Multi-Level Gene/MiRNA Feature Selection using Deep Belief Nets and Active Learning

Rania Ibrahim, Noha A. Yousri, Mohamed A. Ismail and Nagwa M. El-Makky[1]

***Abstract*— Selecting the most discriminative genes/miRNAs has been raised as an important task in bioinformatics to enhance disease classifiers and to mitigate the dimensionality curse problem. Original feature selection methods choose genes/miRNAs based on their individual features regardless of how they perform together. Considering group features instead of individual ones provides a better view for selecting the most informative genes/miRNAs. Recently, deep learning has proven its ability in representing the data in multiple levels of abstraction, allowing for better discrimination between different classes. However, the idea of using deep learning for feature selection is not widely used in the bioinformatics field yet.**

In this paper, a novel multi-level feature selection approach named MLFS is proposed for selecting genes/miRNAs based on expression profiles. The approach is based on both deep and active learning. Moreover, an extension to use the technique for miRNAs is presented by considering the biological relation between miRNAs and genes. Experimental results show that the approach was able to outperform classical feature selection methods in hepatocellular carcinoma (HCC) by 9%, lung cancer by 6% and breast cancer by around 10% in F1-measure. Results also show the enhancement in F1-measure of our approach over recently related work in [1] and [2].

I. INTRODUCTION

Previous studies ([3], [4]) have shown a strong relation between gene expression profiles and disease types/subtypes by using gene expression sets to detect or discriminate cancer or its subtypes. Gene expression sets suffer from a dimensionality curse problem, which exists because of the large number of genes and a few number of samples. To deal with this problem, several feature selection methods were proposed in literature to select a subset of genes to use in classifying cancer or its subtype like ([1], [2], [5], [6]). In addition, feature selection methods are useful for discarding irrelevant genes which can cause noise in the classification phase and therefore enhances the classification model in terms of execution time, size and accuracy.

In this paper, deep learning and active learning are integrated together. Our objective is to enhance the classification accuracy using the least number of most discriminative genes. Deep learning [1] has been widely used recently in several fields like image and audio applications. Deep Belief Net (DBN) is one of the algorithms used in applying deep learning. It has also shown its ability to detect high level features and enhance the classification accuracy. Moreover, to the best of our knowledge, it has been used for the first time in [1] to enhance cancer diagnosis and classification based on gene expression profiles. On the other hand, original feature selection methods ([1], [5], [6]) choose genes to include in detecting cancer or its subtype based only on the gene expression profile, regardless of how groups of genes perform together. Our approach mitigates this problem using active learning. Traditional active learning [7] is a semi-supervised machine learning approach. It is usually used when the unlabeled data is abundant but manually labeling is expensive. The original active learning approach starts by training a classifier with a small set of labeled samples. In this paper, an extension for an active learning is introduced by using it in an unsupervised manner.

Moreover, our approach is extended to deal with miRNA feature selection problem. MicroRNAs (miRNAs) are short (1925 nucleotides) noncoding single-stranded RNA molecules [8], which regulate gene expression either at the transcriptional or translational level [8]. In this paper, our approach is extended to select miRNAs features using the multi-level gene feature selection approach and by exploiting the biology relation of miRNA target genes.

The paper is organized as follows: section II discusses the related work, while section III describes the proposed approaches in detail and section IV shows experimental results. Finally section V concludes the paper.

II. Related Work

Many gene feature selection methods have been proposed in literature recently, e.g. ([1], [2], [5], [6]). Authors in [5] uses relief-f filtering feature selection method and use SVM classifier to discriminate the samples. The work in [6] uses an ensemble of feature selection methods to enhance feature selection stability and classification accuracy. However, genes are evaluated independently regardless of how well they perform together; this problem is solved in our proposed approach at the active learning phase as shown in the next section. Also, this problem was addressed in [2]. Authors in [2] did not consider the gene performance alone but evaluated its performance jointly with other genes. However, our approach evaluates joint gene performance differently by using DBN features and active learning.

Deep learning has been widely used recently in several fields like image and audio applications ([9], [10]). DBN is adopted in this paper to find high level better representations of gene expression sets. [1] has used deep learning by construing an auto-encoder to perform dimensionality reduction and then use DBN with the new dimensions to classify

[1] Rania Ibrahim, Noha A. Yousri, Mohamed A. Ismail and Nagwa M. El-Makky are with Computer and Systems Engineering Department, Alexandria University, Alexandria 21544, Egypt `rania.ibrahim.salama@gmail.com`, `noha.yousri@alexu.edu.eg`, `drmaismail@gmail.com` and `nagwamakky@alexu.edu.eg`

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

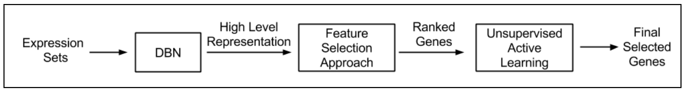
Expression Sets
DBN
High Level Representation
Feature Selection Approach
Ranked Genes
Unsupervised Active Learning
Final Selected Genes

Fig. 1. Multi-level Feature Selection Approach for Gene Expression Sets

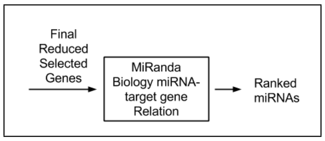
Final Reduced Selected Genes
MiRanda Biology miRNA-target gene Relation
Ranked miRNAs

Fig. 2. miRNAs Feature Selection based on Multi-level Gene Feature Selection Approach

the samples. In our approach, genes were not mapped into principal components first as we want to keep track of the most discriminative genes for the biologist to be able to find their association to the disease in question.

In the next section, the proposed approach MLFS is described in detail, showing its main components and how it is extended to work with miRNAs.

III. The Proposed MLFS Approach

A. System Overview

As shown in figure 1, the approach for gene feature selection is composed of three main components which are:

- Deep Belief Net: The DBN [11] is used to generate a high level representation of the genes that capture their interaction and behavior.
- Feature Selection: Feature selection is then applied on the high level representation generated by the DBN to select a subset of the genes.
- Unsupervised Active Learning: Unsupervised active learning is then applied to reduce the subset of genes by selecting the genes that increase the accuracy of classification.

In addition, the approach is extended for miRNA feature selection as shown in figure 2. The next subsections explain each component of the proposed approach in detail.

B. Deep Belief Nets and Feature Selection

DBNs are defined as graphical models which can learn to extract a deep hierarchical representations of the training data. We use the training procedure proposed in [11] and the open source library implementation in [12]. The deep learning library was used in this paper with two hidden layers and the number of neurons at each layer is equal to the input layer original features.

Feature selection step is applied on the high level features generated by the DBN. In this step, any appropriate feature selection method as statistical t-test or relief-f can be applied. Statistical t-test and relief-f were used in the experimental results section.

C. The Overall MLFS Approach

The proposed multi-level feature selection approach is used to perform feature selection for sample classifiers that are used to discriminate cancer types/subtypes. Our objective is to enhance classification accuracy using the least number of genes. The steps of the approach are described as follows:

1) Use DBN to extract high level representations (e') of the gene expression profiles (e).
2) Apply any classical feature selection method on the high level representations and rank the genes based on their scores (e.g., their p-values in statistical t-test).
3) Let n denote an initial random number of genes to use, g denote the whole set of genes, λ denote certain percentage, c denote a classifier, a denote the accuracy, $Step$ denote number of genes to add to n at each iteration and $bestGenes$ denote the best set of genes in terms of accuracy. Select the number of genes to use as follows:

```
• $x = n$
• $bestGenes$ = top ranked $x$ genes
• While($x <= \lambda * |g|$):
    – $c \leftarrow$ TrainClassifier($bestGenes$, $trainingSet$)
    – $a \leftarrow$ MeasureAccuracy($bestGenes$, $testSet$)
    – If ($a > bestAccuracy$):
        ∗ $bestGenes \leftarrow$ top ranked $x$ genes.
        ∗ $bestAccuracy \leftarrow a$
    – $x \leftarrow x + Step$.
```

4) Apply a reduction phase to reduce the number of genes in the $bestGenes$ set using the proposed unsupervised active learning approach to choose the most informative genes from the $bestGenes$ set, as follows:
   a) Label all genes in the $bestGenes$ set as either related to the cancer type if they have p-value $<$ 0.05. Otherwise, label them as not related.
   b) Construct a gene classifier using a random small subset of $bestGenes$. The classifier is used to tell if the given gene is related to the cancer type or not, based on its expression profile.
   c) Let the classifier label the $bestGenes$ set. Then, get the genes that are most informative to the classifier, which have a classification confidence close to 0.5, for example between 0.4 and 0.6.
   d) Label the most informative genes using statistical t-test. Human judgment is usually used at this stage, however, no human judgment with strong biological background was available to us. So, statistical t-test labels were used.

e) Add the most informative genes to the training set, re-train the classifier and go back to step c.
f) Record the accuracy using the test set and stop when reaching close to or higher than the one produced by using all bestGenes genes.

D. MLFS-miRNA: MiRNAs Feature Selection Extension

For efficient use of DBN, it should be given a huge training set of unlabeled data. However, the number of miRNAs is usually very small. That is why a new extended approach was proposed instead of using the same multi-level feature selection method. Our approach is extended by utilizing the biological relation of miRNA-target gene, as in databases as miRanda [15]. miRNAs are ranked based on the number of their target genes that exist in final gene feature selection set generated by the pervious active learning phase. Finally, the number of miRNAs to select is tuned in the same way as in the previous approach.

IV. Experimental Results

The results of our experiments are given in four subsections to show the effect of each of the proposed components on the classification accuracy. SVM classification is used to construct a gene classifier in the unsupervised active learning phase and a Random Forests (RFs) classifier is used to construct the sample classifier. The SVM and RF implementations were both used from Weka repository [13]. The proposed approaches were evaluated using six cancer types, namely Breast Cancer, Hepatocellular Carcinoma (HCC), Lung Cancer, Prostate Cancer, Colon Cancer and Ovarian Cancer. Table I, II and III show the sizes of the datasets.

TABLE I
Training and testing sample size using gene expression (BC = Breast Cancer, NM = non-metastatic, M = metastatic, A = adenocarcinoma and S = squamous)

Type	**BC (GSE20713)**		**BC (GSE20713)**		**HCC (GSE36376)**		**Lung (GSE41271)**	
Sub Type	**ER+**	**ER-**	**HER2+**	**HER2-**	**NM**	**M**	**A**	**S**
Train	21	23	31	13	98	118	87	41
Test	21	22	31	12	98	120	86	40

TABLE II
Training and testing samples size using miRNA expression. (NM = non-metastatic and M = metastatic)

Type	**Breast Cancer (GSE15885)**		**Breast Cancer (GSE15885)**		**HCC (GSE6857)**	
Sub Type	**ER+**	**ER-**	**HER2+**	**HER2-**	**NM**	**M**
Train	9	7	13	3	193	62
Test	7	6	12	2	162	65

TABLE III
Related Work [1] and [2] Dataset Sizes

Dataset Name	**Number of Genes**	**Training Samples**	**Testing Samples**
Prostate Cancer	12600	102	34
Colon Cancer	2000	32	30
Ovarian Cancer	15154	153	100
SRBCT	2308	63	20
MLL	12582	57	15

A. MLFS Evaluation

The first two phases of the proposed approach (will be referred to as the DBN) are first compared to five feature selection methods. These are random (by randomly selecting k genes), statistical t-test, information gain, relief-f and chi-square, which are applied on the original expressions. Statistical t-test was used from [14] while information gain, chi-square and relief-f were used from Weka repository [13]. F-measure was used to compare the DBN approach to the previous approaches. As shown in figure 3, the DBN approach was able to achieve the highest F-measure among all other methods.

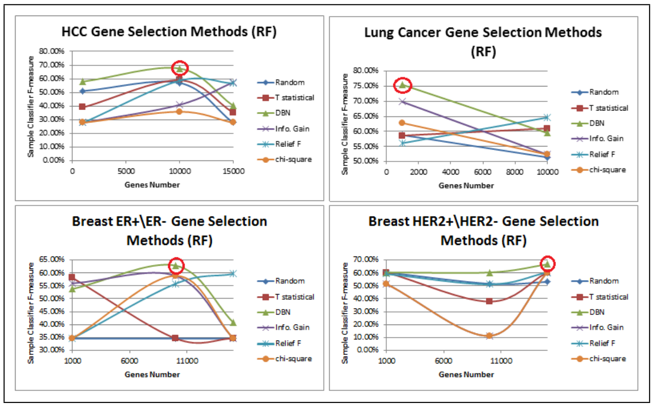

Fig. 3. Comparison to Classical Feature Selection using Gene Expression Sets (Highest approach is marked with a red circle)

B. Unsupervised Active Learning

Unsupervised active learning was used to reduce the *bestGenes* set, which was obtained from the first two phases of our approach. Table IV and V show the results after applying the active learning phase. The tables show that active learning was able to reduce the *bestGenes* set by 60% in lung cancer, 20% in HCC and 50% in Breast Cancer ER+/ER- while increasing or at least maintaining the same classification accuracy.

C. MLFS-miRNA Evaluation

The same five feature selection methods used for gene expression sets are applied to miRNA expression profiles and compared to our approach MLFS-miRNA. Figure 4 shows the comparison between them. As shown in the figure, miRNA extension has performed better in all curves.

TABLE IV

HCC AND LUNG CANCER ACTIVE LEARNING RESULTS

	Precision	Recall	F1-measure
HCC Baseline(1k)	51.70%	50.46%	51.07%
HCC AL Iteration 2 (8k)	**78.73%**	**59.63%**	**67.86%**
Lung Baseline (100)	53.89%	59.84%	56.71%
Lung AL Iteration 2 (400)	**79.75%**	**71.65%**	**75.48%**

TABLE V

BREAST CANCER ER+/ER- AND HER2+/HER2- ACTIVE LEARNING RESULTS

	Precision	Recall	F1-measure
ER+/ER- Baseline (1k)	54.10%	53.49%	53.79%
ER+/ER- AL Iteration 2 (5k)	**78.15%**	**60.47%**	**68.18%**
HER2+/HER2- Baseline (1k) & AL Iteration 1 & 2	51.97%	72.09%	60.40%
HER2+/HER2- AL Iteration 3 (14k)	**66.70%**	**72.09%**	**69.29%**

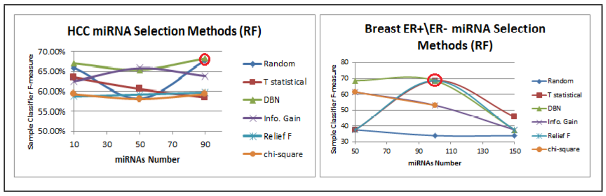

Fig. 4. Comparison to Classical Feature Selection using MiRNA Expression Sets (Highest approach is marked with a red circle)

D. Comparison to Related Work

Table III shows the datasets sizes used to compare our work to [1] and [2]. While, table VI shows the 10-fold cross-validation classification accuracy results compared to [1]. Moreover, two datasets, namely, SRBCT and MLL [16], were used to compare our approach to the approach proposed in [2]. As statistical t-test is limited to 2 classes, it was replaced by relief-f feature selection in Weka [13]. Tables VII show the 10-fold cross validation results.

TABLE VI

COMPARISON TO [1] USING 10-FOLD CROSS-VALIDATION

	Stack Autoencoder	Stacked Autoencoder with Fine Tuning	DBN (200 genes)
Prostate Cancer	73.33%	73.33%	**97.06%**
Colon Cancer	66.67%	**83.33%**	73.33%
Ovarian Cancer	55.03%	99.00%	**100%**

V. CONCLUSION

In this paper, a multi-level feature selection approach is proposed which integrates two unsupervised machine learning approaches, DBN and active learning, to select the least number of the most discriminative genes that will enhance the sample classification accuracy. The experimental results show that the proposed feature selection approach was able to outperform classical feature selection methods in F1-measure by 9% in HCC and 6% in Lung Cancer. In addition, experimental results show the enhancement in F1-measure of our approach over recent related work.

TABLE VII

COMPARISON TO [2] USING 10-FOLD CROSS-VALIDATION

		DBN using top-4 genes in *bestGenes* set	[2]
SRBCT	**Precision**	**85.3%**	74.4%
	Recall	**84.1%**	73.0%
	F1-measure	**84.7%**	73.7%
MLL	**Precision**	**77.9%**	64.3%
	Recall	**66.7%**	**66.7%**
	F1-measure	**71.8%**	65.5%

REFERENCES

[1] R. Fakoor, F. Ladhak, A. Nazi and M. Huber. "'Using deep learning to enhance cancer diagnosis and classification". In Proceedings of the 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), pp. 754-764, 2012.

[3] 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), 2013.

[4] L. Marisa, A. de Reynis, A. Duval, J. Selves, M. P. Gaub, L. Vescovo., M.-C. Etienne-Grimaldi, R. Schiappa, D. Guenot and M. Ayadi. "Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value". PLoS Medicine, DOI: 10.1371/journal.pmed.1001453, 2013.

[5] 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, 2004.

[6] 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), pp. 392-398, 2010.

[7] B. Settles. "Active Learning Literature Survey". Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[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): 817823. October 2012.

[9] S. Zhong, Y. Liu, F. Chung and G. Wu. "Semiconducting bilinear deep learning for incomplete image recognition". In proceedings of the 2nd ACM International Conference on Multimedia Retrieval (ICMR), 2012.

[10] D. Yu, G. E. Hinton, N. Morgan, J.-T. Chien and S. Sagayama. "Introduction to the Special Section on Deep Learning for Speech and Language Processing". IEEE Transactions on Audio, Speech & Language Processing 20(1), pp. 4-6, 2012.

[11] G.E. Hinton and R.R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". Science, (313):5786, pp. 504 507, 2006.

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

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

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

[15] 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.

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