Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

# Sparse least squares support vector machines based on Meanshift clustering method

**X. Wang, H. Liu, and W. L. Ma**

*School of Automation and Information Engineering, Xi'an University of Technology, Xi'an, China*
*(E-mail: liuhan@xaut.edu.cn)*

**Abstract:** Aim to non-sparsity solution problem of least squares support vector machines and not zero support vector value, a Meanshift Clustering Algorithm to select training samples is proposed, in which keeps the samples with large contribution value and removes the samples with small contribution value. Sparse solution and effective classification model could be obtained. The experimental results on UCI test data set show that the proposed algorithm in this paper can effectively reduce the classification error and improve training efficiency by sparsing training samples of least squares support vector machines. The feasibility and effectiveness of proposed method are proved to be valid.

*Keywords:* LS-SVM, sparse, Meanshift clustering, K-means clustering

## 1. INTRODUCTION

Support vector machines (SVM) is a well-established learning machine based on statistical learning theory [Vapnik (1995)], which is widely used due to its strong nonlinear processing ability and good promotion ability. The classical SVM training algorithm solves a convex quadratic programming problem, but the solution to the quadratic programming problem makes the training process complicated and takes up more resources. Suykens and Vandewalle proposed Least Squares SVM (LS-SVM) algorithm [Suykens and Vandewalle (1999)], which obtains the optimal classification surface by modifying the inequality constraints into equality constraints to reduce the computational complexity. However, improvement of LS-SVM method also leads to lack of sparseness, which will bring some restrict to the processing of large-scale samples. For LS-SVM non-sparse solution problem, Suykens had proposed pruning algorithm [Suykens and Vandewalle (1999)], which needs to be modelled by complete data set, and then pruning is implemented based on the value of support vector. So there is no advantage in computational complexity and time. Another solution is to select the larger contribution value of training samples through some way, model with small-scale samples. This can greatly reduce the number of training samples, while ensuring the accuracy and ability to promote the algorithm at the same time, improve the efficiency of the algorithm. [Hou, Yang, and An (2009); Yu, Zou, and Zhao (2012)].

Clustering algorithm can effectively gather similar data, not only an important data analysis method, but also one of the data pre-processing methods. The current study has been able to effectively deal with the large-scale, high-dimensional data with K-means clustering algorithm [Guo and Lin (2017)]. K-means clustering algorithm is simple and easy to understand, but its adaptive is not good, which need to set initial cluster center and number of clusters at first, making the clustering effect very unstable. The larger the number of samples, the convergence rate is slow and even does not converge, which makes the K-means clustering algorithm cannot be applied to sparseness of large-scale data samples. Meanshift algorithm is a nonparametric algorithm proposed by Fukunaga and Hostetler [Fukunaga and Hostetler (1975)], which is used to deal with complex multimodal feature space analysis and feature clustering. It does not require the knowledge of clusters number and shape, in which is applicable to large-scale data samples that do not know the number of clusters in advance.

For the combination of clustering algorithm and LS-SVM, the common research is mainly to use LS-SVM for clustering results respectively. As in [Li, Xiong and Liu (2013)], the data clustering is done first, and then the sub-LSSVM model is established for each class. But in this algorithm sample sparseness is not used to establish the model. On this basis, a method based on Meanshift clustering is proposed in this paper, in which the sample with large contribution in the training sample is selected as the support vector of model to solve the problem of LS-SVM solution sparseness.

## 2. LSSVM AND MEANSHIFT CLUSTERING

### *2.1 Least squares support vector machines*

Given the classification training set $(x_i, y_i)$, $i = 1, 2, \ldots, N$, $x_i \in R^N$ is the training samples, $y_i \in \{-1, +1\}$ is the label, least squares support vector machines optimization problem is

$$\min_{\omega, b, e} J(\omega, e) = \frac{1}{2}\omega^T \omega + \gamma \frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{1}$$

Subject to

$$y_i[\omega^T \phi(x_i) + b] = 1 - e_i, i = 1, \ldots, N \tag{2}$$

Where $\gamma$ is the regularization parameter, $e_i$ is the error. The Lagrange function is introduced to solve the optimal problem

$$L(\omega, b, e, \alpha) = \frac{1}{2}\omega^T \omega + \gamma \frac{1}{2}\sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N} \alpha_i \{y_i[\omega^T \phi(x_i) + b] - 1 + e_i\} \tag{3}$$

The optimal conditions are obtained by partial differential for each variable

$$\begin{aligned} &\frac{\delta L}{\delta \omega} = 0 \rightarrow \omega = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) \\ &\frac{\delta L}{\delta b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0 \\ &\frac{\delta L}{\delta e_i} = 0 \rightarrow \alpha_i = \gamma e_i \\ &\frac{\delta L}{\delta \alpha_i} = 0 \rightarrow y_i[\omega^T \phi(x_i) + b] - 1 + e_i = 0 \end{aligned} \tag{4}$$

It can be seen from (4) that each component of $\alpha_k$ is proportional to the error of sample, and the components are almost nonzero. All training samples have a role in the classification decision function, thus losing the sparseness of SVM solution, which will lead to the decision-making speed reduction. Therefore, in large-scale samples, it is necessary to set out the LS-SVM with sparse solution from the sample selection, which can effectively reduce the computational complexity of solving.

### *2.2 Meanshift Clustering*

Based on the basis of feature space analysis, the well-known Meanshift algorithm is proposed in [Comaniciu and Meer (2002)], which uses the kernel density estimation, also called Parzen window method in [Duda and Hart (1973)].

Given $n$ data points $(i = 1, 2, \ldots, n)$ in the $d$ dimension space $R^d$. In the space optional $x$, then the basic form of Meanshift vector is defined as

$$M_h = \frac{1}{k}\sum_{x_i \in S_k} (x_i - x) \tag{5}$$

Where $S_k$ is a high-dimensional sphere with radius $h$, and the set of $y$ points satisfying the following relationship

$$S_h(x) = \left\{ y : (y - x_i)T(y - x_i) < h^2 \right\} \tag{6}$$

Where $k$ indicates that there are $k$ points falling into the $S_k$ region at $n$ sample points $x_i$ and $T$ means transform. Added the kernel function into basic Meanshift vector, the Meanshift algorithm is transformed into

$$\hat{f}_{h,k}(x) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \tag{7}$$

Where $h$ is the radius, $\frac{c_{k,d}}{nh^d}$ is the unit density. To make $f$ on the maximum, the easiest to think is on the type of derivative

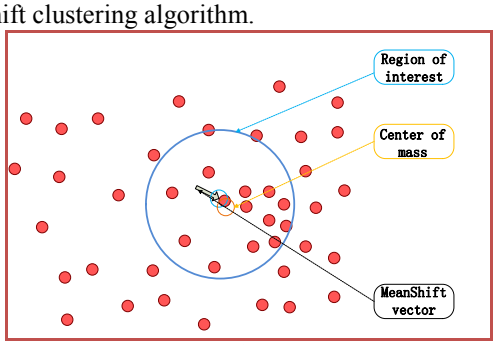
$$\nabla f_{h,k}(x) = \frac{2c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n} (x - x_i)k'\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \tag{8}$$

Let $g(x) = -k'(x)$ and $\hat{\nabla} f_{h,k}(x) = 0$, it can be drawn a new center coordinates

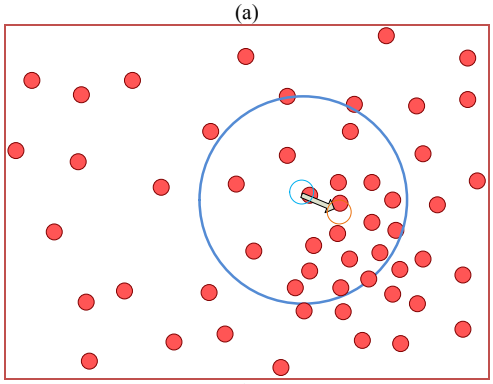
$$x = \frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} \tag{9}$$

As can be seen from (9), Meanshift vector always points to the direction in which the density increases, and the local mean always drifts to the high density region. When the shift vector is zero, the drift ends and its corresponding point is the local density maximum point.

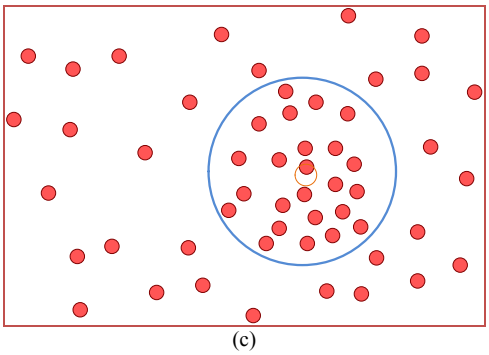
So the Meanshift clustering algorithm consists of three steps: clustering center search, clustering center clustering/merging similarity region and merging small area (optional). The purpose of clustering center search is to find cluster center of all data points. But the clustering center search process will usually get many center points, whose distance is relatively close. If each center is adopted as a class, the classification will be too much to lead to over-segmentation. So region merging is necessary. Fig.1 shows an intuitive description of Meanshift clustering algorithm.




(a)



(b)



(c)

Fig.1 Meanshift process

Fig.1 (a) shows the initial interest area which can be randomly selected and determined the initial center of mass. The Meanshift process is shown in Fig.1 (b) to calculate the new Meanshift vector. Fig.1 (c) means the Meanshift vector is stable and the final position is local maximum of the probability density function, also the clustering center.

## 3. SPARSE LSSVM BASED ON MEANSHIFT CLUSTERING

Considering solution sparseness, the idea of sparse LS-SVM based on Meanshift clustering is proposed to solve the sparse problem of LS-SVM. First, Meanshift clustering algorithm is used to reduce the training samples, find out the data of support value with the higher contribution value in the large samples and remove the data with less influence on the classification result. And then LS-SVM model is established by using the reduced samples as training samples to construct the LS-SVM sparse model based on the Meanshift clustering algorithm.

### *3.1 Determining support vectors*

In SVM, most of the samples near the optimal classification surface are concerned, because support value is not equal to zero are the data points near the classification surface. So it is possible to greatly reduce the amount of computation and the resources that need to be consumed as long as the appropriate method is taken to remove the unsupported vector before training. However, the support vector values in LS-SVM are nonzero, and the absolute value of support vector values indicates the degree to which each data point contributes to the model. The data points with smaller $\alpha_k$ are less relevant to the establishment of classifiers, and it is similar that in standard SVM training samples with $\alpha_k = 0$ do not contribute to the model. Taking the Breast Cancer (BC) data set as an example, the respective descending order sequences are obtained by SVM and LS-SVM. The corresponding relationship between the sequence and the data samples is shown in Fig.2.

It can be clearly seen from Fig.2 that in SVM, when the number of data samples is N = 64, the value of $\alpha_k$ in the subsequent data samples is zero. But there is no explicit zero nor zero boundary in LS-SVM, $\alpha_k$ are just infinitely close to zero. It can be considered that each data points are related to establishment of the model and some of data points are more important than the other, which also verified non-sparsity problem of LS-SVM solution. In addition, the size of $\alpha_k$ is related to the distribution of data points. In [Song, Cui, and Hu (2008)], the relationship between $\alpha_k$ and data points distribution on the two sides of classification surface is explained by a linear separable two-class problem in two-dimensional space. It shows that the support values of data samples with very close or very far from the decision-making boundary are relatively large, which is slightly different from the non-zero support values of the data points near the decision boundary in SVM.

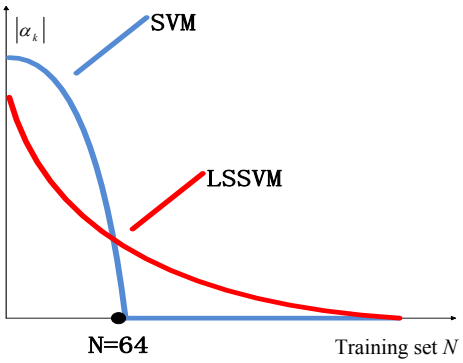



Fig.2 Support vector values in descending order in LSSVM and SVM

From the perspective of spatial geometry, the distance from the data point $x_k$ to the classification sub-plane is

$$d_k = \frac{(\omega^T x_k + b)}{\|\omega\|} \tag{10}$$

And according to the LS-SVM equation constraint condition (2) and $\alpha_k = \gamma e_k$, it can be obtained

$$d_k = \frac{(\omega^T x_k + b)}{\|\omega\|} = \frac{(1 - e_k)}{y_k \|\omega\|} = \frac{(1 - \frac{\alpha_k}{\gamma})}{y_k \|\omega\|} \tag{11}$$

$$\alpha_k = \gamma(1 - d_k y_k \|\omega\|) \tag{12}$$

So from (12) we can see that if $y_k = +1$, it means $d_k$ to be as much possible as small in order to take a large $\alpha_k$ value. If $y_k = -1$, it means $d_k$ to be as much possible as large in order to take a large $\alpha_k$ value. It also means that the data points close to the decision-making boundary or very far from the decision both have a large support vector values.

### *3.2 Clustering support vectors*

Meanshift clustering algorithm is used to cluster the training set first, and the samples are located near the classification surface and far samples are modeled as the training set. It can

be found out the data of large support value with higher contribution value are obtained, and the data with less influence on the classification result are removed to reduce complexity of time and space. Fig.3 shows a sparse LS-SVM algorithm based on Meanshift clustering.

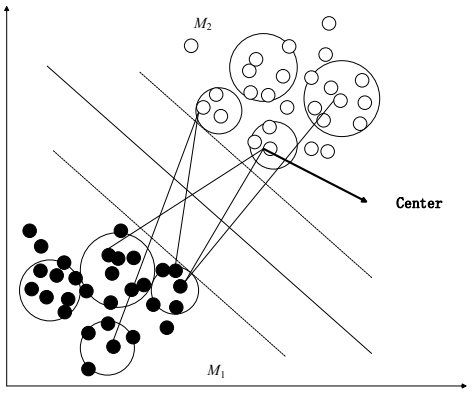



Fig.3 Sparse LSSVM based on Meanshift Clustering

The sparse LSSVM algorithm based on Meanshift clustering is described as follows.

Let $N$ be the training sample data set, $M_1$ is the number of class A sample and $M_2$ is the class B sample with $M_2$=$N$-$M_1$. Parameter $m$, $n$ control sample selection number, $T$ is the test set.

Step1: Apply Meanshift clustering algorithm to $M_1$ and $M_2$ data sets respectively, and record all kinds of cluster centers.

Step2: Calculate the distance $d_i$ of $M_1$ to $M_2$ in each class and the distance $d_j$ of $M_2$ to $M_1$ by (11).

Step3: Select $M_1$ in the distance from the nearest $m$ class and the furthest $n$ categories. Similarly, select the $M_2$ from the $M_1$ nearest $m$ class and the furthest $n$ categories, marked.

Step4: Remove the $M_1$, $M_2$ marked samples as new training set $N'$ for LS-SVM.

Step5: Enter $N'$ to model and get classifier.

Step6: Use the classifier to make a decision output for $T$.

The flow charts of algorithm are shown in Fig.4.

It can be seen from Fig.4 that the algorithm is divided into sparseness stage and LS-SVM stage. Meanshift clustering algorithm is used to get the reduced samples in the training set in sparseness stage, and the samples are modeled in the LS-SVM stage.

## 4. EXPERIMENTAL VERIFICATION

In order to test the effectiveness of algorithm, the experiments are carried on with K-means clustering [Zhou, Zhang and Pan (2017)] and the proposed algorithm under the same conditions. The kernel function is selected by RBF radial basis function. Experimental hardware environment include the processor Inter i5, 3.30GHz, 4.00GB memory on the computer, using Matlab R2014a version. The data in this paper are selected from four UCI datasets including Breast Cancer (BC), Cmc, Kr-vs-Kp (KK), and MAGIC Gamma Telecope (MGT).

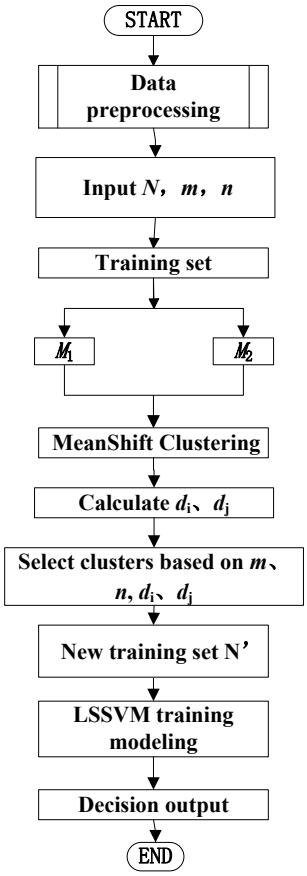



Fig.4 The flow chart of training process

The characteristics of each data set are summarized in Table.1.

**Table.1 A summary of the data set**

| Dataset | Classification number | Number of instances | Dimension |
|---|---|---|---|
| BC | 2 | 424 | 30 |
| Cmc | 2 | 1000 | 9 |
| KK | 2 | 3000 | 36 |
| MGT | 2 | 10000 | 10 |

The size of each data sample set after Meanshift clustering algorithm is shown in Table.2.

**Table.2 MEANSHIFT clustering algorithm for each data set reduction results**

| Dataset | Train | Train' | Reduction ratio |
|---|---|---|---|
| BC | 424 | 185 | 56.37 |
| Cmc | 1000 | 427 | 57.3 |
| KK | 3000 | 1475 | 50.83 |
| MGT | 10000 | 4790 | 52.10 |

Table.2 shows the four data sets after reduction. "Train" indicates the number of training subsets obtained using the Meanshift clustering algorithm, "Reduction ratio" is used to indicate the degree of data reduction, which means that the higher the rate of reduction, the greater the magnitude of the reduction. The reduction ratio is calculated by

$$\text{R}eduction\ ratio = \frac{Train - Train^{'}}{Train} \times 100\% \tag{13}$$

The training time and classification accuracy of each data set by LSSVM training is shown in Table.3.

**Table.3 Classification accuracy and training time**

| Dataset | Test | Classification accuracy (%) | | Training time(s) | |
|---|---|---|---|---|---|
| | | K-means | Meanshift | K-means | Meanshift |
| BC | 200 | 96.7 | 98.2 | 0.039 | 0.023 |
| cmc | 500 | 96.0 | 96.3 | 0.052 | 0.021 |
| kk | 1500 | 96.2 | 96.7 | 0.069 | 0.025 |
| MGT | 5000 | / | 94.2 | / | 1.045 |

In Table.3, "Test" is a randomly selected test samples, "Classification accuracy" is used to determine the performance of the model and "training time" indicates the time spent in the model training phase. "K-means" refers to the classification result obtained by using the K-means clustering. “Meanshift” is used to express the sparseness LSSVM model established by Meanshift clustering sparse algorithm.

As can be seen from Table.3, MGT is not convergent when the K-means clustering is used due to the large amount of data. It is verified that the K-means clustering is not suitable for the thinning of large-scale data samples, but also proved the effectiveness of the proposed algorithm. For each data set, this algorithm compared to the general LS-SVM model, in the classification accuracy and training time are reflected in the advantages. This shows that the key of model establishment is to use decisive role data points for the optimal classification plane other than the more data samples and the better. In order to assure the accuracy of classification, delete the sample which has no effect on the performance of the model or has little effect on the model, otherwise, the excessive training sample will increase the redundancy of the model and interfere with the model, not only the model is not concise enough , It may also reduce the model's performance. The experimental results verify the feasibility and effectiveness of the proposed LS-SVM algorithm based on Meanshift clustering.

## 5. CONCLUSIONS

In this paper, a sparse LS-SVM algorithm based on Meanshift clustering is proposed. The Meanshift clustering algorithm is used to select the large contribution samples. The accuracy and efficiency of algorithm are improved. The numerical experiments for four real data sets verify the feasibility and effectiveness of proposed algorithm.

## ACKNOWLEDGMENT

The authors would like to thank the Key Project of Shaanxi Key Research and Development Program, under Grant 2018YFZDGY0084, the Research Program of Shaanxi Modern Equipment Green Manufacturing Co-innovation Center under Grant 304-210891704, the Distinctive Research Program of Xi’an University of Technology under Grant 2016TS023 and Research project of Shaanxi Provincial Education Department under Grant 2017JS088 for providing the fund.

## REFERENCES

Comaniciu, D. and Meer, P. (2002) Meanshift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5): 603-619.

Duda, R. O. and Hart, P. E.(1973) *Pattern Classification and Scene Analysis*. Wiley.

Fukunaga, K. and Hostetler, D. (1975) The estimation of the gradient of a density function with applications in pattern recognition. IEEE Transactions on Information Theory, 21(1): 32-40.

Guo, Z. Y. and Lin, T. (2017) Research on K-means algorithm for fast clustering of large scale data. Journal of Computer Applications and Software, (05), 43-47+53.

Hou, L., Yang, Q., and An J. (2009) An improved LSSVM regression algorithm, 2009 International Conference on Computational Intelligence and Natural Computing, Wuhan, China, IEEE, 138-140.Vapnik, V. N.(1995) *The nature of statistical learning theory*, New York：Spinger Verlag.

Li, L. J., Xiong, L., and Liu, J. (2013) Multi-model predictive control based on AP-LSSVM. Journal of Zhejiang University (Engineering Science), 47(10), 1741-1746.

Suykens, J. A. K. and Vandewalle, J. (1999) Least squares support veator machine classifiers. Neural Processing Letters, 9(3), 293-300.

Song, Z. Q., Cui, H., and Hu, Y. N. (2008) Research and development of support vector machine theory. Journal of Naval Aeronautical Engineering Institute, 23(2), 143-148.

Yu, Z. T., Zou, J. J., and Zhao X. (2012) Sparseness of least squares support vector machines based on active learning. Journal of Nanjing University of Science and Technology, 36(1), 12-17.

Zhou, T. T., Zhang, H. B., and Pan, L. J. (2017) Application of Three Kinds of Clustering Algorithm in Building Image Segmentation, Modern Computer (Professional Edition), (02):76-80.