Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 11, November 2018, pg.52 – 56

Classification Techniques in Gene Expression Microarray Data

Sarah M. Ayyad[1], Ahmed I. Saleh[2], Labib M. Labib[3]

[1]Department of computer engineering and systems, Mansoura University, Egypt
[2]Department of computer engineering and systems, Mansoura University, Egypt
[3]Department of computer engineering and systems, Mansoura University, Egypt
[1] sarah.aiyad@gmail.com; [2] aisaleh@yahoo.com; [3] labibm@hotmail.com

Abstract— Cancer nowadays is a common and heterogeneous disease affecting all people of all ages. Gene expression data can serve to understand cancer or other types of disease well. Building classification system using gene expression dataset that can properly classify new samples is a challenging task due to the nature of gene expression data that is usually composed of dozens of samples characterized by thousands of genes. This paper put a light on different classification methods used in classifying gene expression data including SVM, NB, C4.5 and some of the state-of-the-art techniques.

Keywords— Gene expression microarray; data mining techniques; cancer classification; classification; microarray data analysis

I. INTRODUCTION

Machine learning (ML) and data mining have established a plenty of effective applications in gene expression analysis. ML is a discipline that use automatic and intelligent learning techniques to resolve various complex and real world-problems [1, 2]. In a simplest terms, it is the study of algorithms that can learn from experience and then predict. Data mining is a fast-growing subarea of machine learning that handle discovering meaningful knowledge from large datasets [3]. Nowadays, DNA microarray technology has formed a new line of research in both data mining and bioinformatics. For example, detection of the hidden patterns in the expression profiles has formed an opportunity for precise cancer classification. These types of dataset suffer from the little number of samples and the immense number of features "genes" as they used in measuring gene expression level. In microarray analysis, different machine learning methods have been effectively employed to many of the main problems, such as gene selection, gene ontology, and gene expression data classification.

In this paper, we have displayed a wide suitability of classification techniques to gene expression microarray data.

II. GENE EXPRESSION MICROARRAY

The advent of DNA microarray technology has produced a broad pattern of gene expression data recorded in a single experiment "sample" to scientists. This expression value is such as a signature effective in diagnosing diseases, identifying tumors and selecting the appropriate remedy to resist illness and discover mutations [4]. In the last years, various datasets have become publicly available online. These datasets face several challenges, for instance, a huge number of gene expression values for each sample, and a comparatively small number of samples. Genes are made up of deoxyribonucleic acid (DNA) that contains the genetic data used in encoding

© 2018, IJCSMC All Rights Reserved

proteins and specific cellular ribonucleic acid (RNA) [1]. Where gene expression level is responsible for turning the gene to form RNA and protein. While some genes could be mutated and this cause tumor occurrence and it is reflected in the variation of the expression level of these particular genes, that implies the genes are expressed extraordinarily in certain cells [1].

A DNA chip is a tiny chip onto which a huge number of DNA molecules are connected to a solid surface. Each DNA cell of the chip identifies a DNA sequence. [5]. Fig. 1 shows the overall operation of obtaining the gene expression microarray from a DNA chip. These gene expression profiles could be operated as inputs to large-scale data analysis. The dataset is generally organized in the form of a matrix of x rows and y columns, which is termed as a gene expression profile. Where n refer to genes and m refer to the test sample [1, 6]. More and more researches have proved that many genes calculated in a DNA microarray experiment are not appropriate in the correct classification of various classes of the problem. To overcome this issue, gene "feature" selection acts a vital role in mining DNA microarray that is defined as the procedure of determining genes or set of genes that can distinguish between normal samples and diseased samples [5].

Fig. 1 Process of obtaining gene expression level from DNA microarray

III. DATA MINING CLASSIFICATION TECHNIQUES

Classification is an important field that focused on allocating a sample to one of a group of classes, based on its features [5]. In this research, we handle the classification problem that concentrates on dividing the samples of cancer patients into two classes. Generally, classification techniques could be split into two groups: supervised and unsupervised. Supervised techniques are executed on a set of training samples. Each sample is given with its label (the corresponding output). The unsupervised techniques (clustering) classifies the samples by finding clusters based upon a particular criterion. As increasingly gene expression microarray data becomes public available, classification techniques for microarray data analysis become a significant task. Different common classification techniques are discussed in the next subsections.

A. Support Vector Machine

Support vector machine (SVM) was firstly presented by Vapnik [7]. This method was introduced based on the statistical learning theory. It has been applied broadly in various classification and regression tasks due to its efficiency in dealing with linearly non-separable and high dimensionality dataset [8]. SVM constructs the hyperplane that most ideally splits the two classes of the training samples in feature space. The feature vectors at the edge of each class are named as the support vectors. The distance between the dividing hyperplane and support vectors is named as the margin. The best hyperplane is the one that maximizes the margin. This is referred to as margin maximization. Fig. 2 shows how data of two classes are divided in the feature space. If training samples for one class are surrounded with samples of the other class in feature space, it is very difficult to bisect the samples by a hyperplane (e.g., linear classification). For nonlinear classification, feature vectors are transformed into a higher dimension, in which samples could be divided by a hyperplane [9]. SVM uses a kernel function that transform the data to a different feature space. The most common kernel functions that used along with SVM are, as follows.

(i) Linear Kernel [10]

$$K(x_i, x_j) = x_i x_j \qquad (1)$$

© 2018, IJCSMC All Rights Reserved

(ii) Polynomial Kernel [10]

$$K(x_i, x_j) = (1 + x_i x_j)^d \quad (2)$$

(iii) Gaussian Kernel [10]

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (3)$$

(iv) Sigmoid Kernel [10]

$$K(x_i, x_j) = \tanh(\sigma x_i^T x_j + C) \quad (4)$$

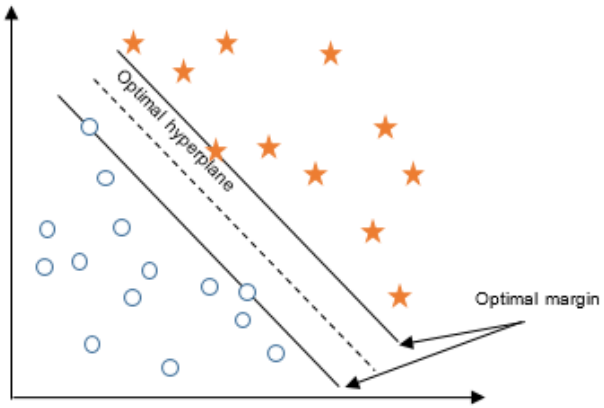

Fig. 2 Applying SVM in a two-dimensional space

B. Naïve Bayes

Naive Bayes (NB) is one of the most common classification methods that is based on Bayesian theorem. It is especially useful for learning with high dimensional data like gene expression data [11]. The key advantage of NB is that it is very simple to build and robust to outlier and irrelevant features. In spite of its simplicity, NB can often outperform more machine learning algorithms. Bayesian theorem for classification is to allocate the test item to the class that has the highest conditional probability. Generally, one set of training samples with a class label is given, a classifier should be experienced to anticipate the class distribution of a sample with its class label undefined. The classifier represented by Bayesian theorem uses Eq. (5) to classify s [2]:

$$c(s) = arg\max_{c \in C} P(c) \prod_{j=1}^{n} P(f_j|c) \quad (5)$$

Where s is the test sample, c(s) denotes the predicted class of sample s, and c is the class label. The set of all class labels is denoted by C, and n is the number of features, and fj is the value of each feature, fj (j = 1, 2… n). P(c) is the prior probability of class c and P (fj|c) is the conditional probability of each feature in the training set.

C. C4.5 Decision Tree

C4.5 employs a divide-and-conquer strategy for developing decision trees that was introduced by Hunt [12]. It is one of the most common methods for classification in different machine learning applications that help in the process of pattern recognition. C4.5 is an enhanced model of ID3 tree learning algorithm, which is capable of learning pre-defined classes from labelled instances. A decision tree is a tree where every node represents the values of a feature, and the leaves depict the class of a sample that fulfils the tests. The tree output will be a 'yes' or 'no' when the full set of samples are tested on it. Classification rules are acquired from the root to a leaf to determine which class a leaf belongs to. The rules can be pruned to reduce the total tree size and avoid over-fitting. C4.5 algorithm employs an enhanced splitting criterion, named gain ratio [13]. Fig. 3 provides a decision tree derived from colon cancer dataset where oval denotes decision node "genes", square denotes leaf node "decision outcome", and the branch denotes the expression level conditions.

© 2018, IJCSMC All Rights Reserved

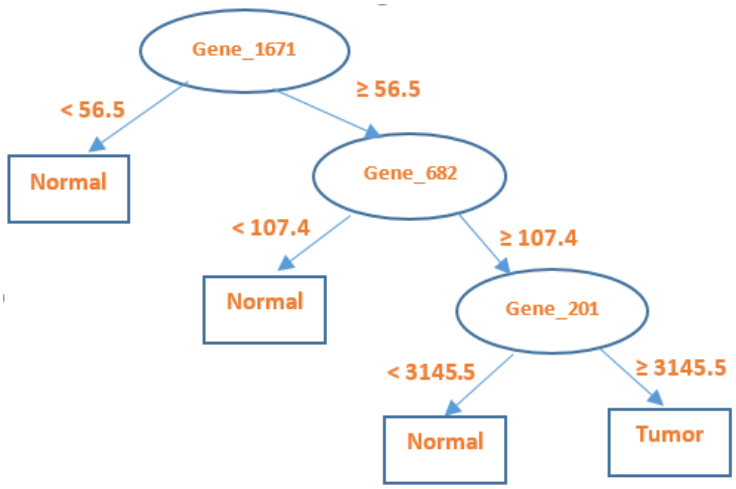

Fig. 3 Illustration of decision tree

D. Ensemble Classifier

Ensemble classifier involves merging decisions of different classifiers to produce a final decision and is often used for obtaining highly accurate results. Ensemble classifiers are quite common in machine learning problems, and they have also been employed in bioinformatics field [14, 15]. Fig. 4 shows how ensemble classifier works. The final classification decision is determined by combining the decision of each classifier. There are two common approaches for combining the decision, namely majority voting and weighted majority voting [16]. In the majority voting, the classification is done based on the class that gets the largest number of votes, while in the weighted majority voting, every classifier is assigned a weight.

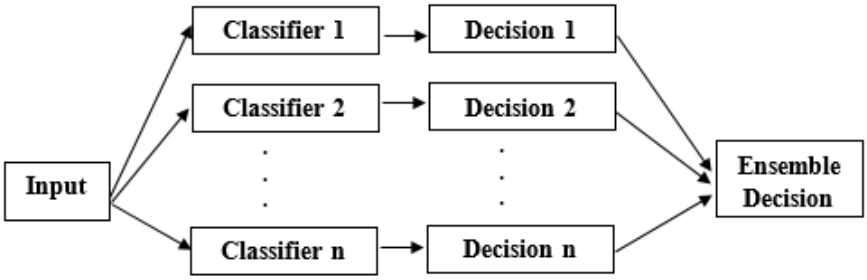

Fig. 4 Ensemble of classifiers

IV. COMPARISONS OF CLASSIFICATION ALGORITHMS

TABLE I
VARIOUS GENE EXPRESSION CLASSIFICATION TECHNIQUES

Year	Technique	Contribution	Dataset	Accuracy
2012	Multiclass classification of microarray data samples with Flexible Neural Tree [17]	A novel a flexible neural tree technique applied to multiclass cancer classification	MLL	98.6
			Lymphoma	100
2014	Gene expression microarray classification using PCA–BEL [18]	A new classification method based on brain emotional learning and principal component analysis to classify gene expression microarray	SRBCT	100
			HGG	95
			Lung	97.5
			Colon	87
2015	Hidden Markov models for cancer classification using gene expression profiles [19]	A new method for cancer classification is suggested by using supervised learning hidden Markov models	DLBCL	98.83 ± 1.33
			Leukemia	98.26 ± 1.68
			Colon	89.11 ± 4.47
			Prostate	92.01 ± 2.59
2017	Microarray Data Classification Using Dual Tree M-Band Wavelet Features [20]	A new classification technique for both gene selection and classification is suggested, where the classic KNN is used for classification	Breast	90.72
			Colon	90.3
			Ovarian	91.7
			CNS	93.3
			Leukemia	91.67

© 2018, IJCSMC All Rights Reserved

2018	Cancer Gene Detection Using Neuro Fuzzy Classification Algorithm [21]	Cancer classification system based on neuro-fuzzy is proposed, but it has some limitations with high dimensional data	Lymphoma	51
2018	Improving Classification of Cancer and Mining Biomarkers from Gene Expression Profiles Using Hybrid Optimization Algorithms and Fuzzy Support Vector Machine [22]	An improved classification system that developed a fuzzy support vector machine classifier	ALL/MLL	100
			Colon	96.67± 7.03
			Breast	98

V. CONCLUSIONS

In our daily life, we are constantly surrounded with different types of big data. Data are becoming larger not only in respect of the number of samples but also the huge number of features. High dimensionality has been reasoned one of the complications of Big Data. This situation has made data mining jobs such as data classification suffering from lower performance and higher computational time issues. Moreover, it becomes challenging to employ feature selection on high dimensionality data or big data. This case is particularly acute in bioinformatics, especially with DNA microarray analysis. In this paper, we outlined the classification techniques applied to the high dimensionality data and made a comparison of the most popular classification algorithms.

REFERENCES

[1] H. Abusamra, "A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data", Master thesis, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia, 2013.

[2] J. Song, K. T. Kim, B. Lee, S. Kim, H. Y. Youn, "A novel classification approach based on Naïve Bayes for twitter sentiment analysis", KSII Transactions on Internet and Information Systems, vol. 11, issue 6, pages 2996-3011, 2017.

[3] K. Yamunadevi, R. Nagaraj, "An Optimized Classification of Human Cancer Disease for Gene Expression Data", International Journal of Advance Research, Ideas and Innovations in Technology, Vol. 4, Issue 2, pages 8-15, 2018.

[4] B. A. Garro, K. Rodríguez, R. A. Vázquez, "Classification of DNA microarrays using artificial neural networks and ABC algorithm", Applied Soft Computing, vol. 38, pages 548-560, 2016.

[5] M. Khashei, A. Z. Hamadani, M. Bijari, "A fuzzy intelligent approach to the classification problem in gene expression data analysis", Knowledge-Based Systems, vol. 27, pages 465-474, 2012.

[6] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter,V. d. Schaetzen, R. Duque, H. Bersini, A. Nowe, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, issue 4, pages 1106 – 1119, 2012.

[7] V.N. Vapnik, "Statistical learning theory", Wiley, 1998.

[8] B. Kalantar, B. Pradhan, S. A. Naghibi, A. Motevalli, S. Mansor, "Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN)", Geomatics, Natural Hazards and Risk, vol. 9, issue 1, pages 49-69, 2018.

[9] H. Ishida, Y. Oishi, K. Morita, K. Moriwaki, T. Y. Nakajima, "Development of a support vector machine based cloud detection method for MODIS with the adjustability to various conditions", Remote Sensing of Environment, vol. 205, pages 390-407, 2018.

[10] A. A. Aburomman, M. B. I. Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system", Applied Soft Computing, vol. 38, pages 360-372, 2016.

[11] J. WU, Z. CAI, "Attribute weighting via differential evolution algorithm for attribute weighted naive bayes (WNB)", Journal of Computational Information Systems, vol. 7, issue 5, pages 1672-1679, 2011.

[12] E. B. Hunt, J. Marin, P. J. Stone, "Experiments in induction", New York: Academic Press, 1996.

[13] M. Ture, F. Tokatli, I. Kurt, "Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients", Expert Systems with Applications, vol. 36, issue 2, pages 2017-2026, 2009.

[14] M. Dettling, P. Buhlmann, "Boosting for tumor classification with gene expression Data", Bioinformatics, vol. 19, issue 9, pages 1061-1069, 2003.

[15] A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, D. Geman, "Simple decision rules for classifying human cancers from gene expression profiles", Bioinformatics, vol. 21, issue 20, pages 3896-3904, 2005.

[16] R. Aguilar, R. Zurita-Milla, E. Izquierdo-Verdiguier, R. de By, "A Cloud-Based Multi-Temporal Ensemble Classifier to Map Smallholder Farming Systems", Remote sensing, vol.10, issue 5, 2018.

[17] X. Lei, Y. Chen, "Multiclass classification of microarray data samples with Flexible Neural Tree", Engineering and Technology (S-CET), 2012 Spring Congress on. IEEE, pages 1-4, 2012.

[18] E. Lotfi, A. Keshavarz, "Gene expression microarray classification using PCA–BEL", Computers in Biology and Medicine, Vol. 54, pages 180-187, 2014.

[19] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, "Hidden Markov models for cancer classification using gene expression profiles", Information Sciences, Vol. 316, pages 293-307, 2015.

[20] J.M. Sonawane, S.D. Gaikwad, G. Prakash, "Microarray Data Classification Using Dual Tree M-Band Wavelet Features", International Journal of Advances in Signal and Image Sciences, Vol. 3, Issue 1, pages 19-24, 2017.

[21] S. Parvathavardhini, S. Manju, "Cancer Gene Detection Using Neuro Fuzzy Classification Algorithm", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol. 3, Issue 3,pages 1223-1229, 2018.

[22] N.Y. Moteghaed, K. Maghooli, M. Garshasbi, "Improving Classification of Cancer and Mining Biomarkers from Gene Expression Profiles Using Hybrid Optimization Algorithms and Fuzzy Support Vector Machine", Journal of medical signals and sensors, Vol. 8, Issue 1, pages 1-11, 2018.

© 2018, IJCSMC All Rights Reserved