International Journal of Database Theory and Application
Vol.9, No.12 (2016), pp.217-230
http://dx.doi.org/10.14257/ijdta.2016.9.12.20

Research on Spatial Clustering Algorithm based on Data Mining

Runtao Lv [1], Jin Kao Zhao [2] and Yu Li [3]

[1,2] Baotou light industry professional technology institute college of electronic commerce, Baotou 014030, china
[3]Baotou city bureau of education test center, Baotou 014030, china
[1]2538155109@qq.com, [2]1223882175@qq.com and [3]2538155109@qq.com

Abstract

We extended the online learning strategy and scalable clustering technique to soft subspace clustering, and propose two online soft subspace clustering methods, OFWSC and OEWSC. The proposed evolving soft subspace clustering algorithms can not only reveal the important local subspace characteristics of high dimensional data, but also leverage on the effectiveness of online learning scheme, as well as the ability of scalable clustering methods for the large or streaming data. Furthermore, we apply our proposed algorithms to text clustering of information retrieval, gene expression data clustering, face image classification and the problem of predicting disulfide connectivity.

Keywords*: data mining, feature weighting, clustering analysis*

1. Introduction

Over the past few decades, with rapid development of Internet and information technology, people are engaged in large-scale data and streaming data in daily life [1-5]. The research on clustering algorithm for large-scale data or streaming data has become one of the important topics in current data mining and machine learning field. One of the common solutions is to raise incremental clustering or online clustering algorithm by combining traditional batch processing clustering algorithm and incremental learning or online learning strategy [6-7]. So far, people have proposed lots of online clustering learning algorithms based on competitive learning theory [8].

Banerjee and Ghosh employed frequent sensitive competitive learning theory to propose an effective frequent sensitive globular K-means clustering algorithm [9-10]; further, based on WTM(Winner-Take-More) competitive learning rules, Borgelt et al. improved learning vector quantization algorithm. With fuzzy membership function of each sample to clustering center, they introduced iterative learning equation of clustering center based on soft competitive learning theory, extending studies on online clustering technology in fuzzy clustering [11].

Soft space clustering algorithm means during clustering, to assign every feature of each aggregate of data into relative feature weighted coefficient and get the importance of every feature to related aggregate cluster in the clustering process. Soft subspace clustering algorithm includes fuzzy weighted soft subspace (FWSC) and entropy weighted soft subspace (EWSC) algorithm.

Inspired by the above competitive theory. It proposed to improve two kinds of online subspace clustering algorithms: online fuzzy weighted soft subspace clustering algorithm (OFWSC) and online entropy weighted soft subspace clustering algorithm (OEWSC).

Experimental results reveal that OFWSC and OEWSC algorithm realized better clustering results than FWSC and EWSC. However during the clustering, OFWSC and OEWSC algorithm need to traverse several times the whole data sample, which often can’t realize reasonable satisfaction in the case of actual large-scale data stream storage. So with scalable clustering framework, it’s an effective data stream clustering processing

ISSN: 2005-4270 IJDTA

technique to divide large-scale dataset into multiple sub-blocks and consecutively do treatment of every data sub-block.

For example, Bradley et al first proposed a scalable clustering algorithm (ScaleKM), in the process of clustering, for large scale data flow of the sample classification, selective retention of important samples, the general data samples were compressed, while eliminating the importance of the sample.

Further, based on the ScaleKM algorithm, Farnstrom et al proposed a simplified version of ScaleKM algorithm for large scale data sets, and also obtained good clustering results. It is easy to see that the above scalable clustering algorithm is based on Crisp Case of Scalable Clustering. Recently, Hall et al used fuzzy membership function, and proposed two "soft partition" scalable clustering algorithm: (Single-Pass Fuzzy C-Means，SPFCM) and (Online Fuzzy C-Means，OFCM).

2. Online Soft Subspace Clustering Algorithm

2.1 Online Learning Strategy based on Competitive Learning Theory

So far, based on competitive learning theory, lots of online learning clustering algorithms were presented; also online learning clustering algorithm can effectively analyze and understand the distribution change of data sample along with time, which has very important application for real data mining problem.

The learning rule of competitive learning theory includes WTA (Winner Take All) and WTM(Winner Take More), which is called hard competitive learning and soft competitive learning. Under WTA rule, there is only one competitive node of new input samples in the dataset. With WTA rule, people put forward a few online learning clustering algorithms, like frequent sensitive globular K-means clustering algorithm (FS-SpKmeans). The central regression equation of online learning clustering algorithm based on WTA rule can be expressed as:

$$\begin{aligned} & v_i(t) = v_i(t-1) - \eta^{(t)} \times D(v_i(t-1), x_{Nt}) \\ & i^* = \arg\min_i d(v_i(t-1), x_{Nt}) \end{aligned} \tag{1}$$

Where, $v_i(t-1)$ represents t-1 time i cluster center, x_{Nt} represents the t time of the data sample Nt, $d(v_i(t-1)), x_{Nt})$ represents the distance between the sample x_{Nt} and the center $v_i(t-1)$. Due to various clustering algorithms have their own metrics, so the distance between x_{Nt} and its nearest cluster center $v_i(t-1)$ can be expressed as: $D(v_i(t-1), x_{Nt}) = \rho(d(v_i(t-1), x_{Nt}))$. Similar to most of the gradient descent learning algorithm. $\eta^{(t)}$ is learning rate, with time continues to decrease, the result can be avoided effectively and the convergence of the algorithm is guaranteed.

WTA rule has the problem: for new input data sample, competitive node is only one; in the case of node with initial value, there may be dead node or insufficient utilization in the learning process. Hence researchers loosen the limits of WTA rule and proposed WTM rule. By introducing fuzzy membership method, it impairs the dependence on initial value of node in the learning course. According to WTM rule, soft competitive learning strategy adjusts each clustering center as per the measured difference between input sample x_{Nt} and several clustering centers $v_i(t-1)$. The central iterative formula of online learning clustering algorithm based on WTM rule can be expressed as:

$$v_i(t) = v_i(t-1) - \eta^{(t)} \times u_{i(Nt)} D(v_i(t-1), x_{Nt}) \tag{2}$$

2.2Online Fuzzy Weighted Soft Subspace Clustering

Used "soft" competitive learning theory, this paper first defines objective function of online fuzzy weighting subspace clustering.

$$\begin{aligned} J_{OFW}(t) &= \sum_{j=1}^{Nt}\sum_{i=1}^{C}u_{ij}^{m}\sum_{k=1}^{D}w_{jk}^{\tau}(x_{jk}-v_{ik})^{2} \\ &= J_{OFW}(t-1)+\sum_{i=1}^{C}u_{i(Nt)}^{m}\sum_{k=1}^{D}w_{jk}^{\tau}(x_{(Nt)k}-v_{ik})^{2} \end{aligned} \qquad (3)$$

At t time, when the first Nt sample x_{Nt} arrives, in this paper, we can get the following formula of fuzzy membership degree:

$$u_{i(Nt)}=\frac{(d_{i(Nt)})^{-1/m-1}}{\sum_{s=1}^{D}(d_{s(Nt)})^{-1/m-1}} \qquad (4)$$

Where, $w_{ik}(t-1)$ represents the weighted coefficients of the individual data clusters obtained at $t-1$ time.

By comparing equation (3) and (4), we can find:

(1) The iterative formula of OFWSC algorithm clustering center is consistent with the online learning center iterative equation based on WTM rule; so OFWSC algorithm utilizes fuzzy membership function to iteratively update online the clustering center;

(2) In Equation (3), $(v_i(t-1)-x_{Nt})$ is used to calculate the difference between the Nt th sample x_{Nt} which arrives at time t and the ith clustering center $v_i(t-1)$, suggesting that OFWSC is a kind of stochastic gradient descent algorithm;

(3) In equation (4), $\eta^{(t)}$ is learning speed, decreasing along with increasing number of arrived sample data; so the feature of $\eta^{(t)}$ declining with time delay ensures convergence of OFWSC algorithm.

It is worth noting that, in practical applications, due to the formula (4) in the learning rate $\eta^{(t)}$ for the fuzzy membership degree $u_{i(Nt)}$ initialization is very sensitive, easy to cause the instability of clustering results. For the sake of convenience, this paper uses the method of Exponentially Decreasing Rate, which is similar to the literature [11]. Based on the Effect Annealing, the learning rate $\eta^{(t)}$ can be defined as:

$$\eta^{(t)}=\eta_0(\eta_f/\eta_0)^{\frac{t}{NM}} \qquad (5)$$

Based on the above description, Online Fuzzy Weighting Soft Subspace Clustering algorithm is as follows:

OFWSC algorithm steps:

Input: Given data set $X=\{x_1,x_2,...,x_N\}\subset R^{D}$, the number of clusters C and traverse the number of words M.

Initialization:

Random initialization feature weighting factor $w_{ik}(0)$, using K-MEANS++ algorithm from the data set to select the C initial clustering center $v_i(0)$, set the iteration index $itr=1$.

Repeat:

```
    For t=1 to N
      (1)For the t time to arrive at the first $Nt$ data sample $x_{Nt}$, using the formula 4 to calculate the fuzzy membership of each cluster center $u_{i(Nt)}$;
      (2)According to the iterative formula 5 update C cluster center $V_{ik}(t)$;
      (3)According to the iterative formula 6 update the weighting coefficients of each data cluster $w_{ik}(t)$;
      (4) (4)t+t+1
      End
      Itr=itr+1;
Until: Iteration index ITR to reach the number of traversal M
Output: Final fuzzy membership matrix U. Cluster center V and data cluster feature weighting coefficient matrix W
```

2.2 Online Entropy Weighted Soft Subspace Clustering

Used "soft" competitive learning theory, this paper first defines objective function of online entropy weighted soft subspace clustering:

$$\begin{aligned} J_{OEW}(t) &= \sum_{j=1}^{Nt}\sum_{i=1}^{C}u_{ij}^{m}\sum_{k=1}^{D}w_{jk}(x_{jk}-v_{ik})^2+\gamma\sum_{i=1}^{C}\sum_{k=1}^{D}w_{ik}\log w_{ik} \\ &= J_{OEW}(t-1)+\sum_{i=1}^{C}u_{i(Nt)}^{m}\sum_{k=1}^{D}w_{ik}(x_{(Nt)k}-v_{ik})^2 \end{aligned} \quad (6)$$

Can be obtained as follows the fuzzy membership degree iterative formula:

$$u_{i(Nt)}=\frac{(d_{i(Nt)})^{-1/m-1}}{\sum_{s=1}^{C}(d_{s(Nt)})^{-1/m-1}} \quad (7)$$

Based on the above description, online entropy weighted soft subspace clustering algorithm is as follows:

OEWSC algorithm steps:

```
Input: Given data set $X=\{x_1,x_2,...,x_N\}\subset R^D$, the number of clusters C and traverse the number of words M.
Initialization:
      Random initialization feature weighting factor $w_{ik}(0)$, using K-MEANS++ algorithm from the data set to select the C initial clustering center $v_i(0)$, set the iteration index $itr=1$.
Repeat:
    For t=1 to N
      (1)For the t time to arrive at the first $Nt$ data sample $x_{Nt}$, using the formula 5 to calculate the fuzzy membership of each cluster center $u_{i(Nt)}$;
      (2)According to the iterative formula 6 update C cluster center $V_{ik}(t)$;
      (3)According to the iterative formula 7 update the weighting coefficients of each data cluster $w_{ik}(t)$;
      (4)t+t+1
      End
      Itr=itr+1;
Until: Iteration index ITR to reach the number of traversal M
Output: Final fuzzy membership matrix U. Cluster center V and data cluster feature weighting coefficient matrix W
```

3 Experiment Design and Discussion

We make testing analysis of online soft subspace clustering algorithm. By choosing six groups of artificial dataset and real dataset, we do comparative experiment. Firstly, introduce parameter setting of every algorithm and experimental arrangement; then, describe the evaluation standard of three groups of clustering performance; further, for OEWSC and OFWSC algorithm, give out artificial data set I, UCI and gene expression data set for testing.

3.1. Parameter Setting and Experimental Arrangement

For the online soft subspace clustering algorithm, we compare OEWSC, OFWSC and five other clustering algorithms, including two soft subspace clustering methods EWSC [12] and FWSC [13], one kind of online globular K-means clustering algorithm OSKM and two kinds of batch processing clustering algorithm SPKM [14-15] and FCM [16].

For the seven different clustering algorithms, relative parameter settings are adopted. For OEWSC and EWSC algorithm, their entropy weighted index γ is set to 5; for OFWSC and FWSC algorithm, their fuzzy weighted index τ is set to 2. Similar as the paper [17-18], fuzzy index m is chosen in a unique manner, i.e. assume N and D represent the size of respectively data sample and feature dimension. If it meets $\min(N, D-1) > 4$, high-dimensional data set m is set to $m = \dfrac{\min(N, D-1)}{\min(N, D-1) - 3}$; otherwise, m is set to 2. Likewise, for all online clustering algorithm, all data samples traverse five times on average; for batch processing technique, here we set the maximum iteration times of each algorithm to 100.

In this paper, we select six sets of experimental data sets to compare the test results. For all the data sets, all the features are normalized, so that the data of each dimension are in $[0, 1]$ range.

In order to ensure the fairness of the experimental comparison, for all of clustering algorithm was carried out 20 times of repeated experiments, the average and variance of the test results of each algorithm are compared. All experiments are run on the Xeno CPU (R) 2.53-GHz Intel working platform, and Using MATLAB software to simulate.

3.2. Evaluation Criteria

To compare the results of all experimental data sets, we take three evaluation indicators [19]: clustering accuracy (CA) [20], mutual information (NMI) [21]and RAND index (RI) [22]. Clustering accuracy (CA) is percentage of correctly divided sample by clustering algorithm in counting all data samples, which is usually defined as:

$$Clustering\ \ Accuracy = \sum\nolimits_{l=1}^{C} n_l \,/\, N \tag{8}$$

Where, n_l represents the number of samples in the data sample is correctly classified as class l, N is the number of data contained in the entire data set.

NMI calculates the average size of mutual information acquired through pairwise coupling of clustering result and actual class label, defined as below:

$$NMI = \frac{\sum\nolimits_{i=1}^{C}\sum\nolimits_{j=1}^{C} n_{ij}\log((N.n_{ij})/(n_i.n_j))}{\sqrt{\sum\nolimits_{i=1}^{C} n_i \log(n_i / N).\sum\nolimits_{j=1}^{C} n_j \log(n_j / N)}} \tag{9}$$

Where, n_{ij} indicates that the clustering results are i and the class is labeled as the number of data samples of j, n_i indicates the number of samples i , n_j indicates the actual class labeled as the number of j samples, N is the total number of data sets of the entire data set.

RI measures the consistency of two division results when samples belong or not belong to the same class, which are obtained through cluster partition of data set and true division of data set, which is often defined like:

$$RI = \frac{n_{00} + n_{11}}{N(N-)/2} \qquad (10)$$

n_{00} represents the data sample pair, which is different from the true class label, and is divided into a number of samples of different clustering results. n_{11} indicates that the data sample pair has the same true class label, and is divided into the same clustering result. N is the total number of sample data contained in the entire data set.

3.3. Comparison of Online Soft Subspace Clustering Algorithms

3.3.1. Artificial data set I. In the paper, we compare OEWSC and OFWSC algorithm with existing EWSC, FWSC, OSKM, SPKM and FCM algorithm on artificial data set I. The generative process of required artificial data set for the experiment accords with [23].

The artificial data set I contains three parameters: subspace ratio ε, used to control the percentage of data cluster's all subspace dimension in the entire feature space dimension; feature overlapping proportion p, used to adjust the percentage of overlapping subspace dimension of each data cluster; data overlapping ratio α, used to control the overlapping between two data clusters of Gaussian distribution. In artificial data set I, parameter p and α is set to {0.5, 0.8} and {0.2, 0.5, 2}, producing six groups of experimental data sets. In accordance with [23], parameter ε is fixed to 0.375.

Each group of data set contains 500 numbers of 100-dimension data samples; the clustering number is 10; each data cluster contains 50 samples. For every single data cluster, the feature of sample in relative subspace complies with the Gaussian distribution: mean value [0,100] and variance is 10; the feature of sample in irrelevant subspace complies with uniform distribution of [0,100].

Table 1,Table2 and Table3 list out the mean value and standard deviation of respectively CA, NMI and RI of the above seven clustering algorithms on artificial data set I. OEWSC and OFWSC algorithms achieved better experimental results than existing clustering methods. Of them, OEWSC algorithm got the highest clustering evaluation index on six groups of artificial data sets. Meanwhile we find FCM algorithm assigns the same significance to sample's feature, so during clustering, it's not difficult to observe that worse clustering results exist in the subspace structure of each data cluster.

Table 1. Clustering Results in Terms of CA for Synthetic Datasets I

p	α		OEWSC	OFWSC	EWSC	FWSC	OSKM	SPKM	FCM
0.5	0.2	Mean	0.922	0.851	0.683	0.727	0.731	0.708	0.308
		Std	0.058	0.072	0.061	0.104	0.072	0.098	0.031
	0.5	Mean	0.980	0.891	0.776	0.791	0.749	0.091	0.471
		Std	0.041	0.116	0.097	0.089	0.083	0.730	0.064
	2	Mean	0.967	0.914	0.822	0.768	0.768	0.730	0.377
		Std	0.047	0.110	0.096	0.087	0.087	0.080	0.062
0.8	0.2	Mean	0.870	0.780	0.545	0.685	0.617	0.659	0.337
		Std	0.104	0.053	0.102	0.073	0.977	0.105	0.055
	0.5	Mean	0.868	0.753	0.684	0.753	0.681	0.662	0.409
		Std	0.579	0,071	0.086	0.103	0.062	0.049	0.087

	2	Mean Std	0.917 0.035	0.908 0.027	0.717 0.097	0.813 0.089	0.710 0.062	0.684 0.086	0.325 0.066

Table 2. Clustering Results in Terms of NMI for Synthetic Datasets I

p	α		OEWSC	OFWSC	EWSC	FWSC	OSKM	SPKM	FCM
0.5	0.2	Mean Std	0.96 0.023	0.935 0.022	0.859 0.029	0.843 0.087	0.878 0.022	0.863 0.048	0.443 0.015
	0.5	Mean Std	0.993 0.0144	0.931 0.054	0.902 0.042	0.896 0.061	0.902 0.024	0.888 0.041	0.512 0.044
	2	Mean Std	0.980 0.018	0.944 0.048	0.922 0.036	0.944 0.048	0.903 0.027	0.086 0.035	0.047 0.059
0.8	0.2	Mean Std	0.939 0.019	0.968 0.027	0.832 0.525	0.860 0.085	0.036 0.875	0.786 0.026	0.394 0.044
	0.5	Mean Std	0.097 0.009	0.968 0,027	0.865 0.052	0.913 0.047	0.087 0.021	0.847 0.043	0.520 0.035
	2	Mean Std	0.917 0.035	0.908 0.027	0.717 0.097	0.860 0085	0.830 0.036	0.786 0.026	0.394 0.044

Table 3. Clustering Results in Terms of RI for Synthetic Datasets I

p	α		OEWSC	OFWSC	EWSC	FWSC	OSKM	SPKM	FCM
0.5	0.2	Mean Std	0.983 0.012	0.969 0.015	0.897 0.030	0.921 0.025	0.920 0.018	0.884 0.019	0.579 0.012
	0.5	Mean Std	0.995 0.008	0.975 0.020	0.926 0.028	0.935 0.020	0.918 0.022	0.891 0.020	0.751 0.032
	2	Mean Std	0.992 0.009	0.980 0.020	0.943 0.021	0.952 0.022	0.928 0.0013	0.886 0.018	0.658 0.040
0.8	0.2	Mean Std	0.973 0.021	0.952 0.016	0.851 0.032	0.911 0.091	0.890 0.023	0.870 0.022	0.605 0.042
	0.5	Mean Std	0.970 0.011	0.948 0,020	0.905 0.023	0.926 0.024	0.877 0.021	0.867 0.019	0.685 0.048
	2	Mean Std	0.928 0.006	0.981 0.015	0.904 0.040	0.940 0.020	0.906 0.020	0.876 0.021	0.592 0.034

To compare comprehensively clustering result of OEWSC and OFWSC algorithm, Figure1 shows the mean value and standard deviation of resultant CA, NMI and RI. Vertical coordinate in left graph refers to mean value of experimental evaluation indicator; vertical coordinate in right picture means standard deviation of related evaluation index. As indicated, (1) both OEWSC and OFWSC algorithms realized the best clustering result on artificial data set I; (2) compared with traditional batch processing clustering algorithms like EWSC, FWSC and SPKM, the new OEWSC, OFWSC and OSKM based on online learning strategy acquired higher and more stable clustering result. In short, compared with current batch processing clustering method, online soft subspace clustering technique can do cluster partitioning more effectively of data set with high-dimension subspace structure.

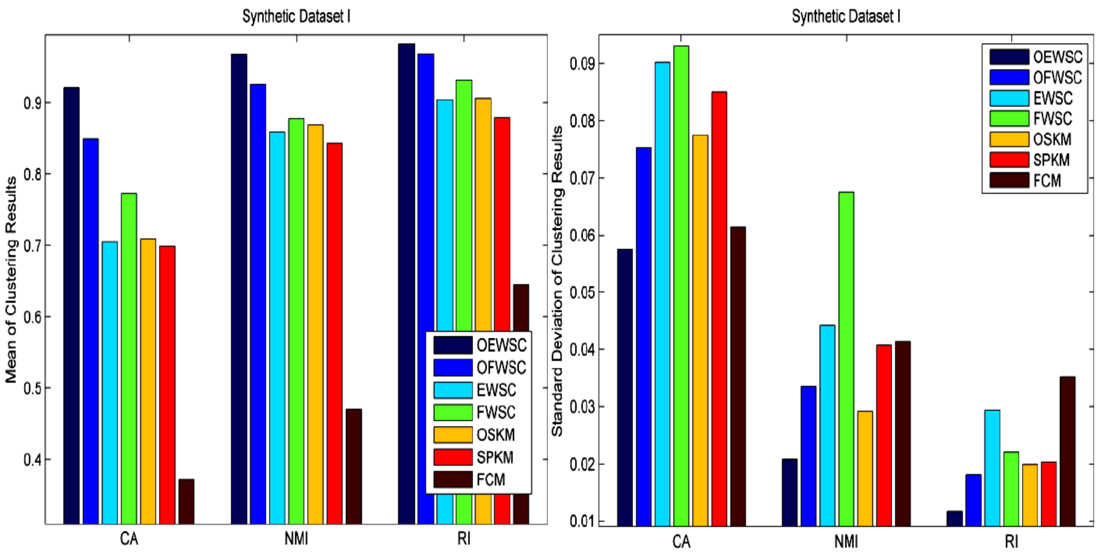


Figure 1. The Averages Clustering Results for the Synthetic Datasets I

The above 7 kinds of clustering algorithm in average running time of I the artificial data OEWSC is as follows:: 0.4063 seconds, OFWSC:0.5353 seconds, EWSC:0.3407 seconds, FWSC:0.4764 seconds, OSKM:0.1278 seconds, SPKM:0.0587 seconds, FCM:0.2951 seconds

3.3.2. UCI data set. To perform comparative experiment of various clustering algorithms on UCI data set, we choose from UCI data set six groups of data to compare OEWSC and OFWSC algorithm [24]. Data set is listed as table4:

Table 4. UCI Datasets Used in the Experiment

Data sets information	Size of data set	Number of dimensions	Number of clusters
Class	214	9	2
Iris	150	4	3
Wine	178	13	3
Lonosphere	351	33	2
Vehicle	846	18	4
Breast-diagnostic	569	30	2

Table5, Table6 and Table7 give the mean value and standard deviation of CA, NMI and RI got by seven clustering approaches on UCI data set. Test results demonstrate that OEWSCSC and OFWSCSC gained better clustering result in most cases. FCM algorithm got the highest clustering evaluation index on Breast-diagnostic data set. Since FCM algorithm is suitable to do clustering division of data in globular data cluster, it’s assumed that Breast-diagnostic data’s geometric distribution would be the fitter for FCM algorithm. The test findings also suggest that no anyone clustering algorithm can get the best clustering result for any data set.

Table 5. Clustering Results in Terms of CA for UCI Datasets

Dataset	OEWSC	OFWSC	EWSC	FWSC	OSKM	SPKM	FCM
Glass: Mean Std	0.918 0.006	0.784 0.137	0.755 0.165	0.700 0.170	0.890 0.002	0.833 0.140	0.876 0.001
Iris: Mean Std	0.888 0.008	0.931 0.018	0.849 0.127	0.809 0.180	0.654 0.744	0.695 0.066	0.889 0.003
Wine:	0.933	0.915	0.921	0.881	0.891	0.880	0.090

Mean Std	0.004	0.050	0.005	0.075	0.007	0.059	0.001
Lonosphere: Mean Std	0.703 0.005	0.721 0.026	0.692 0.039	0.669 0.058	0.607 0.025	0.599 0.067	0.659 0.002
Vehicle: Mean Std	0.422 0.028	0.425 0.028	0.388 0.023	0.419 0.031	0.393 0.001	0.392 0.006	0.341 0.003
Breast-diagnostic: Mean Std	0.886 0.008	0.874 0.008	0.909 0.066	0.866 0.046	0.728 0.014	0.771 0.006	0.927 0.004

Table 6. Clustering Results in Terms of NMI for UCI Datasets

Dataset	OEWSC	OFWSC	EWSC	FWSC	OSKM	SPKM	FCM
Glass: Mean Std	0.532 0.21	0.313 0.147	0.025 0.182	0.190 0.137	0.405 0.010	0.360 0.0113	0.459 0.001
Iris: Mean Std	0.768 0.009	0.809 0.036	0.757 0.056	0.737 0.100	0.623 0.039	0.606 0.030	0.735 0.006
Wine: Mean Std	0.815 0.013	0.788 0.009	0.804 0.014	0.710 0.130	0.06 0.013	0.693 0.044	0.011 0.002
Lonosphere: Mean Std	0.125 0.008	0.017 0.069	0.150 0.053	0.104 0.053	0.011 0.007	0.052 0.079	0.119 0.002
Vehicle: Mean Std	0.166 0.030	0.182 0.028	0.148 0.029	0.171 0.022	0.161 0.002	0.156 0.021	0.098 0.004
Breast-diagnostic: Mean Std	0.559 0.022	0.493 0.023	0.586 0.141	0.467 0.141	0.028 0.023	0.292 0.010	0.615 0.007

Table 7. Clustering Results in Terms of RI for UCI Datasets

Dataset	OEWSC	OFWSC	EWSC	FWSC	OSKM	SPKM	FCM
Glass: Mean Std	0.849 0.010	0.706 0.117	0.681 0.142	0.633 0.127	0.804 0.004	0.756 0.112	0.819 0.001
Iris: Mean Std	0.875 0.006	0.918 0.019	0.867 0.063	0.845 0.097	0.759 0.025	0.773 0.022	0.876 0.003
Wine: Mean Std	0.932 0.005	0.893 0.051	0.923 0.006	0.860 0.067	0.863 0.008	0.857 0.038	0.091 0.002
Lonosphere: Mean Std	0.572 0.004	0.607 0.023	0.581 0.040	0.563 0.036	0.581 0.040	0.531 0.036	0.576 0.001
Vehicle: Mean Std	0.654 0.022	0.671 0.012	0.612 0.053	0.659 0.014	0.650 0.001	0.648 0.010	0.642 0.003
Breast-diagnostic:	0.798 0.013	0.780 0.012	0.843 0.074	0.772 0.056	0.648 0.015	0.658 0.006	0.866 0.004

Mean Std							

3.3.3. Gene expression data set. Five groups of gene expression data sets are selected to test the proposed OEWSC and OFWSC algorithms. The most distinctive characteristic of gene expression data is its high dimensionality [25-29]. Similar as the data set of most biological information, high-dimension data can usually cause the problem of curse of dimensionality. It is shown in table8.

Table 8. Gene Expression Datasets Used in the Experiment

Data sets information	Size of data set	Number of dimensions	Number of clusters
DLBCL	88	4026	6
Pstate3r	33	12626	2
Leukemia	72	7129	2
CNS	34	7129	2
Breast tumours	84	9216	5

Table 9, Table 10 and Table 11present mean value and standard deviation of CA, NMI and RI obtained by those clustering algorithms on gene expression data set. OEWSC and OFWSC methods had better clustering results than existing ones. In the meantime, we note when one algorithm got the highest value for some specific evaluation indicator, it may not achieve the best clustering result for other evaluation indicators. So it's necessary to do synthetic comparison of experimental results with the use of many clustering measurement criteria.

Table 9 Clustering Results in Terms of CA for Gene Expression Datasets

Dataset	OEWSC	OFWSC	EWSC	FWSC	OSKM	SPKM	FCM
DLBCL: Mean Std	0.762 0.050	0.751 0.099	0.642 0.075	0.058 0.102	0.741 0.101	0.684 0.112	0.613 0.012
Pstate3: Mean Std	0.784 0.081	0.792 0.054	0.762 0.162	0.683 0.148	0.734 0.102	0.763 0.131	0.764 0.033
Leukemia: Mean Std	0.731 0.043	0.727 0.046	0.652 0.063	0.645 0.079	0.707 0.026	0.702 0.045	0.670 0.012
CNS: Mean Std	0.624 0.050	0.648 0.053	0.626 0.070	0.615 0.062	0.610 0.060	0.625 0.070	0.600 0.023
Breast tumours : Mean Std	0.477 0.036	0.481 0.051	0.492 0.055	0.391 0.057	0.498 0.049	0.470 0.079	0.536 0.012

Table 10. Clustering Results in Terms of NMI for Gene Expression Datasets

Dataset	OEWSC	OFWSC	EWSC	FWSC	OSKM	SPKM	FCM
DLBCL: Mean Std	0.729 0.048	0.714 0.179	0.615 0.061	0.500 0.077	0.699 0.083	0.682 0.079	0.501 0.021
Pstate3: Mean Std	0.014 0.056	0.371 0.118	0.437 0.370	0.191 0.246	0.211 0.239	0.372 0.284	0.388 0.088

Leukemia: Mean Std	0.145 0.056	0.171 0.068	0.053 0.055	0.083 0.115	0.144 0.032	0.134 0.051	0.092 0.013
CNS: Mean Std	0.137 0.078	0.072 0.073	0.049 0.068	0.076 0.048	0.053 0.048	0.057 0.048	0.067 0.027
Breast tumours: Mean Std	0.425 0.037	0.445 0.054	0.390 0.077	0.180 0.078	0.471 0.051	0.338 0.095	0.372 0.009

Table 11. Clustering results in terms of RI for gene expression Datasets

Dataset	OEWSC	OFWSC	EWSC	FWSC	OSKM	SPKM	FCM
DLBCL: Mean Std	0.850 0.035	0.846 0.048	0.779 0.048	0.719 0.056	0.822 0.049	0.807 0.056	0.730 0.018
Pstate3: Mean Std	0.664 0.095	0.666 0.060	0.712 0.196	0.597 0.144	0.618 0.110	0.696 0.147	0.666 0.045
Leukemia: Mean Std	0.605 0.034	0.601 0.037	0.547 0.038	0.548 0.062	0.599 0.021	0.597 0.034	0.552 0.009
CNS: Mean Std	0.522 0.024	0.539 0.030	0.502 0.040	0.504 0.035	0.515 0.040	0.508 0.032	0.049 0.008
Breast tumours: Mean Std	0.660 0.044	0.640 0.053	0.645 0.062	0.589 0.057	0.607 0.052	0.636 0.053	0.628 0.006

The paper conducted obvious analysis of experimental results of various clustering algorithm. By the t testing method based on 5% significance level, we got P value between OEWSC, OFWSC and other clustering methods. As seen, Table 12 and Table13 gave out P value of NMI result by OEWSC and OFWSC algorithms. As null hypothesis, the paper holds that clustering results of NMI by the two methods don't have significant difference. The alternative hypothesis is obvious significant difference exists between such clustering results. From Table 12 and Table13, it's learned that P value of experimental results by most methods is below 0.05, implying that apparent significant difference does appear between OEWSC and OFWSC algorithms and other peer algorithms.

Table 12. P-values Produced by t-test Comparing OEWSC about NMI Results

Datasets	P-values				
	EWSC	FWSC	OSKM	SPKM	FCM
DLBCL	4.941	4.957	same	0.008	2.378
Prostate3	0.004	0.001	0.001	same	0.047
Leukemia	4.406	0.007	0.266	0.045	5.09
CNS	0.006	0.001	0.037	0.407	0.003
Breast tumours	0.008	3.121	same	0.008	3.718

Table 13. P-values Produced by t-test Comparing OFWSC about NMI Results

Datasets	P-values				
	EWSC	FWSC	OSKM	SPKM	FCM
DLBCL	2.975	4.956	same	0.233	4.424
Prostate3	0.014	0.024	0.016	Same	0.029
Leukemia	1.2445	0.453	0.028	0.031	4.212
CNS	0.042	0.014	0.026	Same	0.034
Breast tumours	0.016	1.125	0.007	0.027	0.006

All in all, the comparison and analysis of the above mentioned online soft subspace clustering algorithm. This paper finds that using online learning strategies, OEWSC and OFWSC algorithms can get better experimental results than the traditional batch clustering algorithms.

4 Conclusion

In practical applications, we need to cluster the high dimensional data or stream data. In this paper, we use the online learning strategy and fuzzy clustering technology and the existing soft subspace clustering algorithm, and proposed two kinds of online soft subspace clustering algorithm (OFWSC, OEWSC). The algorithm can effectively utilize the online learning strategy and the fuzzy scalable clustering technique for large scale data clustering analysis. The test results of artificial data sets and real data sets show the effectiveness of the proposed algorithm.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under grant No. 61163025 and the natural science foundation of the Inner Mongolia autonomous region No. 2010BS0904

References

[1] Z. Jianpeng, C. Fucai, L. Shaomei and L. Lixiong, “Data stream clustering algorithm based on density and nearest neighbor communication”, Journal of automation, vol. 2, **(2014)**, pp. 277-288.

[2] L. Hua, P. Yu and P. Xiyuan, “A multi dimension uncertain data stream clustering algorithm”, Journal of instrumentation, vol. 6, **(2013)**, pp. 131-139.

[3] W. Renhong, “Data stream processing algorithm based on clustering analysis”, Chongqing Jiaotong University, **(2013)**.

[4] H. Ying. Research on distributed data stream clustering algorithm. Beijing Jiaotong University, 2015

[5] H. Decai and W. Tianhong, “Control and decision based on density based data stream clustering algorithm for mixed attribute”, vol. 3, **(2010)**, pp. 416-421.

[6] H. Wei, X. Fuyuan and M. Qingguo, “Artificial Immune Principle Study on data stream clustering algorithm”, computer science, vol. 2, **(2012)**, pp. 195-197.

[7] X. Huyin, “Density and grid data stream clustering algorithm”, Northwest Normal University, **(2012)**.

[8] W. Dongmian, “Dongmian bucket density clustering algorithm of data stream based on the research and application”, Xi'an Electronic and Science University, **(2014)**.

[9] A. Banerjee and J. Ghosh, “Frequency-sensitive competitive learning for scalable balanced clustering on high-dimensional hyperspheres”, IEEE Transactions on Neural Networks, vol. 15, no. 3, **(2004)**, pp. 702-719.

[10] S. Zhong, “Efficient online spherical k-means clustering”, In IEEE International Joint Conference on Neural Networks, Montréal, Québec, Canada, **(2005)**, pp. 3180-3185.

[11] C. Borgelt and A. Nürnberger, “Fast fuzzy clustering of web page collections”, In Proc. PKDD Workshop on Statistical Approaches for Web Mining, Pisa, Italy, **(2004)**.

[12] L. Jing, M. K. Ng and J. Z. Huang, “An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data”, IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 8, **(2007)**, pp. 1026-1041.

[13] G. Gan and J. Wu, "A convergence theorem for the fuzzy subspace clustering (FSC) algorithm", Pattern Recognition, vol. 41, no. 6, **(2008)**, pp. 1939-1947.

[14] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering", Machine Learning, vol. 42, no. 1, **(2001)**, pp. 143-175.

[15] S. Zhong and J. Ghosh, "A unified framework for model-based clustering", Journal of Machine Learning Research, vol. 4, no. 1, **(2003)**, pp. 1001-1037.

[16] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms", New York: Plenum Press: **(1981)**.

[17] Z. Deng, K. S. Choi and D. L. Chung, "EEW-SC: Enhanced Entropy-Weighting Subspace Clustering for high dimensional gene expression data clustering analysis", Applied Soft Computing, vol. 11, no. 8, **(2011)**, pp. 4798-4806.

[18] J. Yu, Q. Cheng and H. Huang, "Analysis of the weighting exponent in the FCM", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 34, no. 1, **(2004)**, pp. 634-639.

[19] N. I. O, T. Boongoen and S. Garrett, "A Link-Based Approach to the Cluster Ensemble Problem", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 99, **(2011)**, pp. 2396-2409.

[20] N. Nguyen and R. Caruana, "Consensus clustering", In International Conference on Data Mining, Omaha, NE, USA, **(2007)**, pp. 607-612.

[21] A. Strehl and J. Ghosh, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions", Journal of Machine Learning Research, vol. 3, **(2003)**, pp. 583-617.

[22] W. M. Rand, "Objective criteria for the evaluation of clustering methods", Journal of the American Statistical association, vol. 66, no. 1, **(1971)**, pp. 846-850.

[23] Y. Lu, S. Wang and S. Li, "Particle swarm optimizer for variable weighting in clustering high-dimensional data", Machine Learning, vol. 82, no. 1, **(2011)**, pp. 43-70.

[24] C. Blake and C. J. Merz, "UCI Repository of machine learning databases. In University of California, Irvine", School of Information and Computer Sciences: http://archive.ics.uci.edu/ml/, **(1998)**.

[25] A. A. Alizadeh, M. B. Eisen and R. E. Davis, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", Nature, vol. 403, no. 6769, **(2000)**, pp. 503-511.

[26] T. R. Golub, D. K. Slonim and P. Tamayo, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", science, vol. 286, no. 5439, **(1999)**, pp. 531-537.

[27] C. M. Perou, T. Sørlie amd M. B. Eisen, "Molecular portraits of human breast tumours", Nature, vol. 406, no. 6797, **(2000)**, pp. 747-752.

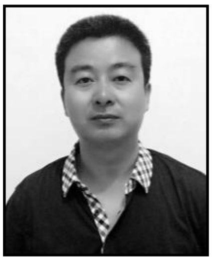
[28] S. L. Pomeroy, P. Tamayo and M. Gaasenbeek, "Prediction of central nervous system embryonal tumour outcome based on gene expression", Nature, vol. 415, no. 6870, **(2002)**, pp. 436-442.

[29] J. B. Welsh, L. M. Sapinoso and A. I. Su, "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer", Cancer research, vol. 61, no. 16, **(2001)**, pp. 5974-5978.

Authors

Runtao Lv. She received her B.S degree from Inner Mongolia normal university computer science education. She is an associate professor in baotou light industry professional technology institute college of electronic commerce. Her research interests include computer network or database.

Jin Kao Zhao. He received his B.S degree from Inner Mongolia normal university computer science education. He is an associate professor in baotou light industry professional technology institute college of electronic commerce. His research interests include computer network.



Yu li. He received his B.S degree from Inner Mongolia normal university computer science education. He is a lecturer in Baotou city bureau of education test center. His research interests include Computer application.