Network Kernel SVM for Microarray Classification and Gene Sets Selection

Bing Yang, Junyan Tan, Naiyang Deng, Ling Jing*
Department of Applied Mathematics
College of Science, China Agricultural University
100083, Beijing, P.R. China

***Abstract*—The importance of network-based approach to identifying biological markers has been increasingly recognized. Lots of papers indicated that genes in a network tend to function together in biological processes, so taking full advantage of the biological observation can improve the performance of microarray classification. However, lots of SVM methods don't consider this situation during their classifier building. The main idea of this paper intends to embed the information of gene networks into a new SVM learning framework. Based on a new regularization, we propose a novel method, Network Kernel SVM (NK-SVM), for binary classification problem and gene sets selection. By constructing some special kernel matrixes from the prior information of gene network, the new NK-SVM method makes the genes in the same set to be selected (or eliminated) together. The numerical experiments on a real microarray application show that the proposed method tends to provide a better performance than other methods on gene sets selection.**

Keywords- Support vector machine, Feature selectio, Gene networ, Gene expressio, Network regularization

I. INTRODUCTION

Lots of studies proved that the microarray technology is a powerful tool for biological and medical research [5]. We can get plenty of valued information by detecting thousands of gene expression levels simultaneously. However, microarray datasets usually contain only a small number of samples. It poses great challenges for sample classification and gene selection. Studies that seek to identify gene markers to refine diagnostic classification and improve prognostic prediction on gene expression data have enriched the literatures [1,4,21,23]. Biological observations show that genes in a network tend to function together in biological processes, researchers recently realize that a possibly more effective means to resolve this problem is to employ a network-based approach, that is, to identify informative gene markers as gene subnetworks, defined as sets of functionally related genes based on the gene network, instead of treating individual gene as completely independent and identical a priori as in most existing approaches [6,10,14,19,20,24]. It has been shown that such network-based approaches not only improve predictive performances, but also put biological insights into molecular mechanisms underlying the clinical outcome.

The L_2-norm SVM is one of the most effective methods for microarray classification [2,3,25,27]. Previous studies have proved its superior ability in terms of classification accuracy. As an important task in machine learning, the researchers often focus on how to identify a subset of features, which contribute the most to classification. The L_1-norm SVM [4,21,26], a version of the standard SVM, makes a good performance in gene selection. Previous researches proved that by training the L_1-norm SVM, we can get a more sparse gene markers than by the L_2-norm SVM. Nevertheless, both L_2-norm and L_1-norm SVM don't take gene networks into account. And the L_1-norm SVM has a drawbacks: when there are several highly correlated genes, the L_1-norm SVM tends to pick only a few of them, and remove the rest. In fact, in the view of biologist, genes in the same set often mean they are highly correlated, so they should selected or eliminated together.

Zou and Yuan [7] applied the concept of grouped variable selection and developed an F_∞-norm penalized SVM to realize simultaneous selection or elimination of all the features derived from the same categorical factor (or a group of variables). Their numerical examples showed that the F_∞-norm SVM outperformed the L_1-SVM in factor-wise variable selection. Reference [7] extended the idea of variable grouping to gene networks: rather than grouping all the dummy variables created from the same categorical factor, they treated two neighboring genes in a network as one group. The network-based penalty is constructed as the sum of the F_∞-norms being applied to the groups of neighboring-gene pairs. However the F_∞-norm SVM is only useful in simply grouped gene selection where one gene must be only in one gene set, but actually, most of gene networks are complex, here the gene sets overlap, a gene might be in different grouped gene sets. This means the F_∞-norm SVM is void when the gene sets have intersections [7]. The figure 1 shows the difference between the grouped gene sets and the network gene sets. In Figure 1(a), there are 17 genes that belong to 4 grouped gene sets; the different grouped sets have no same genes. In Figure 1(b), there are 15 genes that belong to 4 network gene sets; the different sets may have same genes, for example, the Set 1 and Set 2 have the same Gene 8, the Set 1 and Set 3 have the same Gene 4, and the Set 2 and Set 4 have two same genes: Gene 10 and Gene 14, but Set 3 and Set 4 have no intersections.

* Corresponding author: jingling@cau.edu.cn (L. Jing)

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

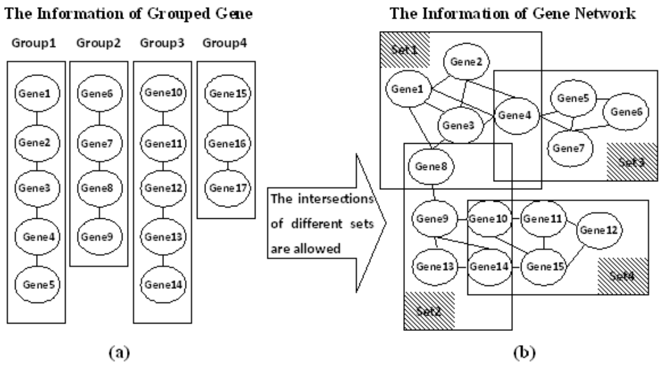

Figure1: The Grouped Genes and The Gene Network

In this paper, we propose a new Network-based kernel SVM (NK-SVM) method. We obtain our new method by building skillfully several corresponding kernels with regard to different network gene sets. The NK-SVM has several major benefits:

(1) It extends the idea of grouped gene sets to network gene sets which means the intersections of different sets are allowed.

(2) Similar to the L$_1$-norm SVM, it could select gene sets automatically.

(3) The genes in the same set can be selected or eliminate together.

The paper is organized as follows. In section 2 we give a brief introduction of the versions of SVM at first. And then we describe our new Network based Kernel SVM method (NK-SVM) and discuss its effectiveness. In section 3, we evaluate our new method's performance by simulation studies in a real world data. The last section concludes the paper and discusses the development trend of network gene sets selection.

II. Method

A. Existing Method

First of all, we briefly introduce some versions of SVM including the standard SVM, L$_1$-norm SVM and F$_\infty$-norm SVM.

In this paper, let's consider as $F=\{1,2,...,n\}$ is the set of all the gene subscripts. According to the prior information of gene networks, the n genes in gene subscripts set F are divided into several subsets:

$$F = P_1 \cup \quad \cup \quad \cup \qquad \text{, where } P_i \subset F \text{ , } i = 1,2,...,g \text{ .}$$

Generally, the standard Support Vector Machine (SVM) is L$_2$-norm SVM which is the basic formulation of SVM used for classification. For a binary classification problem, the essential idea of L$_2$-SVM is to search a linear separating hyperplane: $f(x) = w^T x + b, (w \in R^n, b \in R)$ which maximizes the distance between two classes of data:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t. \quad y_i[(w^T \cdot x_i) + b] \ge 1 - \xi_i \\ &\qquad\quad \xi_i \ge 0, i = 1, \cdots \end{aligned} \qquad (2\text{-}1)$$

where $\|\cdot\|$ denotes the 2-norm of vector, the predefined parameter C is a trade-off between training accuracy and generalization, ξ_i is the slack variable, $w \in R^n$ is a weight vector which defines a direction perpendicular to the hyperplane of the decision function, while b is a bias which moves the hyperplane parallel to itself. The decision function is presented as: $f(x) = \text{sgn}(w^T x + b)$

The above L$_2$-norm SVM forces all nonzero coefficient estimates, which leads to the problem of its inability to conduct variable selection. The L$_1$-norm SVM was proposed to accomplish the goal of variable selection. It is formulated as

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|_1^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t. \quad y_i[(w^T \cdot x_i) + b] \ge 1 - \xi_i \\ &\qquad\quad \xi_i \ge 0, i = 1, \cdots \end{aligned} \qquad (2\text{-}2)$$

where $\|\cdot\|_1$ denotes the 1-norm of vector.

The L$_1$-norm SVM wins over the L$_2$-norm SVM when the true model is sparse, while the L$_2$-norm SVM is preferred if there are not many redundant noise features [26].

Based on the SVM method, Zou and Yuan [7] proposed the F$_\infty$-norm SVM which discussed the above-mentioned grouped gene selection problem. Because the genes were divided into g groups, the n weights of hyperplane's normal vector w were divided into g groups too. The F$_\infty$-norm of *w's i-th* group weight is defined as follows:

$$\|w_{(i)}\|_\infty = \max_{j \in \{1,...,i_{n_i}\}} \{|w_j^{(i)}|\} \qquad (2\text{-}3)$$

Where $w_{(i)}$ is group P_i 's weight vector, $w_{(i)} = (w_1^{(i)},...,w_j^{(i)},...,w_{n_g}^{(i)})^T = \{[w]_j \mid j \in P_i\}$, $n_i = |P_i|$.

Then we can obtain the following F$_\infty$-norm SVM:

$$\begin{aligned} &\min_{w,b,\xi} \sum_{i=1}^{g}\|w_{(i)}\|_\infty + C\sum_{i=1}^{l}\xi_i \\ &s.t. \quad y_i[(w^T \cdot x_i) + b] \ge 1 - \xi_i \\ &\qquad\quad \xi_i \ge 0, i = 1, \cdots \end{aligned} \qquad (2\text{-}4)$$

This optimizing problem is non-differentiable, because the objective function contains $\|w_{(i)}\|_\infty$. When the variable $\beta_+ \ge 0$, $\beta_- \ge 0$, and $M \ge 0$,are brought in, the problem (2-4) can be transformed into a liner programming problem:

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

$$
\begin{aligned}
&\min_{w,b,\xi} \sum_{i=1}^{g} M_i + C\sum_{i=1}^{l} \xi_i \\
s.t.\quad & y_i((\beta_+ - \beta_-)\cdot x_i + b) \ge 1-\xi_i, i=1,...,l \\
& \xi_i \ge 0, i=1,\cdots \qquad\qquad (2\text{-}5)\\
& [\beta_+]_i + [\beta_-]_i \le M_j, \forall i \in P_j, j=1,2,...,g \\
& [\beta_+]_i \ge 0, [\beta_-]_i \ge 0, i=1,2,...,n
\end{aligned}
$$

Solving this linear programming problem can get the optimum solution $(\beta_+^*, \beta_-^*, b^*, \xi^*)$. On this basis, we can get the primary problem' s optimum solution $w^* = \beta_+^* - \beta_-^*$. The numerical experiments show the genes in the same group can get similar weights by solving F$_\infty$-norm SVM [7].

B. Our Method: The Network Kernel SVM

To solve network gene sets selection problem, especially when the intersections of different gene sets are allowed, means $\exists P_i \cap$, $i \neq j$, we must take its characteristic into account, in order to improve existing methods. When a gene belongs to different sets simultaneously, its corresponding weights in classification should be constituted by several parts, each part indicates its corresponding set's contribution to classification. For example, gene *i* belongs to set P_a and P_b simultaneously, its corresponding weight $w_i = w_{ia} + w_{ib}$. It indicates gene *i*'s contribution to classification in set *a* and *b* respectively. So we propose a new regularization for gene network and gave follow optimization problem to solve network gene selection problem:

$$
\begin{aligned}
&\min_{\{f_g\},b,\xi,d} \frac{1}{2}\sum_{m=1}^{g}\left(\frac{1}{d_m}\sum_{i=1}^{n_m}[w_i^{(m)}]^2\right) + C\sum_{i=1}^{l}\xi_i \\
s.t.\quad & y_i \sum_{m=1}^{g}\sum_{i=1}^{n_m} w_i^{(m)} \cdot x_{[P_m]_i} + y_i b \ge 1-\xi_i \qquad (2\text{-}6)\\
& \xi_i \ge 0, \quad i=1,\cdots \\
& \sum_{m=1}^{g} d_m = 1, \quad d_m \ge 0 \quad (m=1,2,...,g)
\end{aligned}
$$

Where $n_m = |P_m|$, $\sum_{i=1}^{n_m}[w_i^{(m)}]^2$ is a regularization which represents the maximal margin. d_m is a parameter which to some extent, measures the contribution of set P_m to classification. These constraints, $\sum_{m=1}^{g} d_m = 1, \quad d_m \ge 0 \quad \forall m$, can ensure the solution we get would be sparse, like the L$_1$-norm SVM.

However there isn't an effective way to solve the problem (2-6). Therefore we consider its dual problem:

$$
\begin{aligned}
&\min_{d} \ \max_{\alpha} \quad -\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \sum_{m=1}^{g} d_m K_m + \sum_{i=1}^{l}\alpha_i \\
& with \quad \sum_{i=1}^{l}\alpha_i y_i = 0 \qquad\qquad (2\text{-}7)\\
& d_m \ge 0, \sum_{m=1}^{g} d_m = 1 \\
& C \ge \alpha_i \ge 0 \quad \forall i
\end{aligned}
$$

Where K_m is a kernel matrix produced from set P_m . For two samples *i* and *j*, $[K_m]_{ij} = \sum_{l \in P_m}[x_i]_l \cdot [x_j]_l$.

This optimization problem is named Network Kernel SVM, because the key of Network Kernel SVM is the several kernel matrixes K_m . Different kernel matrix contains the information of different gene set which form the gene network. And this problem is a multiple kernel learning problem. An effective method to solve this kind of optimization problem is proposed by Alain Rakotomamonjy et al. in 2008 [9].

The figure2 shows the flow-process diagram of NK-SVM.

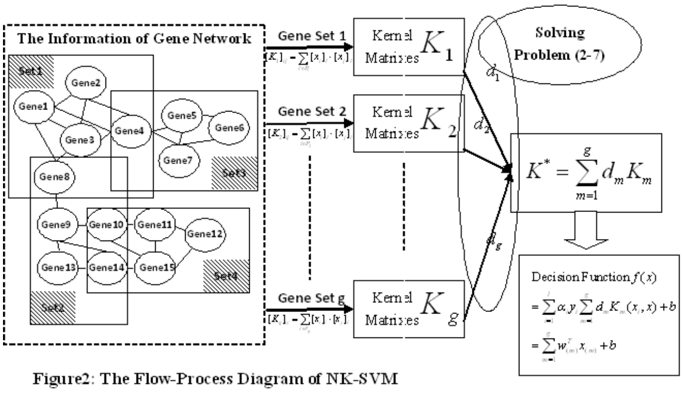

Figure2: The Flow-Process Diagram of NK-SVM

Solving the NK-SVM programming (2-7), we can get the optimum solution (α^*, d^*) . On this basis we can get the solution of the primary problem (2-6), for gene t, its corresponding weight $w_t = \sum_{i=1}^{l}\alpha_i y_i \sum_{m=1}^{g} d_m x_{it} \cdot u(t,m)$, where $u(t,m) = \begin{cases} 1 & ,t \in P_m \\ 0 & ,t \notin P_m \end{cases}$.

Therefore $w_t^{(m)} = \sum_{i=1}^{l}\alpha_i y_i d_m x_{it} \cdot u(t,m)$ expresses gene *t*'s weight in set P_m . Besides set P_m 's weight vector, $w_{(m)} = (w_1^{(m)}, ..., w_j^{(m)}, ..., w_{n_m}^{(m)})^T = \{[w]_j \mid j \in P_m\}$ is easy to be obtained.

$$
f(x) = \sum_{m=1}^{g} w_{(m)}^T x_{(m)} + b = \sum_{i=1}^{l}\alpha_i y_i \sum_{m=1}^{g} d_m K_m(x_i, x) + b
$$

is decision function. It shows we can select genes and its set by using $w_{(m)}$ or d_m under different conditions or needs. We can observe each gene's weight by using $w_{(m)}$. And d_m expresses the set's contribution to classification. When we consider network gene sets selection, d_m may be a better choice.

The NK-SVM algorithm is summarized as follow:

```
Algorithm1  NK-SVM Algorithm
Input: Training data $T = \{(x_1, y_1), ..., (x_l, y_l)\} \in (X \times Y)^l, C$
       Gene network $G = \{P_1, P_2, ..., P_g\}$. $\varepsilon$ is a cutoff given by users
Step1: Construct g kernel matrixes, $[K_m]_{ij} = \sum_{l \in P_m} [x_i]_l \cdot [x_j]_l$ , $m = 1, 2, ..., g$;
Step2: Let K as $K = \sum_{m=1}^{g} d_m K_m$ ,solve the problem(2-7) to obtain $d_m, \alpha, b$
Step3: Select informative gene sets by finding $d_m \geq \varepsilon$ ,
       Construct decision function:
           $f(x) = \sum_{i=1}^{l} \alpha_i y_i \sum_{m=1}^{g} d_m K_m(x_i, x) + b$
           for new instances classification
```

III. Numerical Experiments

In this section, the new method's capabilities in network gene sets selection problem will be examined. To evaluate the new method's performance in the real world, we applied it to one microarray gene expression data sets related to the colon cancer. This data was also tested in [1,12,13]. This dataset consists of 62 samples (40 colon cancer tumors and 22 normal tissues). Each sample consists of 2000 genes. We obtain the information of gene network from KEGG dataset [8]. Each pathway will be constructed one set. The genes in the same pathway are in the same set. Therefore, this problem is a network problem, because the pathways stuck to one another. For the genes which are not in any pathway, considering the correlation between each genes. We make the highly correlated genes (Pearson correlation coefficient $\rho > 0.8$) in one set. For rest genes, each gene formed a set independently. NK-SVM has been compared with three classical methods, the standard L_2-norm SVM, SVM-RFE and L_1-norm SVM.

The result was reported in Table 1. The result of SVM-RFE is from [1], The rest result is from our experiments. The test error was measured by averaging ten ten-fold cross validation runs. The parameter *C* was chosen from the set {2^i |i =−5, . . . , 4} by 10-fold cross-validation on each training fold. The numbers of selected genes and set is gained by the same method but training in entire data. The numbers in the parentheses are the corresponding standard errors.

Table 1 Results on the original real world datasets: the colon cancer dataset

Method	Test Error(%)	Numbers of Genes
L_2-norm SVM	14.52% (±1.32%)	All
SVM-RFE	17.74% (±0.87%)	128
L_1-norm SVM	16.13% (±1.02%)	15
NK-SVM	14.52% (±0.36%)	43

The result shows NK-SVM is available in real world data. It got higher predication accuracy in colon cancer dataset. At the same time, NK-SVM uses the information of gene network sufficiently and can gave us a suggestion in identifying the informative gene and its set. The genes and its sets which selected by NK-SVM, are gave in Table 2.

Table 2 The selected genes and its set (pathway id) by the NK-SVM from the colon cancer dataset

Gene Sets	Gene number	Pathway id
1	245,249,267,765	highly correlated
2	49,70,116,346,432,526,632, 792,952,982,986,1033,1173, 1474,1830,1831,1968	hsa 04514
3	26,94,190,271,783, 1194,1235,1923,1942	hsa 04670
4	1771,1772	highly correlated
5	642,823,1224,1530,1679,1730	hsa 00500
6	144,679,914,1023,1326,1620	hsa 03420

IV. Conclusion

At present, researchers are going deeper and deeper in analyzing the classification and genes selection problem. Some researchers have proposed analyzing results in different situations. As usual priori information, gene networks are very common in real world problems. So to speak that the network-based gene selection method make a preferably description about the basis of these problems which have such priori information. It can be predicted that researchers would keep interest in it for a while. Nowadays, the research about network gene sets selection is on the early stage of development. Based on the new network regularization, we propose a novel method, Network Kernel

SVM (NK-SVM), for gene classification and gene sets selection problem, and obtain some significative conclusions. Meanwhile our numerical experiments on the real data indicate that the proposed NK-SVM method is able to identify informative genes with the information of gene network, and make accurate predictions. But as an iterative algorithm, NK-SVM has its weakness. We must spend plenty of time to solve the complex optimization problem. Therefore, other methods which based on SVM are still to be further investigated.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (10971223, 11071252), Chinese Universities Scientific Fund (2012YJ130, 2011JS039).

References

[1] L. Wang, J. Zhu and H. Zou, "Hybrid huberized support vector machines for microarray classification and gene selection." Bioinformatics. vol. 24, pp. 412–419, 2008

[2] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[3] B. Scholkopf, and A. Smola, Learning with Kernels–Support Vector Machines, Regularization, Optimization and Beyond, MIT Press, Cambridge, 2002.

[4] P. Bradley, and O. Mangasarian, "Feature selection via concave minimization and support vector machines." In Proceedings of the 15th International Conference on Machine Learning, 1998.

[5] I. Guyon, et al. "Gene selection for cancer classification using support vector machines. " Machine Learning, vol. 46, pp. 389–422, 2002.

[6] H.Y. Chuang, E.J. Lee, Y.T. Liu, D.H. Lee, T. Ideker, "Network-based classification of breast cancer metastasis." Mol Syst Biol vol. 3, pp. 140, 2007.

[7] H. Zou, M. Yuan, "The F_∞-norm Support Vector Machine." Stat. Sin., vol. 18, pp. 379-398, 2008.

[8] KEGG: Colon Cancer, http://www.kegg.com

[9] A. Rakotomamonjy, F.R. Bach, S. Canu, Y. Grandvalet, "SimpleMKL." Journal of Machine Learning Research, vol. 9, pp. 2491-2521, 2008.

[10] Y. Zhu, X. Shen, W. Pan, "Network-based support vector machine for classification of microarray samples." BMC Bioinformatics, 10(Suppl 1):S21, 2009

[11] C. Cortes, V. Vapnik, "Support-vector networks." Machine Learning, vol. 20, pp.273-297, 1995.

[12] T. Yang, "The simple classification of multiple cancer types using a small number of significant genes." Mol Diagn Ther, vol. 11, pp.265-275, 2007.

[13] L. Ein-Dor, O. Zuk, E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer." Proc Natl Acad Sci USA,vol. 103, pp. 5923-5928, 2006.

[14] M. Liu, A. Liberzon, S. Kong, W. Lai, P. Park, I. Kohane, S. Kasif, "Network-based analysis of affected biological processes in type 2 diabetes models." PLoS Genet vol. 3, pp. 96, 2007.

[15] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Ares, D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines." Proc Natl Acad Sci USA, vol. 97, pp.262-267, 2000.

[16] T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer, D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics vol. 16, pp. 906-914, 2000.

[17] G. Wahba, Y. Lin, H. Zhang, "GACV for support vector machines." Advances in Large Margin Classifiers, Edited by: A. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans, Cambridge, MA, MIT Press, pp. 297-311, 2000.

[18] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, New York, Springer; 2001.

[19] C. Li, H. Li, "Network-constrained regularization and variable selection for analysis of genomic data." Bioinformatics, vol. 24, pp. 1175-1182, 2008.

[20] H. Zou, T. Hastie, "Regularization and variable selection via the elastic net." J R Statist Soc B, vol. 67, pp. 301-320, 2005.

[21] L. Wang, J. Zhu, H. Zou, "The doubly regularized support vector machine." Stat Sin, vol. 16, pp. 589-615, 2006.

[22] P. Bradley, and O. Mangasarian, "Feature selection via concave minimization and support vector machines." In Proceedings of the 15th International Conference on Machine Learning, 1998.

[23] C. Burges, "A tutorial on support vector machines for pattern recognition." Data Mining Knowl. Discov., vol. 2, pp. 121–167, 1998.

[24] T. Evgeniou, et al. "Regularization networks and support vector machines." Edited by: A. Smola, et al., Advances in Large Margin Classifiers. MIT Press, 1999.

[25] S. Mukherjee, et al. "Support vector machine classification of microarray data." Technical report, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 2000.

[26] J. Zhu, et al. "1-norm SVMs." In Proceedings of the Neural Information Processing Systems, 2004.

[27] Naiyang Deng, Yingjie Tian, Chunhua Zhang. Support Vector Machines: Theory, Algorithms, and Extensions, CRC Press, 2012.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE