KERNEL LOGISTIC REGRESSION-LINEAR FOR LEUKEMIA CLASSIFICATION USING HIGH DIMENSIONAL DATA

S.P. Rahayu [1,2] S.W. Purnami[2] A. Embong[2] Jasni Mohammad Zain[2]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Insititut Teknologi Sepuluh Nopember, Surabaya - Indonesia
[2] Faculty of Computer System & Software Engineering (Data Mining Research Group), Universiti Malaysia Pahang
Email: santi_pr@statistika.its.ac.id

Kernel Logistic Regression (KLR) is one of the statistical models that has been proposed for classification in the machine learning and data mining communities, and also one of the effective methodologies in the kernel–machine techniques. Basely, KLR is kernelized version of linear Logistic Regression (LR). Unlike LR, KLR has ability to classify data with non linear boundary and also can accommodate data with very high dimensional and very few instances. In this research, we proposed to study the use of Linear Kernel on KLR in order to increase the accuracy of Leukemia Classification. Leukemia is one of the cancer types that causes mortality in medical diagnosis problem. Improving the accuracy of Leukemia Classification is essential for more effective diagnosis and treatment of Leukemia disease. The Leukemia data sets consists of 7120 (very high dimensional) DNA micro arrays data of 72 (very few instances) patient samples on the state of Leukemia types. In Leukemia classification based upon gene expression, monitoring data using DNA micro array offer hope to achieve an objective and highly accurate classification. It can be demonstrated that the use of Linear Kernel on Kernel Logistic Regression (KLR–Linear) can improve the performance in classifying Leukemia patient samples and also can be shown that KLR–Linear has better accuracy than KLR–Polynomial and Penalized Logistic Regression.

Keywords: *Leukemia diagnosis, Kernel Logistic Regression, High dimensional data, Linear Kernel, Improving Accuracy*

In the last decade, it was found that the use of classifier system is one of the most important factors in cancer diagnosis and treatment, besides evaluating data that taken from patient and decision of medical expert [1]. Classification system can achieve an objective and highly accurate cancer classification by minimizing errors due to fatigued or inexperienced expert.

As many authors have pointed out, problem domain such as medical diagnosis does require transparent reasoning (interpretable) as well as accurate classification method [2]. KLR approach is particularly well suited for this type of situation. Kernel Logistic Regression (KLR) is one of the classification methods in the machine learning and data mining communities that has ability to explain the reasoning for the classification/decision process (KLR provides probability of classification membership).

There are some previous research in classifying Leuke–mia that used transparent classification method. Zhu and Hastie used Penalized Logistic Regression – RFE (classification accuracy : 95.9%) [3], while Rahayu and Embong applied KLR (Polynomial Kernel) in classifying Leukemia patient samples (classification accuracy : 90.3%) [2].

Trust in a system is developed by the clear description of how they were derived (transparent/interpretable) and also by quality of the results (accuracy). In this research, we proposed to use Linear Kernel on Kernel Logistic Regression (KLR), in order to improve the accuracy of KLR–Polynomial in classifying Leukemia patient samples. Hsu et all [4] suggested to use Linear Kernel when the number of features is very large. If the number of features is large, one may not need up data to a higher dimensional space. That is, the non linear mapping (like Polynomial Kernel) does not improve the performance. Hence, using the Linear Kernel is good enough.

This paper is organized as follows. In section 2, we give a description to KLR, the theory and the design of experiment that will be conducted. Section 3 reports the numerical results of experiment, and finally, we conclude in section 4.

KERNEL LOGISTIC REGRESSION

Kernel Logistic Regression (KLR), a non–linear form of Logistic Regression (LR), can be achieved via the so–called "kernel trick", whereby a familiar LR model is developed in a high–dimensional feature space, induced by a Mercer kernel.

Logistic Regression

Suppose we have a classification problem with c classes ($c \geq 2$), with a training set $\{(x_i, y_i)\}_{i-1}^{n}$ of n input samples independent and identically distributed (i.i.d) **x**, $\mathbf{X} \in \mathbf{R}^d$, and corresponding label **y**. The problem of classification consists of assigning input samples vector $\mathbf{X}$ into one of c classes label.
In Logistic Regression, we define a linear discriminant function or logit model for class k as [5]

$$\begin{aligned} g_k(x) &= \ln \frac{P(y=k|x)}{P(y=c|x)} \\ &= \beta_k^T X, \quad k = 1, ..., c-1 \end{aligned} \tag{1}$$

The conditional or posterior probability that x_i belongs to class c via the linear discriminant function is written as

$$P(y=k|x) = \frac{exp(\beta_k^T x)}{1 + \sum_{i=1}^{e-1} exp(\beta_l^T x)} \tag{2}$$

The class of membership of new point x can be given by this classification rule.

Considering a binary or two class problem with labels $y_i \in \{0, 1\}$. The success probability of the sample x_i belonging to class 1 (y_i = 1) is given by $P(y=1|x)$, since

$P(y = 0|x) = 1 - P(y = 1|x)$ that it belong to class 0 ($y_i = 0$). Then, we define a linear discriminant function (logit model) for two class problem based on equation (1) as

$$\begin{aligned} g(x) &= \ln \frac{P(y=1|x)}{P(y=0|x)} \\ &= \ln \frac{\pi(x)}{1-\pi(x)} \\ &= \beta^T X \end{aligned} \qquad (3)$$

where β denotes the weight vector with size $(d + 1) \times 1$ including the intercept, while the first element of $\mathbf{X}$ is 1. Via the logit model in equation (3), we can write the posterior probability of the class membership as

$$\begin{aligned} \pi(x) &= \frac{exp(\beta^T X)}{1 + exp(\beta^T X)} \\ &= \frac{1}{1 + exp(-\beta^T X)} \end{aligned} \qquad (4)$$

and

$$1 - \pi(x) = \frac{1}{1 + exp(\beta^T X)} \qquad (5)$$

The logit link function constraint the output of the model to lie in the range {0,1}.

Assuming the label, y_i, represent an i.i.d sample drawn from a Bernoulli distribution conditioned on the input vector $\mathbf{X}$,

$$\xi(x_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}, \qquad (6)$$

$y_i = 0, 1; i = 1, 2, 3, ..., n$

The likelihood of the data is given by

$$l(\beta) = \prod_{i=1}^{n} (\pi(x_i))^{y_i}(1 - \pi(x_i))^{1-y_i}, \qquad (7)$$

The optimal model parameter β, are then determined by maximizing the conditional log likelihood,

$$\begin{aligned} L(\beta) &= \log l(\beta) \\ &= \sum_{i=1}^{n} \{y_i \log[\pi(x_i)] + \\ &\quad (1 - y_i) log[1 - \pi(x_i)]\} \end{aligned} \qquad (8)$$

or equivalently, by minimizing the negative logarithm of the likelihood

$$\begin{aligned} L(\beta) &= -\log l(\beta) \\ &= -\sum_{i=1}^{n_i} \{y_i \log[\pi(x_i)] + \\ &\quad (1 - y_i) log[1 - \pi(x_i)]\} \\ &= \sum_{i=1}^{n} -y_i \beta^T X + \log(1 + exp(\beta^T X)) \end{aligned} \qquad (9)$$

we wish to solve the equation system $\frac{\partial L(\beta)}{\partial L(\beta_j)} = 0$, in order to find the optimizing weight vector β. Since the $\pi(x_i)$ depend nonlinearly on β, this system cannot be solved analytically and an iterative technique must be applied. The optimal model parameters can be found using Newton's method or equivalently an iteratively re–weighted least squares procedure [6]. *The Newton Raphson's Method*

$$\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1} q^{(t)} \qquad (10)$$

Where the Hessian

$$\begin{aligned} H^{(t)} &= \frac{\partial^2 L(\beta)}{\partial(\beta_i)\partial(\beta_u)} \\ &= \sum_{i=1}^{n} x_{ij} x_{iu} [\pi_i^{(t)}](1 - [\pi_i^{(t)}]) \\ &= -X^T W^{(t)} X \end{aligned}$$

The gradient of L,

$$\begin{aligned} q^{(t)} &= \frac{\partial L(\beta)}{\partial(\beta_i)} \\ &= \sum_{i=1}^{n} x_{ij}(y_i - [\pi_i^{(t)}]) \\ &= X^T W[W^{-}1(y - p)] \\ &= X^T (y - p) \end{aligned}$$

The Newton's method can be restated as an Iteratively Re–weighted Least Squares (IRLS) problem [7] .

Iteratively Re–weighted Least Squares Procedure Forming a variable that states a generalized linear model, [8]

$$Z^{(t)} = X\beta^{(t)} + \varepsilon \qquad (11)$$

the normal form equations of least squares problem with input matrix $(W^{(t)})^{\frac{1}{2}} X$ and dependent variables $(W^{(t)})^{\frac{1}{2}} Z^{(t)}$ can be written as,

$$(X^T W^{(t)} X)\beta^{(t+1)} = X^{(t)} W^{(r)} Z^{(r)} \qquad (12)$$

At each iteration, the model parameters are given by the solution of a weighted least–squares problem, such that

$$B^{(t+1)} = (X^{(T)} W^{(t)} X)^{-1} X^T W^{(t)} Z^{(t)} \qquad (13)$$

$$w_i = \pi_i(1 - \pi_i), \qquad (14)$$

and

$$Z^{(t)} = X\beta^{(t)} + (W^{(t)})^{-1}(y - p) \qquad (15)$$

The algorithm proceeds iteratively, updating the weights according to (13) and then updating W and Z according to (14) and (15) until convergence is achieved.

Kernelized Logistic Regression

Consider Kernel Logistic Regression, a non–linear form of Logistic Regression. Logistic Regression is a linear classifier and well known classification method in the field of statistical learning and also is the model of choice in many problem domains. However, this method has limitation to classify the data with nonlinear boundaries [8]. Kernel Logistic Regression, may overcome this limitation via the so–called "kernel trick". The "kernel trick" [9, 10, 11] provides a general mechanism for constructing nonlinear generalizations of familiar linear Logistic Regression by mapping of original data **X** into a high–dimensional Hilbert space *F*, usually called feature space, and then by using linear pattern analysis to detect relations in the feature space. Mapping is performed by specifying the inner product between each pair of data. The inner product in the feature space is often much more easily computed than the coordinates of the points (notably when the dimensionality of the feature space is high). The linear nature of the underlying model means that the parameters of a kernel model typically given by the solution of a convex optimization problem [12], with a single global optimum, for which efficient algorithm exist. Given an input space **X** and a feature space *F*, we consider a function $\phi : \mathbf{X} \to F$. The Kernel Logistic Regression model implements a well known linear Logistic Regression model in the feature space (appears as nonlinear model in the input space). According to logit model in equation (3), after mapping into feature space the logit model can be written as

$$g(x) = \beta^T \phi(x) \tag{16}$$

where $\phi(.)$ represent a nonlinear mapping of the original data **X** into feature space.

Kernel Function

Rather than defining the feature space explicitly, it is instead defined by a kernel function that evaluates the inner product between the images of input vectors in the feature space,

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{17}$$

For the interpretation of the kernel function as an inner product in a fixed feature space to be valid, the kernel must obey Mercer's condition [13], that is the kernel must be positive (semi) definite. There are usually the following choices for kernel function: $K(x_i, x_j) = x_i^T x_j$ (linear kernel), $K(x_i, x_j) = (x_i^T x_j + h)^b$ (polynomial of degree *b*, with h ≥ 0 a tuning parameter, $K(x_i, x_j) = exp(\frac{-\|x_i - x_j\|^2}{2\sigma^2})$ (radial basis function, RBF, where σ is a tuning parameter. In this work, we use Linear Kernel as suggested by Hsu et al [4].

Globally, there are two reason when we use Linear Kernel. The first reason is about number of features that are much larger than number of instances. The second one is because both number of features and instances are large.

Generally, in other situation, Hsu et al [4] suggest that RBF Kernel is reasonable first choice. The RBF Kernel has less numerical difficulties and can handle the case when relation between class labels and attributes is nonlinear (unlike Linear Kernel).

Kernel Logistic Regression Modelling

When constructing a statistical model in a high dimensional space, it is necessary to take steps to avoid over fitting the training data, that is to impose a penalty on large fluctuations of the estimated parameters β. The most popular method is ridge penalty $\frac{\lambda}{2}\|\beta\|^2$ that was introduced by [14]. As a result, the Kernel Logistic Regression model is trained by adding a quadratic regularized to negative log likelihood,

$$\begin{aligned} L(\beta)_{ridge} &= L(\beta) + \frac{\lambda}{2}\|\beta\|^2 \\ &= \sum_{i=1}^{n} -y_i \beta^T X + \\ &\qquad \log(1 + exp(\beta^T X)) + \frac{\lambda}{2}\|\beta\|^2 \end{aligned} \tag{18}$$

where λ is regularization parameter that must be set in order to obtain a good bias– variance trade off and avoid over fitting [6, 15]. Furthermore, $L(\beta)_{ridge}$ represents a convex optimization problem. The representer theorem [10, 16] states that the solution of an optimization of the equation (19) can be written in the form of an expansion over training pattern, (x_i is replaced by $\phi(x_i)$) where W = diag($\{w_i, w_2, ..., w_n\}$) is a diagonal weight matrix with non–zero elements given by

$$\beta = \sum_{i=1}^{n} \alpha_i \phi(x_i) \tag{19}$$

and so from (17) we have the so–called kernel machine,

$$\begin{aligned} g(x) &= \sum_{i=1}^{n} \alpha_i \phi(x_i)^T \phi(x_j) \\ &= \sum_{i=1}^{n} \alpha_i K(x_i, x_j) \end{aligned} \tag{20}$$

With the usual kernel trick, the inner product can be substituted by kernel functions satisfying Mercers condition. Substituting the expansion of β in (19) into (12), this lead us to nonlinear generalization of Logistic Regression in kernel feature spaces which we call Kernel Logistic Regression [8]. We can write,

$$(KW^{(t)}K + \lambda K)\alpha^{(t+1)} = KW^{(t)}(S^{(t)})^T \tag{21}$$

$$\alpha^{(t+1)} = (K + \lambda(W^{(t)})^{-1})^{-1}(S^{(t)})^T \tag{22}$$

with $(S^{(t)})^T = K\alpha^{(t)} + W^{-1}(y - p)$

Like LR, KLR also produce posterior probability of the class membership. Bartlett and Tewari (2004) [17] proved that KLR can be used to estimate all conditional probabilities.

Experiment

Data Description

The Leukemia data set used in this study come from `http://www.genome.wi.mit.edu/cancer`. This

Table 1: Representation of Confusion Matrix

Actual/Observed	Prediction	
	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Table 2: Confusion matrix

	Actual/Observed	Prediction	
		ALL	AML
KLR-Polynomial	ALL	47	0
	AML	7	18
KLR-Linear	ALL	47	0
	AML	2	23

data set consists of 72 samples of two types of acute leuke–mias, Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) [16]. Each sample is a vector corresponding to 7129 genes. The Leukemia patient (1=AML, 2=ALL) will be classified according to those genes. There are 25 samples of data set fit in to AML, and remaining 47 data is ALL.

Methodology

The goal of this experiment is to study the classification performance of applying KLR to classify Leukemia patient samples. In order to achieve this goal, the data set was conducted with k-fold cross validation (cv) method [1]. *k*–Fold cross validation (cv) is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated *k* times. Each time, one of the *k* subsets is used as the test set and the other $k-1$ subsets are put together to form a training set. Then the average error across all *k* trials is computed. In this work, we are using 10–fold cross validation.

In this experiment, we pre processed the data set so that the mean is 0 and standard deviation is 1 and used linear kernel with λ = 0.06 [3] to perform classification task. Then, in order to know the performance of KLR–Linear, we also compared the result of KLR–Linear with KLR–Polynomial [2] and Penalized Logistic Regression RFE, PLR–RFE [3].

Performance Evaluation Method

We have used four indicators to evaluate the classification performance of leukemia diagnosis. These indicators (accuracy, sensitivity and specificity analysis) are based on confusion matrix and Receiver Operating Characteristic (ROC) curve (area under the curve).

Confusion Matrix

A confusion matrix [1] contains information about actual and predicted classifications done by a classification system. Table 1 shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study:

(a) The number of **correct** predictions that a patient is **ALL (True Negatives, TN)**
(b) The number of **incorrect** predictions that a patient is **AML (False Positives, FP)**
(c) The number of **incorrect** predictions that a patient is **ALL (False Negatives, FN)**
(d) The number of **correct** predictions that a patient is **AML (True Positives, TP)**

Table 3: The result of performance evaluation

Indicator	KLR-Polynomial	KLR-Linear
Accuracy	90.3%	97.2%
Sensitivity	72%	92%
Specificity	100%	100%
AUROC curve	97.4%	99.4%

Three standard terms have been defined for the two class matrix [1]:

i) The *accuracy* is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$accuracy = \frac{(TN+TP)}{TN+FP+FN+TP}(\%)$$

ii) The *sensitivity* is the proportion of AML cases that were correctly identified, is calculated using the equation:

$$sensitivity = \frac{(TP)}{TP+FN}(\%)$$

iii) The *specificity* is the proportion of ALL cases that were correctly classified as ALL, is calculated using the equation:

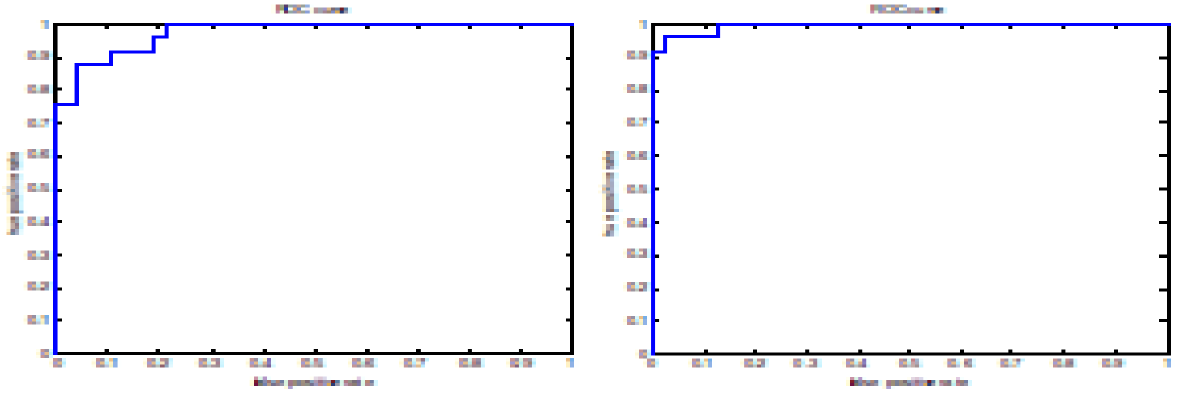
$$specificity = \frac{(TN)}{FP+TN}(\%)$$

Receiver Operating Characteristics Curve

A Receiver Operating Characteristic (ROC) curve [18] shows the relationship between False Positives (FP) and True Positives (TP). In the ROC curve the horizontal axis has the percentage of FP and vertical axis has the percentage of TP for a database sample. The final performance of this work is assessed using the Area Under the ROC (AUROC) curve.

RESULT

In this experiment, we create confusion matrix based on classification prediction result of applying KLR–Linear in classifying Leukemia patient samples, calculate the total accuracy, sensitivity, and specificity classification prediction. Then, we draw the ROC curve and calculate the AUROC curve of classification prediction. The Confusion matrix is shown in Table 2. At the same time, we compare the results of KLR-Linear with previous research that used KLR-Polynomial [2]



(a)

(b)

Figure 1: KLR-Polynomial

Table 4: Comparison of leukemia classification method

Method	Accuracy (%)
PLR-RFE [3]	95.9
KLR-Polynomial [2]	90.3
KLR-Linear	97.2

According to Table 2, we see that the ALL cases were perfectly identified (47;100%) (by using KLR–Linear and KLR–Polynomial) while the AML cases were 92% (23/25) correctly classified by using KLR–Linear.
The results of ROC curve is drawn, as shown in Figure 1.

Table 3 summaries the classification performance indicators of applying KLR (Polynomial and Linear) to classify Leukemia patient samples, according to confusion matrix and ROC curve above.

All indicators (accuracy, sensitivity, specificity, AUROC curve) of KLR–Linear display high values. It shows that the classification performance of KLR–Linear is better than KLR-Polynomial.

In addition, we compared the accuracy result of KLR (Linear and Polynomial) in classifying leukemia patient with PLR–RFE. The result shows that the accuracy of KLR–Linear is higher than KLR–Polynomial and PLR RFE.

CONCLUSION

We have proposed Kernel Logistic Regression with Linear Kernel (KLR–Linear) for high dimensional data problem to classify Leukemia patient samples. It can be shown that Kernel Logistic Regression with Linear Kernel (KLR–Linear) has better classification performance as compared with KLR–Polynomial and PLR–RFE.

ACKNOWLEDGMENT

We wish to acknowledge Universiti Malaysia Pahang for providing the grant for this research.

REFERENCES

[1] Polat, K., Gunes, S.: *Breast Cancer Diagnosis using Least Square Support Vector Machine*. Digital Signal Processing **17** (2007) 694–701

[2] Rahayu, S.P., Embong, A.: *Kernel Logistic Regression For Leukemia Classification Using High Dimensional Data*. In: Proceeding of The 4th International Conference on Information & Communication Technology and Systems (ICTS). (2008)

[3] Zhu, J., Hastie, T.: *Classification of gene microarrays by penalized logistic regression*. Biostatistics **5**(3) (2004) 427–443

[4] Hsu, C., Chang, C., Lin, C.: *A Practical Guide to Support Vector Classification*. Department of Computer Science and Information Engineering, National Taiwan University (2008)

[5] McCullagh, P., Nelder, J.A.: *Generalized Linear Models*. Monographs on Statistics and Applied Probability **37** (1989)

[6] Cawley, G.C., Talbot, N.L.C.: *Efficient Model Selection for Kernel Logistic Regression*. In: Proceeding of the 17th International Conference on Patten Reconition. (2004)

[7] Nabney, I.T.: *Efficient training of RBF networks for classification*. In: Proceedings of the Ninth International Conference on Artificial Neural Networks. Volume 1. (1999)

[8] Roth, V.: *Pobabilistic Discriminative Kernel Classifiers for Multi–Class Problems*. Lecture Notes in Computer Science **2191** (2001) 246–253

[9] Schölkopf, B., Smola, A.J.: *Learning With KernelŮ-Support Vector Machines Regularization Optimization and Beyond*. MIT Press, Cambridge, MA (2002)

[10] Cortes, C., Vapnik, V.: *Support vector network*. In: Mach. Learning 20. (1995) 273–297

[11] Shawe-Taylor, J.N., Cristianini: *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge (2004)

[12] Boyd, S., Vandenberghe, L.: *Convex Optimization*. Cambridge University Press, Cambridge (2004)

[13] Mercer, J.: *Functions of positive and negative type and their connection with the theory of integral equations*. Philos. Trans. Roy. Soc. London A **209** (1990) 415–446

[14] Hoerl, A., Kennard, R.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12** (1970) 55–67

[15] Geman, S., Bienenstock, E., R, D.: *Neural networks and the biasvariance dilemma*. Neural Computation **4**(1) (1992) 1–58

[16] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M.: *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. Science **286** (1999) 531–536

[17] Bartlett, P., Tewari, A.: *Sparseness versus Estimating Conditional Probability: Some Asymtotic Results*. Proceedings of the 17th Annual Conference on Learning Theory, Lecture Notes in Computer Science **3120** 564–578

[18] Schölkopf, B., Herbrich, R., Smola, A.J.: *A Generalized Reprenter Theorem*. In: Poceeding of the Fourteenth International Conference on Computational Learning Theory. (2002)