Ensemble Clustering Classification Applied to Competing SVM and One-Class Classifiers Exemplified by Plant MicroRNAs Data

Malik Yousef[1,*], Waleed Khalifa[2] and Loai AbdAllah[2,3]

[1]Community Information Systems, Zefat Academic College, Zefat, 13206, Israel

[2]Computer Science, The College of Sakhnin, Sakhnin, 30810, Israel

[3]Management Information Systems, The Max Stern Yezreel Valley College, 19300, Israel

Summary

The performance of many learning and data mining algorithms depends critically on suitable metrics to assess efficiency over the input space. Learning a suitable metric from examples may, therefore, be the key to successful application of these algorithms. We have demonstrated that the k-nearest neighbor (kNN) classification can be significantly improved by learning a distance metric from labeled examples. The clustering ensemble is used to define the distance between points in respect to how they co-cluster. This distance is then used within the framework of the kNN algorithm to define a classifier named ensemble clustering kNN classifier (EC-kNN). In many instances in our experiments we achieved highest accuracy while SVM failed to perform as well. In this study, we compare the performance of a two-class classifier using EC-kNN with different one-class and two-class classifiers. The comparison was applied to seven different plant microRNA species considering eight feature selection methods. In this study, the averaged results show that EC-kNN outperforms all other methods employed here and previously published results for the same data. In conclusion, this study shows that the chosen classifier shows high performance when the distance metric is carefully chosen.

1 Introduction

MicroRNAs represent a recently discovered class of non-coding RNAs that play key roles in the post-transcriptional regulation of gene expression in animals and other species. Mature miRNAs are ~22 nucleotides in length and are harbored within pre-miRNAs which in turn are excised from primary microRNAs (pri-miRNAs). For plants, the latter are highly heterogeneous in length [1].

A number of computational prediction tools have been developed to detect miRNAs and their targets based on the parameterization of pre-miRNAs and target duplex structures [2][3][4]. Most of these approaches are using two-class machine learning algorithms (TCC) and depend on a negative class artificially created using randomly generated or selected sequences.

However, some recent studies have consider using one-class classifiers (OCC) that don't rely on a negative class of unknown quality [5–10]. We recently analyzed the use of OCC for miRNA detection in plants and found that it was competitive in comparison to TCC although the analysis was biased towards TCC [7]. Machine learning depends on parameterization and many features describing a pre-miRNA have been proposed. Feature selection has been

* To whom correspondence should be addressed. Email: malik.yousef@gmail.com

investigated before, but mostly for TCC [11–13], while only little has been done for OCC [14–16]. In this study, we used the same feature selection approaches from the study of Khalifa and colleagues [17] and compared their effectiveness with our new EC-kNN classifier. Different selection approaches provide different set of features that we consider in current study.

The performance of many machine learning algorithms depends critically on being given a good metric representing the complete input space. Moreover, in many cases the geometric distance does not reflect the actual similarity between points. The following example illustrates this situation (Figure 1). Considering the dataset in Figure 1a with two labelled points a classifier using the Euclidean distance metric would perform rather poor and many points belonging to the magenta class would be classified to be red as shown in Figure 1b. It is important to note that in this case the problem is not in selecting the labelled points and any other two points will bring to us to a poor result because the geometric distance does not reflect the actual similarity between the points as shown in this example. On the other hand, we can see that points with the same cluster may share some common features even though their geometric distance is large. As a result, in this research we decided to measure the similarity between the points according to the clustering results. Returning to the example in Figure 1a, we can see that it worked very well (Figure 1c).

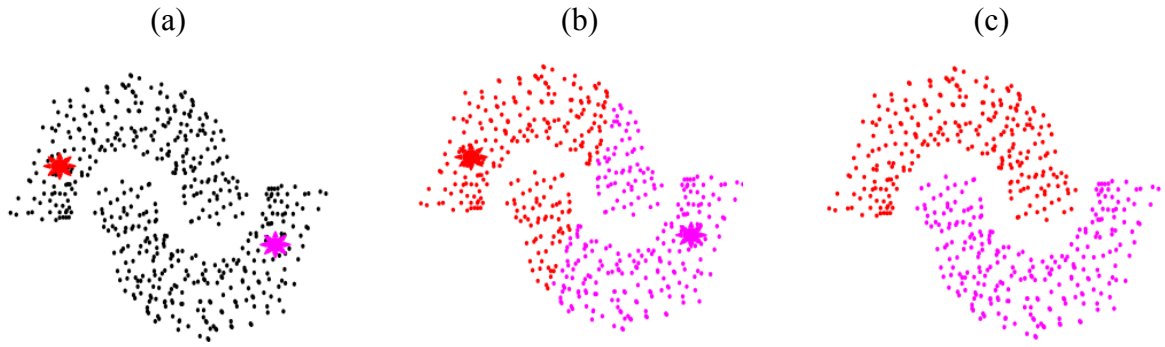


Figure 1. Example of Euclidean distance and Clustering based distance methods to differentiate the points into clusters.

As a result, we conclude that learning a "good" metric from examples may therefore be the key for a successful application of these algorithms. For instance, many researchers have demonstrated that k-nearest neighbor (kNN) [18] classification can be significantly improved by learning a distance metric from labeled examples. Especially, in the field of bioinformatics many researches have used the kNN classifier based on Euclidean distance that may not reflect the actual similarity between the data samples. To this end, we her propose the use of the EC-kNN classifier [19] which offers a novel way of computing the metric for the kNN algorithm.

Khalifa and colleagues [17] showed that feature selection is essential for OCC and a difference of about 30% accuracy can be observed, the maximum difference for TCC is ~10%. Khalifa et al. (2016) conclude that feature selection is essential for OCC, but it does not affect TCC as much.

In this study we consider all the data used in Khalifa et al. (2016) study and apply our EC-kNN method. The results clearly show that EC-kNN outperformed the two-class SVM and also outperformed results from other published studies on the same data.

This paper is organized as follows: Related work and previous methods are discussed in Section 2. The distance metric using ensemble clustering is described in Section 3. Experimental results are presented in Section 4. Finally, our conclusions are presented in Section 5.

2 Related work

2.1 Motif extraction

Here a sequence motif is a short stretch of nucleotides that is widespread among plant hairpins. Motif discovery in turn is the process of finding short sequences within a larger sequence; here we are searching for sequence motifs among plant precursor sequences. The distribution of the length of animal miRNA precursors is in the range of 45-215 nt with mean about 87 nt, while precursors from plants show large heterogeneity, lying in the range of 55-930 nt with a mean of ~146 nt [20]. The MEME (Multiple EM for Motif Elicitation) [21] suite web server is used in our study to discover sequence motifs from our input data which consist of plant pre-microRNA (positive sequences) and plant pseudo hairpins (negative sequences). The MEME algorithm for motif discovery is based on [22] which works by searching for repeated, ungapped sequence motifs that occur in the DNA or protein sequences. MEME provides the results as regular expressions like in the following example:

```
[GA]A[GAC][AC][GC]A[AG]A[CG][AG][GA][ACG][AC][AGC][AC][CG][GAC][AGC]
AAA
```

Nucleotides within brackets represent alternatives for the given position in the sequence; without brackets, only the given nucleotide occurs abundantly within all collected sequences representing the motif.

2.2 Sequence-based and motif features for plant pre-miRNA parameterization

Simple sequence-based features have been described and used for *ab initio* pre-miRNA detection in numerous studies. These simple features, also called words, k-mers, or n-grams, describe a short sequence of nucleotides of length k or n. For example a 1-gram over the alphabet {A,T,C,G} can produce the words A,T,C,G; while a 2-gram over {A,U,C,G} can generate: AA, AC, AG, AU, CA, CC, CG, CU, GA, GC, GG, GU, UA, UC, UG, and UU. Higher n have also been used [23] but selectively for interesting 3-grams.

Motif features are different from n-grams in that they are not exact and allow some degree of error tolerance. In this study motifs are represented as regular expressions (see above). Regular expressions are widespread in approximate pattern matching and many programs allow searching with regular expressions (e.g.: most Linux tools such as grep). Here we use PatMatch [24] to analyze whether a pattern is within a hairpin (1) or not (0). The hairpin is analyzed using the following algorithm:

```
Let w be the length of the given motif
Let max be 0
For i:0 to len(sequence)-w+1
     Align w sized window with i^th position of sequence
     Let ls be the calculated match score
     updateMax(ls,max)
Report max and return corresponding motif
```

Different studies have used this representation such as [25], [26]. Moreover, MEME provides the sequence profile additionally to the regular expression that can be consider for representation.

2.3 One class classification

For one-class classification the DDtools[27] implementation of an OCC was utilized. 100 fold Monte Carlo cross validation [28] was performed using randomly sampled 90% of the positive data for training and 10% for testing. Moreover, the pseudo negative sequences were injected as unknown class during testing. We employed k-means in this study as previously described [29] since it performed well in respect to OCC although it is a clustering algorithm.

Kmeans is a well-known clustering algorithm which can partition data into k clusters. Using OC-kMeans we divide the data into k clusters. For an unknown sample z the distance d(z) to all clusters is calculated. Generally, the class is assigned by returning the label of the closest cluster. In this case, learned clusters are from the target class and thus if the unknown example is closer to the clusters than a threshold, they are assigned the target class or otherwise receive the label 'unknown'.

2.4 Two class classification

Support Vector Machines (SVMs) are used in machine learning and were first proposed by [30]. SVMs have been used in bioinformatics [31–35] and in the field of pre-miRNA detection. Here, the WEKA library [36] SVM implementation, which is based on LibSVM [37], was utilized. The radial basis function was set to a gamma value of 0.7 and the cost parameter was chosen to be 4.0 and the normalization option was set to true. Any machine learning algorithm needs initial training and we performed a 100 fold Monte Carlo cross validation [28] during learning, by employing random sampling using 90% of the data for training and 10% for testing.

2.5 Feature selection strategies

Feature selection has been shown to be an NP-hard problem and, therefore, other approximate feature selection strategies have been developed. In machine learning for pre-miRNAs more than 1000 features have been proposed which makes feature selection especially hard [38]. To investigate the impact of feature selection on model performance for OCC and TCC, four negative and four positive feature selection methods were designed. Previously Sacar and Allmer [38] found that a set of 50 to 100 features may be sufficient for successful pre-miRNA detection. Using more than 50 features increases the likelihood that the feature set contains some features which may conceal differences among feature selection methods. Therefore, a feature set size of 50 was selected for model training in this study. Previously, Yousef et al. [29] performed feature selection for OCC using similar feature selection methods as we consider here. It is also important to compare the impact of OCC and TCC. We have considered the eight feature selection methods that were used in the study by [17]. Briefly, the

methods are: selecting features with low information gain (LIG), random feature selection (RFS), selecting random features from feature clusters (RFC), and selecting features from clusters (SFC). The latter were selecting features with high information gain (HIG), selecting the highest information gain from feature clusters (HIC), zero-norm feature selection (ZNF), and Pearson correlation-based feature selection (PCF).

All feature selection methods except for the last two were performed using KNIME [39]. The workflows for the eight feature selection methods, developed in KNIME, are available for download from the following website: http://bioinformatics.iyte.edu.tr/supplements/featsel. For more detailed information see [17].

2.6 k nearest neighbor using ensemble clustering

Classification can be significantly improved by learning a distance metric from labeled examples. However, like any classifier, kNN has some drawbacks. One of its main drawbacks is that most implementations of kNN use only the geometric distance to measure the similarity and the dissimilarity between the objects without using any statistical regularity in the data.

In this study, we used clustering for defining a better metric rather than using only the geometric distance to measure the similarity and the dissimilarity between the objects. As there is no optimal clustering algorithm with optimal parameter values, several clustering runs were performed yielding an ensemble of clustering results. The distance between points was defined by how many times the points were not clustered together. This distance is then used within the framework of the kNN algorithm (EC-kNN). Applying this method solved the distance definition problem. Moreover, points that are always clustered together in the same cluster (distance= 0) are defined as members of an equivalence class. As a result, the algorithm now runs on equivalence classes instead of single points. In our experiments the number of equivalence classes is usually less than one tenth to one fourth of the number of points. This equivalence class representation is in effect a novel data reduction technique which can have a wide range of applications. It is complementary to other data reduction methods such as feature selection and methods for dimensionality reduction such as the well-known principal component analysis (PCA). In the following a more formalized description of the algorithm as pseudocode:

1. $Cls_{mat} \leftarrow \{empty\}$ //Cls is the clusters matrix
2. repeat until satisfied
   2.1. Run a clustering algorithm with different parameters and store the clustering label for each point in c, $c \leftarrow alg_{clustering}(data)$.
   2.2. Add the clustering results to the clustering matrix, $Cls_{mat} \leftarrow Cls_{mat} \cup \{c\}$.
3. Return Cls_{mat}

For a given dataset the well-known k-means clustering algorithm was run with k = 2 .. 30. The results are stored in the clusters matrix C. The equivalence relation was employed to build the equivalence matrix C'. The new space is usually smaller than the original space. Our algorithm assumes that points belonging to the same equivalence class have the same label.

At the first stage a training set of size 80% of the labelled dataset is randomly drawn. Then the kNN algorithms with k = 3 is applied on the test part. The results are averaged over 100 different runs on each dataset.

3 Application

3.1 Data

Positive examples for pre-miRNAs from selected plant species were downloaded from miRBase [40] (Releases 20 and 21). Glycine max (gma), Zea mays (zma), Sorghum bicolor (sbi), Physcomitrella patens (ppt), Arabidopsis thaliana (ath), Populus trichocarpa (ptc), and Oryza sativa (osa) make up the positive dataset. Negative examples for miRNAs consisted of 980 pseudo pre-miRNAs from the PlantMiRNAPred dataset [35]. For these data, all pre-miRNA features were calculated as described previously [7], [38], [41] and also repeated in the next sections.

Table 1: plant species and their number of pre-miRNA sequences available in miRBase.

Dataset	Number of Pre-miRNA
gma	83
zma	97
sbi	131
ath	180
ppt	211
ptc	233
osa	397

3.2 Evaluation methods

Positive data from miRBase and negative data from PlantMiRNAPred were used to evaluate the models derived via training algorithm. We calculated that the performance of the classifier with the known sensitivity (SE) and specificity (SP) and accuracy (ACC) statistics as follows (TP refers to true positives, FP to false positives, TN to true negatives, and FN to false negatives):

$$SE = \frac{TP}{TP+FN}, SP = \frac{TN}{TN+FP} \text{ and } ACC = \frac{TP+TN}{TP+TN+FP+FN}$$

Additionally, we present the standard deviation (SD) in relevant locations.

4 Results

We present the results from a previous study in Table 2 in order to serve as a foundation for the following comparisons with our new suggested method using ensemble clustering with k-nearest neighbor (EC-kNN). We compare the performance of EC-kNN to the other methods. For more information see Xuan, Guo, Liu, et al. 2011 [35] and Khalifa et al. 2016 [17].

Table 2. Results of different tools and methods from previous published studies.

Dataset	Size	PlantMiRNAPred	Triplet-SVM	microPred	MotifmiRNAPred		
		Accuracy			ACC	SE	SP
ath	180	92.22	76.06	89.44	93.30	88.90	97.80
gma	83	98.59	74.12	86.75	89.80	84.30	95.20
osa	397	94.21	75.54	90.43	90.30	88.20	93.70
ppt	211	92.42	71.49	89.57	90.20	87.20	93.40
ptc	233	91.85	75.21	84.98	92.20	90.60	94.00
sbi	131	98.47	69.51	94.66	93.50	89.30	97.70
zma	97	98.31	66.97	93.81	94.80	94.80	94.80
Average	**180**	**95.76**	**73.64**	**90.62**	**92.19**	**89.11**	**95.48**

In order to evaluate the performance of EC-kNN we ran experiments on several datasets from previous research; Table 2 shows that the tools PlantMiRNAPred and MotifmiRNAPred are very close in terms of average performance while they are much better compared to the other listed tools.

Table 3 and Table 5 are representative examples of tables for each feature selection method over the seven miRNA plant species. See the supplementary file that contains all result tables. The OCC k-Means and SVM columns are taken from the study of Khalifa and colleagues [17]. The EC-kNN columns are new data generated using EC-kNN classifier.

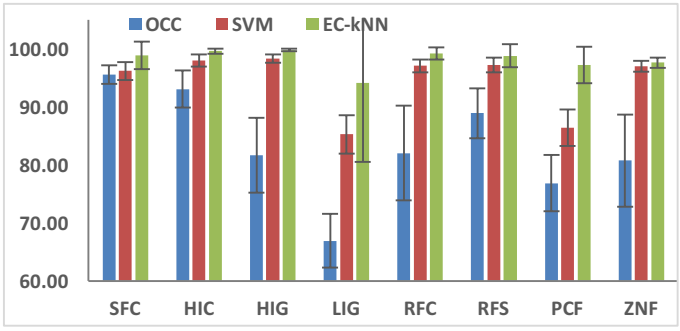
Table 3. Performance of three classification methods for the plant miRNA species considered in this study. The feature selection method is LIG as an example. k is the number of clusters in the one-class k-means algorithm.

		OCC *k*-Means			SVM			EC-*k*NN		
	k	SP	SE	ACC	SP	SE	ACC	SP	SE	ACC
LIG_ath	55	70.37	48.44	69.42	90.70	60.60	81.28	73.44	17.86	60.83
LIG_ gma	30	75.60	51.12	75.11	96.70	49.40	88.54	100.00	100.00	100.00
LIG_osa	70	57.90	69.77	58.96	83.90	82.40	83.12	95.00	100.00	97.80
LIG_ppt	45	66.27	63.29	66.12	86.90	68.20	80.42	100.00	100.00	100.00
LIG_ptc	55	69.02	59.09	68.47	91.20	78.50	86.50	100.00	100.00	100.00
LIG_sbi	35	64.41	55.15	64.11	95.00	65.60	87.68	100.00	100.00	100.00
LIG_zma	25	66.51	55.67	66.27	97.20	56.70	89.27	100.00	100.00	100.00
Average	**45**	**67.15**	**57.50**	**66.92**	**91.66**	**65.91**	**85.26**	**95.49**	**88.27**	**94.09**

Table 4 and Table 6 shows the average performance of each feature selection method over the seven plant miRNA species for OCC, SVM and EC-kNN classifiers. The right side of Table 4 and Table 6 is the visualization of the results including bars represent the standard deviation over the 100 iteration for each experiments. Table 4 shows that in most cases EC-kNN is better than SVM except the LIG_ath results. If we examine individual results, EC-kNN accuracy is better than SVM with more than 12% in most cases.

Table 4. Average performance of each feature selection method over the 7 plant miRNA species. The histogram is a visualization of the table with standard deviation bar.

FS Method	OCC	SVM	EC-*k*NN
SFC	95.59	96.19	98.84
HIC	93.07	98.04	99.63
HIG	81.63	98.33	99.81
LIG	66.92	85.26	94.09
RFC	82.03	97.07	99.23
RFS	88.90	97.25	98.82
PCF	76.84	86.41	97.27
ZNF	80.75	97.02	97.63
Average	**83.22**	**94.45**	**98.17**



In Table 4, the first column is the name of the feature selection method. OCC is for one-class classification results, SVM is for two-class SVM while the last column is for our new suggested method EC-kNN. The graph is a visualization of table 4 with standard deviation bar.

Table 5. Performance of three classification methods over the 7 plant miRNA species. The combined feature selection method is LIG as an example.

		OCC k-Means			SVM			EC-*k*NN		
	k	SP	SE	ACC	SP	SE	ACC	SP	SE	ACC
LIG_comb_ath	55	69.62	51.61	68.84	75.00	95.50	89.08	88.89	48.48	77.46
LIG_comb_ gma	25	61.29	49.25	61.05	96.70	48.20	88.33	100.00	88.14	98.30
LIG_comb_osa	65	57.46	70.05	58.58	90.40	88.20	89.29	99.39	96.91	98.02
LIG_comb_ppt	50	71.03	57.57	70.36	97.00	83.90	92.43	100.00	99.66	99.83
LIG_comb_ptc	60	78.92	61.39	77.96	91.40	80.30	87.30	99.84	98.34	99.03
LIG_comb_sbi	25	69.49	60.92	69.22	95.77	77.10	91.10	99.83	98.14	99.25
LIG_comb_zma	19	78.66	61.11	78.27	98.20	69.10	92.50	99.87	96.11	98.82
Average	**43**	**69.50**	**58.84**	**69.19**	**92.07**	**77.47**	**90.00**	**98.26**	**89.40**	**95.82**

Similarly, to Table 4, Table 5 shows that in most cases EC-kNN is better than SVM except the LIG_comb_ath results. If we examine individual results, EC-kNN accuracy is better than SVM with more than 10% in most cases.

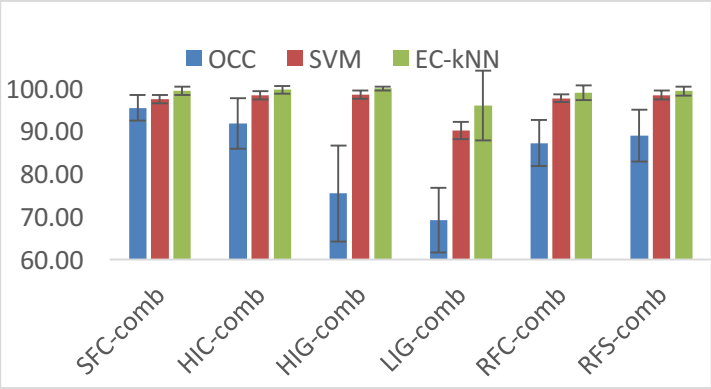
In Table 6, the first column is the name of the combined feature selection method. OCC is for one-class results, SVM is for two-class SVM, while the last column is for our new suggested method EC-kNN. The graph is a visualization of table 6 with standard deviation bar.

The results clearly show that EC-kNN is superior comparing with the other classification methods. On average, EC- KNN is better than SVM in the individual feature selection method of about 4% while with about 2% in the combined method. But clearly it is much better than the OCC method in 14% and more.

Comparing with Table 1, EC-kNN performs better than Triplet-SVM, microPred and PlantMiRNAPred see [35], also better than MotifmiRNAPred [26].

Table 6. Average performance of each combined method over the 7 plant miRNA species. The histogram is a visualization of the table with standard deviation bar.

FS Method	OCC	SVM	EC-*k*NN
SFC-comb	95.32	97.30	99.29
HIC-comb	91.67	98.26	99.51
HIG-comb	75.37	98.36	99.84
LIG-comb	69.19	90.00	95.82
RFC-comb	87.12	97.57	98.84
RFS-comb	88.80	98.26	99.24
Average	**84.58**	**96.62**	**98.76**



5 Conclusions

In this work, we have presented a new unsupervised distance metric learning based on ensemble clustering and use it within the kNN classifier. Each data point is characterized by the identity of the clusters that it belongs to in several clustering runs. The distance metric is defined as the Hamming distance between these clustering results.

This new distance has two important contributions. The first contribution is that this distance is more meaningful than the Euclidean distance between points resulting in a better kNN classifier. The second and more general contribution results from the observation that all points which always belong to the same cluster form an equivalence relation. Thus, the algorithm only has to consider one member of each equivalence class. This reduces the complexity of the algorithm considerably (by at least two orders of magnitude in our case). Our algorithm however is only a private case of a more general concept, the concept of data reduction

The performance of many learning and data mining algorithms depends critically on giving them a good metric over the input space. Learning a "good" metric from examples may therefore be the key of a successful application of these algorithms. We have demonstrated that k-nearest neighbor classification can be significantly improved by learning a distance metric from labeled examples. The ensemble of clustering is used to define the distance between points as many times the points were not clustered together. This distance is then used within the framework of the kNN algorithm to define the classifier named EC-kNN. For a large part of the experiments we performed, we succeeded to get full accuracy while the SVM failed to reach that (See Supplementary File). This concept is orthogonal to other methods of data reduction such as feature selection or PCA which reduce the size of the representation of the data points but not their number. In conclusion, this study shows that a simple classifier can reach very good results when that the distance metric is wisely and carefully chosen.

Acknowledgements

The work was supported by the Zefat Academic College to MY and by The Max Stern Yezreel Valley College to LA.

References

[1] M. Ha and V. N. Kim. Regulation of microRNA biogenesis. Nature Reviews. Molecular Cell Biology, 15(8):509–524, 2014.

[2] C. P. C. Gomes, J.-H. Cho, L. Hood, O. L. Franco, R. W. Pereira, and K. Wang. A Review of Computational Tools in microRNA Discovery. Frontiers in genetics, 4(May):81, Jan. 2013.

[3] M. Yousef, L. Showe, and M. Showe. A study of microRNAs in silico and in vivo: Bioinformatics approaches to microRNA discovery and target identification. FEBS Journal, 276(8):2150–2156. 2150–2156, 2009.

[4] H. Hamzeiy, J. Allmer, and M. Yousef. Computational methods for microRNA target precition. in miRNomics: MicroRNA Biology and Computational Analysis, Methods in Molecular Biolog, 1107:279–302, 2014, 279–302.

[5] L. B. Koski, M. W. Gray, B. F. Lang, and G. Burger. AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics, 6:151, 2005.

[6] M. Yousef, S. Jung, L. C. Showe, and M. K. Showe. Learning from positive examples when the negative class is undetermined--microRNA gene identification. Algorithms for molecular biology : AMB, 3:2, Jan. 2008.

[7] M. Yousef, J. Allmer, and W. Khalifa. Sequence Motif-Based One-Class Classifiers Can Achieve Comparable Accuracy to Two-Class Learners for Plant microRNA Detection. Journal of Biomedical Science and Engineering, 8(10):684–694, 2015.

[8] M. Yousef, N. Najami, and W. Khalifa. A Comparison Study Between One-Class and Two-Class Machine Learning for MicroRNA Target Detection. Journal of Biomedical Science and Engineering, 2010.

[9] H. T. Dang, H. P. Tho, K. Satou, and B. H. Tu. Prediction of microRNA hairpins using one-class support vector machines. in *2nd International Conference on Bioinformatics and Biomedical Engineering, iCBBE 2008*, 2008, :33–36.

[10] C. Wang, C. Ding, R. F. Meraz, and S. R. Holbrook. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. Bioinformatics, :btl441, 2006.

[11] S. Paul, M. Magdon-Ismail, and P. Drineas. Feature selection for linear SVM with provable guarantees. Journal of Machine Learning Research, 38:735–743, 2015.

[12] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning. Machine Learning, 46(1–3):389–422, 2002.

[13] M. E. Ahsen, N. K. Singh, T. Boren, M. Vidyasagar, and M. A. White. A new feature selection algorithm for two-class classification problems and application to endometrial cancer. in *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, 2012, :2976–2982.

[14] L. H. N. Lorena, A. C. P. L. F. Carvalho, and A. C. Lorena. Filter Feature Selection for One-Class Classification. Journal of Intelligent & Robotic Systems, :1–17, Sep. 2014.

[15] P. Xuan, M. Guo, Y. Huang, W. Li, and Y. Huang. MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs. PloS one, 6(11):e27422, Jan. 2011.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software. ACM SIGKDD Explorations Newsletter, 11(1):10, Nov. 2009.

[17] W. Khalifa, M. Yousef, M. D. Sacar Demirci, and J. Allmer. The impact of feature selection on one and two-class classification performance for plant microRNAs. PeerJ, 4:e2135, 2016.

[18] T. C. and P. Hart. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1):21–27, 1967.

[19] L. AbedAllah and I. Shimshoni. k Nearest Neighbor Using Ensemble Clustering. in Data Warehousing and Knowledge Discovery: 14th International Conference, DaWaK 2012, Vienna, Austria, September 3-6, 2012. Proceedings, :265–278, A. Cuzzocrea and U. Dayal, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, 265–278.

[20] V. Thakur, S. Wanchana, M. Xu, R. Bruskiewich, W. P. Quick, A. Mosig, and X.-G. Zhu. Characterization of statistical features for plant microRNA prediction. BMC genomics, 12(1):108, 2011.

[21] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME SUITE: tools for motif discovery and searching. Nucleic acids research, 37(Web Server issue):W202-8, Jul. 2009.

[22] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology, 2:28–36, Jan. 1994.

[23] M. V. Cakir and J. Allmer. Systematic computational analysis of potential RNAi regulation in Toxoplasma gondii. in *Health Informatics and Bioinformatics (HIBIT), 2010 5th International Symposium on*, 2010, :31–38.

[24] T. Yan, D. Yoo, T. Z. Berardini, L. A. Mueller, D. C. Weems, S. Weng, J. M. Cherry, and S. Y. Rhee. PatMatch: a program for finding patterns in peptide and nucleotide sequences. Nucleic acids research, 33(Web Server issue):W262-6, Jul. 2005.

[25] M. Yousef, J. Allmer, and W. Khalifaa. Plant MicroRNA Prediction employing Sequence Motifs Achieves High Accuracy. 2015.

[26] M. Yousef, J. Allmer, and W. Khalifa. Accurate Plant MicroRNA Prediction Can Be Achieved Using Sequence Motif Features. Journal of Intelligent Learning Systems and Applications, 8(1):9–22, 2016.

[27] D. M. J. Tax. DDtools, the Data Description Toolbox for Matlab. 2015.

[28] Q.-S. Xu and Y.-Z. Liang. Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems, 56(1):1–11, Apr. 2001.

[29] M. Yousef, M. D. Saçar Demirci, W. Khalifa, and J. Allmer. Feature Selection Has a Large Impact on One-Class Classification Accuracy for MicroRNAs in Plants. Advances in Bioinformatics, 2016:1–6, 2016.

[30] V. N. Vapnik. The nature of statistical learning theory. New York, New York, USA: Springer-Verlag, 1995.

[31] J.-W. Nam, K.-R. Shin, J. Han, Y. Lee, V. N. Kim, and B.-T. Zhang. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res, 33(11):3570–3581, 2005.

[32] J. Ding, S. Zhou, and J. Guan. MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. BMC bioinformatics, 11 Suppl 1(Suppl 11):S11, Jan. 2010.

[33] K. L. S. Ng and S. K. Mishra. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics, 23(11):1321–30, Jun. 2007.

[34] C. Xue, F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics, 6:310, 2005.

[35] P. Xuan, M. Guo, X. Liu, Y. Huang, W. Li, and Y. Huang. PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. Bioinformatics (Oxford, England), 27(10):1368–76, May 2011.

[36] J. E. Gewehr, M. Szugat, and R. Zimmer. BioWeka--extending the Weka framework for bioinformatics. Bioinformatics (Oxford, England), 23(5):651–3, Mar. 2007.

[37] C.-C. Chang and C.-J. Lin. LIBSVM. ACM Transactions on Intelligent Systems and Technology, 2(3):1–27, Apr. 2011.

[38] M. D. Sacar and J. Allmer. Data mining for microrna gene prediction: On the impact of class imbalance and feature number for microrna gene prediction. in *2013 8th International Symposium on Health Informatics and Bioinformatics*, 2013, :1–6.

[39] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. KNIME: The Konstanz Information Miner. in SIGKDD Explorations, 11(1):319–326, 2008, 319–326.

[40] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. 36(suppl_1):D154-158. D154-158, 2008.

[41] M. D. Saçar, C. Bağcı, and J. Allmer. Computational Prediction of MicroRNAs from Toxoplasma gondii Potentially Regulating the Hosts' Gene Expression. Genomics, proteomics & bioinformatics, 12(5):228–238, Oct. 2014.