Evaluation of CML Therapeutic Efficacy Based on Compressed Sensing Classification Algorithm

Xinyan Pu, Kangning Yang, Tingxuan Cheng, Jingxue Wang, Yanbin Zhou, Jie Gao *

School of Science, Jiangnan University, Wuxi, Jiangsu, China
* Correspondence: gaojie@jiangnan.edu.cn

Abstract: Imatinib is an effective therapeutic agent for Chronic myeloid leukemia (CML). However, imatinib resistance makes the treatment of patients with CML complex and diverse. In order to judge the effect of treatment and the prognosis, a method based on compressed sensing classification algorithm is proposed. Data set GSE33075 from GEO DataSets was analyzed by using the classification algorithm based on compressed sensing. Every time training samples were randomly selected to construct a redundant dictionary. The projection of test samples on the redundant dictionary was computed with Orthogonal Matching Pursuit (OMP) algorithm. The projection errors were used to determine the category of test samples. The average correct rate of repeated experiments was over 90%. Based on this classification algorithm, the probability of classification to normal people was estimated as the therapeutic effect of imatinib on CML patients. The results demonstrate that the proposed method can judge the efficacy of imatinib.

Keywords: compressed sensing; chronic myeloid leukemia (CML); orthogonal matching pursuit; imatinib

1. Introduction

Chronic myeloid leukemia (CML) is a common malignant disease caused by Philadelphia chromosome, which is the result of a reciprocal chromosome translocation between chromosomes 9 and 22. BCR-ABL fusion protein [1] with sustained activation of tyrosine kinase (TK), which is encoded by Philadelphia chromosome, leading to abnormal proliferation of hematopoietic stem cells. According to the molecular mechanism of CML, tyrosine kinase inhibitors (TKIs) are mainly used for molecular targeted therapy [2]. Imatinib, the first-line TKI that greatly improves the prognosis of CML [3]. However, only a few patients completely alleviate symptoms after treatment [4]. Some patients discontinue medication due to intolerance [5]. And some patients developed imatinib resistance [1].

Today, the actual efficacy of drugs is mainly based on clinical symptoms, but there are few studies to judge the efficacy of drugs based on gene expression data. Gene expression data generated by DNA microarray experiments can diagnose diseases at the gene level and select appropriate drugs for treatment. However, due to the large amount of redundancy and noise in gene expression data [6], how to extract effective information from massive data mining had become a hot issue. In 2006, D.L. Donoho [7], E. Candes et al. [8] proposed compressed sensing based on related research. In recent years, it has been used in classification research. W.L. Tang et al. [9] proposed a novel compressive sensing (CS) based approach for the subtyping of leukemia. W. Shao et al. [10] applied compressed sensing to radar classification system, in which observation matrix was obtained by Fourier transform. C. R. Jane and X. Y. Chen proposed a classification method based on sparse representation and least squares regression [11]. Y.L. Liang [12] presented a classification method based on dictionary learning algorithms, which implemented medical data classification.

Based on the theory of compressed sensing, we propose a classification algorithm for CML patients and normal persons. The untreated CML patients and normal persons were randomly selected as training samples to construct redundant dictionaries. By solving the sparse representation problem of test samples on the redundant dictionary, the category of test samples were determined. In addition, after repeated experiments, the probabilities of samples belong to normal persons can be calculated. The reliability of this method can be tested by comparing the calculated probabilities with the actual situation. According to the probabilities of patients after treatment, we estimate the efficacy of imatinib, and expect to provide some clinical references.

2. The Theory of Compressed Sensing

In the traditional sampling process, a large number of data will be collected to satisfy Nyquist Sampling theorem. The theory of compressed sensing provides a new acquisition method for sparse signal, which combines data compression and acquisition. The theory mainly includes three aspects: sparse representation of signals, design of observation matrix and signal reconstruction algorithm.

Suppose we have a signal *f* of size *N*, which could be represented by some orthogonal transform bases Ψ:

$$\boldsymbol{f} = \Psi \boldsymbol{x} \tag{1}$$

Table 1. Gene expression data from GSE33075.

Number	ID	Gene Symbol				
		'1-Mar'	'1-Mar'	'3-Mar'	'4-Mar'	……
1	GSM817258	4.2181	4.7700	4.4665	4.0422	……
2	GSM818670	4.1374	4.7400	4.9034	3.5762	
3	GSM818671	4.2628	4.4700	5.1857	3.6125	
……	……	……				
26	GSM818824	4.3114	4.7903	4.9826	3.7578	
27	GSM818825	4.9632	4.8037	4.8512	3.4171	

Where x is an N-dimensional vector. When x has k<<N non-zero elements and the remaining elements are zero, f can be sparsely represented. According to the theory of compressed sensing, it is possible to recover a signal that can be sparsely represented. Even if some elements of x are discarded, the reconstruction of the original signal will not be affected.

We assume that f can be sparsely represented. We need to design an observation matrix $\Phi \in R^{M\times N}$, which is not related to Ψ [13]. Projecting the N-dimensional signa f to the observation matrix, then an M-dimensional observation vector y can be obtained by:

$$\boldsymbol{y} = \Phi \boldsymbol{f} \tag{2}$$

From (1) and (2) we get:

$$\boldsymbol{y} = \Phi \boldsymbol{f}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} \tag{3}$$

Where $\boldsymbol{A} = \Phi\Psi$ is a M×N matrix, and A is called a sensor matrix.

How to recover f from observation vector y is the core of compressed sensing. However, the problem can not be solved directly from (2) due to M<<N. But it can be solved by solving the minimum L-0 norm problem [14], which has been proved theoretically, as following:

$$\hat{\boldsymbol{x}} = \arg\min ||\boldsymbol{x}||_0 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \tag{4}$$

This problem is NP-hard, but the suboptimal solution [15] can be obtained by Orthogonal Matching Pursuit (OMP) algorithm. OMP algorithm is a greedy algorithm, which mainly approximates the observed signal through iteration. Choosing the best matching vector from the sensor matrix A for sparse approximation and orthogonalizing vectors of A in each iteration. According to the calculated residual of the observation vector y, selecting the best matching vector to iterate until the residual converges. At this time, the original signal f can be reconstructed accurately by, and the compressed sensing signal reconstruction is realized.

3. Evaluation of CML Therapeutic Efficacy Based on Compressed Perception Classification

3.1. Data

Gene expression data are obtained by microarray hybridization after isolating total RNA from bone marrow. The gene expression data we used was the date set GSE33075 which was downloaded from the online GEO DataSets. There are 27 samples of gene expression profile chip data, nine of which from normal donors and the rest are from nine PH-positive CML patients. In all patients bone marrow was sampled before administration commenced and 4 weeks after receiving a daily dose of 400 mg imatinib mesylate. The dataset we downloaded is shown in Table 1.

Gene expression profile data has the characteristics of few samples, high noise, large dimension and redundancy. By selecting feature genes, we can eliminate redundant or noisy data as many as possible to prevent "dimension disaster" and improve the classification performance of classification models. Feature selection includes feature selection and feature extraction. This paper used the filtering method in feature selection to select features. The Signal to Noise Ratio, Bhattacharyya distance and Fisher discrimination criterion were selected as metrics, and the given threshold was used to screen feature genes.

3.2. Classification Algorithm Based on Compressed Sensing Theory

There are 27 samples in GSE33075. Therefore, we divided them into three categories A, B and C. They were CML patients before administration commenced, CML patients four weeks after the treatment and normal people respectively. After feature selection, sample dimension was reduced to M dimension. a and c samples were selected from Category A and C respectively as training data sets to construct M×N dimension redundant dictionary matrix A:

$$\boldsymbol{A} = [\boldsymbol{y}_{1,1} \quad \boldsymbol{y}_{1,2} \quad \cdots \quad \boldsymbol{y}_{1,a} \quad \boldsymbol{y}_{2,1} \quad \cdots \quad \boldsymbol{y}_{2,c}]$$

where $y_{1,j}$ represented the j^{th} sample of Category A in the training sample, and $y_{2,j}$ represented the j^{th} sample of Category C in the training sample, $y_{i,j} \in R^{M\times 1}$.

Taking all samples of the dataset as the test data set, any test sample can be represented as:

$$\boldsymbol{y} = \boldsymbol{y}_{1,1}x_{1,1} + \boldsymbol{y}_{1,2}x_{1,2} + \cdots + \boldsymbol{y}_{1,a}x_{1,a} + \boldsymbol{y}_{2,1}x_{2,1} + \cdots + \boldsymbol{y}_{2,c}x_{2,c} = \boldsymbol{A}\boldsymbol{x}^T$$

where $\boldsymbol{x} = [x_{1,1} \quad \cdots \quad x_{1,a} \quad x_{2,1} \quad \cdots \quad x_{2,c}] \in R^{1\times N}$.

If the test sample belongs to Category A, then $\boldsymbol{x} = [x_{1,1} \quad \cdots \quad x_{1,a} \quad 0 \quad \cdots \quad 0] \in R^{1\times N}$ theoretically. The total number of samples was larger than the number of samples in each category, thus x was sparse.

Based on the theory of sparse representation and signal reconstruction, if x is sparse, solving equation $y=Ax^T$ can be transformed into solving the minimum L-0 norm problem as following:

$$\hat{\boldsymbol{x}} = \arg\min ||\boldsymbol{x}||_0 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^T .$$

The OMP algorithm was used to solve the minimum L-0 norm problem. The error threshold was set to be 1e-3 and the maximum iteration number was 1e3. Due to the influence of noise, there was usually a certain error in the actual solution. For example, if the test sample belongs to Category A, $\boldsymbol{x}_{2,j}, j \in \{1,2,\cdots,c\}$ may be non-zero which should be zero theoretically. Therefore, in order to widen the gap between classes and improve the classification accuracy, we set:

$$\boldsymbol{A}_1 = [\boldsymbol{y}_{1,1} \quad \boldsymbol{y}_{1,2} \quad \cdots \quad \boldsymbol{y}_{1,a}]$$

$$\boldsymbol{x}_1 = [x_{1,1} \quad x_{1,2} \quad \cdots \quad x_{1,a}]$$

$$\boldsymbol{A}_2 = [\boldsymbol{y}_{2,1} \quad \boldsymbol{y}_{2,2} \quad \cdots \quad \boldsymbol{y}_{2,c}]$$

$$\boldsymbol{x}_2 = [x_{2,1} \quad x_{2,2} \quad \cdots \quad x_{2,c}] .$$

When calculating residuals, only the residuals of the classes belong to are calculated. For each test sample, the residuals to be calculated were as following:

$$r_i(x) = \| \boldsymbol{y} - \boldsymbol{A}_i \boldsymbol{x}_i \| \qquad i = 1,2$$

The category corresponding to the smaller residual was the category of test samples identified by the classifier.

For different measurement criteria and thresholds, five hundred experiments were repeated each time. Training samples were randomly selected in each experiment. The average recognition accuracy of samples in Category A and C was calculated as the performance evaluation index of the classifier.

Choosing thresholds which have a high accuracy for each measurement criteria. Then 500 experiments were repeated with each chosen threshold. Calculating the probabilities of test samples being classified into Category C. The effect of taking medicine can be estimated by the results.

4. Results and Discussion

Each sample in GSE33075 contains 23518 genes. Using Signal to Noise Ratio (SNR), Bhattacharyya distance and Fisher discrimination criterion as measurement criteria. Thresholds were selected by the distribution results of genes under each measurement criterion. Take the case of Bhattacharyya distance, the results are shown in Figure 1.

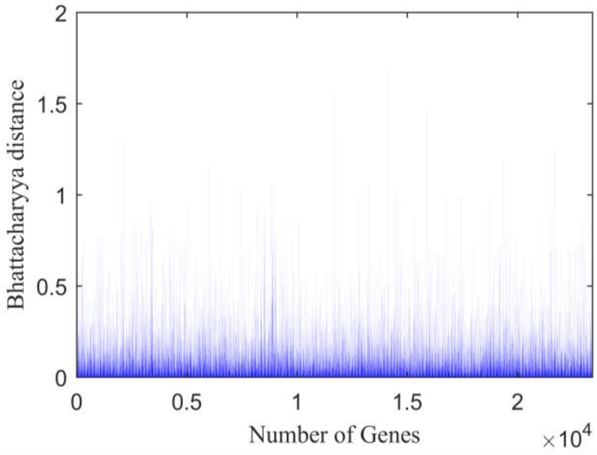

Figure 1. The Bhattacharyya distance of each gene in GSE 33075 is calculated based on the gene expression data of Category A and C samples.

Figure 1 shows that most genes have a Bhattacharyya distance of less than 0.4. Only a few genes have a value of more than 1.2. Therefore, for Bhattacharyya distances, 0.6, 0.9 and 1.1 were selected as thresholds to select characteristic genes. The samples were classified according to the selected characteristic genes. Similarly, 1.0, 1.4 and 1.6 were selected as thresholds for SNR. 3, 3.8 and 4.2 were selected for Fisher discrimination criterion. The final classification results are summarized in Table 2.

Table 2. Experimental results of classification.

Measurement Criteria	Threshold Value	Number of Feature Genes	Accuary (%)
Signal to noise Ratio	1.0	202	89.1667
	1.4	14	87.3778
	1.6	7	75.6222
Bhattacharyya Distance	0.7	121	90.4778
	0.9	38	92.2000
	1.1	10	84.1667
Fisher	2.8	51	93.1667
	3.4	24	92.4889
	4.2	7	76.8556

For the same metrics, the final classification accuracy varied wildly between different thresholds. It can be seen that the selection of appropriate feature genes plays an important role in the classification accuracy. If we select appropriate threshold for different metrics, classification accuracy can basically reach 90%. This means that the classification algorithm based on compressed sensing may be helpful for the diagnosis of CML.The classification accuracy of SNR was lower than the other two metrics, which may be due to the inadequate consideration of size of variance.

In order to accurately judge the efficacy of imatinib on patients in Category B, the threshold with the highest classification accuracy was selected in the next experiment. After 500 experiments, calculating the probabilities of test samples being classified into Category C by the classification model. The results are shown in Figure 2.

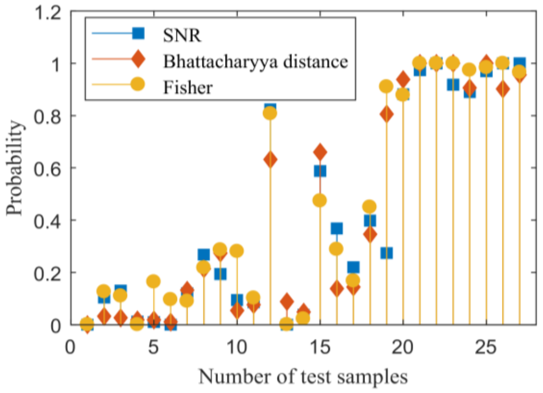

Figure 2. The probabilities of samples being categorized to a normal person in 500 experiments after selecting characteristic genes by three metrics.

Samples numbered 1-9 are from CML patients before administration, and the probabilities of dividing them into Category C are low under three metrics. The probabilities are less than 0.2 except for patients numbered 9. Based on the experimental results, it is supposed that samples numbered 1-9 may be CML patients. The probabilities of dividing samples numbered 19-27 into Category C is close to 1, so it is supposed that samples numbered 19-27 are from normal persons. In fact, these samples are from normal persons. The above experimental results are consistent with the actual situation. It proves that it is also feasible to judge the therapeutic effect of imatinib according to the probability.

Samples numbered 10-18 are CML patients four weeks after administration. From Figure 2, it can be observed that the probabilities of samples numbered 10, 12, 15 and 18 are significantly higher than that before administration. But the rest of patients have a slight increase or even decrease. Therefore, it is speculated that patients numbered 10, 12, 15 and 18 have an obvious improvement after taking the medicine and the rest of patients might be intolerant of imatinib, the treatment effect is not ideal.

5. Conclusion

In this paper, data set GSE33075 was analyzed based on compressed sensing theory classification algorithm. We proposed a method to judge the therapeutic effect of imatinib on CML patients. We compared the accuracy of classification under different metrics, and selected thresholds with high classification accuracy to improve the reliability of experimental results. The experimental results of CML patients before administration commenced and normal people basically coincided with the actual situation, so it is speculated that the experimental results of CML patients after four weeks after administration should also be consistent with the actual situation.

However, due to the small number of experimental samples, experimental results may still have errors. After further validation by more data sets, this method can be applied to clinical diagnosis. The classification method can be used as a reference for clinical diagnosis of CML patients, and can also be used as an indicator for clinical prognosis and comparison of efficacy of different drugs. To improve the classification accuracy, the next step is to study how to select feature genes more effectively and how to improve the signal reconstruction algorithm.

Acknowledgment

This research was supported by the Major Research Plan of National Natural Science Foundation of China (Grant No. 91730301), the Key Projects of National Natural Science Foundation of China (Grant No. 1183 1015) and the Foundation of the Innovation Project of Jiangsu Province (Grant No. 201810295049X).

References

[1] H. Zhou, and R. Xu, Leukemia stem cells: the root of chronic myeloid leukemia. *Protein & Cell*, **2015**, 6(6): 403-412.

[2] Y. Zhou, H. Li, and H. Hui. Update on emerging treatments for chronic myeloid leukemia. *Pharmaceutical Biotechnology*, **2018**, 25(4): 363-367.

[3] S. Claudiani, and J.F. Apperley, The argument for using imatinib in CML. *Hematology Am. Soc. Hematol. Educ. Program*, **2018**, (1): 161-167.

[4] M. Hu, K.F. Yuan, X.M. Li, Y. Chen, T. Mao, and H.Y. Xing. Clinical effect of imatinib, nero imatinib and dasatinib on chronic myeloid leukemia in chronic phase. *The Chinese Journal of Clinical Pharmacology*, **2016**, 32(6): 511-513.

[5] H. Wang. Clinical efficacy of imatinib in the treatment of chronic myelogenous leukemia. *Guide of China Medicine*, **2018**, 16(3): 32-33.

[6] J.C. Ang, A. Mirzal, H. Haron, and H.N.A. Hamed. Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **2016**, 13(5): 971-989.

[7] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, **2006**, 52(4): 1289-1306.

[8] R.G. Baraniuk, E. Candes, R. Nowak, and M. Vetterli. Compressive sampling. *IEEE Signal Processing Magazine*, **2008**, 25(2): 12-13.

[9] W.L. Tang, H.B. Cao, and Y.P. Wang. Subtyping of leukemia with gene expression analysis using compressive sensing method. 2011 First IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology, 2011, pp. 76-80.

[10] W. Shao, A. Bouzerdoum, and S.L. Phung. Compressed sensing-based frequency selection for classification of ground penetrating radar signals. IEEE International Conference on Acoustics, 2012, pp. 3377-3380.

[11] C.R. Jian, and X.Y. Chen. Gene expression data classification model based on sparse representation and least square regression. *Journal of Fuzhou University*, **2015**, 43(6): 738-741.

[12] Y.L. Liang. Classification of medical data based on compressed sensing. Unpublised MA dissertation, Yangzhou University.

[13] R.G. Baraniuk, Compressive sensing. *IEEE Signal Processing Magazine*, **2007**, 24(4): 118-121.

[14] E. Candes, J. Romberg, and T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, **2006**, 52(2): 489-509.

[15] J.A. Becerra, M.J. Madero-Ayora, and J. Reina-Tosina. A doubly orthogonal matching pursuit algorithm for sparse predistortion of power amplifiers. *IEEE Micro-wave and Wireless Components Letters*, **2018**, 28(8): 726-728.

Xinyan Pu was born in 1998. She is a B.S. candidate at Jiangnan University. Her research interests include Information and Computing Science.

Kangning Yang was born in 1998. He is a B.S. candidate at Jiangnan University. His research interests include Information and Computing Science.

Tingxuan Cheng was born in 1998. She is a B.S. candidate at Jiangnan University. Her research interests include Information and Computing Science.

Jingxue Wang was born in 1998. She is a B.S. candidate at Jiangnan University. Her research interests include Information and Computing Science.

Yanbin Zhou was born in 1998. He is a B.S. candidate at Jiangnan University. His research interests include Information and Computing Science.

Jie Gao was born in 1972. She received the Ph.D degree. She is a professor and MA supervisor at Jiangnan University. Her research interests include Bioinformatics, Applied Statistics, etc. She has published over 10 SCI papers and been supported by several funds such as the Major Research Plan of National Natural Science Foundation, General Program of National Natural Science Foundation.