Semi-supervised Gaussian Mixture Models Clustering Algorithm Based on Immune Clonal Selection

Wenlong Huang [1, a], Xiaodan Wang [1,b]

[1] Air and Missile Defense College, Air Force Engineering University, Xi'an 710051,China

[a]wljy_rghymt@163.com, [b]afeu_w@163.com

Keywords: Semi-supervised clustering; Gaussian Mixture Models; immune clonal selection.

Abstract. Semi-supervised Clustering with constraints is an active area of machine learning and data mining research.. Shental used the Expectation Maximization (EM) procedure to handle semi-supervised Gaussian Mixture Models (GMM) estimation, in which positive and negative constraints are incorporated with to improve clustering results. However the conventional EM algorithm only produces solutions that are locally optimal, and thus may not achieve the globally optimal solution, and it is sensitive to initialization, moreover, the number of components of mixture model must be known in advance. This paper introduces the artificial immune clonal selection algorithm into semi-supervised GMM-based clustering techniques, where the EM algorithm is incorporated with the ideas of a clonal selection algorithm. The new algorithm overcomes the various problems associated with the traditional EM algorithm. It can improve the effectiveness in estimating the parameters and determining simultaneously the optimal number of clusters automatically. The experimental results illustrate the proposed clustering algorithm provides significantly better clustering results.

Introduction

One of the most interesting techniques in pattern recognition, data mining and knowledge discovery is clustering. Semi-supervised clustering approach uses additional constraints to guide the clustering process, which has attracted significant research effort in machine learning and data mining communities. Semi-supervised clustering is usually performed by imposing some constraints to an existing clustering method. As K-means algorithm is a popular technique in data clustering for its simplicity and ease implementation, several research work has been done to take into account limited user supervision with K-means. Basu et al. utilized a small number of labeled samples to generate initial centroids for K-means[1]. Wang and Li proposed an active semi-supervised spectral clustering based on actively selecting inform[2]. Abdullin et al. proposed mutual semi-supervision clustering for heterogeneous data[3]. Chen et al. proposed a semi-supervised approach by using spectral clustering[4]. Ahmed et al. proposed a new semi-supervised hierarchical active clustering based on ranking constraints for analysts groupization[5].

Fig.1 shows a simple example of the role pairwise constraints can have when used for semi-supervised clustering. Any of the partitions (b) and (c) of the data items in (a) can be solutions to an unsupervised clustering algorithm, and for some algorithms the choice will depend on random factors (such as the initialization of the prototypes). By providing pairwise constraints like the ones pictured in (d), the user can guide clustering to the solution he prefers.

Among various clustering methods, GMM are one of the more widely used methods for unsupervised clustering of data, where clusters are approximated by Gaussian distributions, often performing better than the hard-partitioning clustering algorithms such as K-means, as well as hierarchical clustering methods. However, there are some challenges associated with mixture modeling, for example, estimation of the parameters of the mixture models and choosing the optimal number of components. Recent literature shows better performance of these methods with respect to totally unsupervised ones even with a small amount of side information. Shental et al.[9] propose a constrained Expectation-Maximization procedure that fits a Gaussian mixture model to a data set.

They provide an EM algorithm for using only must-link constraints, and a generalized EM algorithm for use with both must-link and cannot-link constraints.

As we know, the standard EM has some inherent drawbacks, such as requirement the number of components in advance, sensitivity to initialization and possible convergence to the boundary of the parameter space. In view of such conditions, a novel global search mechanism based the clonal selection algorithm into semi-supervised GMM clustering techniques is proposed in this paper. The novel algorithm can improve the effectiveness in estimating the parameters and determine the optimal number of clusters automatically.

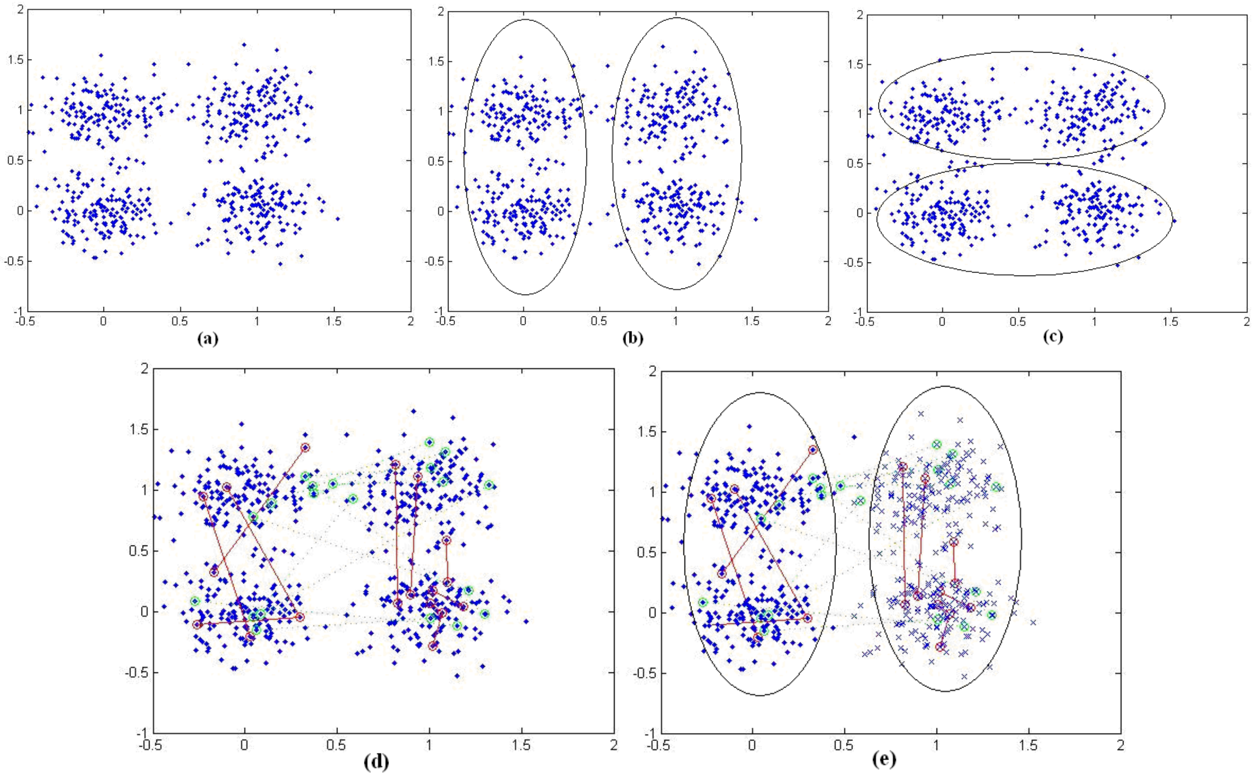

Fig.1 Influence of pairwise constraints on clustering: (a) data items to cluster, (b) and (c) alternative potential solutions for unsupervised clustering, (d) specification of pairwise constraints (red continuous line for the must-link and green dashed line for the cannot-link), (e) solution obtained by semi-supervised clustering using these constraints.

Semi-supervised GMM

Standard EM Algorithm

GMM are often used in generative clustering algorithms, where each Gaussian source is interpreted as a different cluster. A GMM is usually computed in an unsupervised manner using the Expectation Maximization algorithm. Shental et al. present a closed form EM algorithm for handling positive constraints, and a generalized EM algorithm using a Markov network for the incorporation of negative constraints[6].

First, consider the EM algorithm for a standard mixture with K components. For a set of samples $\tilde{X} = (X_1, X_2, \cdots, X_N)$, assumed to be generated independently, we define the potential as the negative complete data log likelihood, that is,

$$U(M,\Theta) = -\sum_{i=1}^{N}\sum_{k=1}^{K} M_{ik} \log[\alpha_k p(x_i|\theta_k)] \tag{1}$$

where α_k denotes the weight of each Gaussian, θ_k its respective parameters, and K denotes the number of Gaussian sources in the GMM. $p(\cdot)$ is the component density specified by parameter set θ_k,

$\Theta = \{\{\theta_k\}, \{\alpha_k\}\}$ and $M = [M_{ik}]$is the assignment matrix with M_{ik} if sample $x_i$is assigned to component k; else$M_{ik} = 0$.

Constrained EM: the update rules

1) Incorporating positive constraints

Let a chunklet denote a small subset of data points that are known to belong to a single unknown class. Chunklets may be obtained by applying the transitive closure to the set of "is-equivalent" constraints.

The expectation of the log likelihood is the following[12]:

$$E\left[\log\left(p\left(X,Y\middle|\Theta^{new},E_\Omega\right)\right)\middle|X,\Theta^{old},E_\Omega\right] = \sum_Y \log\left(p\left(X,Y\middle|\Theta^{new},E_\Omega\right)\right)\cdot p\left(Y\middle|X,\Theta^{old},E_\Omega\right) \tag{2}$$

we can obtain

$$E\left(LogLikehood\right) = \sum_{l=1}^{M}\sum_{j=1}^{L}\sum_{x_i\in X_j}\log\left(p\left(x_i\middle|l,\Theta^{new}\right)\right)\cdot\left(p\left(Y_j=l\middle|X_j,\Theta^{old}\right)\right)+\sum_{l=1}^{M}\sum_{j=1}^{L}\log\alpha_l\cdot p\left(Y_j=l\middle|X_j,\Theta^{old}\right) \tag{3}$$

We differentiate (3) with respect toμ_l, Σ_l andα_l. We get the following rules:

$$\alpha_l^{new} = \frac{1}{L}\sum_{j=1}^{L} p\left(Y_j = l\middle|X_j,\Theta^{old}\right),\quad \mu_l^{new} = \frac{\sum_{j=1}^{L}\bar{X}_j p\left(Y_j=l\middle|X_j,\Theta^{old}\right)\left|X_j\right|}{\sum_{j=1}^{L}p\left(Y_j=l\middle|X_j,\Theta^{old}\right)\left|X_j\right|},\quad \Sigma_l^{new} = \frac{\sum_{j=1}^{L}C_{jl}^{new}p\left(Y_j=l\middle|X_j,\Theta^{old}\right)\left|X_j\right|}{\sum_{j=1}^{L}p\left(Y_j=l\middle|X_j,\Theta^{old}\right)\left|X_j\right|} \tag{4}$$

where $\bar{X}_j$denotes the sample mean of the points in chunklet *j*, $\left|X_j\right|$ denotes the number of points in chunklet *j* and C_{jl}^{new} denotes the sample covariance matrix of the *j*th chunklet of the *l*th class.

2) Incorporating negative constraints

Assume we have a group $\Omega = \left\{\left(a_i^1, a_i^2\right)\right\}_{i=1}^{P}$ of index pairs corresponding to *P* pairs of points that are negatively constrained, and define the event E_Ω={Y complies with the constraints}. We can get

$$p\left(X,Y\middle|\Theta,E_\Omega\right) = \frac{1}{Z}\prod_{\left(a_i^1,a_i^2\right)}\left(1-\delta_{y_{a_i^1},y_{a_i^2}}\right)\prod_{i=1}^{N}p\left(y_i\middle|\Theta\right)p\left(x_i\middle|y_i,\Theta\right) \tag{5}$$

We derived an EM procedure which maximizes $\log\left(p\left(X\middle|\Theta,E_\Omega\right)\right)$. The update rules are

$$\alpha_l = p\left(y_i = l\middle|\Theta^{new},E_\Omega\right),\quad \mu_l^{new} = \frac{\sum_{i=1}^{N}x_i p\left(y_i=l\middle|X,\Theta^{old},E_\Omega\right)}{\sum_{i=1}^{N}p\left(y_i=l\middle|X,\Theta^{old},E_\Omega\right)},\quad \Sigma_l^{new} = \frac{\sum_{i=1}^{N}\hat{\Sigma}_i l p\left(y_i=l\middle|X,\Theta^{old},E_\Omega\right)}{\sum_{i=1}^{N}p\left(y_i=l\middle|X,\Theta^{old},E_\Omega\right)} \tag{6}$$

$\hat{\Sigma}_i l = \left(x_i-\mu_l^{new}\right)\left(x_i-\mu_l^{new}\right)^T$ denotes the sample covariance matrix. Note, however, that now the vector of probabilities $p\left(y_i = l\middle|X,\Theta^{old},E_\Omega\right)$ is inferred using the net.

Semi-supervised Immune Clonal Selection EM Algorithm(IEM)

The immune system can be considered to be a remarkably efficient and powerful information processing system which operates in a highly parallel and distributed manner[7]. Immune Clonal selection remembers by stimulating the growth of cells that bind to the antigen. When an antigen is detected, those antibodies that best recognize this antigen will proliferate by cloning. During the process of cell division (reproduction), individual cells suffer a mutation that allows them to become more adapted to the antigen recognized: the higher the affinity of the parent cell, the lower the mutation they suffer.

Antibody Encoding

In our IEM algorithm, antigen represents a problem, and antibodies represent candidates of the problem. The limited-length character string $\tilde{a} = a_1a_2\cdots a_l$ is the antibody coding of variable x, denoted by $\tilde{a} = h(x)$□, and x is called the decoding of antibody $\tilde{a}$, expressed as $x = h^{-1}(\tilde{a})$. Set I is called antibody space, namely $\tilde{a} \in I$. The antibody population $A = \{\tilde{a}_1, \tilde{a}_2, \cdots, \tilde{a}_n\} \in I^n$ is an n-dimensional group of antibody $\tilde{a}$, namely, $I^n = \{A : A = \left(\tilde{a}_1,\tilde{a}_2,\cdots,\tilde{a}_n\right), \tilde{a}_k \in I,\ 1 \le k \le n\}$, where the positive integer n is the antibody population size.

In our algorithms, each mixture model is coded as an antibody to represent a possible solution of the Gaussian mixture model: each antibody gene is composed of two parts. The first gene segment is binary which is used to encode the number of clusters, where the length of this part is determined by the maximal number of allowed components $K_{\max}$. Each of these bits is related to a particular component. If a bit is set to zero, then its associated component is omitted for modeling the mixture, while setting the bit to one includes the component. The second gene segment consists of floating point values representing the parameters of the models, which length is $L = D + D(D+1)/2$. The parameters for each cluster include the mixing proportion π_k, the mean vector μ_k and the covariance matrix Σ_k.

Antibody-Antigen Affinity

The minmum description length (MDL) criterion is used as the antibody-antigen affinity for model selection. The best individual is the one that has the lowest MDL value. The MDL criteria is a consistent estimator of model order that is expressed as

$$MDL(K,\theta_K) = -\sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \phi(X_i|\theta_k)\right) + \frac{1}{2} L \log(ND) \tag{7}$$

The fitness of the antibody $\tilde{a}_i$ is $f(\tilde{a}_i) = \frac{MDL_{\max} - MDL(\tilde{a}_i)}{MDL_{\max} - MDL_{\min}}$, where $MDL_{\max}$ and $MDL_{\min}$ are the highest and the lowest MDL value respectively. The fitness of an antibody is evaluated by invoking an R-iteration EM algorithm. Starting from the mixture model indicated by an antibody, the algorithm runs the E-step and the M-step for n iterations.

Algorithm description

We integrate the process of EM into the clonal selection algorithm, thus obtaining a method which is able to simultaneously perform estimation of the model parameters and determining the optimal number of clusters automatically. The total process of the proposed IEM is in the following:

Step 1: Give the antibody population n. Randomly generate the original antibody population $A(0) = \{\tilde{a}_1(0), \tilde{a}_2(0), \cdots \tilde{a}_n(0)\} \in I^n,\ t := 0$;

Step 2: Perform R EM steps on each antibody $\tilde{a}_i(t)$, $i = 1,2,\cdots n$. Compute the antibody-antigen affinities MDL of all the antibodies in $A(t)$, and get the antibodies fitness of $A(t)$;

Step 3: Antibody Clone:

$$Y(t) = T_c^C(A(t)) = \left[T_c^C(\tilde{a}_1(t)),\ \ T_c^C(\tilde{a}_2(t)), \cdots, T_c^C(\tilde{a}_n(t))\right]^T$$

Step 4: Clonal Mutation: $Z(k) = T_m^C(Y(k))$

Step 5: Affinity maturation: Perform R EM steps on antibody population $Z(t)$. Compute the antibodies fitness of $Z(t)$.

Step 6: Clonal selection: $A'(t+1) = T_s^C(Z(t))$

Step 7: Clonal death: $A(t+1) = T_d^C(A'(t+1))$

Step 8: t=t+1； If the stop criterion is achieved, then stop, otherwise go to step 2.

Experiment Results

We test our IEM algorithm on both artificial data and real-world data, and compare the results with the K-means algorithm and five prevalent semi-supervised methods: (1) The hard-clustering method based on K-means (COPK) [1]; (2) The constraint-based approach of a semi-supervised clustering scheme has been used for initial seeding of the clusters, which. keeps the grouping of the labeled data fixed throughout the clustering process (CK)[2]; (3) The constrained hierarchical agglomerative clustering with integrated metric learning(CL)[4]; (4) Two semi-supervised Gaussian mixture models clustering algorithms (PEM and FEM)[9].

In order to evaluate the results of the different methods, we use the Rand index (RI), the accuracy (ACC) and the variation of information (VI)[8], to assess the quality of the clustering algorithms.

Synthetic datasets

Two 2-dimensional artificial data sets are designed to highlight the problems that cannot be effectively solved by centroid-based clustering algorithms. The distribution of data points in these data sets and the ideal clustering results that we hope to obtain can be seen in Fig. 2. Different levels of constraint information are considered: 0%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50% of constraints are considered relative to the total number of samples in the data set.

Fig.2. shows the two-moons synthetic data where each moon consists of 100 data points (balanced data). Every point should be similar to points in its local neighborhood, and points in one moon should be more similar to each other than to points in the other moon. Fig.3 to Fig.4 present the dependence between the performance evaluations and the number of pairwise constraints considered for the data set.

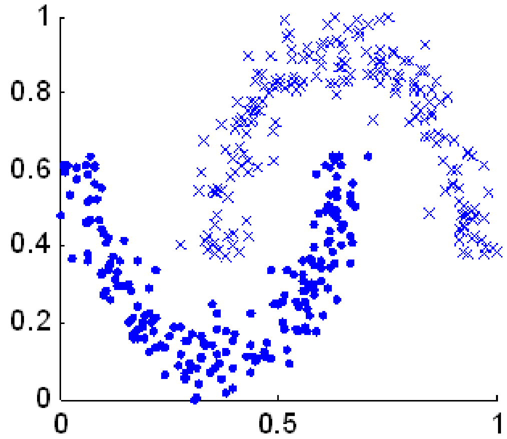

Fig.2 The distribution of data points in synthetic datasets (Class are denoted by symbols)

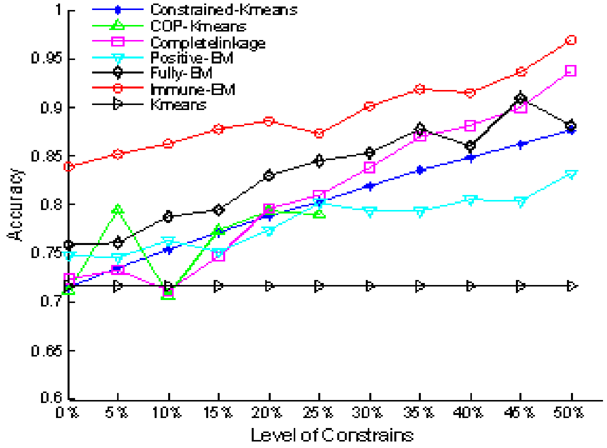

Fig.3 ACC evaluations results on the Two-moons.

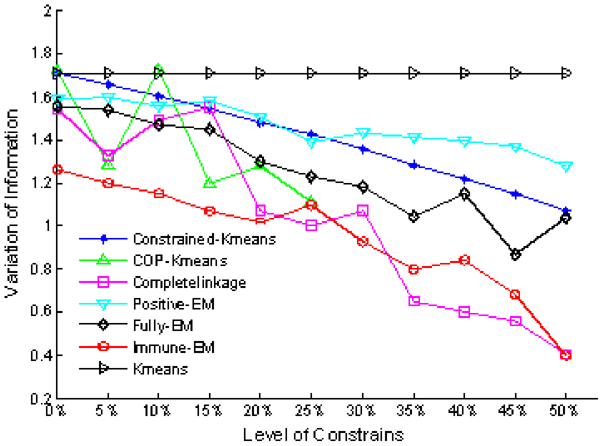

Fig.4 VI evaluations results on the Two-moons.

As we easily notice in Fig.3 to Fig.4, a centroid-based clustering algorithm, like Kmeans, do not respond to the two data sets at all. On two moons data sets our method outperforms the other six methods when pairwise relations are more than 15% and our method also gives the highest clustering accuracy after 35% relations for XOR data set. In addition, COP-Kmeans cannot find a satisfying resolution for XOR and Two moons data sets when relations are more than 25%. Moreover, all standard deviation values of the total performances in our method are smaller than values of the other algorithms, which shows that our method has more stable results.

Real-World datasets

We also present results on the well-known Iris data set in UCI repository. Table1 presents the clustering performances of the different clustering algorithms on the Iris data set, according to various levels of constraint information respectively. The bold face number is the best result among all seven methods.

Table 1 RI evaluations results on the Iris data sets

Methods	0% Constrains	10% Constrains	30% Constrains	50% Constrains
Kmeans	**0.8759±0.0029**	0.8759±0.0029	0.8759±0.0029	0.8759±0.0029
CK	0.8759±0.0029	**0.8836±0.0085**	0.9014±0.0137	0.9190±0.0152
COPK	0.8759±0.0029	0.8762±0.0030	0.9198±0.0282	0.9517±0.0254
CL	0.8368±0.0000	0.8454±0.0578	0.8762±0.0719	0.8744±0.0656
PEM	0.8455±0.0226	0.8620±0.0260	0.8849±0.0214	0.8902±0.0301
FEM	0.8455±0.0226	0.8643±0.0247	0.9436±0.0307	0.9532±0.0485
IEM	0.8595±0.0190	0.8802±0.0207	**0.9571±0.0274**	**0.9792±0.0147**

As shown in Table 1, in most occasions, our method gives the best results among all seven methods. In other occasions, CL gives slightly better or comparable results. Balance scale data sets have high dimensional and sparse feature vectors, which makes COP-Kmeans inapplicable. Therefore on these two data sets we only present six algorithms results. Moreover, all standard deviation values of the total performances in our method are smaller than values of the other algorithms. That is, we can notice that the performance of our method is more stable and prominent.

Conclusions

We have presented a novel immune clonal selection semi-supervised mixture model clustering algorithm, where we introduce a novel global search mechanism based the clonal selection algorithm into semi-supervised Gaussian Mixture Models clustering techniques. The pairwise constraints are used for semi-supervised clustering tasks by modifying the standard Gaussian mixture models clustering algorithm to take into account ML and CL information. The novel algorithm can find globally optimal solutions for model-based clustering problems based on artificial immune evolution, and determine the optimal number of clusters automatically. It is seen that significant performance improvement is achieved over the existing prevalent semi-supervised clustering algorithms, for both artificial data and real-world data sets. Moreover, the novel algorithm shows a best robust behavior among all methods.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (61273275, 61573375), the Aviation Science Foundation of China (20151996015), the Open Research Fund of State Key Laboratory of Astronautic Dynamics (2012ADL-DW0202), and General Financial Grant from China Postdoctoral Science Foundation (2015M572778).

References

[1] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised Clustering by Seeding," in Proceedings of the Nineteenth International Conference on Machine Learning San Francisco, CA, USA, 2002, pp. 27-34.

[2] N. Wang, X. Li, “Active semi-supervised spectral clustering based on pairwise constraints”, Acta Electronica Sinica, Vol.38, No.1, pp.172–176, 2010. (in Chinese).

[3] A. Abdullin, O. Nasraoui, Clustering heterogeneous data with mutual semi-supervision, SPIRE 2012,LNCS 7608,pp.18-29

[4] W.F. Chen, G.C. Feng, Spectral clustering: A semi-supervised approach, Neurocomputing, 77(1), 2012, pp.229–242

[5] E. B. Ahmed, A. N., F. Gargouri, A new semi-supervised hierarchical active clustering based on ranking constraints for analysts groupization, Applied Intelligence, 39(2), 2013, pp.236–250

[6] N. Shental, A. Bar-Hillel, T. Hertz, and D. Weinshall, "Computing Gaussian mixture models with EM using equivalence constraints," Advances in Neural Information Processing Systems, vol. 16, 2004, pp. 465–472.

[7] E. Hart, "Immunology as a Metaphor for Computational Information Processing: Fact or Fiction," Doctor thesis, University of Edinburgh, 2002.

[8] M. Halkidi, D. Gunopulos, M. Vazirgiannis, N. Kumar, and C. Domeniconi, "A clustering framework based on subjective and objective validity criteria," ACM Transactions on Knowledge Discovery from Data, 1(4) , 2008, pp.1-25.