Adopting a new sample strategy to predict miRNA-disease associations

Xing Zhong, Tian Lei, RongHui Rong*
College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
*wrh@hnu.edu.com

Abstract. Exploring unknown miRNA-disease associations by computational tools is a new way to study the correlations between genes and diseases. In this paper, we proposed a new model named ABPUSVM, which consisted of a new sample strategy based on Positive-Unlabeled learning and a prediction model that combined AdaBoost and SVM. When ABPUSVM was applied to predict unknown associations of miRNA-disease, the AUC of 0.9383 improved greatly based on 5 folds cross-validation and showed its excellent performance, which indicated ABPUSVM was significantly better than other classic models. Afterward, a case study confirmed that 46 out of the top predicted 50 miRNAs of breast cancer by ABPUSVM were supported by three databases. Whose results showed the dataset by ABPUSVM were significantly better than that of the other methods. All results have shown that ABPUSVM is a promising and potential tool for exploring the associations of miRNA-disease.

Keywords: miRNA-disease interactions, bioinformatics, Sample Strategy, prediction model.

1. Introduction
A computational method for exploring miRNA-disease associations has gradually gained attention due to its high efficiency and speed and is mainly categorized into two types[1], which includes similarity calculation and machine learning[2][3][4]. The core component of machine learning models to predict miRNA-disease association includes sample strategy and prediction model[5]. However, the positive sample data can be obtained completely by experiments but many bioinformatics-related datasets have no way to obtain the true negative sample. As for a negative sample of miRNA-disease it means that the association of a miRNA and a disease is unclear by now[6]. Therefore, it’s particularly important to choose sample strategy to combine positive samples and negative samples when training machine learning models to predict unknown associations of miRNA-disease[7], the results may change greatly by different strategies to combine positive samples and negative samples[8].

To address the above problems, in this paper, we proposed a new model named ABPUSVM to improve performance from two aspects. First, ABPUSVM adopted a new strategy to combine positive samples and negative samples based on the Positive-Unlabeled (PU) Learning[9]; Second, ABPUSVM combined AdaBoost[10] and SVM[11] into an integrated model to predict the associations of potential miRNA-disease. Performance evaluation, case study, and comparisons with other method show that ABPUSVM is significantly better than other methods and has huge potential to explore miRNA-disease associations.
2. MATERIAL AND METHOD

2.1. Materials
5430 known miRNA-disease associations matrix $A$ between 495 miRNAs and 383 diseases were collected from Chen[7]. Matrix $A$ consists of 495 rows in which the $i-th$ row represents miRNA$_i$ ($i\in\{1,2,...,495\}$) and $j-th$ column represents disease$_j$($j\in\{1,2,...,383\}$). $A_{ij}$ signifies the association between miRNA$_i$ and disease$_j$ and its value is 1 if there exists validated association otherwise 0; Meanwhile the integrated similarity matrix $SM$ between 495 miRNAs and the integrated similarity matrix $SD$ between 383 diseases was obtained( please refer to Chen for more details of $SM$ and $SD$).

2.2. Sample Strategy
Based on a concept that among all unknown associations of miRNA-disease sample the probability that a miRNA or a disease sample are correlated with known associations of other diseases or miRNAs is more likely to a negative sample[12]. According to matrix $A$ between 495 miRNAs and 383 human diseases, in which 1 of $A_{ij}$ signifies the association between miRNA$_i$ and disease$_j$. The forms of $i-th$ row in $A$ are [disease$_2$, disease$_2$, disease$_3$,...,disease$_{383}$] and the count of 1 represents the known associations of miRNA$_i$ with 383 diseases. The number of known associations confirmed by the literature or
experiments was defined as research hot. The 495 miRNAs were ranked according to the research hot and the larger research hot represented the greater priority to be a negative sample. Similarly, 383 diseases were conducted by the same means.

Then we combined the integrated similarity matrix $S_M$ and $S_D$ into a feature matrix $S$, which contained 495*383 feature vectors and then were divided into matrix $U$ according to the associations between the corresponding miRNA and disease was known. Then other feature vector of $U$ were collected into $K$; Referred to Rayhan[13], the matrix $U$ was clustered into 23 sub-categories ($U_1, U_2... U_{23}$) by k-mean[14]; And then a total of 5430 feature vectors were extracted from according to the proportion of $U_i$ to $U$ and miRNA or disease of each feature vector must rank top 120 of research hot. Next 5430 feature vectors and $K$ were combined to training dataset (dataset5) for ABUPSVM. And then the trained model was obtained and the unknown associations of miRNA-disease were predicted and ranked.

2.3. ABPUSVM model

We proposed an improved machine learning model which consisted of an improved strategy to sample and an improved model to explore potential associations of miRNA-disease. The improved strategy to sample was described in the above section. As for the improved model which was inspired by Li[15], we combined the AdaBoost and SVM to explore potential association for a miRNA-disease pair[16], in which SVM and AdaBoost were integrated into a classifier to boost performance. So, our model was called ABPUSVM. Fig. 1 shows the pipeline of ABPUSVM.

**Algorithm of ABPUSVM**

Input: (1) miRNA-disease interactions matrix $A$ for 495 miRNAs and 383 diseases
(2) Integrated similarity matrix $S_M$ of 495 miRNAs.
(3) Integrated similarity matrix $S_D$ of 383 diseases.
Output: potential miRNA-disease s ranking score.

Step 1: let $S$ as feature matrix with the row of (495*383) and the column of (495+383)
for $i$ ($i = 1, 2...495$) do
  for $j$ ($j = 1, 2...383$) do
    for $m$ ($m = 1, 2...495$) do
      $S_{(i,j)} \leftarrow S_M_{(i,m)}$
      end for
    for $n$ ($n = 1, 2...383$) do
      $S_{(i,j,495+n)} \leftarrow S_M_{(i,n)}$
      end for
  end for
for each $sample_i$ in $S$
  if the value of $miRNA_i - disease_j$ for $sample_i$ is 0 in $A$ then
    add $sample_i$ to $U$
  if the value of $miRNA_i - disease_j$ for $sample_i$ is 1 in $A$ then
    add $sample_i$ to $K$

Step 2: Clustering $U$ into 23 clusters ($U_1, U_2... U_{23}$) by k-means
Step 3: Based on miRNA-disease interactions $A$ to make miRNA hot matrix $HM$ (1*495) and disease hot matrix $HD$ (1*383)
  for $row_i$ in $A$
    $HM_i \leftarrow \text{sum} (A_{(i,j)})$

end for
for \textit{column}_i in A
\begin{align*}
\text{HD}_i &\leftarrow \text{sum} (A(i,j)) \\
\end{align*}
end for

Rank \textit{HM} and \textit{HD}

Step 4: \texttt{dataset5$\leftarrow$U} (separately select the same proportion of \textit{U}_i to \textit{U} feature vector from \textit{U}_i and miRNA or disease of feature vector rank in the top 120 in 23 clusters) $\cup$ \textit{K}

\texttt{dataset4$\leftarrow$U} (separately select the same proportion which depends on the number of each cluster account for \textit{U} in 23 clusters) $\cup$ \textit{K}

\texttt{dataset3$\leftarrow$U} (select the same number with \textit{K} by turn) $\cup$ \textit{K}

\texttt{dataset2$\leftarrow$U} (select the same number with \textit{K} by random) $\cup$ \textit{K}

\texttt{dataset1$\leftarrow$U} (separately select the 240 samples in 23 clusters by random) $\cup$ \textit{K}

Step 5: Applying \texttt{dataset5} to train ABUPSVM to predict the scores of the potential of miRNA-disease and potential miRNA-disease rank by predictive score.

\section*{3. Result}

\subsection*{3.1. Validation of improved sample strategy}

In step 4, \textit{U} and \textit{K} were respectively collected into \texttt{dataset1}, \texttt{dataset2}, \texttt{dataset3}, \texttt{dataset4} and \texttt{dataset5} according to different strategies. For evaluation of the sample strategy of ABUPSVM, five datasets were applied to the PBMDA\cite{17} to train predicted classifiers, each predictive result was evaluated by 5-folds cross-validation. ROC and P-R curve\cite{18} was drawn to compare performance intuitively. As shown in Fig. 2, the \texttt{dataset5} by ABPUSVM could improve the overall performance effectively, which directly proved the effectiveness of the sample strategy of ABUPSVM.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{ROC and P-R curve of \texttt{dataset1}, \texttt{dataset2}, \texttt{dataset3}, \texttt{dataset4} and \texttt{dataset5} by PBMDA}
\end{figure}

\subsection*{3.2. Performance evaluations of ABPUSVM}

In step 3, based on the known miRNA-disease associations from the HMDD V2.0\cite{19}, 5-fold cross-validation was executed to ABPUSVM to evaluate the performance. And the predictive performance was evaluated by AUC, AUPR, precision, recall and F1-Score\cite{20}. The AUC, AUPR, precision, recall and F1-Score of ABPUSVM were 0.9383, 0.9478, 0.8313, 0.8803 and 0.8551. Then we compared...
ABPUSVM with EKRRMDA [21] and KBMFMDA [22] for those two methods used the same datasets and also was verified by 5-fold cross-validation. The overall performance of ABPUSVM improved greatly compared with EKRRMDA and KBMFMDA. The AUC of ABPUSVM was 0.9383 which was higher than that of EKRRMDA (0.9282) and KBMFMDA (0.9052). In terms of AUC, the results of EKRRMDA were better than, HDMP (0.8342 ± 0.0010), MaxFlow (0.8579 ± 0.001), NCPMDA (0.8763 ± 0.0008), PBMDA (0.9127 ± 0.0007), LRSSLMDA (0.9181 ± 0.0004), WBSMDA (0.8185 ± 0.0009), MCMMDA (0.8767 ± 0.0011), RLSMDA (0.8569 ± 0.0020), BNPMDA (0.8980 ± 0.0013), MDHGI (0.8794 ± 0.0021) and IMCMDA (0.8367 ± 0.0005) [21]. So, those facts confirmed that ABPUSVM was better than those methods.

3.3. Case study

To further validate the prediction performance, the top 50 miRNAs of breast cancer were extracted from the prediction results of ABPUSVM, and then the prediction results were compared and analyzed with the DBDEMC [23], miRCancer [24] and HMDD3.0 [25]. In Table 3, 46 miRNAs of the top 50 predicted miRNAs of breast cancer were supported by DBDEMC, miRCancer and HMDD3.0; and 9 out of the top 10 predicted miRNAs of breast cancer were supported by those three databases. Especially has-mir-503 and has-mir-498 were included in the top 10 predicted miRNAs, Long [26] et al has verified that has-mir-503 was significantly down-regulated in breast cancer tissues and cells and its overexpression would reduce cell proliferation [27]. Chai [28] showed that has-mir-498 in breast cancer cells promoted cell proliferation and migration by targeting the tumor suppressor PTEN, and inhibiting the overexpression of has-mir-498 could weaken its effect on cell proliferation and migration [29].
4. Conclusion

In recent years research involving cancer-related fields has increasingly and has become a research focus. With the deepening of cancer research, a type of single-stranded endogenous short non-coding RNA (miRNA) has entered the research field, and more increasing research evidence shows that miRNAs are closely associated with many human diseases[30][31]. In this paper, a new model called ABPUSVM was proposed, which improved the predicting performance of miRNA-disease associations from two aspects. First, ABPUSVM adopted a new sample strategy based on research hot ranking to collect positive samples and negative samples. Second, BPUSVM combined AdaBoost and SVM into an integrated model to predict unknown association scores of miRNA-diseases. Datasets by different sample strategies were applied to PBMDA and the result showed dataset by ABPUSVM’s sample strategy boosted performance greatly. Performance evaluation showed the AUC value by ABPUSVM improved greatly and outperformed than many other methods, which meant prediction accuracy improved greatly. Then the case study confirmed that three databases supported 46 miRNAs in the top 50 predicted miRNA-disease associations. Conclusively ABPUSVM paved the way for future research to obtain better performance in exploring miRNA-disease associations.

References

[1] Chen X, Xie D, Zhao Q and You ZH 2019 M. Brief Bioinform. 20 Oxford University Press 515–39
[2] YS LI, XY Jin, ZS Wang, LL Li, H Chen, XY Lin, S Yi, YP Zhang and J Xu 2019 J. Brief Bioinform. 20 Oxford University Press 1193–204
[3] Z Huang, LB Liu, YX Gao, JC S, QH Cui, JW Li, Y Zhou 2019. Genome Biol,20 B BioMed Central Ltd 1–13
[4] X Chen , CC Yan, X Zhang, ZH You 2017 J,Brief Bioinform.18. Oxford University Press 558–76
[5] Fu L and Peng Q 2017 M. Sci Rep.7 Nature Publishing Group 1-13
[6] Mueller M, Juan Antonio Vizcaíno PJ and Côté R 2008. Proteomics. 7 Wiley-VCH Verlag 1138–48
[7] Shi J-Y, Li J-X, Mao K-T, Cao J-B, Lei P and Lu H-M 2019. Comput Methods Programs Biomed. 168 Elsevier Ireland Ltd 1–10.
[8] Carroll NM, Jaeger EEM, Choudhury S, Dunlop AAS, Matheson MM and Adamson P 2020. J Clin Microbiol. 38 American Society for Microbiology 1753–7
[9] Yang P, Li X, Chua HN, Kwoh CK and Ng SK 2014. PLoS One .9 Public Library of Science 1-10
[10] Zhong L, Wang JTL, Wen D, Aris V, Soteropoulos P and Shapiro BA 2013. Omi A J Integr Biol. 17 Mary Ann Liebert Inc 486–93
[11] Liu H, Yue D, Chen Y and Huang Y 2009. IEEE Int Work Genomic Signal Process Stat GENSIPS, IEEE 2009 1-4
[12] Manter DK, Weir TL and Vivanco JM 2010. Appl Environ Microbiol. 76. 2010 American Society for Microbiology 2086–90
[13] Rayhan F, Ahmed S, Shatabda S, Farid DM, Mousavian Z and Dehzangi A 2017. Sci Rep.7 Nature Publishing Group 1–18
[14] McCall MR, Mehta T, Leathers CW and Foster DM 1992. Am J Clin Nutr. 56 american journal of clinical nutrition 385–93
[15] Li X, Wang L and Sung E 2008. Eng Appl Artif Intell. 21 Elsevier Ltd 785–95
[16] Li X, Wang L and Sung E 2015. Proc Int Jt Conf Neural Networks. 1:196–201.
[17] You ZH, Huang ZA, Zhu Z, Yan GY, Li ZW and Wen Z2017. PLoS Comput Biol. 13 Public Library of Science 1-10
[18] Li J, Zhang S, Liu T, Ning C, Zhang Z and Zhou W 2020. Bioinformatics. 36 Oxford University Press 2538–46
[19] Li Y, Qiu C, Tu J, Geng B, Yang J and Jiang T 2014. Nucleic Acids Res. 42 Oxford University Press 1070–4
[20] Yang Y, Fu X, Qu W, Xiao Y and Shen H Bin 2018. Bioinformatics. 34 Oxford University Press 3547–56
[21] Peng LH, Zhou LQ, Chen X and Piao X 2020. Front Bioeng Biotechnol. 8 Frontiers Media S.A 1–15.
[22] Chen X, Li SX, Yin J and Wang CC 2020. Genomics. 112 Academic Press Inc
[23] Yang Z, Ren F, Liu C, He S, Sun G and Gao Q 2010. BMC Genomics. 11 SUPPL. 4 BioMed Central Ltd S5
[24] Xie B, Ding Q, Han H and Wu D 2013. Bioinformatics. 5 Oxford University Press 638–644.
[25] Huang Z, Shi J, Gao Y, Cui C, Zhang S and Li J 2019. Nucleic Acids Res. 47 Oxford University Press D1013–7.
[26] Long J, Ou C, Xia H, Zhu Y and Liu D 2015. Tumor Biol. 36 SAGE Publications Inc 8697–702
[27] Jiang L, Zhao Z, Zheng L, Xue L, Zhan Q and Song Y 2017. Proteomics & Bioinformatics. 15 Beijing Genomics Institute 208–17
[28] CS Chai, H Wu, BF Wang, DD Eisenstat and RP Leng 2018. Carcinogenesis. 39 Oxford University Press 1185–96
[29] Epis MR, Giles KM, Candy PA, Webster RJ and Leedman PJ 2014. Journal of Neuro-Oncology. 116 Kluwer Academic Publishers 67–75
[30] Esquela-Kerscher A and Slack FJ 2006. Nat Rev Cancer. 6 Nature Publishing Group 259–69
[31] Ambros V 2004. Nature. 431 Nature Publishing Group 350–5