters, et al. (eds.), vol. 228, pp. 278–296.
[23] Radzikowska, A., Kerre, E.E.: A comparative study of fuzzy rough sets. Fuzzy Sets and Systems 126, 137–155 (2002)
[24] Roweis, S., Ghahramani, Z.: A unifying review of Linear Gaussian Models. Neural Computation (1999)
[25] Roy, A., Pal, S.K.: Fuzzy discretization of feature space for a rough set classifier. Pattern Recogn. Lett. 24, 895–902 (2003)
[26] Skowron, A., Stepaniuk, J.: Tolerance approximation spaces. Fundam. Inf. 27, 245–253 (1996)
[27] Slowinski, R., Vanderpooten, D.: Similarity relation as a basis for rough approximations. In: Wang, P. (ed.) Advances in Machine Intelligence and Soft Computing, vol. 4, pp. 17–33. Duke University Press, Duke (1997)
[28] Slowinski, R., Vanderpooten, D.: A Generalized Definition of Rough Approximations Based on Similarity. IEEE Trans. on Knowl. and Data Eng. 12, 331–336 (2000)
[29] Torkkola, K.: Feature extraction by non-parametric mutual information maximization. Journal of Machine Learning Research 3, 1415–1438 (2003)
[30] Vakarelov, D.: A modal characterization of indiscernibility and similarity relations in Pawlaks information systems. In: Slezak, et al. (eds.), vol. 300, pp. 12–22 (plenary talk)
[31] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
[32] Wu, Y., Zhang, A.: Feature selection for classifying high-dimensional numerical data. CVPR 2, 251–258 (2004)
[33] Xu, L., Jordan, M.: On convergence properties of the EM algorithm for Gaussian mixtures. Neural Computation 7 (1995)
[34] Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Fawcett, T., Mishra, N. (eds.) ICML, pp. 856–863. AAAI Press, Menlo Park (2003)
[35] Yuille, A.L., Stolorz, P., Utans, J.: Statistical physics, mixtures of distributions and the EM algorithm. Neural Computation 6, 334–340 (1994)

Appendix: EM Algorithm

Maximum Likelihood (ML) is a famous method for finding the model parameters for complete data. In case of incomplete data, the EM algorithm can be used for determining the parameters. Assume X is some observation data which is incomplete, and $Z = (X, Y)$ be a complete data with the density function,

$$p(\mathbf{z}|\Theta) = p(\mathbf{x}, \mathbf{y}|\Theta) = p(\mathbf{y}|\mathbf{x}, \Theta)p(\mathbf{x}|\Theta) \tag{19}$$

The complete data likelihood is defined as,

$$L(\Theta|Z) = L(\Theta|X, Y) = p(X, Y|\Theta). \tag{20}$$

The problem is to find Θ which makes the maximume likelihood for the complete data. In this case the unknown-random variable Y leads to have a variable likelihood. The EM algorithm is used for finding the parameters in 2 steps namely Expectaion Step (E-step) and Maximization Step (M-step).

In the E-step, the expected value of log-likelihood of the complete data is determined as follow,

$$Q(\Theta, \Theta^{(i-1)}) = E[\log(p(X, Y|\Theta)|X, \Theta^{(i-1)})] = \int_{\mathbf{y}\in\Upsilon} \log p(X, \mathbf{y}|\Theta) f(\mathbf{y}|X, \Theta^{(i-1)}) d\mathbf{y} \tag{21}$$

where the notations are as follow,

X	Observed-incomplete data and is constant.
$\Theta^{(i-1)}$	Current estimation of the parameter Θ and is constant.
Y	Unlnown-Random variable with a presumably governed by an underlying distribution $f(\mathbf{y}\|X, \Theta^{(i-1)})$.
Θ	Normal variable. The objective is to adjust Θ to obtain the maximum likelihood for the complete data Z.

Then, the M-step is applied to determine the value of the Θ in the iteration i that maximizes the expected value of log-likelihood of the complete data,

$$\Theta^{(i)} = \arg\max_{\Theta} \; Q(\Theta, \Theta^{(i-1)}). \tag{22}$$

The EM algorithm iterates both steps alternatively till converge.