

基於文化計算的SVM技術於癌症分類

Nga, Nguyen Thi、吳幸珍

E-mail: 360072@mail.dyu.edu.tw

摘要

癌症是可怕的疾病之一，也是21世紀科學家的研究挑戰之一。在癌症的診斷和治療中，癌症分類扮演著非常重要的角色。隨著DNA晶片技術的出現，為不同類型的癌症，建構其相對基因表達譜，已成為一個具前瞻性的癌症分類手段；然而，它卻也為目前機器學習研究帶來挑戰。晶片技術的資料特點是，高維度且樣本數少，過穩合是一個重大的問題；因其高維度，而樣本數少的數據使情況變得更糟。支持向量機（SVM）統計分類算法，藉著超平面的幫助將數據分離。SVM對受雜訊干擾之高維數據(如晶片技術)的分類性能表現良好，然，使用支持向量機的一個主要缺點是，分類器的性能取決於參數設置。在這篇論文中，我們使用SVM分類對基因表達數據作癌症分類。一個粒子群優化（PSO）和模擬退火（SA）的混合方法，以確定最佳的SVM參數設置，提高SVM模型的性能；為了有效的提高癌症分類方法，混合各種技術，以彌補單一技術的弱點。該方法的驗證，則以6種不同的癌基因表達數據集；結腸癌，白血病，肺癌，卵巢癌，前列腺癌和乳腺癌。實驗結果證明，該方法的分類準確率，相較於其他現有的方法，非常具競爭力。

關鍵詞：癌症分類，支持向量機，參數優化

目錄

中文摘要.....	iii	Abstract	iii
.....	iv	Acknowledgement	iv
.....	v	Contents	v
.....	vi	List of Figures	vi
.....	vii	List of Tables	vii
.....	viii	Chapter 1 Introduction	viii
.....	1	Chapter 2 Support Vector Machine Classification	1
.....	4	2.1 Overview	4
.....	4	2.1.1 Linear SVM classifier	4
.....	4	2.1.2 Nonlinear SVM classifier	4
.....	9	2.2 Effects of setting parameters on SVM model	9
.....	12	Chapter 3 Hybrid PSO-SA-SVM for Cancer Classification	12
.....	15	3.1 The hybrid PSO-SA method	15
.....	15	3.1.1 Frankenstein ' s PSO	15
.....	18	3.1.2 Refine the personal best position using SA algorithm	18
.....	18	3.2 SVM parameter optimization	18
.....	19	3.3 Implementation.....	19
.....	21	Chapter 4 Conclusion	21
.....	26	References	26
.....	28	-vii- List of Figures	28
.....	5	Figure 2-1. Binary classification problem	5
.....	8	Figure 2-2. Soft margin classifier for linearly non-separable case	8
.....	10	Figure 2-3. Mapping from input space to feature space	10
.....	13	Figure 2-4. Effect of the soft margin constant C on SVM model	13
.....	13	Figure 2-5. Effect of kernel parameters on SVM model.....	13
.....	17	Figure 3-1. Topology change process.	17
.....	20	Figure 3-2. Tuning SVM model parameters	20
.....	20	-viii- List of Tables	20
.....	11	Table 2-1. Compilation of the most common kernels	11
.....	22	Table 3-1. Main characteristics of the microarray datasets used	22
.....	23	Table 3-2. The scale range of four datasets	23
.....	23	Table 3-3. Parameter setting summary	23
.....	24	Table 3-4. The classification accuracies on testing data sets	24
.....	24	Table 3-5. The comparison of tuned SVM classifiers with default SVM classifiers on the classification accuracy (%)	24
.....	25	Table 3-6. The comparison of our method with others	25

參考文獻

[1] M. Eisen and P. Brown, " DNA Arrays for Analysis of Gene Expression " , Methods Enzymology, vol. 303, pp. 179-205, 1999.

- [2] R. Lipshutz, S. Fodor, T. Gingeras, and D. Lockhart, "High Density Synthetic Oligonucleotide Arrays", *Nature Genetics*, vol. 21, pp. 20-24, 1999.
- [3] G. Sophia Reena, P. Rajeswari, "A Survey of Human Cancer Classification using Micro Array Data", *Int. J. Comp. Tech. Appl.*, vol. 2 (5), pp. 1523-1533, 2011.
- [4] F. Chu, L. Wang, "Application of Support Vector Machines to Cancer Classification with Microarray Data", *International Journal of Neural Systems*, vol. 15, no. 6, pp. 475 – 484, 2005.
- [5] S. Wang, J. Wang, H. Chen and B. Zhang, "SVM-Based Tumor Classification with Gene Expression Data", *ADMA 2006, LNAI 4093*, pp. 864-870.
- [6] J. Phan, R. Moffitt, J. Dale, J. Petros, A. Young, M. Wang, "Improvement of SVM Algorithm for Microarray Analysis Using Intelligent Parameter Selection", *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4, 2005*.
- [7] O. Okun, H. Priisalu, "Ensembles of Nearest Neighbors for Gene Expression Based Cancer Classification", *Studies in Computational Intelligence (SCI) 126*, pp. 115 – 134, 2008.
- [8] R. Diaz-Uriarte, S. Alvarez-de Andres, "Gene Selection and Classification of Microarray Data using Random Forest", *BMC Bioinformatics* 7: 3, 2006.
- [9] A. Hasan, "Evaluation of Decision Tree Classifiers and Boosting Algorithm for Classifying High Dimensional Cancer Datasets", *International Journal of Modeling and Optimization*, vol. 2, no. 2, April 2012.
- [10] J. Khan, J. Wei, M. Ringner, L. Saal, M. Ladanyi, "Classification and Diagnostic Prediction of Cancers using Expression Profiling and Artificial Neural Networks", *Nat Med* 7, pp. 673 – 679, 2001. 29 [11] P. Rajeswari and G. Sophia Reena, "Human Liver Cancer Classification using Microarray Gene Expression Data", *International Journal of Computer Applications (0975 – 8887)*, vol. 34, no. 6, November 2011.
- [12] Lee JW, Lee JB, Park SH M Song, "An Extensive Comparison of Recent Classification Tools Applied to Microarray Data", *Computational Statistics and Data Analysis* 2005, 48:869 – 885.
- [13] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, "A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis", *Bioinformatics* 21 (5), pp. 631 – 643, 2005.
- [14] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., New York: Springer-Verlag, 1999.
- [15] B. Scholkopf, C. J. C. Burges and A. Smola, *Advances in Kernel Methods: Support Vector Learning*, Cambridge, MA: MIT Press, 1999.
- [16] K.M. Chung, W.C. Kao, C.L. Sun, L.L. Wang, C.J. Lin, "Radius Margin Bounds for Support Vector Machines with RBF Kernel", *Neural Comput.* 15 (11), November 2003.
- [17] F. Friedrichs, C. Igel, "Evolutionary Tuning of Multiple SVM Parameters", *Neurocomputing* 2004, 64:107 – 17.
- [18] V. N. Vapnik. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
- [19] J. Kennedy, R.C. Eberhart, "Particle swarm optimization", In: *Proc IEEE International Conference on Neural Networks*, IEEEservice Center, Piscataway, N.J., 4:1942 – 1948, 1995.
- [20] M. A. M. de Oca, T. Stutzle, M. Birattari, and M. Dorigo, "Frankenstein's PSO: A Composite Particle Swarm Optimization Algorithm", *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 5, October 2009. 30 [21] C-C. Chang and C-J. Lin. LIBSVM: A Library for Support Vector Machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [22] H. Hu, J. Li, A. Plank, H. Wang, G. Daggard, "A Comparative Study of Classification Methods for Microarray Data Analysis", *Proc. Fifth Australasian Data Mining Conference (AusDM)*, 2006.
- [23] L. Guo-Zheng, Z. Xue-Qiang, Y. Y. Jack, Y. Mary Qu, "Partial Least Squares Based Dimension Reduction with Gene Selection for Tumor Classification", *Bioinformatics and Bioengineering (BIBE)*, 2007.
- [24] P. Yanxiong, L. Wenyuan, and L. Ying, "A Hybrid Approach for Biomarker Discovery from Microarray Gene Expression Data for Cancer Classification", *Cancer Informatics*, pp. 301 – 311, 2006.
- [25] Y. Jieping, L. Tao, X. Tao, and J. Ravi, "Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 4, Oct-Dec 2004.