Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

Integration of clinical and microarray data with kernel methods

Anneleen Daemen, Olivier Gevaert and Bart De Moor

Abstract— **Currently, the clinical management of cancer is based on empirical data from the literature (clinical studies) or based on the expertise of the clinician. Recently microarray technology emerged and it has the potential to revolutionize the clinical management of cancer and other diseases. A microarray allows to measure the expression levels of thousands of genes simultaneously which may reflect diagnostic or prognostic categories and sensitivity to treatment. The objective of this paper is to investigate whether clinical data, which is the basis of day-to-day clinical decision support, can be efficiently combined with microarray data, which has yet to prove its potential to deliver patient tailored therapy, using Least Squares Support Vector Machines.**

I. INTRODUCTION

Since the rise of microarray technology approximately one decade ago, few microarray models have reached the clinic. Decisions concerning diagnosis, prognosis or treatment of cancer are still based on clinical data, such as patient history, laboratory analysis or ultrasound parameters. But, since cancer is thought to be caused by genetic aberrations, microarray technology has the potential to revolutionize the clinical management of cancer. This technology however has its disadvantages. Microarray data is noisy and high dimensional and publicly available microarray data sets suffer from small sample sizes. When considered together with clinical data however, possibly complementary data sources can be combined in one model.

In this paper we will investigate whether clinical data and microarray data from breast cancer patients can be efficiently combined. In most microarray studies on cancer the focus is on the microarray analysis while the clinical data is not modeled in the same manner. When integrating both heterogeneous data sources, we can take advantage of the strengths of both data sources. Previously we have investigated the use of Bayesian networks to integrate these data sources. However, this model is not tuned for classification. Therefore we investigate in this paper the use of kernel methods, more specifically Support Vector Machines (SVMs).

SVMs, introduced by Vapnik in [15], are popular for classification because, contrary to most other classification methods, they can handle high dimensional data such as microarray data. An SVM maps data into a vector space where it determines a linear discriminant boundary with maximum distance between members of the positive and the negative class. This corresponds to a non-linear separation hyperplane in the space of the original data. A modified version of the Vapnik SVM classifier formulation was introduced by Suykens *et al.* in [10] as the Least Squares Support Vector Machine (LS-SVM) which is much faster on microarray data than SVM. We will use LS-SVMs in combination with a publicly available breast cancer data set containing both clinical and microarray data. The developed models will be evaluated using prediction accuracy, sensitivity, specificity and Receiver Operator Characteristics (ROC) curves. The best model will also be compared to conventional prognostic markers.

All authors are with ESAT, Department of Electrical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium {anneleen.daemen, olivier.gevaert, bart.demoor}@esat.kuleuven.be

II. METHOD

A. Materials

The data set used in this study consists of 295 breast cancer patients from the Netherlands Cancer Institute [14]. This group of patients can be divided into 2 classes according to the appearance of distant subclinical metastases based on the primary tumour: 88 patients with and 207 patients without distant metastases. All patients were younger than 53 years old at diagnosis. The tumours, all smaller than 5 cm, were primary invasive breast carcinoma. The microarray data set contained 24188 gene expression values and was already normalized and background corrected. Missing values were estimated using K-nearest neighbours with K=15 [13].

The clinical data contained the following 13 variables [2][14]: diameter, T ($\leq$2cm or $>$2cm), N (pN0, 1-3, $\geq$4), number of positive lymph nodes, mastectomy (yes or no), estrogen receptor (positive or negative), grade (poorly differentiated, intermediate or well differentiated), age, chemotherapy (yes or no), hormonal therapy (yes or no), St. Gallen criteria (chemotherapy or no chemotherapy), National Institutes of Health (NIH) consensus criteria (chemotherapy or no chemotherapy) and NIH risk (low, intermediate or high). These variables were all known for the 295 patients.

The complete data set (clinical and microarray data) was divided into a training set (44 with and 104 without distant metastases) to develop the models and an independent validation set (44 with and 103 without distant metastases) to assess the performance of the models on data not used for training.

B. Kernel methods and LS-SVMs

Kernel methods are a group of algorithms that, instead of representing data entities by their properties, represent data entities through a set of pairwise comparisons called the kernel matrix. In this manner the representation of the data is independent of the nature of the data allowing integration of heterogeneous data in a uniform way. Moreover the size of the matrix is determined only by the number of data

entities that are modeled. For example a set of 100 patients each characterized by 24188 gene expression values is still represented by a 100 x 100 matrix [9]. This kernel matrix can be geometrically expressed as a mapping into a high dimensional feature space. However, there is no need to have an explicit representation of the mapping function $\Phi(x)$ in the feature space for each data point x. The transformation of two data points x_k and x_l is defined implicitly by their inner product, $\langle \Phi(x_k), \Phi(x_l) \rangle$ via a kernel function $K(x_k, x_l)$. Any symmetric, positive semidefinite function is a valid kernel function, resulting in many possible kernels, e.g. linear, polynomial and diffusion kernels. They all correspond to a different transformation of the data, meaning that they extract a specific type of information from the data set. Therefore, the kernel representation can be applied to many different types of data and is not limited to vectorial form.

A supervised classification algorithm belonging to the kernel methods is the Support Vector Machine (SVM) developed by Vapnik and others [15]. The SVM forms a linear discriminant boundary in feature space with maximum distance between samples of the two considered classes. This corresponds to a non-linear discriminant function in the original input space. A modified version of SVM, the Least Squares Support Vector Machine (LS-SVM), was developed by Suykens *et al.* [10][11]. For classification this modification leads to solving a linear system instead of a quadratic programming problem, which makes LS-SVM much faster than SVM on microarray data sets. The optimization problem in the LS-SVM formulation with the corresponding dual problem are described in [5]. In the next section we describe the use of LS-SVMs with a normalized linear kernel to predict the prognosis of breast cancer patients based on clinical and microarray data.

C. Data fusion

Three ways exist to learn simultaneously from multiple data sources with kernel methods: early integration, intermediate integration and late integration [7]. Fig. 1 gives an overview of these three methods.

- *Early integration*: the clinical and microarray data set are considered as one big data set. An LS-SVM is trained directly on the single kernel computed for the concatenated data set.
- *Intermediate integration*: A kernel is computed for each data set separately. An LS-SVM is trained on the explicitly heterogeneous kernel function (as the (weighted) sum of the separate kernels).
- *Late integration*: For each data set separately, a kernel is computed and an LS-SVM is trained. The outcomes of the multiple models are combined with a decision function to become a single outcome.

In this paper, the combination of both the clinical and microarray data is done with intermediate integration because this type of data fusion seemed to perform better than early and late integration [7]. Intermediate integration has the advantage that the nature of each data set is taken into account when compared to early integration. The separate

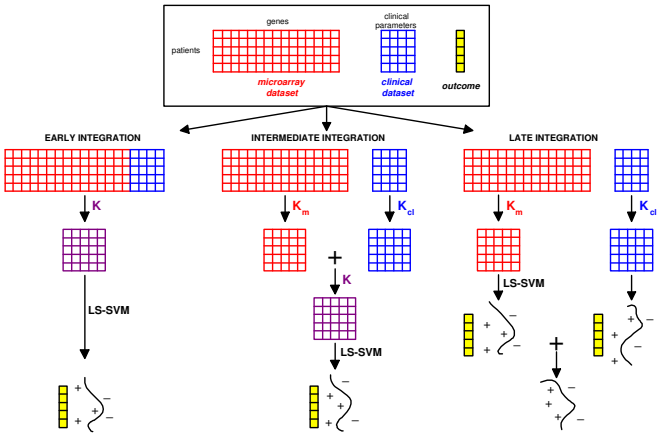

Fig. 1. *Three methods to learn from multiple data sources.* In early integration, an LS-SVM is trained on the kernel matrix, computed from the concatenated data set. In intermediate integration, a kernel matrix is computed for both data sets and an LS-SVM is trained on the sum of the kernel matrices. In late integration, two LS-SVMs are trained separately for each data set. A decision function results in a single outcome for each patient.

kernel functions can be better adapted to each of the data sets. The polynomial relations between inputs in the same data sets (clinical parameters in the clinical data set, genes in the microarray data set) can be modeled more accurately for the construction of the decision boundary. A disadvantage however is that polynomial relations between inputs from different data sources are ignored when using intermediate integration. On the other hand, when compared to late integration, intermediate integration has the advantage that a model is trained by weighing both data sources simultaneously through the use of kernels. This results into one prediction for each patient and only one hypothesis has to be formed instead of two independent hypotheses that have to be combined afterwards.

D. Model building

Because each data set is represented by a kernel matrix, data sources can be integrated in a straightforward way by adding the multiple kernel matrices according to the intermediate integration explained previously. In this combination, each of the matrices is given a specific weight μ_i. Positive semidefiniteness of a linear combination of kernel functions

$$K = \sum_{i=1}^{m} \mu_i K_i \qquad (1)$$

with m the number of data sources is guaranteed when the weights μ_i are constrained to be non-negative.

In this paper, we studied the normalized linear kernel function

$$K(x_k, x_l) = K(x_k, x_l) / \sqrt{K(x_k, x_k) K(x_l, x_l)} \qquad (2)$$

with $K(x_k, x) = x_k^T x$ instead of the linear kernel function $K(x_k, x_l) = x_k^T x_l$. The difference between both is that in a normalized linear kernel the data points are projected onto the unit sphere. With the normalized version, the values in

the kernel matrix will be bounded, while these elements can take very large values without normalization. Normalizing is thus required when combining multiple data sources. In this way, the kernel matrices of both data sets have the same order of size and the weights with which the matrices are combined can be interpreted as the relative importance of the corresponding data set.

As shown in Fig. 2(a), a leave-one-out cross-validation (LOO-CV) strategy is performed on the training data set to optimize the weights μ_i and the regularization parameter γ. In each LOO-CV iteration, the Wilcoxon rank sum test is used to rank the genes. In this manner the microarray data set is reduced to 1000 genes most significantly different between the 2 considered classes.

On this reduced data set, two parameters (the weight μ of the clinical data set and the regularization parameter γ of the LS-SVM) have to be optimized. To accomplish this, we defined a two-dimensional grid as shown in Fig. 2(a) on which the parameters are optimized by maximizing a criterion on the training set. The possible values for γ on this grid range from 10^{-10} to 10^{10} on a logarithmic scale. The weights for the kernels were optimized with a linesearch at each possible γ by increasing the weight of one kernel from 0 to 1 in steps of 0.02 and decreasing the weight of the other kernel equally.

The different models with the instantiated parameters are evaluated on the left out sample of the training set by maximizing the sensitivity with an as high as possible specificity.

This whole procedure is repeated for all samples in the training data set. The result is a fully specified classifier with as parameters the combination that has globally the highest sensitivity with an as high as possible specificity on the training data set.

Fig. 2(b) shows how the best fully specified LS-SVM is trained on the training set after selecting the 1000 most significant genes. This final LS-SVM is subsequently tested on the independent validation set.

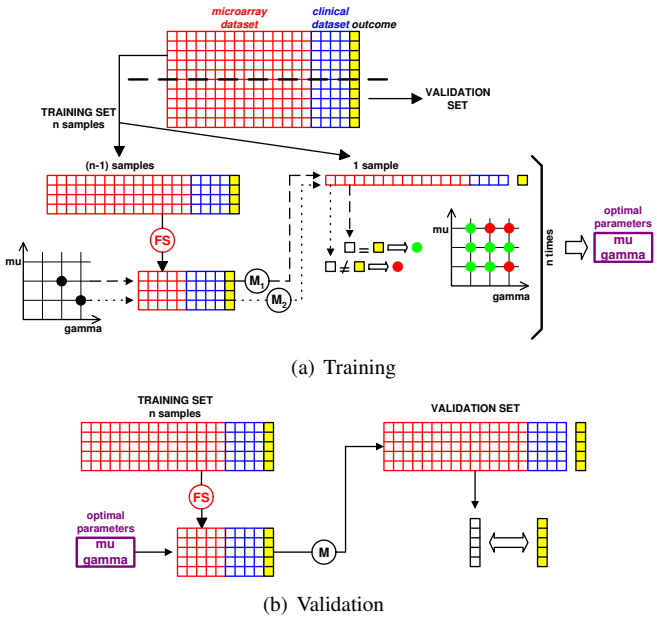

Fig. 2. *Methodology for developing and validating a classifier.* **(a)** The initial data set is divided into a training set and a validation set. The weights μ_i with which the kernel matrices are combined and the regularization parameter γ are determined with a leave-one-out cross-validation strategy on the training set. In each leave-one-out iteration, an LS-SVM model is trained on the 1000 most significant genes for all possible combinations of μ_{cl} and γ. This gives a globally best parameter combination (μ_{cl}, γ). **(b)** With the optimal parameter combination (μ_{cl}, γ), an LS-SVM is trained on the complete training set after selecting the 1000 most significant genes. This model is subsequently evaluated on the independent validation set.

III. Results

We evaluated our methodology as described in section II-D on a publicly available breast cancer data set [14]. This data set was split up into a training set (n=148) and a validation set (n=147), similarly as in [14]. The model, where each data source was represented by a normalized linear kernel, showed the highest sensitivity with an as high as possible specificity on the training set for γ equal to 5.878, giving a weight of 0.28 to the clinical data set and 0.72 to the microarray data set. From now on, we refer to this model as CMKIM (Clinical and Microarray Kernel Integration Model). For comparison, we also developed a model build only on the clinical data set, only on the microarray data set and finally a model with equal weights for both the clinical and microarray data. CMKIM is compared with the performance of these three other models on the validation set. Fig. 3 shows the ROC curves of these four models, together with the area under the ROC curve (AUC) and the standard error. In table I the four models are compared on the basis of prediction accuracy, sensitivity, specificity, positive and negative predictive value and Kappa coefficient. CMKIM had a weight of 0.28 for the clinical data and predicted the presence of distant metastases correctly in 104 of the 147 patients (=70.75%) in the validation set. It identified 36.4% of the patients who developed distant metastases while 85.4% of the patients who did not develop metastases would have been spared from exposure to some form of therapy.

Next we compared the classification of CMKIM with the following conventional prognostic markers: the St Gallen consensus [5], the National Institutes of Health (NIH) consensus [3] and the Nottingham Prognostic Index (NPI) [4]. The St Gallen and the NIH prognostics were taken from the clinical data set published in [2] as described in [5] and [3], respectively. For the NPI we took the original formula [4], although many authors have proposed changes to or completion of the NPI using additional prognostic factors [1]. We used the conventional threshold of 3.4 to distinguish between good prognosis (below the threshold) and moderate and poor prognosis (above the threshold) [12].

Table II shows the sensitivity, specificity, prediction accuracy and Kappa coefficient of CMKIM compared to the three considered indices. The St Gallen and the NIH consensus criteria have a very high sensitivity, but an untolerated low specificity which would lead to an overuse of some form of therapy. Their prediction accuracy is not better than random

TABLE I

THE PERFORMANCE ON THE VALIDATION SET OF CMKIM. THIS PERFORMANCE IS COMPARED TO THE PREDICTION QUALITY WHEN USING BOTH DATA SETS WITH EQUAL WEIGHT, WHEN ONLY USING THE CLINICAL DATA SET AND ONLY THE MICROARRAY DATA SET.

	μ_{cl}	μ_m	TP	FP	FN	TN	Sens	Spec	PPV	NPV	Acc	Kappa	γ
CMKIM	0.28	0.72	16	15	28	88	0.3636	0.8544	0.5161	0.7586	104/147 (70.75%)	0.2382	5.878
Equal weight	0.5	0.5	15	17	29	86	0.3409	0.8349	0.4687	0.7478	101/147 (68.71%)	0.1908	5.878
Clinical data	1	0	10	11	34	92	0.2273	0.8932	0.4762	0.7302	102/147 (69.39%)	0.1417	62.35
Microarray data	0	1	13	15	31	88	0.2954	0.8544	0.4643	0.7395	101/147 (68.71%)	0.1672	5.878

μ_{cl}, weight for the kernel matrix computed on the clinical data set; μ_m, weight for the kernel matrix computed on the microarray data set; TP, true positive; FP, false positive; FN, false negative; TN, true negative; Sens, sensitivity; Spec, specificity; PPV, positive predictive value; NPV, negative predictive value; Acc, predictive accuracy; Kappa, kappa coefficient; γ, regularization parameter.

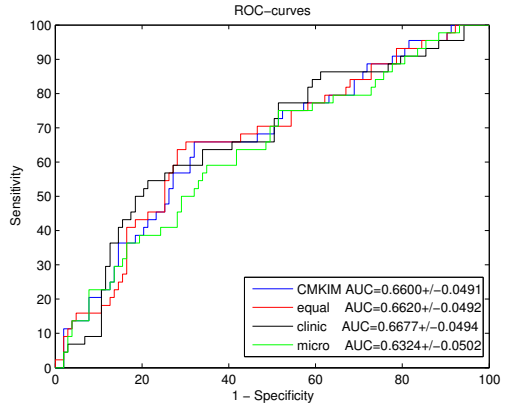

Fig. 3. ROC curves and area under the ROC curves (AUC) on the validation set for CMKIM (blue), the model with equal influence of both data sets on the outcome (red), the model build on the clinical data set (black) and the model build on the microarray data set (green).

as shown by the Kappa coefficient. The NPI has a higher specificity than the other two indices, but can only predict distant metastases correctly in less than half of the patients (47.62%). Thus integration of clinical and microarray data outperforms the pure clinically indices (St Gallen, NIH and NPI).

TABLE II

SENSITIVITY, SPECIFICITY, PREDICTION ACCURACY AND KAPPA COEFFICIENT FOR CMKIM COMPARED TO THREE CLINICAL PROGNOSTIC MARKERS ON THE VALIDATION SET.

	TP	FP	FN	TN	Sens	Spec	Acc	Kappa
○	16	15	28	88	0.3636	0.8544	70.75	0.2382
†	41	95	3	8	0.9318	0.0777	33.33	0.0059
♮	44	100	0	3	1	0.0291	31.97	0.0176
♢	36	69	8	34	0.8182	0.3301	47.62	0.1061

○ CMKIM
† St Gallen: recommends chemotherapy when one of the following criteria holds: ER negative, lymph node positive, pathological tumour size $>$ 2cm, grade poorly (3) or intermediate (2) differentiated or age $<$ 35.
♮ NIH: chemotherapy when lymph node positive or tumour size $>$ 1cm.
♢ NPI: sum of 0.2 times the tumour size in cms, the tumour grade and the lymph node stage.

IV. CONCLUSIONS AND FUTURE WORK

A. Conclusions

We have developed a framework for the integration of multiple data sources in disease management (see Fig. 2). We represent each data set with a kernel matrix, based on a normalized linear kernel function. These matrices are combined according to the intermediate integration method suggested by [7] and illustrated in Fig. 1. An LS-SVM is trained on the combined kernel matrix. This gives a step in the right direction to improve predictions for an individual patient about prognosis, metastatic phenotype and therapy response.

In this paper, we evaluated our method on the breast cancer data set in [14]. Patients included in this data set belonged to two classes according to the presence of distant subclinical metastases based on the primary tumour.

Because two parameters had to be optimized (weight for the clinical data set and regularization parameter γ), all possible combinations of these parameters were investigated with a LOO-CV strategy, illustrated in Fig. 2. CMKIM with the highest sensitivity and an as high as possible specificity on the training set had a γ equal to 5.878, giving a weight of 0.28 to the clinical data set and 0.72 to the microarray data set. This model correctly classifies 70.75% of the patients in the validation set. Table I and Fig. 3 illustrate that this performance accuracy is slightly but not significantly better than each of the three other models we compared to. However, the Kappa statistic takes into account the prediction accuracy that is expected by chance. Comparing the Kappa coefficient of CMKIM with the other three models indicates that the prediction accuracy of CMKIM is more different from the accuracy expected by chance than the other models. CMKIM could identify three patients who developed distant metastases which were missed by the model build only on the microarray data, resulting in a higher sensitivity, PPV and NPV. This indicates that it seems worth to include extra data sources beside microarray data. When giving the clinical data and microarray data an equal weight, the performance accuracy is the same as when only considering microarray data. The weighing of the kernels can be specified beforehand, but, since one data set can contain more information than another, the weights can be learned from data. An advantage of this optimization is that a redundant data source or a data source with much noise will receive a weight close to zero and its

influence on the outcome will be kept small [6]. Moreover the weights of the kernel functions reflect the relative importance of the different data sources.

We finally showed in Table II that integration of clinical and microarray data outperforms the pure clinically indices (St Gallen, NIH and NPI).

B. Future Work

In the future, this algorithm for data fusion in disease management will be tested on other publicly available data sets. Moreover since our framework does not put any restrictions on the number of data sources being integrated, it will be expanded to data sets where more than two data sources are available such as proteomics and metabolomics.

V. ACKNOWLEDGMENTS

AD is research assistant of the Fund for Scientific Research - Flanders (FWO-Vlaanderen). OG is research assistant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. This work is partially supported by: **1.** Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, several PhD/postdoc & fellow grants. **2.** Flemish Government: **a.** FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (ontologies in bioi), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); **b.** IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame. **3.** Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011). **4.** EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

REFERENCES

[1] Balslev I, Axelsson C, Zedeler K, Rasmussen B, Carstensen B, Mouridsen H (1994). The Nottingham Prognostic Index applied to 9149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG). *Breast Cancer Res Treat*, **32**, 281-90.

[2] Chang H, Nuyten D, Sneddon J, Hastie T, Tibshirani R, Sorlie T, Dai H, He Y, van't Veer L, Bartelink H, van de Rijn M, Brown P, van de Vijver M (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl. Acad. Sci. USA*

[3] Eifel P, Axelson J, Costa J, et al (2001). National Institutes of Health Consensus development conference statement: adjuvant therapy for breast cancer. *J Natl Cancer Inst*, **93**, 979-89.

[4] Galea M, Blamey R, Elston C, Ellis I (1992). The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat*, **22**, 207-19.

[5] Goldhirsch A, Glick J, Gelber R, Coates A, Senn H (2001). Meeting highlights: International consensus panel on the treatment of primary breast cancer: Seventh international conference on adjuvant therapy of primary breast cancer. *J Clin Oncol*, **19**, 3817-27.

[6] Lanckriet G, De Bie T, Cristianini N, Jordan M, Noble W (2004). A statistical framework for genomic data fusion. *Bioinformatics*, **20**(16), 2626-35.

[7] Pavlidis P, Weston J, Cai J, Grundy W (2001). Gene functional classification from heterogeneous data. *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology*, 242-48.

[8] Pochet N, De Smet F, Suykens J, De Moor B (2004). Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics*, **20**(17), 3185-95.

[9] Schölkopf B, Tsuda K, Vert J-P. *Kernel methods in computational biology.* MIT Press.

[10] Suykens J, Vandewalle, J (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, **9**(3), 293-300.

[11] Suykens J, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J (2002). Least Squares Support Vector Machines. *World Scientific Publishing Co., Pte Ltd. (Singapore)*.

[12] Todd J, Dowle C, Williams M, Elston C, Ellis I, Hinton C, Blamey R, Haybittle J (1987). Confirmation of a prognostic index in primary breast cancer. *Br J Cancer*, **56**, 489-92.

[13] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**(6), 520-25.

[14] van de Vijver M, He Y, van't Veer L, Dai H, Hart A, Voskuil D, Schreiber G, Peterse J, Roberts C, Marton M, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers E, Friend S, Bernards R (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, **347**(25), 1999-2009.

[15] Vapnik V (1998). Statistical Learning Theory. *Adaptive and learning systems for signal processing, communications and control. Wiley, New York.*