International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 7 4476 - 4480

An Unsupervised Based Stochastic Parallel Gradient Descent For Fcm Learning Algorithm With Feature Selection For Big Data

Jayapratha. T[1]
Assistant Professor,
Department of Information Technology
Sri Eshwar College of Engineering,
e-mail: itzbtechengineer@gmail.com

Vanitha. M[2]
Assistant Professor,
Department of Information Technology
Sri Eshwar College of Engineering,
e-mail: vanitham87@gmail.com

Pradeepa. T[3]
Research Scholar,
Department of Computer Science
Sri Ramakrishna College of Arts and Science for Women
e-mail:tdeepu1991@gmail.com

Priyanka. B[4]
Tata consultancy services
System Engineer
e-mail: b.priyanka2@tcs.com

***Abstract*—** Huge amount of the dataset consists millions of explanation and thousands, hundreds of features, which straightforwardly carry their amount of terabytes level. Selection of these hundreds of features for computer visualization and medical imaging applications problems is solved by using learning algorithm in data mining methods such as clustering, classification and feature selection methods .Among them all of data mining algorithm clustering methods which efficiently group similar features and unsimilar features are grouped as one cluster ,in this paper present a novel unsupervised cluster learning methods for feature selection of big dataset samples. The proposed unsupervised cluster learning methods removing irrelevant and unimportant features through the FCM objective function. The performance of proposed unsupervised FCM learning algorithm is robustly precious via the initial centroid values and fuzzification parameter (m). Therefore, the selection of initial centroid for cluster is very important to improve feature selection results for big dataset samples. To carry out this process, propose a novel Stochastic Parallel Gradient Descent (SPGD) method to select initial centroid of clusters for FCM is automatically to speed up process to group similar features and improve the quality of the cluster. So the proposed clustering method is named as SPFCM clustering, where the fuzzification parameter (m) for cluster is optimized using Hybrid Particle Swarm with Genetic (HPSG) algorithm. The algorithm selects features by calculation of distance value between two feature samples via kernel learning for big dataset samples via unsupervised learning and is especially easy to apply. Experimentation work of the proposed SPFCM and existing clustering methods is experimented in UCI machine learning larger dataset samples, it shows that the proposed SPFCM clustering methods produces higher feature selection results when compare to existing feature selection clustering algorithms , and being computationally extremely well-organized.

Keywords- *Clustering, Classification, unsupervised learning, Stochastic Parallel Gradient Descent (SPGD), Particle Swarm Optimization (PSO), Fuzzy c means (FCM) clustering, Genetic algorithm operators, parallel processing, big-data.*

*****

I. INTRODUCTION

With the advanced growth of the Internet and WWW, mining of big data in larger dataset and reduction of curse dimensionality problem have become known emerging technologies in various applications, such as data mining, text mining and information recovery [1]. For example consider [2] studies the problem of larger dataset mining with collaborative filtering for email spam application along with 16 trillion (1013) exclusive features are presented there. The solving the curse dimensionality issue [3-4] is not easy task for larger dataset because of storage and time complexity. Providentially, for several data sets by curse dimensionality, several numbers of the features are inappropriate to the output. Consequently, reducing the inappropriate features and choosing the small amount significant features can greatly enhance the performance of the system.

Feature selection is a successful method designed for solving curse dimensionality problem and importance discovery [5]. It progress the performance of methods in terms of their correctness, effectiveness, and model interpretability [6] in a several number of applications such as text mining, image processing, genetic analysis and bio medical applications. The massive increase of large-scale data sets fetches new challenges in feature selection for big dataset samples [7].

Real-time systems cannot have enough time to perform this feature selection process. Most of the paper focus on applying and solving this problem by proposing data mining methods and techniques some of the based on the statistical model and supervised learning systems, but this method is applied to smaller set of features effects [8] only. On other hand some of them methods are difficult or impracticable to distinguish the features are relevant or not, so it is complex task.

Some of the feature selection methods are spectrum-based feature selection with kernel [9] , histogram based feature extension [10] with intersection kernel , and so on. Boosting algorithms is also applied for feature in recent years [11-12]. These algorithms are designed based on the greedy selection process in each iteration with current coefficients feature value. The major issue of the method is that it depends on structure procedure for boosting and computationally complexity for larger dataset samples with millions of features.

This paper presents a novel unsupervised cluster based feature selection methods larger dataset samples and big dataset samples. The methods select important features and remove irrelevant features based on fuzzy c means clustering algorithm .The proposed SPFCM clustering method centroid values are automatically selected via SPGD algorithm and fuzzy membership function (m) are optimized using HPSG algorithm,

ISSN: 2321-8169
4476 - 4480

so it becomes less computational complexity and removes irrelevant features for larger dataset and big dataset samples when compare to earlier learning performance [13]. The proposed SPFCM clustering algorithm selects features through estimation of feature subsets as a result handle and removes unnecessary features efficiently.

II. BACKGROUND KNOWLEDGE

Supervised feature selection simply implemented to labeled training data based on the several number of scoring and distance function, thus results higher accuracy and best feature election results. Though, supervised feature selection are exclusive and time consumption task in real time applications [14] . Thus motivates to select semi-supervised learning methods researchers to perform feature selection.

Zhao et al[15] introduces a feature selection method based on semi-supervised learning algorithm , this method is applicable to only single label dataset. To address this problem unsupervised feature selection have been focused in some papers. The unsupervised learning method for feature selection chose important features based on the data similarity functions such as laplacian Score [16], Spectral score function [15] and it is applied to several number of the fields including bioinformatics, text examination and image explanation [17-18].

Online Feature Selection (OFS) have been also used in recent years to perform feature selection process in sequential manner for time series dataset samples [19]. This method features are implicit to appear individual at a time whereas the remaining features are presented before performing learning process to select important feature and removes irrelevant features at each time for real-world applications.

Parallel based feature selection algorithm is also developed in [20] for logistic regression classification algorithm . It follows the procedure of MapReduce framework for parallel learning criterion objective function which is chosen from logistic regression model. After the completion of Parallel feature selection process, it needs to retrain in iterative manner therefore it should be more capable.

III. PROPOSED STOCHASTIC GRADIENT FOR FCM ALGORITHM BASED ON ONLINE KERNEL LEARNING

The proposed unsupervised feature selection approach is based on a hybrid FCM clustering method to select important features for big dataset samples. The proposed hybrid FCM clustering is varied from usual FCM clustering which overcomes initial centroid selection and random selection of fuzzification parameter (m).The select best initial centroid a Stochastic Parallel Gradient Descent (SPGD) is proposed ,so it is named as SPFCM clustering. The proposed SPFCM clustering method is used to group the most important features and removes irrelevant features. Though, in this SPFCM clustering algorithm, the fuzzification parameter(m) also affects the feature selection performance to manage this issue; fuzzification parameter (m) is optimized using Hybrid Particle Swarm with Genetic (HPSG) algorithm. It uses ring point crossover(RC) operation to update the position of fuzzification parameter thus results optimized fuzzification parameter results , the working procedure of the proposed work is illustrated in Figure.1.

Let us consider $X_i = (x_1, \ldots$ be the number of big data samples with features $F_i = (f_1, \ldots$, C be the number of cluster $C = \{C_1^{\phi}, \ldots,$ found from big dataset samples , where randomly selected initial centroid value is represented as ,which is optimized using SPGD algorithm.

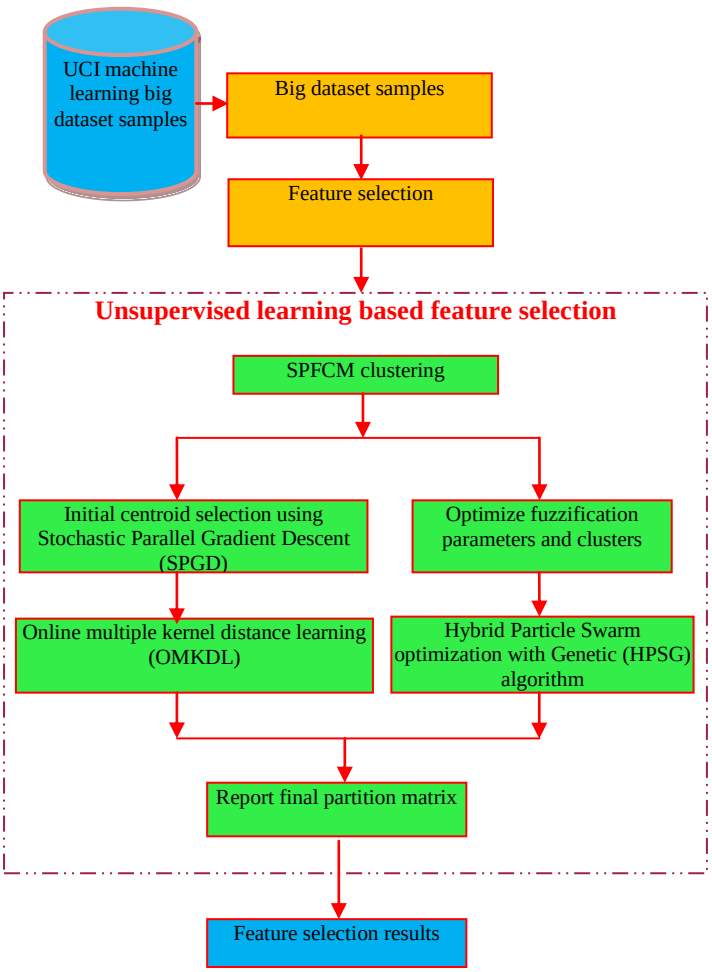

Figure 1. Overall illustration of the proposed work

A. *Stochastic Parallel Gradient Descent (SPGD) algorithm for initial centroid selection*

The proposed SPGD algorithm is applied to the automatic selection of centroid values for all features cluster from big data samples which is collected from benchmark dataset samples (UCI and text classification data samples) .The proposed SPGD evaluated the initial centroid values with the help of gradient function . It is measured using intensity distributions to focus initial centroid value for cluster in SPFCM algorithm and updated as:

$$f^{k+1} = f^k - \gamma \delta CV^k \delta p^k \quad (1)$$

where k is the total number of iteration to complete initial centroid selection process, $f = (f_1, f_2, ..$ be the number of features vector values for big data samples, is the total number of features from big data samples; is the which is positive gain distance value between two features vectors from big dataset samples for minimizing CV, denotes small random initial centroid perturbations; δ is the selected initial centroid value.

$$\delta CV = CV(f + \delta f) - CV(f) = CV(f_1 + \delta f_1, \ldots f_n + \delta f_n, \ldots f \quad (2)$$

The estimated cluster centroid value for ℓ is used as follows :

$$\gamma^{k+1} = \gamma^{k}.CV^{k}(\text{to minimize CV}) \quad (3)$$

In our case, feature samples of big dataset samples determination are plan on the centroid value selection . Here the initial centroid value for FCM clustering of big dataset samples is chosen based on the centroid value metric (CV) :

$$CV = \frac{\iint \sqrt{(x - cvx')} + (y - cvy')dxdy}{\iint I(x,y)\,dxdy} \quad (4)$$

where $B(x$ is the feature vector intensity distribution for big dataset samples ; $(cvx', cv$ is the selected initial feature vector centroid for SPFCM clustering. Fuzzification parameter (m) is optimized using hybrid particle Swarm optimization with Genetic (HPSG) algorithm

B. HPSG algorithm to optimize fuzzification parameter

Hybrid PSO based methods is proposed by Xin et al [21] and PSO-DE is also proposed by Akbari and Ziarati [22] . Proposed work is extended to optimize the fuzzification parameter (m) for SPFCM algorithm**.** Genetic algorithm creates new fuzzification parameter (m) values through combining existing fuzzification parameter (m) values via Ring crossover (RC) operation. The two different randomly initialized fuzzification parameter (m) from PSO is given to genetic RC operation thus generates new fuzzification parameter. The proposed work RC is performed two different fuzzification parameter values which follows four major steps is illustrated in Figure.2 and summarized as follows.

Step-1: In this step, two fuzzification parameter (m) (particles) is given for crossover process, as shown in Figure. 2(a).

Step-2: The two different fuzzification parameters (particles) are initially merged through a form of ring, as shown in Figure. 2(b).

Step-3: The children of two different fuzzification parameter (particles) is created ,similarly for all fuzzification parameters is chosen with the help clockwise direction and anti-clockwise direction, as shown in Figure.2(c).

Step-4: In final step of the work swapping and reversing process is performed for newly selected fuzzification parameter (particles) in the step 3, as shown in Figure. 2(d). In swapping process, a number of fuzzification parameters are swapped in crossed fuzzification parameters. In reversion process, the remaining fuzzification parameters are reversed in crossed fuzzification parameters particles.

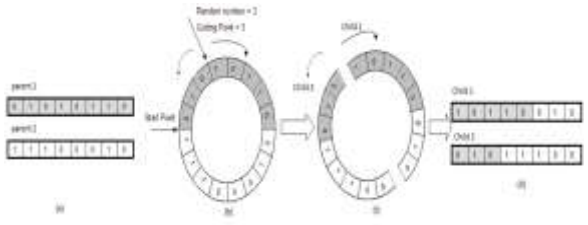

Figure 2. Ring crossover

The basic operation of genetic and crossover operation is modify the position of fuzzification parameter in the PSO algorithm. The Fuzzification particles parameter positions are updated via (5) and (6).

$$v_i^t = m_{i1}^t + F(m_{i2}^t - m_{i3}^t) \quad (5)$$

$$mp_{i1}^t = \begin{cases} v_{i,j}^t & rand() < pm_{cr} \\ mp_{i,j}^t & \end{cases} \quad (6)$$

where , represent the positions of three individual fuzzification parameters in t^{th} generation; is the j^{th} element of fuzzification parameter in t^{th} generation; F is scaling factor ( ), and is called the crossover probability for fuzzification parameter m. If new fuzzification parameter position is best than older one ,it save as the global position in HPSG. The local search algorithm for fuzzification parameter is described in Algorithm 1(at step 3) , where is the t^{th} generation of fuzzification parameter m update their position through ring crossover and swapping in step 3 and step 4 performs until the optimization of fuzzification parameter .

Algorithm 1: HPSG algorithm to optimize fuzzification parameter

1. Initialize fuzzification parameter m as particles with local and global position
2. Use the ring crossover operator to update the fuzzification parameter for current particles using (5) and (6)
3. Randomly select fuzzification parameters and perform local search of fuzzification parameter
   - I. Let $\Delta mp_i^0 = mp_i^{t+1} - mp_i^t$ and iteration =0
   - II. If f(mp_i^{t+1})> f(mp_i^t) ,let $\Delta mp_i^0 = -\Delta mp_i^0$
   - III. If $itr < n_{itr}$ go to next step or else final step VII
   - IV. $mp_i^{'} = mp_i^{t+1} + \Delta mp_i^{itr}$
   - V. if f($mp_i^{'}$)> f(mp_i^{t+1}) ,let $\Delta mp_i^{itr+1} = \frac{\Delta mp_i^{itr}}{2}$ orelse $mp_i^{t+1} = mp_i^{'}$
   - VI. $itr = itr + 1$ go to step III
   - VII. Finish the process
4. Update the local and global position of fuzzification parameter (m) for particles
5. If the maximum number of iterations is met ,then go to next step or else move to step 2
6. Output optimized global fuzzification parameter value from PSO

Similarity measure among features and selected feature for each cluster is determined via kernel learning space function $\mathcal{K}($ which is denoted using Hilbert Space function via extending the work to linear function $L:\mathcal{H} \rightarrow$ that maps between features and selected feature. The distance between the selected cluster centroid feature and feature is denoted as $d_L(f,$,

$$d_L(f, cv) = < k(cv, .), L(x, .) >_{\mathcal{H}} \quad (7)$$

is the feature of the big dataset samples and be the selected cluster center results from SPGD algorithm with linear space operator which is represented as follows,

$$L_{tr} = arg \min_{L \in \mathcal{L}} \frac{1}{2} \|L - L_{tr-1}\|_{HS}^2 + C l_L(f_{tr}, f_{tr}^+, f_{tr}^-) \quad (8)$$

Where $\|L - L_{tr-1}\|$ function represents linear operator follows the procedure of Hilbert Space function and $l_L(f_{tr}, f_{tr}^+, f_{tr}^-) = max\ (0, 1 - d_L(f, f_{tr}^+) + d_L(f, f_t$, denotes the positive selected feature value of cluster and denotes the selected data points belongs to negative selected feature value for each iteration of clustering process.

$$L_{tr} = L_{tr-1} + \tau_{tr} \quad (9)$$

$$\tau_{tr} = \min\left\{Cl, \frac{\max\ (0, 1 - d_L(f, f_{tr}^+) + d_L(f, f_{tr}^-))}{k(f_{tr}, f_{tr})\left(k(f_{tr}^+, f_{tr}^+) - 2k(f_{tr}^+, f_{tr}^-) + k(f_{t}^-\right.\right. \quad (10)$$

The major objective of this work is to expand distance similarity value of SPFCM clustering method into learning procedure based on kernel function as objective fucntion is specified as follows : $mk = (mk_1, .. m$ and it is changed in FCM function is articulated as follows :

$$Df(x, cv) = \sum_{g=1}^{G} mk_g\, d_{Lg}(f, cv) = \sum_{g=1}^{G} mk_g\left(k_g(cv, .), L_g\right. \quad (11)$$

Algorithm 2:Online multiple kernel distance learning(OMKDL)

Input kernel parameters kernel space function and the cluster center

1. Initialize $L_0 = 1$
2. For $tr = 1..TR$ do
3. Receive a training feature samples big dataset samples $(f_{tr}, f_{tr}^+, f_{tr}^-)$
4. Compute τ_{tr} from equation (10)
5. Update L_{tr} from equation
6. Compute distance function with kernel space in (11)
7. End for

Using this $Df(x,$ similarity measure, the method calculates a membership degree for each features of cluster by (12).

$$u_{ij} = \frac{d_{ij}^{-1/(m-1)}}{\sum_{l=1}^{c} d_{ij}^{-1/(m-1)}} \quad (12)$$

The convergence speed of the proposed SFCM is restricted via the fuzzy membership function (m) so the proposed work have higher convergence speed since it is optimized using **HPSG**. The proposed SPFCM method for clustering big data samples features is summarized as follows:

Algorithm 3:Proposed SPFCM clustering algorithm

1. Select number of clusters nc ,number of iterations ,fuzzification parameter from HPSG algorithm and convergence threshold ε
2. Select number of cluster centroid cv from SPGD
3. Run the procedure of FCM by one iteration ,to calculate initial partition matrix
4. For $itr = 1$ to maximum number
   I. Apply OMKDL to member of each cluster and calculate $Df(x, cv)$ in (11) ,support vectors
   II. Calculate fuzzification matrix U and assign each feature vector samples corresponding clusters
   III. Compute $E = \|U_{itr} - U_{itr-1}\|$
   IV. If $E < \varepsilon$ stop
   V. Else next itr
5. Report final partition matrix and clustering the selected feature samples as final result.

IV. EXPERIMENTATION WORK

The proposed unsupervised fuzzy based cluster learning and existing clustering methods for feature selection is implemented with the help of mat lab environment and high performance accuracy is also achieved. In the experiment to measure the feature selection result of unsupervised learning methods the following three feature selection algorithms is chosen from base work: trace-ratio [23], HSIC [24], SPFS [25] and proposed SPFCM clustering is also compared among them. The experimentation work of these selected methods is applied to four benchmark data sets ,two of them be text data and remaining two of them be the UCI data .text dataset samples were extracted from 20-newsgroups data [26] RELATH (BASEBALL vs. HOCKEY) and PCMAC (PC vs. MAC). Two are UCI data: Com-munities and Crime Unnormalized (CRIME) and SLICELOC [27]. The CRIME (u10mf5k) consists of 5,000 features and 10 million instances and SLICELOC (s25mf5k) consists of 5,000 features and 25 million instances. A detail on the dataset samples is specified in TABLE I .

TABLE I. SUMMARY OF THE BENCHMARK DATA SETS

Dataset	Features	Instances	Classes
RELATH	4322	1427	2
PCMAC	3289	1943	2
S25mf5k	5000	25000000	-
U10mf5k	5000	10000000	-

TABLE II shows the results of the clustering methods in terms of data variance with p-val and experimented by different algorithms. The result shows that SPFCM achieved achieves higher clustering accuracy when compared to base work (3 methods) on all four data sets. The result shows the strong ability of the proposed SPFCM algorithm is designed for maintaining variance in feature selection.

TABLE II. UNSUPERVISED FEATURE SELECTION: EXPLAINED VARIANCE WITH P-VAL

Algorithm	RELATH	PCMAC	S25mf5k	U10mf5k
trace-ratio	0.44	0.43	0.32	0.38
HSIC	0.44	0.42	0.45	0.42
SPFS	0.47	0.45	0.46	0.48
SPFCM	0.63	0.61	0.65	0.56

Redundancy rate: TABLE III shows that clustering results of the various feature selection methods based on the average redundancy rate. It shows that SPFCM and SPFS clustering methods achieve higher results than other two algorithms. This is to be predictable, because they are considered to manage unnecessary features, while the remaining methods don't handle this.

TABLE III. Unsupervised feature selection: redundancy rate with p-val

Algorithm	RELATH	PCMAC	S25mf5k	U10mf5k
trace-ratio	0.27	0.20	0.32	0.21
HSIC	0.25	0.17	0.23	0.17
SPFS	0.11	0.07	0.16	0.12
SPFCM	0.03	0.05	0.05	0.06

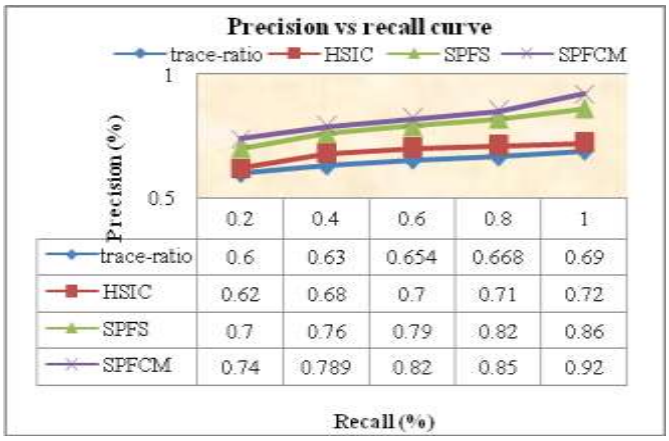

Figure 3. Precision-recall for feature selection vs. clustering methods

Fig.3 shows the precision-recall results for various feature selection algorithms on the four benchmark dataset samples. It shows that the proposed SPFCM achieves higher precision and recall results than the baseline three methods , since the proposed clustering methods is based on the unsupervised fashion with optimized fuzzification parameter and automatic selection of centroid value which improves the accuracy of cluster to handle larger and big dataset samples.

V. CONCLUSION AND FUTURE WORK

This paper presents a novel unsupervised learning based FCM clustering method for feature selection of big dataset samples. The proposed unsupervised cluster learning methods eliminate the inappropriate and insignificant features of big dataset samples via FCM clustering .The proposed unsupervised cluster learning methods is varied from general FCM clustering methods, since the proposed cluster centroid value is automatically selected via SPGD method and fuzzification parameter (m) is also optimized automatically via HPSG algorithm to enhance feature selection for larger dataset samples. The proposed unsupervised learning algorithm types a combined method for feature selection. For unsupervised feature selection, it sustains the learning procedure of FCM clustering method. It is extremely appropriate designed for big dataset samples due to its ease and capability to decrease the number of features via the number of iterations. As demonstrated by a wide-ranging experimental learning, the proposed SPFCM Clustering method can attain higher performance in the favor of unsupervised feature selection. The present work will be extended to the sustain of semi-supervised ,hybrid semi-supervised feature selection leaning algorithm and feature extraction methods is also via PCA ,ICA ,LDA ,Fisher based LDA and LDA.

REFERENCES

[1] J. Deng, A. C. Berg, and F. Li, "Hierarchical semantic indexing for large scale image retrieval, " In CVPR,. IEEE, 2011 ,pp.785-792.

[2] K. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg, " Feature hashing for large scale multitask learning, " In Proceedings of the 26th Annual International Conference on Machine Learning ,2009, pp. 1113-1120.

[3] B . Blum, M. I. Jordan, D. E. Kim, R. Das, P. Bradley, and D. Baker, " Feature selection methods for improving protein structure prediction with rosetta," In Advances in Neural Information Processing Systems, 2007, pp. 137-144.

[4] A. Dasgupta, P. Drineas, and B. Harb, " Feature selection methods for text classification," In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining ,2007,pp. 230-239.

[5] QinbaoSong, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data, " IEEE transaction on Knowledge and Data Engineering , vol.25,no.1,2013,pp. 1-14.

[6] I. Guyon, A.Elissee, " An introduction to variable and feature selection", Journal of Machine Learning Research,vol.3,2003,pp. 1157-1182

[7] M.J. Zaki ,C.T. Ho, eds.: Large-scale parallel data mining, Springer ,2000, No. 1759.

[8] Z. Zhao and H.Liu, "Searching for Interacting Features in Subset Selection" ,Journal Intelligent Data Analysis, vol.13,,no.3,2009, pp. 207-228.

[9] S. Sonnenburg, G. Ratsch, and K. Rieck. Large scale learning with string kernels. MIT Press, Cambridge, MA , 2007, pp. 73-103.

[10] J.Wu, "Efficient hik svm learning for image classification," IEEE Trans. Image Process, vol.21, 2012, pp.4442-4453.

[11] P. Li, Robust logitboost and adaptive base class (abc) logitboost, In UAI, 2010.

[12] S. Li and Z. Zhang, "Floatboost learning and statistical face detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol.26, 2004,pp.1112–1123.

[13] C. Ding, H.Peng, "Minimum redundancy feature selection from microarray gene expression data," In: Proceedings of the CSB, 2003,pp. 523-529

[14] Y. Luo, D. Tao, C.Xu, , D. Li and C.Xu, "Vector valued multi-view semi-supervised learning for multi-label image classification," In Proc. AAAI,2013.

[15] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," In Proc. ICML, 2007, pp.1151– 1157.

[16] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection, " In NIPS, 2005,pp. 507-514.

[17] Z. Ma, Y. Yang, F. Nie, J. R. R. Uijlings, and N. Sebe, " Exploiting the entire feature space with sparsity for automatic image annotation," In ACM Multimedia, 2011, pp.283–292.

[18] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis,"In SDM, 2007, pp. 641-646.

[19] Wu, K. Yu, H. Wang, and W. Ding, " Online streaming feature selection," In ICML, 2010, pp.1159–1166,.

[20] Singh, S., et al, "Parallel large scale feature selection for logistic regression," In: Proc. of SDM. 2009, pp. 1172-1183.

[21] B. Xin, , J. Chen, Z. Peng, F. Pan, "An adaptive hybrid optimizer based on particle swarm and differential evolution for global optimization," Science China Information Sciences ,vol.53,2010,pp. 980–989

[22] R. Akbari, K .Ziarati , " Combination of particle swarm optimization and stochastic local search for multimodal function optimization ," In: IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application (PACAIIA), 2008, pp. 388–392.

[23] F. Nie, et al, "Trace ratio criterion for feature selection," In Proceedings of the 24th international conference on Machine learning, 2008, pp. 823-830.

[24] L. Song, et al , "Supervised feature selection via dependence estimation," In: Proceedings of ICML, 2007, pp. 823-830.

[25] Z. Zhao, L. Wang, H. Liu, J. Ye, "On similarity preserving feature selection," IEEE Transactions on Knowledge and Data Engineering 99 2011, pp.198-206.

[26] http://people.csail.mit.edu/jrennie/20Newsgroups/.

[27] http://archive.ics.uci.edu/ml/index.html.