Chapter 5

Conclusion and Future Work

Chapter 5
Conclusion and Future Work

5.1 Conclusion

In this thesis, a multilevel feature selection approach (MLFS) and two semi-supervised machine learning approaches adaptation are proposed. MLFS integrates deep and active learning to select the best genes that will enhance the classification accuracy. The proposed feature selection approach was also extended to apply for miRNAs feature selection. The experimental results show that the proposed feature selection approach was able to outperform classical feature selection methods in terms of F1-measure by 9% in HCC, 6% in lung cancer and 10% in breast cancer. In addition, experimental results show the enhancement in F1-measure of our approach over recent related work in [1] and [2].

Also, in this thesis, two semi-supervised machine learning approaches were adapted to classify cancer subtypes based on miRNA and gene expression profiles. They both exploit the expression profiles of unlabeled samples to enrich the training data. The miRNA-gene relation is additionally used to enhance the classification in co-training. Both self-learning and co-training approaches improved the accuracy compared to Random Forests and SVM as baseline classifiers. The results show up to 20% improvement in F1-measure in breast cancer, 10% improvement in precision in metastatic HCC cancer and 3% improvement in F1-measure in squamous lung cancer. Co-Training also outperforms LDS approach by around 25% improvement in F1-measure for breast cancer.

5.2 Future Work

Our future work can be summarized in the following points:

- Integrate both MLFS feature selection method with cancer classifiers resulted from semi-supervised machine learning techniques (self-learning and co-training).
- Explore different mapping functions for mapping miRNA expression profiles to gene expression profiles and vice versa in the co-training approach.
- Explore different feature selection techniques to use with the DBN high level representations.
- Explore different active learning strategies other than the uncertainty sampling used in the MLFS approach.
- Integrate other biology relations like gene ontology and gene pathways.
- Explore Single Nucleotide Polymorphism (SNP) effect on genes and miRNAs and how can this information be integrated to enhance cancer classifiers.

Bibliography

[1] R. Fakoor, F. Ladhak, A. Nazi and M. Huber. Using deep learning to enhance cancer diagnosis and classification. In Proceedings of the International Conference on Machine Learning (ICML) Workshop on the Role of Machine Learning in Transforming Healthcare (WHEALTH). Atlanta, GA, June 2013.

[2] A. Sharma, S. Imoto, and S. Miyano. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Transcation on Computational Biology and Bioinformatics, 9(3):754-764, 2012.

[3] `http://deeplearning.net/tutorial/DBN.html` [Last Seen 16/3/2014]

[4] B. John, AJ. Enright, A. Aravin, T. Tuschl, C. Sander and D. Marks. MiRanda application: Human microRNA targets. PLoS Biol. Jul., 3(7):e264, 2005.

[5] `http://en.wikipedia.org/wiki/Gene`[Last Seen 3/8/2014]

[6] `http://en.wikipedia.org/wiki/Gene_expression`[Last Seen 3/8/2014]

[7] `http://en.wikipedia.org/wiki/RNA`[Last Seen 3/8/2014]

[8] Y. Katayama, M. Maeda, K. Miyaguchi, S. Nemoto, M. Yasen, S. Tanaka, H. Mizushima, Y. Fukuoka, S. Arii and H. Tanaka. Identification of pathogenesis-related microRNAs in hepatocellular carcinoma by expression profiling. Oncol. Lett., 4(4):817-823, October, 2012.

[9] Chang. Liu, G. Adrian Calin, S. Volinia and C. M Croce. MicroRNA expression profiling using microarrays. Nature Protocols 3, 563-578, 2008.

[10] S. Volinia, G. Calin and C. Liu. A microRNA expression signature of human solid tumors defines cancer gene targets. Proceedings Natl. Acad. Sci. USA, 103:2257-2261, 2006.

[11] N. Yanaihara, N. Caplen and E. Bowman. Unique micro RNA molecular profiles in lung cancer diagnosis and prognosis. Cancer Cell, 9:189-198, 2006.

[12] EJ. Lee, Y. Gusev and J. Jiang. Expression profiling identifies microRNA signature in pancreatic cancer. Int. J. Cancer, 120:1046-1054, 2007.

[13] Y. Murakami, T. Yasuda and K. Saigo. Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and nontumorous tissues. Oncogene. 25:2537-2545, 2006.

[14] A. Budhu, H. Jia and M. Forgues. Identification of metastasis-related microRNAs in hepatocellular carcinoma. Hepatology, 47:897-907, 2008.

[15] H. Varnhort, U. Drebber and F. Schulze. MicroRNA gene expression profile of hepatitis C virus-associated hepatocellular carcinoma. Hepatology. 47:1223-1232, 2008.

[16] A. Mohamed, G. E. Dahl and G. E. Hinton. Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech & Language Processing, 20(1):14-22, 2012.

[17] T. N. Sainath, B. Kingsbury, A. Mohamed and B. Ramabhadran. Learning filter banks within a deep neural network framework. IEEE Workshop on

Automatic Speech Recognition and Understanding (ASRU), pp. 297-302, Olomouc, Czech Republic, 2013.

[18] F. Agostinelli, M. R. Anderson and H. Lee. Robust image denoising with multi-Column deep neural networks. In proceedings of Neural Information Processing Systems (NIPS), pp. 1493-1501, Nevada, USA, 2013.

[19] H. Lee, C. Ekanadham and A. Y. Ng. Sparse deep belief net model for visual area V2. In proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems (NIPS), pp. 873-880, Nevada, USA, 2008.

[20] G.E. Hinton and R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, 313(5786):504-507, 2006.

[21] G. E. Hinton. A practical guide to training restricted boltzmann machines. Lecture Notes in Computer Science, 7700:599-619, 2012.

[22] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[23] M. Wand and X. Hua. Active learning in multimedia annotation and retrieval: a survey. ACM Transactions on Intelligent Systems and Technology (TIST), 2(2), 2011.

[24] B. Yang, J. Sun, T. Wang, Z. Chen. Effective multi-label active learning for text classication. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), pp. 917-926, Paris, France, 2009.

[25] M. Shi and B. Zhang. Semi-supervised learning improves gene expression-based prediction of cancer recurrence. Bioinformatics, 27(21):3017-3023, 2011.

[26] C. Rosenberg, M. Hebert and H. Schneiderman. Semi-supervised self-training of object detection models. 7th IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing (WACV/MOTION), pp. 29-36, Breckenridge, CO, USA , 2005.

[27] R. Mihalcea. Co-training and self-training for word sense disambiguation. In Proceedings of Conference on Natural Language Learning (CoNLL), pp. 33-40, Boston, MA, USA, 2004.

[28] O. Chapelle, B. Scholkopf and A. Zien. Semi-supervised learning. Cambridge, Mass., MIT Press, 2006.

[29] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. Proceedings of the Workshop on Computational Learning Theory (COLT), pp. 92-100, Wisconsin, USA, 1998.

[30] S. Kiritchenko and S. Matwin. Email Classification with co-training. In Proceedings of Conference of the Center for Advanced Studies on Colleborative Research (CASCON), pp. 301-312, Ontario, Canada, 2011.

[31] S. Gilbert Nancy and S. Appavu alias Balamurugan. A comparative study of feature selection methods for cancer classification using gene expression dataset. Journal of Computer Applications (JCA), 6(3):78-84, 2013.

[32] Y. Wanga, I. V. Tetkoa, M. A. Hallb, E. Frankb, A. Faciusa, K. F.X. Mayera and H. W. Mewesa. Gene selection from microarray data for cancer classification-a machine learning approach. Computational Biology and Chemistry, 29:37-46, 2005.

[33] Y. Wang and F. Makedon. Application of relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray

Data. In Proceedings of the IEEE Computational Systems Bioinformatics Conference, pp. 477-478, Stanford, CA, USA, 2004.

[34] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong. Feature selection for gene expression using model-based entropy. IEEE/ACM Transcation on Computational Biology and Bioinformatics, 7(1):25-36, 2013.

[35] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont and Y. Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 26(3):392-398, 2010.

[36] K. Kim and S. Cho. Exploring features and classifiers to classify microRNA expression profiles of human cancer. International Conference on Neural Information Processing (ICONIP), pp. 234-241, Australia, 2010.

[37] Y. Wang and M. H. Dunham. Classifier fusion for poorly-differentiated tumor classification using both messenger RNA and microRNA expression profiles. In Proceedings of the 2006 Computational Systems Bioinformatics Conference (CSB 2006), Stanford, California, 2006.

[38] Y. Zheng and C. Keong Kwoh. Cancer classification with microRNA expression patterns found by an information theory approach. Journal of Computers (JCP), 1(5):30-39, 2006.

[39] H. Liu, J. Li and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. International Conference on Genome Informatics, 13:51-60, Tokyo, Japan, 2002.

[40] `http://www.cs.waikato.ac.nz/ml/weka/` [Last Seen 16/3/2014]

[41] `https://code.google.com/p/symja` [Last Seen 16/3/2014]

[42] http://www.biolab.si/supp/bi-cancer/projections/info/SRBCT.htm [Last Seen 16/3/2014]

[43] L. Breiman. Random foresets. Machine Learning, 45(1):5-32, 2001.

[44] O. Okun and H. Priisalu. Random forest for gene expression based cancer classification. Overlooked Issues. Pattern Recognition and Image Analysis, 4478:483-490, 2007.

[45] M. Klassen, M. Cummings and G. Saldana. Investigation of random forest performance with cancer microarray data. In proceedings of the ISCA 23rd International Conference on Computers and Their Applications (CATA), pp. 64-69, Cancun, Mexico, April 9-11, 2008.

[46] R. Diaz-Uriarte and S. Alvarez de Andres. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7:3, 2006.

[47] http://olivier.chapelle.cc/lds/ [Last Seen 8/7/2014]

[48] http://www.ncbi.nlm.nih.gov/geo [Last Seen 8/7/2014]

[49] H. Chen, B. Yang, J. Liu and D. Liu. MicroRNA signatures predict oestrogen receptor, progesterone receptor and HER2/neu receptor status in breast cancer. Breast Cancer Res., 11(3):R27, 2009.

[50] S. Thorgeirsson and J. Grisham. Molecular pathogenesis of human hepatocellular carcinoma. Nat. Genet., 31:339-346, 2002.

[51] D. Parkin, F. Bray, J. Ferlay and P. Pisani. Global cancer statistics. CA Cancer J. Clin., 55:74-108, 2005.

[52] K.Yuki, S. Hirohashi, M. Sakamoto, T. Kanai and Y. Shimosato. Growth and spread of hepatocellular carcinoma. A review of 240 consecutive autopsy cases. Cancer, 66:2174-2179, 1990.

[53] A. Chambers, A. Groom and I. MacDonald. Dissemination and growth of cancer cells in metastatic sites. Nat. Rev. Cancer, 2:563-572, 2002.

[54] J. A. Bishop, H. Benjamin, H. Cholakh, A. Chajut, D. P. Clark and W. H. Westra. Accurate classification of nonsmall cell lung carcinoma using a novel microRNA-based approach. Clin. Cancer Res., 16(2):610-619, 2010.

[55] R. Ibrahim, N. A. Yousri, M. A. Ismail and N. M. El-Makky. miRNA and gene expression based cancer classification using self-learning and co-training approaches. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 495-498, China, 2013.

[56] R. Ibrahim, N. A. Yousri, M. A. Ismail and N. M. El-Makky. Multi-level gene/miRNA feature selection using deep belief nets and active learning. In proceedings IEEE International Conference of the Engineering in Medicine and Biology Society (EMBC), Chicago, IL, USA, 2014.

ملخص الرسالة

اصبح تحسين مصنفات مرض السرطان مسالة ذات اهمية كبري في مجال المعلومات الحيوية . يمكن تحسين مصنفات مرض السرطان بطريقتين مختلفتين. الطريقة الاولي تكمن في تحسين اختيار المميزات المستخدمة مع مصنفات مرض السرطان بينما الطريقة الثانية تكمن في تحسين مصنفات مرض السرطان ذاتها. تقدم هذه الرسالة طريقة لاختيار المميزات بناءا علي تعبير الجين و المايكروارايه . تعتمد الطريقة المقدمة علي التصرف الجماعي للجينات بدلا من التصرف الفردي و ذلك لاختيار افضل الجينات و المايكروارايز. كما تجمع الطريقة المقدمة بين التعليم العميق و التعليم النشط.

لم يتم استخدام البيانات المسماة و غير المسماة لتدريب مصنفات مرض السرطان من قبل في حالة وجود كلا من تعابير الجينات و المايكروارايز . كما انه يوجد حافز للدمج بين تعابير الجينات و المايكروارايز للحصول علي معلومات اكثر عن مرض السرطان. نقدم في هذه الرسالة اثنين من طرق التعلم شبه المشرف و هما التعلم الذاتي و التعلم المساعد لتحسين مصنفات مرض السرطان.

تظهر النتائج ان الطريقة المقدمة لاختيار المميزات تفوقت علي الطرق التقليدية . كذلك تم تقييم طرق التعلم شبه المشرف علي سرطان الثدي ، سرطان الكبد الوبائي و سرطان الرئة . اظهرت النتائج تحسن بمقدار ٢٠ % عن مصنفات مرض السرطان التقليدية.

الباب الاول: يحتوي علي مقدمة الرسالة و الهدف من الرسالة و يستعرض ابواب الرسالة

الباب الثاني: يحتوي علي شرح للطرق التقليدية و مقارنة بين الطرق التقليدية و الطرق المقدمة في الرسالة

الباب الثالث: يشرح الطريقة المستخدمة لاختيار الجينات و المايكروارايز بناءا علي ملامح التعبير و يعقد تجارب لمقارنة الطريقة المقدمة بالطرق التقليدية

الباب الرابع: يشرح طرق التعلم شبه المشرف وهما التعلم الذاتي و التعلم المساعد و يعقد تجارب لمقارنة الطرق المقدمة بالطرق التقليدية

الباب الخامس: يتعرض لخاتمة الرسالة و الدراسات المستقبلية التي يمكن تطبيقها

جامعة الإسكندرية
كلية الهندسة
قسم الحاسب و النظم

طرق اختيار الميزات بناءا علي تعبير الجين و المايكروارايه و طرق تصنيف مرض السرطان باستخدام طرق التعليم شبه المشرف

رسالة علمية

مقدمة الي الدراسات العليا بكلية الهندسة جامعة الاسكندرية

استيفاءا للدراسات المقررة للحصول علي درجة

الماجستير في العلوم الهندسية

في

هندسة الحاسب و النظم

مقدمة من

المهندسة رانيا محمد محمد ابراهيم

٢٠١٥

طرق اختيار المميزات بناءا علي تعبير الجين و المايكروارانايه و طرق تصنيف مرض السرطان باستخدام طرق التعليم شبه المشرف

مقدمة من

رانيا محمد محمد ابراهيم

للحصول علي درجة الماجستير

في

هندسة الحاسب و النظم

لجنة المناقشة و الحكم علي الرسالة	**موافقون**
الاستاذ الدكتور / مجدي حسين ناجي	.....
الاستاذ الدكتور / محمد عبد الحميد اسماعيل	.....
الاستاذ الدكتور / نجوي مصطفي المكي	.....
الاستاذ الدكتور / صالح عبد الشكور الشهابي	.....

وكيل الكلية للدراسات العليا و البحوث
كلية الهندسة ـ جامعة الاسكندرية

لجنة الاشراف علي الرسالة **موافقون**

الاستاذ الدكتور / محمد عبد الحميد اسماعيل .....

كلية الهندسة ، جامعة الاسكندرية

الاستاذ الدكتور / نجوي مصطفي المكي .....

كلية الهندسة ، جامعة الاسكندرية