21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Iterative Clustering and Support Vectors-based High-Confidence Query Selection for Motor Imagery EEG Signals Classification

Huijuan Yang, Cuntai Guan, Kai Keng Ang, Haihong Zhang and Chuan Chu Wang
*Institute for Infocomm Research, A*STAR, Singapore.*
Email: {hjyang, ctguan, kkang, hhzhang and ccwang}@i2r.a-star.edu.sg

Abstract

This paper proposes a novel active learning method for the classification of motor imagery electroencephalogram (EEG) signals. Specifically, we propose an iterative clustering and support vector-based criterion to select samples of high-confidence to construct a robust training set. The common spatial pattern (CSP)-based features are iteratively clustered till the number of support vectors in the cluster is less than a predefined threshold. A predefined number of samples close to the cluster centers are chosen. When such clusters cannot be found, the samples that are of farthest distances to a group of support vectors of class "0" and "1" are alternately chosen. Experimental results on BCI competition IV dataset IIb show superior performance compared with a baseline method, which is 9% increase in accuracy averaged across subjects and training sizes.

1. Introduction

Motor imagery, i.e., mentally rehearses or simulates a given action, modifies neuronal activity in the similar way as that generated by voluntary movements. The frequency-specific changes such as event-related desynchronization/synchronization patterns can be translated into commands to operate external devices [1]. Machine learning plays an important role in translating brain signals into interpretable commands. However, most supervised classifiers are sensitive to the quality and quantity of training samples and perform poorly when the number of training samples is small or the training samples are not representative of data distributions. Active learning can be employed to select the most informative and representative samples to build a good training set, hence, a robust classifier.

The most commonly used method for query selection is "Query by uncertainty" [2, 3, 4, 5]. It chooses the uncertain samples that lie or are close to the decision hyperplane [2, 3], or chooses those samples with low confidence from the classifier output [5]. The second type of method is "Query by committee" [5, 6], which chooses the samples that are assigned to different class labels by a committee of classifiers. The third type of method is "Query by error reduction" [7, 8], which estimates the expected future error of a model using training set plus the query on unlabeled set. Recently, active learning has found wide applications in text, video and electrocardiographic data classification [4, 5, 9]. Can we use those samples with high-confidence to build a robust classifier? These samples are especially important in exploring the intrinsic data structure of EEG signals. The features are iteratively clustered by choosing the cluster with minimum number of support vectors. This process is iterated till the number of support vectors within cluster is less than a predefined threshold, which leads to the samples of high confidence being selected. If such samples cannot be found, the samples that are of farthest distances to a group of support vectors of class "0" and "1" are alternately selected. Our proposed **I**terative **C**lustering and **S**upport **V**ector-based **H**igh-**C**onfidence **Q**uery **S**election (ICSV-HCQS) is illustrated in Figure 1.

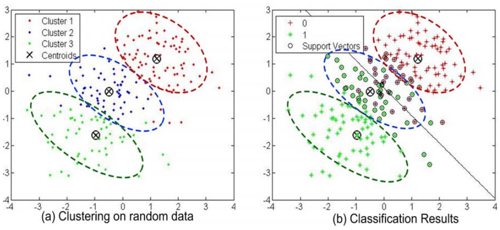


Figure 1: An illustration of query selection criterion. The high-confidence samples are either selected from those clusters with no or few support vectors from iterative clustering, or selected from those that lie far away from a group of support vectors of class "0" and "1" of current classifier.

2 Proposed Active Learning Method

We now describe our proposed active learning scheme. The initial labeled set L contains a small amount of labeled data, denoted as E_l=$(e_{ijk})^{n_c \times n_t \times n_l}$, where n_c, n_t and n_l are the number of channels, samples and trials, respectively. $Y_l \in \{0, 1\}$ are the labels. Set U contains the unlabeled data, denoted as E_u=$(e_{ijk})^{n_c \times n_t \times n_u}$. It is generally assumed that the amount of unlabeled data is much larger than that of labeled ones, i.e., $n_u \gg n_l$ holds. The EEG signals are firstly band-pass filtered, the CSP filters are then computed using labeled data and used to spatially filter the band-pass filtered signals. Finally, the log variance of the spatially filtered signals for the first and last several rows are used as features, denoted as F_l and F_u for labeled and unlabeled data [1]. Note that a sample here refers to a trial which is represented by features.

Proposed **ICSV-HCQS** algorithm.

Input: the labeled and unlabeled set of features.
Output: actively selected labeled set and a classifier.

We firstly formulate the active learning process as:

$$Q_e = \mathcal{L}_a(\mathcal{C}_c, \mathcal{Q}_f, F_l, F_u) \qquad (1)$$

The learner repeatedly call the function $\mathcal{Q}_f$ to obtain the query samples based on the current classifier $\mathcal{C}_c$ and features for labeled and unlabeled data, i.e., F_l and F_u. The criterion for query sample selection by taking support vector machines (SVMs) classifier as an example is illustrated in Figure 2. As is known, SVMs are the hyperplanes that separate the training data in a maximum margin. The support vectors (SV_s) that lie near to the hyperplanes are utilized to select the representative high-confidence queries. The initial set of indexes for clustering is I_{dx}=$L \bigcup U$. The features for both labeled and unlabeled data, i.e., F_s=$[F_l\ F_u]$, are clustered into a predefined number of clusters (n_k), which are given by

$$[I_c, \mathcal{O}_c] = \mathcal{C}_m(F_s, n_k, I_{dx}) \qquad (2)$$

where $\mathcal{C}_m$ are the clustering methods such as K-Means clustering; I_c and $\mathcal{O}_c$ are the indexes of samples for each cluster and that of the cluster centers. The number of support vectors from the ith iteration that fall into the kth cluster (denoted as $N_{sv}(k)$, whose indexes are $I_{sv}(k)$) is firstly calculated. The cluster that has the minimum number of SV_s is likely to lie far away from existing hyperplane, e.g., clusters 1 and 3 in Figure 1. The query samples chosen from these clusters are of high confidence considering existing hyperplane, the inclusion of which will hence boost the performance. The chosen cluster index $\hat{k}$ for (i+1)th iteration is given by

$$\hat{k} = \arg\min_{k} (N_{sv}(k) | I_{sv}(k) \in I_c(k)) \qquad (3)$$

where k=$1, 2, ..., n_k$ denotes the number of clusters, $\hat{k}$ is index of the cluster that has the minimum number of SV_s. If the number of SVs in the cluster (denoted as $N_{sv}(\hat{k})$) is not greater than a predefined threshold, i.e., $N_{sv}(\hat{k}) \leq T_{sv}$ (*threshold condition*), the clustering process stops. Otherwise, the cluster that has the smallest N_{sv} is chosen and the clustering process continues. The iterative clustering process makes it possible to discover the intrinsic data structure, as a result, those samples with high confidence are selected. However, how to choose the instances from the selected cluster to avoid the noisy outliers is still a concern. To address this problem, we propose to search in the selected cluster and choose those samples that are of minimum distances to the cluster centers, which is given by

$$\hat{j} = \arg\min_{j, j \in U \& j \in I_c(\hat{k})} (D_s(F_s(j), F_s(\mathcal{O}_c(\hat{k})))) \qquad (4)$$

where $D_s(F_s(j), F_s(\mathcal{O}_c(\hat{k})))$ measures the Euclidean distances between features of the jth instance and that of the center of selected cluster, i.e., $\mathcal{O}_c(\hat{k})$. $D_s(\mathbf{a}, \mathbf{b})$=$\sqrt{\sum_{i=1}^{l_f} (a(i) - b(i))^2}$ for features $\mathbf{a}$ and $\mathbf{b}$, where l_f is the dimension of the features.

Assume Q_e query samples are required in each iteration, however, only Q_n instances can be selected from the chosen cluster, or no cluster can be found to satisfy the *threshold condition*, i.e., $Q_n < Q_e$. The rest of samples Q_r=Q_e-Q_n will be chosen by measuring the distances between the features of instances in unlabeled set and groups of SVs in ith iteration. Q_r samples of the farthest distances alternately to a group of SVs in class "0" and "1" are selected. Each query qq ($qq \in Q_r$) is chosen by

$$qq = \begin{cases} D_m(k, I_{sv}(1)) & \text{if } mod(m, 2) \\ D_m(k, I_{sv}(0)) & \text{otherwise} \end{cases} \qquad (5)$$

where $mod()$ is modulo function and m=$(i-1)$*Q_r+p, where i and p denote iteration and pth query in ith iteration. $I_{sv}(q)$ (q=0,1) denote indexes of SVs in classes "0" and "1", i.e., $I_{sv}(q)$=$I_{sv}|(Y_l(I_{sv})$=$q)$. $D_m(k, I_{sv}(q))$ (denoted as D_m for brevity) is given by

$$D_m = \arg\max_{k, k \in U} (\sum_{v=1}^{n_v} D_s(F_s(k), F_s(I_{sv}(q, v)))) \qquad (6)$$

where n_v is the number of SVs in class q. The labeled (L) and unlabeled (U) sets are thus updated and the CSP filters are re-computed based on the new labeled set.

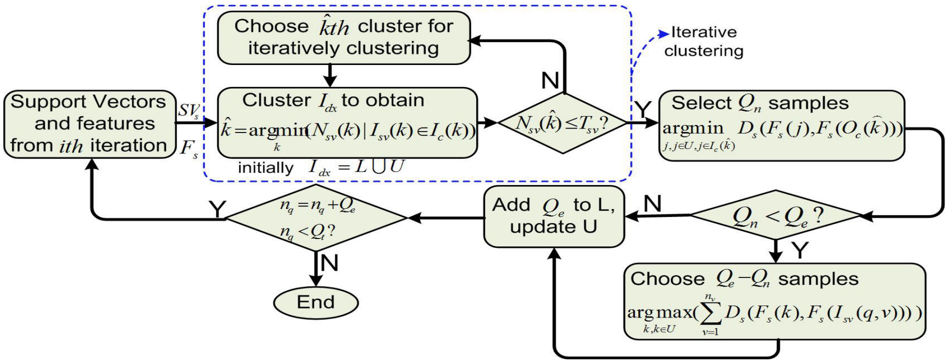
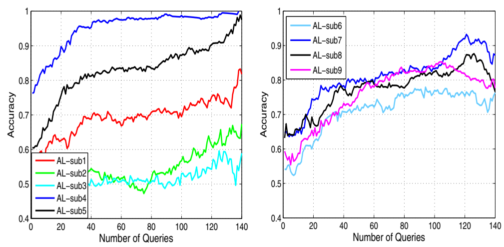


Figure 2: Iterative Clustering and Support Vector-based High-Confidence Query Sample Selection (ICSV-HCQS).

The number of query samples is updated by n_q=n_q+Q_e. If n_q does not reach the total number of queries Q_t, i.e., n_q<Q_t, the query process will continue.

3 Experimental Results

Experiments are conducted to validate our proposed method, BCI competition IV data set IIb, session 3 is used. The data consist of 9 subjects performing motor imagery of left and right hand, with 120 trials per session. Three bipolar recordings were recorded with a sampling frequency of 250Hz. The data were bandpass-filtered between 0.5Hz and 100Hz. The number of initially labeled data (randomly selected) is chosen as 14 and time segments from 0.5s to 2.5s with reference to the onset of cue are used. The average accuracies of 10 runs are used as the results which are shown in Figure 3. It can be observed from the figure that active query selection has boosted the performance especially when the number of training samples is small, e.g., less than 40. However, the performance decreases when there is not enough unlabeled data to select, e.g., the query number is greater than 120. A simplified FBCSP algorithm



(a) Subjects: 1-5. (b) Subjects: 6-9.

Figure 3: Classification accuracy for nine subjects using proposed active query selection.

.

[1] is used as the baseline passive learning algorithm for comparison (PS-sFBCSP). SVM with linear kernel function is used as the classifier. It is worth noting that no selection of subject-specific time segments and frequency bands are carried out for both PS-sFBCSP and our proposed method. Overlapping frequency bands from 4 to 30Hz with a step size of 2Hz are employed. For PS-sFBCSP, a predefined number of training samples (e.g., L=14, 28, 56, 70, 98 and 126) is randomly chosen to train the classifier, which is then used to classify the unlabeled data, with the average classification results of 10 runs shown in Table 1. While the number of samples (L) used for training the classifier is actively selected for our proposed approach. It is easily noticed that active learning can pick up the most informative samples of high-confidence to build a robust classifier. This leads to an accuracy increase of 9% compared with that achieved using the passive learning method. Note that the accuracy is averaged across subjects and training sizes (see the two rows in bold in Table 1). A paired sample t-test is conducted to test the null hypothesis that the difference of accuracies of active and passive learning methods is a random sample from a normal distribution with mean 0. The null hypothesis is rejected at 5% significance level with p=0.0208. This indicates the significant difference in average accuracies of active versus passive learning methods.

Note that recomputing the CSP filters definitely increases the computational load in the calibration. Nevertheless, once the classifier is built, it can be used for online testing, which is fast enough. The reasons for adding a pre-clustering process instead of directly selecting those high-confidence samples lie in: a) clustering is effective to exclude the outliers or noisy samples. b) iterative clustering is effective to explore the intrinsic cluster structures of the data. Hence, more discriminant features can be included in the training set. A

Table 1: Comparisons of Accuracies Using Different Training Sizes (L) for Passive and Active Learning Methods

Methods	Subj.	L(14) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(28) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(56) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(70) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(98) ($\mathcal{A}_c \pm \mathcal{V}_r$)	L(126) ($\mathcal{A}_c \pm \mathcal{V}_r$)
Passive (PS-sFBCSP)	S1	52.05±5.93	55.00±4.60	59.13±4.69	60.89±5.51	59.68±7.53	60.88±6.80
	S2	49.18±2.43	46.36±4.21	46.54±4.51	47.78±4.68	43.55±9.34	41.18±3.40
	S3	49.11±3.86	51.14±3.85	47.50±3.46	44.33±5.28	41.13±4.64	44.12±10.09
	S4	69.11±7.68	80.83±5.46	87.40±4.33	91.22±3.07	91.45±3.88	92.06±3.68
	S5	55.00±3.81	56.74±5.50	64.62±3.39	63.22±5.50	67.42±4.86	65.59±9.61
	S6	52.33±3.10	53.56±5.52	61.06±3.43	59.00±4.43	59.35±4.29	61.47±6.42
	S7	62.12±6.87	66.36±6.13	68.46±7.08	72.11±4.27	72.10±4.10	70.88±5.62
	S8	60.21±7.74	61.36±4.97	66.92±4.18	66.89±5.07	68.06±4.61	71.47±5.20
	S9	56.10±4.17	58.64±4.09	64.23±4.90	66.44±4.18	65.32±3.67	66.18±8.23
	A_{ac}	**56.13±5.07**	**58.89±4.93**	**62.87±4.44**	**63.54±4.67**	**63.12±5.21**	**63.76±6.73**
Active (ICSV-HCQS)	S1	55.70±0.37	59.30±0.39	66.26±0.44	70.13±0.24	72.04±0.44	72.89±0.20
	S2	52.76±0.12	51.08±0.08	49.57±0.08	53.14±0.35	54.99±0.28	58.89±0.20
	S3	51.54±0.09	51.30±0.27	51.75±0.10	50.16±0.24	47.54±0.30	52.69±0.46
	S4	70.68±0.59	81.81±0.83	95.55±0.13	97.59±0.02	98.11±0.004	98.59±0.02
	S5	63.82±0.24	70.95±0.16	79.75±0.11	81.91±0.07	84.34±0.12	88.93±0.18
	S6	54.46±0.06	58.13±0.22	68.51±0.34	72.62±0.41	79.30±0.14	77.15±0.59
	S7	59.84±0.26	67.72±0.22	77.75±0.23	82.45±0.09	86.71±0.14	89.79±0.28
	S8	68.75±0.32	67.21±0.31	77.66±0.12	77.85±0.11	78.93±0.16	81.44±0.49
	S9	59.27±0.11	61.67±0.37	74.46±0.20	77.55±0.17	83.71±0.19	84.36±0.43
	A_{ac}	**59.65±0.24**	**63.24±0.32**	**71.25±0.19**	**73.71±0.19**	**76.19±0.20**	**78.30±0.32**

Note: A_{ac}: Average Accuracy (%) over all subjects (shown in bold in the last row). $\mathcal{A}_c$: Accuracy (%), $\mathcal{V}_r$: Variance.

good tradeoff should be achieved on how to choose the maximum number of support vectors in a cluster (T_{sv}). A larger value will compromise the effectiveness of the algorithm, whereas a small value will increase the computational load due to the iterative clustering in searching for the cluster that satisfies the criterion. On the other hand, the prior knowledge of the feature structure of the EEG signals should be taken into consideration in selecting the number of clusters. In the implementation, T_{sv}=2 and n_k=2 are chosen, where the intrinsic data structure can be exposed by the iterative clustering process. In general, the performance of the proposed approach is not very sensitive to the choice of number of support vectors and clusters.

4 Conclusions

In this paper, we presented a novel iterative clustering and support vector-based active learning method for the classification of motor imagery EEG signals. The features are firstly clustered iteratively and the cluster with the required minimum number of support vectors is selected. The samples that are of minimum distances to the center of selected cluster are chosen as queries. When such clusters cannot be found, the samples that are of maximum distances to a group of support vectors of class "0" and "1" are alternately chosen. The selection of these high-confidence samples based on iterative clustering helps discover the intrinsic data structure, which ensures robustness of the classifier. Using BCI competition IV dataset IIb, an increase in the averaged accuracy of 9% is achieved comparing proposed active with that of a passive learning method, which demonstrates the effectiveness of the proposed method.

References

[1] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface," *30th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC) 2008*, pp. 4178-4181.

[2] S. Tong and D. Koller, "Support Vector Machine Active Learning With Applications to Text Classification," *Journal of Machine Learning Research*, vol. 2, pp. 45-66, 2001.

[3] G. Schohn, D. Cohn, "Less is More: Active Learning with Support Vector Machines," *Proc. of the 17th Int. Conf. on Machine Learning,* 2000.

[4] J. Zhu, H. wang, B.K. Tsou and M. Ma, "Active Learning With Sampling by Uncertainty and Density for Data Annotations," *IEEE Trans. on Audio, Speech and Language Proc.,* vol. 18, no. 6, pp. 1323-1331, Aug. 2010.

[5] E. Pasolli and F. Melgani, "Active Learning Methods for Electrocardiographic Signal Classification," *IEEE Trans. on Information Technology In Biomedicine,* pp. 1405-1415, vol. 14, no. 6, November 2010.

[6] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by Committee," *Proc. ACM Workshop Comput. Learn Theory,* pp. 287-294, 1992.

[7] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," *Proc. of 18th Int. Conf. on Marchine Learning (ICML'2001)*, pp. 441-448, 2001.

[8] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions," *In ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*.

[9] W. Hu, W. Hu, N. Xie and S. Maybank, "Unsupervised Active Learning Based on Hierarchical Graph-Theoretic Clustering," *IEEE Trans. on System, Man, Cybernetics-Part B: Sybernetics,* vol. 39, no. 5, pp. 1147-1161, Oct. 2009.