proceedings

MDPI

Proceedings

Query-By-Committee Framework Used for Semi-Automatic Sleep Stages Classification †

Nela Grimova [1,*] and Martin Macas [2]

[1] Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, 166 36 Praha 6, Czech Republic
[2] Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, 166 36 Praha 6, Czech Republic; martin.macas@cvut.cz
* Correspondence: ngrimova@gmail.com; Tel.: +420-608-361-537
† Presented at the 13th International Conference on Ubiquitous Computing and Ambient Intelligence UCAmI 2019, Toledo, Spain, 2–5 December 2019.

Published: 21 November 2019

Abstract: Active learning is very useful for classification problems where it is hard or time-consuming to acquire classes of data in order to create a subset for training a classifier. The classification of over-night polysomnography records to sleep stages is an example of such application because an expert has to annotate a large number of segments of a record. Active learning methods enable us to iteratively select only the most informative instances for the manual classification so the total expert's effort is reduced. However, the process is able to be insufficiently initialised because of a large dimensionality of polysomnography (PSG) data, so the fast convergence of active learning is at risk. In order to prevent this threat, we have proposed a variant of the query-by-committee active learning scenario which take into account all features of data so it is not necessary to reduce a feature space, but the process is quickly initialised. The proposed method is compared to random sampling and margin uncertainty sampling which is another well-known active learning method. It was shown that, during crucial first iteration of the process, the provided variant of query-by-committee acquired the best results among other strategies in most cases.

Keywords: active learning; uncertainty sampling; query-by-committee; PSG classification; sleep stages classification

1. Introduction

Despite there exist a large amount of various machine learning techniques which can be adopted for numerous applications, in more and more real-world settings we often encounter the problem that it is possible to gather a large number of data, but the process of annotating them (i.e., assigning each instance to a specific class so that they can be used for training of a classifier) is expensive and time-consuming.

The classification of over-night polysomnography (PSG) records to sleep stages is a good example of the mentioned problem. In reality, a doctor or a studied annotator has to walk through the whole several hours-long PSG record, which was previously split to 30 seconds-long segments, and manually classify all segments to one of sleep stages [1]. Nowadays, the resolution of sleep phases provided by the American Association of Sleep Medicine (AASM) is used—sleep is divided into five stages: wake, REM (rapid-eye movement sleep), N1, N2 and N3, where N1, N2 and N3 are specified subsets of non-REM sleep (non-rapid eye movement sleep) [2]. It is clear that the whole process is very time-demanding and it would be appropriate to make it more automatic; on the other hand, the information about sleep stages is used for the patient's diagnosis so the review of an expert is crucial.

The solution is in the adoption of semi-supervised methods as active learning which is used for choosing of the instances that are sufficient for learning an adequately good classifier [3].

2. Active Learning

Let X be the observation space and Y the space of classes. At the beginning of semi-supervised methods, there are two sets of instances: a small set of labeled instances S_L which contains observations that are assigned to some class (i.e., their class is known) – $S_L = \{(x^1, y^1), \ldots, (x^l, y^l)\}$, $x_i \in X, y_i \in Y$, and a large set of unlabeled instances $S_U = \{x^1, \ldots, x^u\}$, $x \in X$, whose classes we have no information about. The active learning process can be divided into a few steps [3]:

1. Learn a classifier c on the set of labeled instances S_L.
2. Assign instances from the set of unlabeled instances S_U to some class by using the learnt classifier c.
3. Use a query strategy in order to select instance from set S_U.
4. Ask an "oracle" for the class which the selected instance belongs to. By "oracle" it is often meant an expert—a human annotator who has an expertise in the given field.
5. Add the newly classified instance to set S_L (and remove it from set S_U).
6. Repeat steps 1–5 until a terminal condition is met (e.g., a given number of iterations is reached, the error attained a specified threshold, etc.).

2.1. Query Strategies

In this section, we would like to discuss the third step in the previously mentioned list about the active learning process. The most crucial part of active learning is to determine how instances will be selected for the classification by the "oracle". Settles et al. [3] have introduced a large number of various methods which are commonly used. We will mention two of them, which are in our opinion the most favourite ones—margin uncertainty sampling (MUS) [4] and query-by-committee (QBC) [5]. These two methods are often compared among the literature [6,7].

In the margin uncertainty sampling scenario, the instance, whose class classifier c is the least certain of, is queried [8]. In order to explain it formally, the observation x^* is chosen, for which holds:

$$x^*_{MUS} = \arg\min_x P_{Y|X}(y_2|x) - P_{Y|X}(y_1|x), \tag{1}$$

where $P_{Y|X}$ is the conditional probability of a class when the instance is observed, $\bar{y_1}$ states for the most probable class of x and $\bar{y_2}$ is the second most probable class of x.

The second approach consists in the utilisation of an ensemble of classifiers which represents competing hypotheses [9]. All models are learnt on the set of labeled instances S_L and the instance, which the classifiers disagree the most about, is selected for assigning to its label. For measuring of the level of disagreement we will use the vote entropy [10]—the instance x^* is queried for which holds:

$$x^*_{QBC} = \arg\max_x -\sum_{i=1}^{n} \frac{v(y_i)}{p} \log \frac{v(y_i)}{p}, \tag{2}$$

where n is the number of classes, p is the number of models in the committee and $v(y_i)$ represents how many classifiers decided that instance x belongs to class y_i.

Now let us mentioned third query strategy we will use in order to compare the performance of described strategies: random sampling (RS). As the name suggests, in each iteration of the algorithm the instance, which was selected at random, is queried.

2.2. Advantages and Disadvantages of Active Learning

In this section we would like to discuss a few pros and cons which can be encountered when active learning is utilised.

- **Advantages of Active Learning**

- Saving of time and money: there is no need to annotate a large amount of data, it is sufficient to label only the most informative instances.
- Online adaptation of the classifier: the classifier is automatically retrained when new unseen instances are available.

- **Disadvantages of Active Learning**
  - Application-dependent selection of the query strategy: the query strategy has to be chosen wisely according, i.e., to a chosen classifier (e.g., margin uncertainty sampling is suitable when the classifier computes posterior probabilities [3]), to some specific relationship among data instances in the observation space (then density-weighted methods are useful [11]), etc.
  - Sensitivity to the initialisation: when the process is not properly initialised, the performance of the chosen classifier is insufficient during several first iterations (the so-called "cold start problem" [12]) which can result in a slower convergence of the learning process.

3. Dataset

In our work, the dataset consisting of 36 full-night PSG recordings was used. 18 healthy individuals and 18 insomniac patients were examined using the standard 10–20 montage EEG [13] in the National Institute of Mental Health, Czech Republic. Although EOG and EMG were also recorded, only EEG signals were used in this study due to varying quality in both of EOG and EMG recordings. The detailed specification of the measured group of patients is provided in Table 1.

Table 1. Information about polysomnography (PSG) recordings of tested group of healthy individuals and insomniac patients.

	Healthy Patients	**Insomniac Patients**
Number of patients	18	18
Males	27.8%	38.9%
Recording duration	7.8 ± 0.9 h	7.4 ± 0.7 h

All records were split to 30 seconds-long segments without overlaps. 21 features were extracted from all of used EEG derivatives (namely: Fp1, Fp2, F3, F4, C3, C4, P3, P4, F7, F8, T3, T4, T5, T6, Fz, Cz, Pz, O1, O2), i.e., the total amount of features was $21 \times 19 = 399$. List of all computed features is shown in Table 2. The Continuous wavelet transformation (CWT) was used for obtaining of the frequency spectrum which was utilised for computing of features 7–21.

Table 2. List of extracted features.

	Feature	Description
1	STD	Standard deviation of the signal in the time domain
2	SWNS	Skewness of the signal in the time domain
3	KRTS	Skewness of the signal in the time domain
4	MBL	Mobility of the signal in the time domain
5	CMPL	Complexity of the signal in the time domain
6	E	Shannon entropy
7	SE	Spectral entropy of CWT spectrum
8	SEF90	Spectral edge frequency below 90% of the total power of the signal is located
9	SEF95	Spectral edge frequency below 95% of the total power of the signal is located
10	PPF	Power peak frequency—frequency of maximum power
11	MDF	Mean dominant frequency
12	SMF	Median frequency
13	HFD	Higuchi fractal dimension
14	CWT0.5-3	Relative PSD for frequency band of range 0.5–3 Hz
15	CWT3-7	Relative PSD for frequency band of range 3–7 Hz
16	CWT7-12	Relative PSD for frequency band of range 7–12 Hz
17	CWT11-13	Relative PSD for frequency band of range 11–13 Hz
18	CWT12-22	Relative PSD for frequency band of range 12–22 Hz
19	CWT13-15	Relative PSD for frequency band of range 13–15 Hz
20	CWT22-30	Relative PSD for frequency band of range 22–30 Hz
21	CWT30-45	Relative PSD for frequency band of range 30–45 Hz

Features listed in Table 2 were aggregated by using their median value over all EEG derivatives. As a result, every 30 s-long segment was described by 21 features.

4. Proposed Method

As we mentioned in Section 1, active learning often suffers from the cold start problem. It is now necessary to denote that at the beginning of active learning process the initial set of labeled instances contains only few instances described by relatively many features and this all can lead to overfitting. In our previous work [14] we successfully proposed a method which can be used for the increase of the initial set of labeled instances by 1-nearest neighbour classifier without any additional information about classes of selected instances.

It is also possible to tackle this problem by reducing the feature space. This can be done e.g., by calculating of the mutual information between features and labels [15], i.e., by detecting the features which described instances' classes the most. Values of the mutual information between individual features and labels for all datasets are shown in the Figure 1. At first sight it is clear that skewness, kurtosis and spectral entropy do not acceptably describe classes of instances (values of the mutual information for these features are approaching zero for all datasets). Furthermore, there is not any obvious pattern that some features gives a better detailed account about labels than other features, so the selection of fewer features is not able to be done.

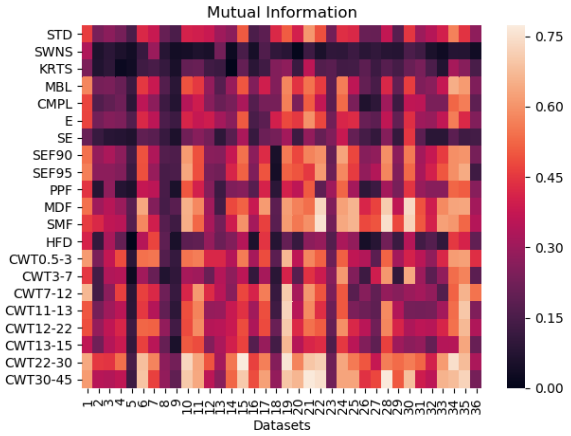


Figure 1. Mutual information between features and corresponding labels for all datasets.

Our idea was to utilise a property of the query-by-committee framework. We created an ensemble of simple linear classifiers, each classifier was learnt only on one feature. The instance, about whose class the classifiers had disagreed the most, was queried, classified, and moved to set S_L. If there are more instances with the same level of disagreement, one of them is randomly selected and queried. As a result of this method, we suppose that the error on testing data will be smaller (i.e., classifier is well adapted to data) in first crucial iterations of the algorithm when the proposed version of query-by-committee will be adopted.

5. Experiments and Results

We split each dataset to a training and a testing subsets, the training sets always contain 60% of all instances from a dataset. Training data were divided to the set of unlabeled data and the set of labeled data in such way that five instances of each class were randomly chosen and were added to set S_L; the set of unlabeled instances was created by the rest of training instances. A linear classifier was chosen for learning on training data and consequently for the estimation of test error E on testing data which is defined as:

$$E = \frac{1}{4} \sum_{i=1}^{4} e_i, \tag{3}$$

where e_i is the percentage of incorrectly classified testing instances of each class.

The whole process followed previously mentioned steps of active learning (see Section 2). Note that E is computed in each iteration.

In order to get more reliable results, the whole process was repeated ten times (each time with different initialisation) and the estimations of E for each iteration were averaged.

We decided to compare three query strategies—random sampling, margin uncertainty sampling and our version of query-by-committee. In Tables 3–5 mean values of E acquired in 5th, 10th and 50th iterations are shown.

Table 3. Values of the average of E obtained in the 5th iteration of the algorithm for all strategies. The smallest value of E within query strategies is in bold.

Dataset	RS	MUS	QBC	Dataset	RS	MUS	QBC
1	0.279	0.241	**0.231**	19	0.309	0.307	**0.276**
2	0.580	0.469	**0.429**	20	0.475	0.399	**0.380**
3	0.539	0.412	**0.351**	21	0.255	0.179	**0.174**
4	0.353	0.347	**0.293**	22	0.312	0.245	**0.220**
5	0.633	0.619	**0.584**	23	0.200	0.154	**0.153**
6	0.354	0.276	**0.225**	24	0.293	0.209	**0.151**
7	**0.339**	0.348	0.343	25	0.277	0.241	**0.210**
8	0.496	0.490	**0.479**	26	0.468	0.423	**0.421**
9	0.621	0.606	**0.592**	27	0.419	0.331	**0.315**
10	0.437	0.279	**0.250**	28	0.182	0.174	**0.168**
11	0.323	0.278	**0.264**	29	0.386	**0.351**	0.380
12	0.472	0.406	**0.399**	30	0.315	0.222	**0.191**
13	0.513	**0.448**	0.461	31	0.405	**0.343**	0.354
14	0.362	0.328	**0.297**	32	0.518	0.477	**0.467**
15	0.395	0.344	**0.334**	33	0.336	0.318	**0.314**
16	0.412	0.335	**0.320**	34	0.287	0.212	**0.211**
17	0.355	0.300	**0.286**	35	0.260	**0.226**	0.229
18	0.614	**0.569**	0.589	36	0.478	0.380	**0.344**

Table 4. Values of the average of E obtained in the 10th iteration of the algorithm for all strategies. The smallest value of E within query strategies is in bold.

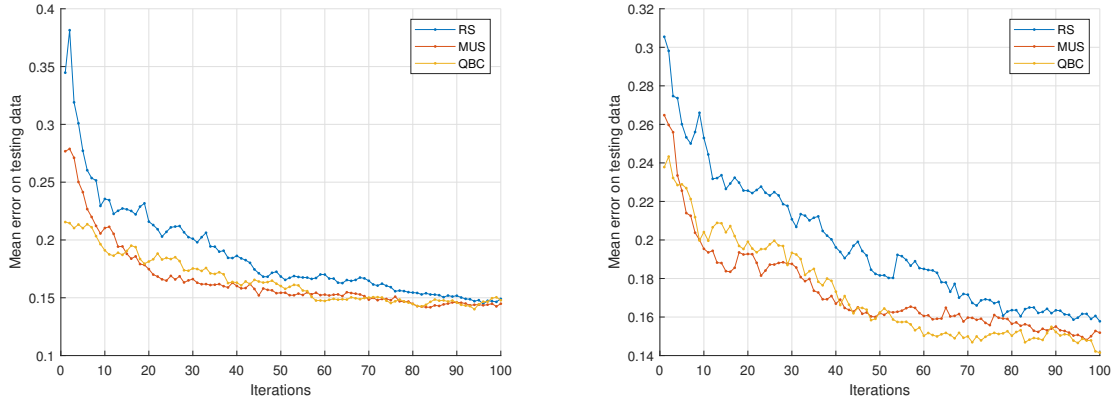
Dataset	RS	MUS	QBC	Dataset	RS	MUS	QBC
1	0.255	0.236	**0.231**	19	0.290	**0.269**	0.290
2	0.522	0.447	**0.399**	20	0.393	**0.363**	0.369
3	0.418	0.368	**0.345**	21	0.238	**0.169**	0.172
4	0.296	0.298	**0.288**	22	0.258	0.229	**0.224**
5	0.636	0.638	**0.605**	23	0.218	**0.128**	0.136
6	0.311	0.244	**0.219**	24	0.239	0.180	**0.171**
7	0.336	**0.322**	0.331	25	0.235	0.210	**0.191**
8	0.480	0.470	**0.467**	26	0.453	0.425	**0.415**
9	0.606	**0.580**	0.596	27	0.373	**0.282**	0.313
10	0.399	0.252	**0.245**	28	0.176	0.169	**0.161**
11	0.286	0.278	**0.253**	29	0.383	**0.329**	0.363
12	0.416	**0.372**	0.382	30	0.212	**0.180**	0.197
13	0.478	**0.427**	0.468	31	0.381	**0.322**	0.326
14	0.360	0.307	**0.295**	32	0.494	0.449	**0.443**
15	0.344	0.345	**0.315**	33	0.331	**0.282**	0.297
16	0.357	**0.298**	0.315	34	0.269	**0.190**	0.195
17	0.307	**0.276**	**0.276**	35	0.253	**0.195**	0.204
18	**0.574**	0.575	0.588	36	0.448	0.320	**0.312**

Table 5. Values of the average of E obtained in the 50th iteration of the algorithm for all strategies. The smallest value of E within query strategies is in bold.

Dataset	RS	MUS	QBC	Dataset	RS	MUS	QBC
1	0.207	**0.191**	0.194	19	0.207	**0.191**	0.243
2	0.390	0.401	**0.375**	20	0.309	**0.256**	0.329
3	0.311	0.297	**0.292**	21	0.154	**0.125**	0.131
4	0.238	**0.216**	0.242	22	0.197	0.175	**0.170**
5	0.639	**0.589**	0.604	23	0.110	**0.084**	0.099
6	0.182	**0.166**	0.188	24	0.167	**0.143**	0.147
7	0.287	0.242	**0.233**	25	0.168	**0.154**	0.160
8	0.426	**0.418**	0.422	26	0.386	**0.331**	0.380
9	0.537	0.530	**0.529**	27	0.227	**0.210**	0.222
10	0.244	0.214	**0.208**	28	0.127	**0.117**	0.127
11	0.244	0.194	**0.187**	29	0.278	**0.244**	0.252
12	0.312	**0.290**	0.349	30	0.128	**0.121**	0.161
13	0.398	**0.354**	0.372	31	0.250	**0.205**	0.253
14	0.265	**0.254**	0.283	32	0.409	**0.362**	0.394
15	0.310	**0.254**	0.287	33	0.245	**0.221**	0.235
16	0.264	**0.230**	0.264	34	0.151	0.137	**0.130**
17	0.250	0.244	**0.228**	35	0.182	**0.162**	**0.162**
18	0.569	**0.549**	0.560	36	0.239	0.235	**0.227**

Let us summarize achieved results. Except Dataset 7, both active learning strategies reached a smaller test error than random sampling in the 5th iteration. Furthermore, the query-by-committee framework overcame margin uncertainty sampling in 30 cases. In the 10th iteration, random sampling acquired the smallest test error only on Dataset 18, the query-by-committee scenario reached the best results in 18 cases, In the end in the 50th iteration, this framework beat other strategies in 9 cases.

Let us show examples of typical results in Figure 2. The averaged test error during first 100 iterations is plotted. In both cases, the error of the query-by-committee achieves smaller values than other strategies in several first iterations, than the results are in favour of margin uncertainty sampling.



(**a**) Dataset 25 (**b**) Dataset 35

Figure 2. Mean test error during 100 iterations of the algorithm for all used strategies.

6. Conclusions and Discussion

In this paper, we adopted the query-by-committee framework which consists in training the ensemble of basic linear classifiers (each classifier is learnt on one feature) on the set of unlabeled data. An instance, which class classifiers disagree the most, is then chosen, annotated and added to the set of labeled instances.

Acquired results showed that the test error in several first iterations is actually smaller when query-by-committee is used in comparison with margin uncertainty sampling, but margin uncertainty sampling is in most cases faster in following iterations. The statement, that the utilisation of the proposed variant of the query-by-committee framework is able to help the classifier with a faster adaptation to high-dimensional data, was fulfilled. This leads to the conclusion that the proposed variant of the query-by-committee scenario leads to preventing the cold start problem. Note that random sampling almost always achieved the worst results, so this validates the usage of active learning strategies.

The contribution of margin uncertainty sampling is invaluable, but the utilisation of this method is often limited, because margin uncertainty sampling is entitled to a proper classifier (as it was mentioned above, only classifiers which estimates posterior probabilities can be used). On the other hand, query-by-committee is more robust which was shown e.g., in [6]). Furthermore, our proposed method handles both selection of the most informative instance and dealing with high-dimensional data.

That raises a question of using the combination of both tested query strategies – the variant of query-by-committee at the beginning and then margin uncertainty sampling in next iterations. This will be tested in the future work as well as the utilisation of the proposed method on different data.

Author Contributions: Conceptualization, N.G. and M.M.; Methodology, N.G.; Software, N.G. and M.M.; Validation, N.N.; Investigation, N.G.; Writing—Original Draft Preparation, N.G.; Supervision, M.M.

Funding: The research has been supported by CVUT institutional resources (SGS grant No. SGS17/216/OHK4/3T/37).

Acknowledgments: We would like to acknowledge Martin Brunovsky, Jana Koprivova, Daniela Dudysova, and Alice Heuschneiderova from National Institute of Mental Health, Czech Republic, and Vaclav Gerla from Czech Institute of Informatics, Czech Technical University in Prague, Czech Rebublic, for their assistance, discussions, and contribution to the acquisition and evaluation of the PSG data.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AASM	American Association of Sleep Medicine
CWT	Continuous wavelet transform
EEG	Electroencephalography
EMG	Electromyography
EOG	Electrooculography
MUS	Margin uncertainty sampling
PSD	Power spectral density
PSG	Polysomnography
QBC	Query-by-committee
RS	Random sampling

References

1. Gerla, V. Automatic Analysis of Long-Term EEG Signals. Ph.D. thesis, Czech Technical University, Prague, Czech Republic, 2012.
2. Duce, B.; Rego, C.; Milosavljevic, J.; Hukins, C. The AASM recommended and acceptable EEG montages are comparable for the staging of sleep and scoring of EEG arousals. *J. Clin. Sleep Med.* **2014**, *10*, 803.
3. Settles, B. *Active Learning Literature Survey*; Computer Sciences Technical Report 1648; University of Wisconsin–Madison: Madison, WI, USA, 2009.
4. Scheffer, T.; Decomain, C.; Wrobel, S. Active hidden markov models for information extraction. In Proceedings of the International Symposium on Intelligent Data Analysis, Springer, 13–15 September 2001; pp. 309–318.
5. Seung, H.S.; Opper, M.; Sompolinsky, H. Query by Committee. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, COLT '92, Pittsburgh, PA, USA, 27–29 July 1992; ACM: New York, NY, USA, 1992; pp. 287–294.

6. Ramirez-Loaiza, M.E.; Sharma, M.; Kumar, G.; Bilgic, M. Active learning: An empirical study of common baselines. *Data Min. Knowl. Discov.* **2017**, *31*, 287–313.
7. Schein, A.I.; Ungar, L.H. Active learning for logistic regression: An evaluation. *Mach. Learn.* **2007**, *68*, 235–265.
8. Lewis, D.D.; Catlett, J. Heterogeneous Uncertainty Sampling for Supervised Learning. In Proceedings of the Eleventh International Conference on Machine Learning, New Brunswick, NJ, USA, 10–13 July 1994; pp. 148–156.
9. Tomanek, K. Resource-aware Annotation Through Active Learning. Ph.D. thesis, Technical University Dortmund, Dortmund, Germany, 2010.
10. Dagan, I.; Engelson, S.P. Committee-Based Sampling For Training Probabilistic Classifiers. In Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; pp. 150–157.
11. Settles, B.; Craven, M. An Analysis of Active Learning Strategies for Sequence Labeling Tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08, Honolulu, HI, USA, 25–27 October 2008; Association for Computational Linguistics: Stroudsburg, PA, USA, 2008; pp. 1070–1079.
12. Attenberg, J.; Provost, F. Why label when you can search? Alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–28 July 2010; pp. 423–432.
13. Klem, G.H.; Lüders, H.O.; Jasper, H.; Elger, C. The ten-twenty electrode system of the International Federation. *Electroencephalogr Clin Neurophysiol* **1999**, *52*, 3–6.
14. Grimova, N.; Macas, M.; Gerla, V. Addressing the Cold Start Problem in Active Learning Approach Used For Semi-automated Sleep Stages Classification. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018; pp. 2249–2253.
15. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *8*, 1226–1238.