J Med Syst (2010) 34:727–733
DOI 10.1007/s10916-009-9287-4

ORIGINAL PAPER

A Multi-Voting Enhancement for Newborn Screening Healthcare Information System

Sung-Huai Hsieh · Po-Hsun Cheng · Chi-Huang Chen · Kuo-Hsuan Huang · Po-Hao Chen · Yung-Ching Weng · Sheau-Ling Hsieh · Feipei Lai

Received: 20 February 2009 / Accepted: 25 March 2009 / Published online: 6 May 2009
© Springer Science + Business Media, LLC 2009

Abstract The clinical symptoms of metabolic disorders during neonatal period are often not apparent. If not treated early, irreversible damages such as mental retardation may occur, even death. Therefore, practicing newborn screening is essential, imperative to prevent neonatal from these damages. In the paper, we establish a newborn screening model that utilizes Support Vector Machines (SVM) techniques and enhancements to evaluate, interpret the Methylmalonic Acidemia (MMA) metabolic disorders. The model encompasses the Feature Selections, Grid Search, Cross Validations as well as multi model Voting Mechanism. In the model, the predicting accuracy, sensitivity and specificity of MMA can be improved dramatically. The model will be able to apply to other metabolic diseases as well.

Keywords Newborn screening · Tandem mass spectrometry · Support vector machines · Methylmalonic acidemia

S.-H. Hsieh · Y.-C. Weng · S.-L. Hsieh · F. Lai
Information Systems Office, National Taiwan University Hospital, Taipei, Taiwan

S.-H. Hsieh · P.-H. Chen · Y.-C. Weng · F. Lai
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan

C.-H. Chen · F. Lai
Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan

S.-L. Hsieh
Network and Computer Centre, National Chiao Tung University, Hsin Chu, Taiwan

F. Lai
Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei City, Taiwan

P.-H. Cheng (✉)
Department of Software Engineering, National Kaohsiung Normal University, Taipei, Taiwan
e-mail: cph@nknu.edu.tw

K.-H. Huang
Department of Computer Science and Engineering, Tatung University, Taipei, Taiwan

Introduction

Tandem Mass Spectrometry (MS/MS) has been used for years to identify and measure inborn errors of metabolism [1, 2]. In National Taiwan University Hospital (NTUH), Taiwan, MS/MS is used to quantify concentrations of up to 35 metabolites simultaneously from a single blood spot. It leads high dimension data of each newborn. The primary markers are selected by clinical experience and then tested against patient and control groups to maximize diagnostic accuracy. According to the most important feature, cut-off value is applied to be a threshold of concentration value of metabolism [3]. The newborn's data, which exceed the cut-off values in the primary tests, are submitted to the confirmatory tests with the same cards. If the result is successively positive, the case will be diagnosed of suspect positive. For some diseases, it has already been demonstrated that sensitivity and specificity can be improved by taking into account combinational features rather than single feature alone. For example, for methylmalonic acidemia (MMA) disease, the features of C3 and C4DC have been proven useful.

Machine learning has been widely and successfully applied to many real world classification problems such as text categorization, face detection, protein secondary structure prediction, etc. In the most cases, the generalization performance of machine learning is outstanding. The basic idea of machine learning, for example, a Support Vector Machine (SVM) which is a classifier that the set of binary labeled training data vectors can be separated by a hyperplane. In the simplest case of a linear hyperplane there may exist many possible separating hyperplanes. There are many possible linear classifiers that can separate the data, but among them, the SVM classifier seeks the separating hyperplane that produces the largest separation margin (maximizes the distance between it and the nearest data point of each class). This linear classifier is termed to the optimal separating hyperplane. The hyperplane with maximal margin is the ultimate learning goal in statistical learning theory, and will probably perform well in classifying the new data.

Finding good features of the data for the learning algorithms [10], in the form of suitable features, is known to have high influence on the performance of the resulting classifiers. This task is addressed by feature selection and feature cross validation techniques, which are directed at optimizing the data representation for subsequent data-driven learning algorithms.

In the following sections of the paper, we first elaborate the method of our proposed adaptive feature selection and SVM. The experiment results as well as comparison with original method in NTUH and published results will be illustrated in 'Experimental results'. Finally, the paper concludes in 'Conclusion'.

Methodology

We propose a proper supervised classification [4] data flow to enhance the accuracy and sensitivity of the Newborn Screening process, as depicted in Fig. 1. In the diagram, the Train Dataset undergoes learning to produce the SVM prediction model; the New Dataset processes the same methods to obtain the prediction result according to the trained model. Before training or predicting, the dataset is preprocessed by the MS/MS machine. The Feature Selection part will generate the most relevant features by a Pearson-like formula [5]. The Scaling method is used to avoid biasing and to improve computing efficiency. The SVM machine learning with Grid Search then generates the prediction model for the New Dataset.

Preprocessing

Tandem mass spectrometry (MS/MS) has been used for years to identify and measure carnitine ester concentrations in blood and urine of children suspected of having inborn errors of metabolism. MS/MS permits very rapid, sensitive and, with internal standards, accurate quantitative measurement of many different types of metabolites by conversion of raw mass spectra into clinically meaningful results.

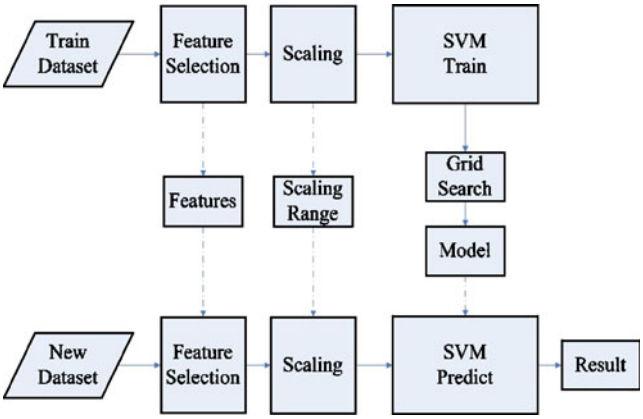


Fig. 1 Data flow model for newborn screening

Feature selection

We apply a Pearson-like Correlation coefficient for the feature selection method. After ranking the features' coefficients, we choose the most relevant features for the continued machine learning model and the new predicted datasets. For a series of n measurements of X and Y, written as x_i and y_i where $i=1\ldots, n$, the Correlation coefficient formula as the form:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \quad (1)$$

Where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, s_x and s_y are the sample standard deviations of X and Y and the sum is from $i=1$ to n.

Scaling

The range of the feature's raw data is not always the same, and can be either too large or too small. Therefore, all features must be scaled to a proper range. The training and predicting processes will be faster, and they can avoid certain large value features so as not to induce a learning model bias. The most common range is [0, 1] or [−1, 1].

SVM training

The central idea of SVM classification [8] is to use a linear separating hyperplane to create a classifier. The vector which can effect the separation is called a "support vector".

Given a training set of instance–label pairs:

$$(x_i, y_i), i = 1, \ldots, l, x_i \in R^n, y \in \{1, -1\}^l \quad (2)$$

The hyper-plane can be expressed as follows:

$$< w^T \cdot x > + b = 0 \quad (3)$$

Where

$$w = [w_1, w_2, \ldots, w_n]^T,\ x = [x_1, x_2, \ldots, x_n]^T$$

Then the definition of a decision function is

$$f(x) = \text{sign}\ \left(w^T \phi(x) + b\right) \quad (4)$$

with the largest possible margin, which apart from being an intuitive idea has been shown to provide theoretical guarantees in terms of generalization ability.

If the data is distributed in a highly nonlinear way, employing only a linear function causes many training instances to be on the wrong side of the hyperplane, which results in underfitting occurs and the decision function does not perform well. Thus SVM non-linearly transforms the original input into a higher dimensional feature space. More precisely, the training data x is mapped into a (possibly infinite) vector

$$\Phi(x) = (\phi_1(x), \phi_2(x) \cdots\cdots, \phi_i(x), \cdots)$$

In this higher dimensional space, it is more possible that data can be linearly separated. We therefore try to find a linear separating plane in a higher dimensional space.

$$\begin{aligned} \min_{w,b,\xi} \quad & \tfrac{1}{2} w^T w + C \sum_{i=l}^{l} \xi_i \\ \text{subject to} \quad & y_i (w^T x_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \quad (5)$$

The parameters ξ_i are called the slack variables, and they ensure that the problem has a solution in case the data is not linear separable. The constraints in (5) contain a penalty term, $C \cdot \sum_{i=1}^{L} \xi_I$, where $C>0$, and the parameter is chosen by the user to assign a penalty to errors. Usually this problem is called a primal problem. After the data are mapped into a higher dimensional space, the number of variables (w, b) becomes very large or even infinite. We handle this difficulty by solving the dual problem [6, 7].

A Kernel for a nonlinear SVM projects the samples to a feature space of higher dimension via a nonlinear mapping function. Among the nonlinear kernels, the radial-based function (RBF) is defined as

$$K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0 \quad (6)$$

Where γ is a kernel parameter.

Parameter optimization and cross validation

There are two parameters while using RBF kernels: a penalty (C) and a gamma (γ). It is not known beforehand which C and γ are best for a given problem; consequently, some kind of model selection (parameter search) must be performed. The goal is to identify a good set of (C, γ) such that the classifier can accurately predict unknown data (i.e., testing data). Note that it may not be useful to achieve a high training accuracy (i.e., the classifiers accurately predict training data whose class labels are already known). Therefore, a common method is to separate the training data into two parts where one part is considered unknown in training the classifier. Then, the prediction accuracy on this set can more precisely reflect the performance in classifying the unknown data. An improved version of this procedure is a k-fold cross-validation.

Experimental results

In the experiment, we intend to build a robust model, as depicted in Fig. 2, to classifier the MMA cases by using the data mining system we proposed. In addition, the model will be implemented in the decision support system to help the NTUH staff of Screening Department identify the MMA cases.

Data set

During the experiment, we collected 350 samples for SVM training. The samples were divided into five parts

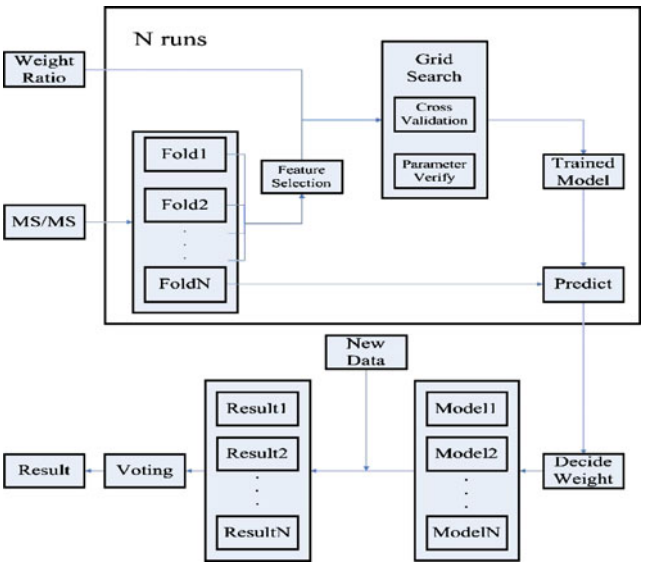


Fig. 2 Experimental processes in robust model

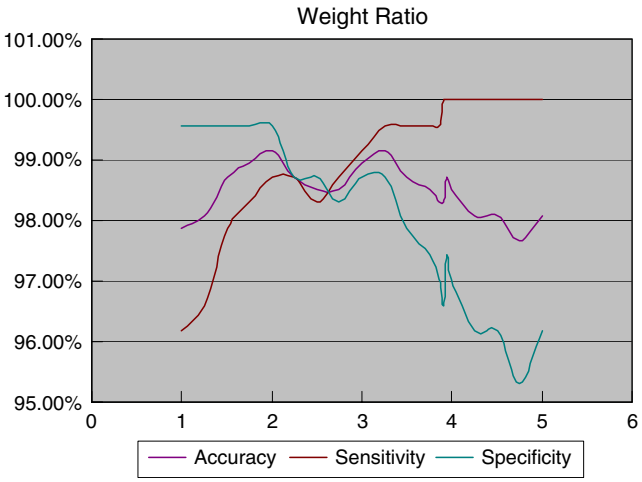


Fig. 3 Weight ratio results

equally; in a part, we randomly selected 35 sick samples and 35 normal samples, i.e., the total data comprising (35+35)×5=350 samples. We utilized the data to build five models and to tune the model weights. Afterwards, we took 60 sick samples and 60 normal samples from the origin MS/MS data set to be the New Data, and we used the new data to evaluate our voting mechanism.

Experimental steps

At beginning, we decide weight ratios [8] and adjust different misprediction penalties (Fig. 3). While a positive case being predicted as a negative one is a very serious problem which should be avoided. However, a negative case being erroneously predicted as a positive one is not imperative. Under the current NTUH Newborn Screening Diagnoses MMA rule, it allows 1.6% false positive ratios [9]. In here, we require a higher specificity and a more accurate classification model.

Secondly, we randomly separate the samples into N parts (N Folds), with each part possessing the same number of sick and normal cases (currently N=5). During a Cross Validation, we choose $N-1$ parts as the training data and the remaining part (i.e., FoldN) as the prediction data. In the Grid Search, we select the different feature number, C and γ values for iteration in order to generate the weight ratio parameter. In the Search, all parameters are using the default values initially. After the $N-1$ parts are trained to generate the Train Model, we then verify the parameters. We utilize the N-th part to be the prediction fold and obtain the weight ratio. After N iterations, N weight ratios are generated to decide the individual model weight for voting mechanism later.

In the third step, we use all models to predict a New Data set and create Results. Finally, based on the Results and the model weights, we apply the voting rule to decide whether the case is positive or negative. The voting algorithm will be illustrated later.

Weight ratio

We randomly chose 235 sick samples and 235 normal samples in order to tune the weight ratio. This was different from the data set which we used to train the models. Using an equal number of sick and normal samples can avoid biasing the learning model. We used five-part cross validation for the weight ratio result.

A cross validation was performed for these experiments. We had 235 abnormal sample data points, so we divided the samples into five parts. Each part included the screening data of the same case, and 200 randomly chosen normal cases were added to each part. One part was used for the prediction, and the others were used as the training data. (i.e., part 1 was used to predict, and parts 2–5 were used to train). We performed the experiment five times, and took the average of the results.

Since we wanted the model to identify the sick cases as best as possible, we needed to check the sensitivity, and we were not overly concerned about the accuracy or specificity. Figures 4 and 5 shows that weight ratios of 3.2 and 3.9 performed well in classifying the sick samples. In this work, we decided to use 3.9 as the weight ratio result.

Grid search with feature numbers

A grid search was conducted on the features selected as indicated in red rectangular of Fig. 4. In the diagram, by the definition of correlation, a coefficient greater than 0.8

0.707452319	F32
0.700834545	F15
0.635640171	F18
0.593935525	F13
0.552183402	F8
0.520412774	F5
0.498286848	F4
0.452625167	F10
0.428105964	F21
0.420345616	F19
0.366240882	F12
0.334937529	F29
.	.
.	.
.	.

Fig. 4 Features ranking and selection

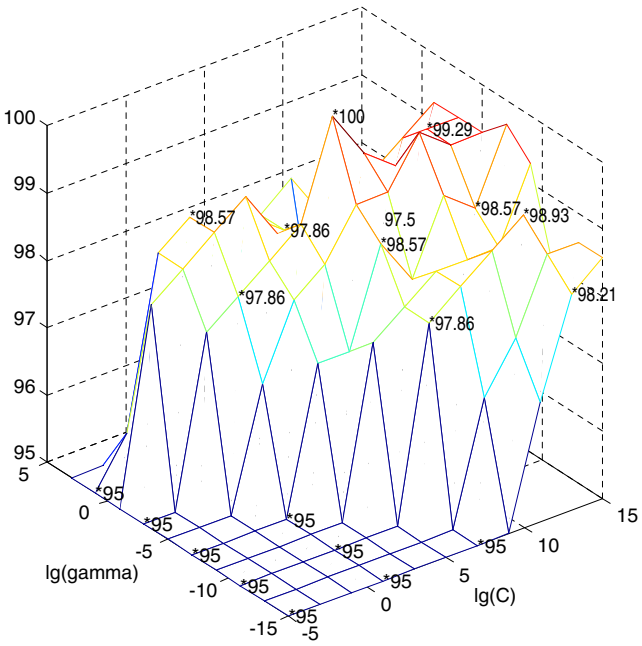


Fig. 5 The accuracy from Grid Search with five features in model1 to obtain the best Trained Model1

means strong correlation; between 0.6 to 0.8 means high correlation; and between 0.4 to 0.6 means intermediate correlation. Therefore, we set the Features Numbers Ranking threshold equal to 0.5. In the experiment, we decide to use the most relevant features to build the model. All 6 features selected are greater than 0.5.

After determined six features, we plan to find the best model in each iteration repeatedly by tuning the parameters: Feature, C and Gamma. For example, we can obtain the best Trained Model1 via the highest accuracy of Grid Search involving the top five correlation coefficient as the selected Features. In Fig. 5, it indicates the best Trained Model of mode1 having the best parameters C=7, γ=−3 and Feature Number=5. In addition, in each Grid Search, the Cross Validation is performed to prevent over fitting problem.

From Fig. 5, we can fix one axis and run through the other axis values. We can then analyze the trend of the log (C) and log(γ) for five feature numbers. We present this figure to show that the best parameter set (denoted by 100%), which is the model parameter chosen in model 1.

Figure 6 shows that in the log(C) and log(γ) space, point accuracy results above 98% are concentrated on the right side, which is the good region [11]. @(model) denotes that the point is the model1 parameters we chose, and #(default) denotes the default parameter of the model. The default parameters gave a poor result, with the accuracy lower than 90%.

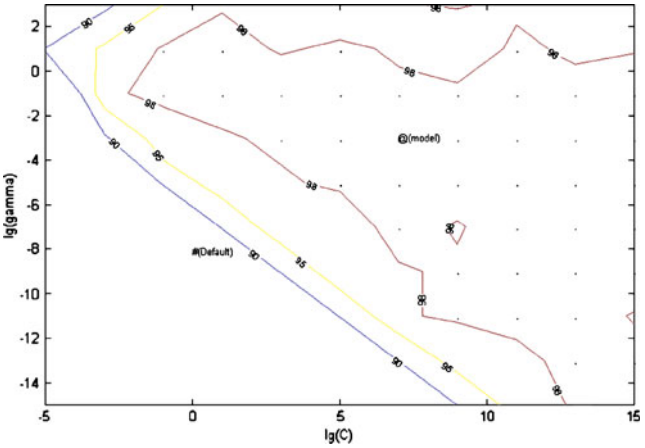


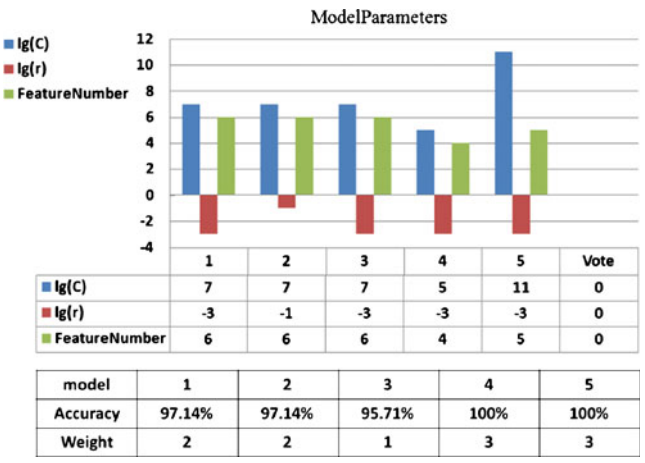
Fig. 6 Comparison of default C & γ parameters # (default) with the best Trained Mode1 ones @(model)

Model parameters and model weight

In Fig. 7, the above graph shows the results for the five models with their parameters respectively, and the table below shows the predicted results. We decided the weighting of the model by the weight results. Models 4 and 5 yielded better results, so they were given higher weights of 3. Model 3 gave the worst result, and were given the lowest weight of 1. Models 1 and 2 were between model 3 and models 4 and 5 in terms of accuracy, so the weight value was chosen to be between the weight values of model 3 and models 4 and 5.

Voting

The last step of the experiment was voting. We proposed a voting algorithm, which is implemented in our system and was introduced below.



	1	2	3	4	5	Vote
lg(C)	7	7	7	5	11	0
lg(r)	-3	-1	-3	-3	-3	0
FeatureNumber	6	6	6	4	5	0

model	1	2	3	4	5
Accuracy	97.14%	97.14%	95.71%	100%	100%
Weight	2	2	1	3	3

Fig. 7 Model parameters and model prediction results

```
Model set M = {M1, M2,...Mn};
   Wi = weight of Mi;
   Ri = predicted results of Mi;
Ss is the sick score with an initial value of 0;
Ns is the normal score with an initial value of 0;
   For i = 1 to n
       If Ri = sick, then Ss = Ss + weight * 1;
       If Ri = normal, then Ns = Ns + weight * 1;
   End of for
       If Ss >=Ns then vote result of Q = sick
       else vote result of Q = normal;
```

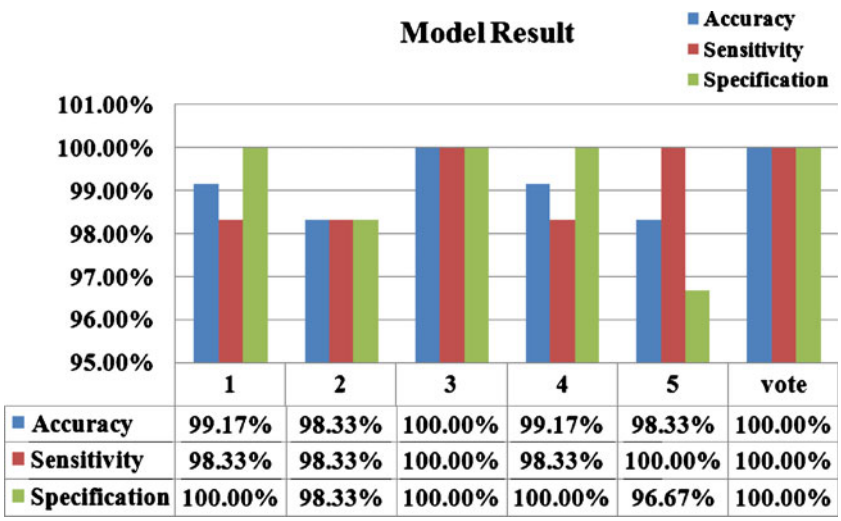
Consider that, if we want to predict a sample Q, the model sets M includes {M1, M2,…Mn}. Wi means the weight of Mi, and Ri equals the predicted result values of Mi (i=1, 2…,n). Then we evaluate the Si (the score of sick score) and Ni (the score of normal case) by a for loop. Finally to determine if the predicted case is sick or normal. When the model weights are determined, we follow the algorithm to decide the voting results based on a sample having 60 sick ones and 60 normal ones.

Figure 8 show that the voting result has a better performance than the original models; 99.17% was the highest accuracy, although the specificity of the vote was poor for models 1 and 3. The sensitivities of model 1 and model 3, however, were poor. The two models missed some sick samples, so that the vote system was still better. The predicted results of the five models are shown as the below part of Fig. 8.

The algorithm is only biased for sick samples when Ss= Ns, i.e., when the sick score equals to the normal score, and we judge the voting results as sick. In our experiment, there are five different models.

Conclusion

In the paper, we design and develop a model for Newborn Screening System. The model provides Support Vector Machines (SVM) classification techniques and enhancements to evaluate, interpret, and determine whether a newborn has MMA metabolic disorders. The model encompasses the Feature Selections, Grid Search, Cross Validations and multi model Voting Mechanism. In the model, the predicting accuracy, sensitivity and specificity of MMA can be improved dramatically. The model will be able to apply to other metabolic diseases as well.



	1	2	3	4	5	vote
Accuracy	99.17%	98.33%	100.00%	99.17%	98.33%	100.00%
Sensitivity	98.33%	98.33%	100.00%	98.33%	100.00%	100.00%
Specification	100.00%	98.33%	100.00%	100.00%	96.67%	100.00%

Fig. 8 The voting results with accuracy, sensitivity and specificity

References

1. Chace, D. H., Kalas, T. A., and Naylor, E. W., Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns. *Clin. Chem.* 49:1797–1817, 2003. doi:10.1373/clinchem.2003.022178.
2. Pinheiro, M., Oliveira, J. L., Santos, M. A. S., Rocha, H., Cardoso, M. L., and Vilarinho, L., NeoScreen: A software application for MS/MS newborn screening analysis. In Biological and Medical Data Analysis (ISBMDA'2004), Lecture Notes in Computer Science—Volume 3337, Barcelona, Spain, 2004.
3. Chien-Ming Tu, Hsing-Yu Chang, Mei-Yu Tang, Faipei Lai, et al., The design and implementation of a next generation information system for newborn screening. HEALTHCOM 2007, June, 2007.
4. Donald Michie, D. J. Spiegelhalter, C. C. Taylor, J. C. (eds), Machine learning, neural and statistical classification. 1995.
5. Forthofer, N., Lee, E. S., and Hernandez, M., Biostatistics, 2nd Edn: A Guide to Design, Analysis and Discovery. 2006.
6. Cortes, C., and Vapnik, V., Support-vector network. 1995.
7. Ward, J. J., McGuffin, L. J., Buxton, B. F., and Jones, D. T., Secondary structure prediction with support vector machine. *Bioinformatics*. 19:1650–1655, 2003. doi:10.1093/bioinformatics/ btg223.
8. Chen, P. H., Fan, R. E., and Lin, C. J., A study on SMO-type decomposition methods for support vector machines. January 2005.
9. Chien-Ming Tu: The New Generation of Information System for Newborn Screening—A Case Study of National Taiwan University Hospital. Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan, Master Thesis, June, 2007.
10. Baumgartner, C., Böhm, C., and Baumgartner, D., Modelling of classification rules on metabolic patterns including machine learning and expert knowledge. *J. Biomed. Inform.* 38 (2)89–98, 2005. doi:10.1016/j.jbi.2004.08.009.
11. S. Sathiya Keerthi and Chih-Jen Lin, Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput.* 15:1667–1689, 2003.