

APPLYING LOGISTIC REGRESSION AND RVM TO ACHIEVE ACCURATE PROBABILISTIC CANCER DIAGNOSIS FROM GENE EXPRESSION PROFILES

Balaji Krishnapuram¹, Alexander Hartemink², Lawrence Carin¹

¹Dept. of Electrical Engineering, Duke University, ² Dept. of Computer Science, Duke University

ABSTRACT

Recent research has clearly demonstrated that accurate cancer diagnosis is indeed possible based on gene expression analysis and a database of stored expression profiles from known cancerous tumors. While current techniques largely succeed in correctly identifying the class membership of a tumor among various classes of cancers from its gene expression profile, they do not provide a probability of membership of the tumor in each class. In clinical settings where false negatives in identifying cancer are far more harmful than false positives, such posterior probabilities must be explicitly computed in order to handle asymmetric misclassification costs in a principled theoretical framework. In this paper we demonstrate the success of two related classifiers that achieve classification performance superior to the state-of-the-art Support Vector Machine (SVM) classifier, while also providing a probability of class membership explicitly.

1. INTRODUCTION

Traditional cancer diagnosis relies on human expert interpretation of clinical and histopathological information. However, clinical information can be incomplete or misleading, and many tumors are atypical or lack morphological features that are useful for differential diagnosis, resulting in diagnostic confusion. While molecular diagnostic techniques offer the hope of more accurate and objective cancer classification, characteristic molecular markers for most solid tumors have yet to be identified [1].

Recently, tumor gene expression profiles have been used for cancer diagnosis [4]. Oligonucleotide arrays and DNA microarrays enable the simultaneous measurement of expression levels of thousands of genes from a tumor locus in a single experiment. While several supervised and unsupervised methods from the pattern recognition literature have been used for cancer diagnosis from this data, classifiers based on the Support Vector Machine (SVM) have proven to be the most accurate methods to date in this context [1,2,3].

While these techniques have been largely successful in identifying from gene expression profiles the class membership of tumors from among various classes (examples of classes may be malignant versus benign tumors, differential response to a course of treatment, or tissue-of-origin for metastatic cancers), current techniques do suffer from several disconcerting shortcomings. Perhaps the biggest drawback associated with their use in clinical settings is that these methods commonly identify the class membership based on the gene expression profile but do not give a probability of membership of belonging

to each class. This problem is particularly important in the context of asymmetric misclassification costs where the misclassification cost associated with some classes may be significantly higher than that of others. For example, while falsely diagnosing a benign tumor as malignant may be undesirable due to the unnecessary time and effort expended on further investigation, falsely diagnosing a malignant tumor as benign would lead to far more disastrous consequences.

Although SVM techniques have been adapted for training the diagnostic classifier on imbalanced datasets that have a large number of example expression profiles from one class and relatively few expression profiles from the other class, this only addresses the issue of imbalanced training datasets and not the issue of asymmetric misclassification costs. Furthermore, attempts to derive posterior class membership probabilities from the SVM have been shown to be theoretically suboptimal [6].

In this paper, we demonstrate the success of logistic regression [5] and Relevance Vector Machine (RVM) [6,7] classifiers that not only outperform an SVM classifier in their classification accuracy in our tests, but also provide a principled estimate of posterior class membership probabilities. Further, the RVM can be extended to automatically determine the relevance of each gene to the identification of class membership. This has the potential to provide us with valuable clues about the genes responsible for causing cancers as well. The extension of these techniques to multi-class problems is also straightforward, and more principled than the current multi-class SVM classification techniques that lack a systematic basis in identifying the class boundaries directly from the quadratic programming problem solved therein.

The rest of this paper is structured as follows. In Section 2, we describe the logistic regression classifier and study its relationship to the SVM. Section 3 outlines the RVM classifier, which can be viewed as an extension of both the logistic regression and the SVM classifiers. We conclude with a discussion of our experimental results on a measured benchmark cancer dataset in Section 4.

2. LOGISTIC REGRESSION

In the traditional pattern recognition literature, the problem of cancer diagnosis using the gene expression profile of a new tumor and a database of known gene expression profiles and their diagnoses falls under the category of supervised pattern recognition. In the general setting, we are presented with a set of M training samples x_i , indexed by $i \in \{1, 2, \dots, M\}$. Each sample x_i is a d -dimensional vector so that $x_i \in R^d$. The class membership of each sample is known and denoted by y_i . In the two class case, we can assume that $y_i \in \{-1, +1\}$ without loss of generality. Thus,

we define the training set D to consist of the M samples and the corresponding class membership labels:

$$D = \{ \langle \mathbf{x}_i, y_i \rangle : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}, \text{ for } i \in \{1, 2, \dots, M\} \}$$

Our objective is to find a function $f: \mathbb{R}^d \rightarrow \{-1, +1\}$ such that the training set is learned by the function f faithfully (i.e., for all i $f(\mathbf{x}_i) = y_i$) and further, that $f(\mathbf{x}')$ correctly identifies the true class label y' , even for new \mathbf{x}' that are not part of the training set. In the specific context of cancer diagnosis, the vectors \mathbf{x}_i represent gene expression profiles taken from a wide range of tissue samples and the values of y_i indicate the various classes to be distinguished, such as malignant versus benign. We briefly note that multi-class classification can be formulated adopting the same approach and constructing C classifiers (for the C class problem) as above, each distinguishing one class C_1 from all the remaining classes. However, in this paper we shall specifically deal with the binary classification subproblem. In this case, the class membership likelihood model for the logistic regression is:

$$P(y = \pm 1) = \sigma(y \mathbf{a}^T \mathbf{x}) = \frac{1}{1 + \exp(-y \mathbf{a}^T \mathbf{x})} \quad (1)$$

where \mathbf{a} is a parameter vector that is learned from the training samples. Specifically, given the dataset D , we want to find the parameter \mathbf{a} that maximizes the log-likelihood $l(\mathbf{a})$ over the training set D :

$$l(\mathbf{a}) = - \sum_{i=1}^M \log(1 - \exp(-y_i \mathbf{a}^T \mathbf{x}_i)) \quad (2)$$

We can show that the gradient of the log-likelihood is:

$$\mathbf{g} = \Delta_{\mathbf{a}} l(\mathbf{a}) = \sum_{i=1}^M (1 - \sigma(y_i \mathbf{a}^T \mathbf{x}_i)) y_i \mathbf{x}_i \quad (3)$$

Gradient descent using (3) resembles the perceptron learning algorithm except that it will always converge (for a suitable step size) regardless of whether the classes are separable. The Hessian of the log-likelihood can be shown to be:

$$H = \frac{d^2 l(\mathbf{a})}{d \mathbf{a} d \mathbf{a}^T} = - \sum_{i=1}^M \sigma(\mathbf{a}^T \mathbf{x}_i) (1 - \sigma(\mathbf{a}^T \mathbf{x}_i)) \mathbf{x}_i \mathbf{x}_i^T \quad (4)$$

$$\text{i.e. } H = -XAX^T$$

where X is a matrix composed of the training vectors:

$$X = [\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_M^T]^T \quad (5)$$

and A is a diagonal matrix with:

$$A_{ii} = \sigma(\mathbf{a}^T \mathbf{x}_i) (1 - \sigma(\mathbf{a}^T \mathbf{x}_i)) \quad (6)$$

Note that the Hessian does not depend on how the \mathbf{x}_i are labeled. It is negative definite which implies that $l(\mathbf{a})$ is convex. However, if we perform maximum likelihood estimation for \mathbf{a} we may run into trouble as explained below. We note that ML estimation is equivalent to MAP estimation under the special circumstance of a uniform prior on \mathbf{a} . If we assume such a uniform prior on \mathbf{a} then when the classes are nonseparable by a hyperplane whose normal is given by the vector \mathbf{a} the posterior is approximately Gaussian:

$$P(\mathbf{a} | X, \mathbf{Y}) \approx N(\mathbf{a}; \hat{\mathbf{a}}, -H^{-1}), \quad (7)$$

$$\hat{\mathbf{a}} = \arg \max \prod_{i=1}^M P(y_i | \mathbf{x}_i, \mathbf{a})$$

so the model likelihood can be approximated by:

$$P(\mathbf{Y} | X) \approx P(\mathbf{Y}, \hat{\mathbf{a}} | X) (2\pi)^{d/2} |H|^{-1/2} \quad (8)$$

where d is the dimensionality of \mathbf{a} . Unfortunately, when the classes are separable (in the training data), the posterior is

improper: the magnitude of \mathbf{a} can be increased without bound. In intuitive terms, this corresponds to a scaling of the projection along the normal to the hyperplane. Since this scaling is important in estimating the posterior class membership probabilities (via the sigmoidal link function $\sigma(\cdot)$), we are left in a situation where we are able to classify the data, but not able to get an accurate estimate of the posterior probability. Further, since an infinite number of hyperplanes can separate linearly separable training data, this situation leads to overfitting the model in the sense that a large margin hyperplane (i.e. a hyperplane that attempts to maximize the distance between the classes along its normal) cannot in general be obtained. This discussion suggests the need for a nonuniform prior on the vector \mathbf{a} such as a mean zero spherical covariance Gaussian prior:

$$P(\mathbf{a}) = N(0, vI) \quad (9)$$

Performing MAP estimation with such a prior (instead of an ML estimation) gives a regularized estimate. In our research, we have used efficient conjugate gradient techniques to estimate the MAP value of \mathbf{a} [5]. The prior is intimately related to the regularization operator (in the context of regularization theory) used to measure the generalization of the decision function on unseen data. In regularization theory, we seek to minimize an objective function that is a weighted sum of the generalization measure, and the accuracy of the classifier on the training data.

The logistic regression model presented in this section has thus far constructed a hyperplane classifier in the space of the gene expression feature vectors. It is trivial to see that it can also be used to construct other nonlinear generalizations by incorporating other basis functions (such as quadratic terms) in the design matrix. In other words, we can always replace the gene expression vector \mathbf{x} throughout the above discussion with a k -dimensional vector function of \mathbf{x} :

$$\mathbf{h}(\mathbf{x}) = [1, \phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots, \phi_k(\mathbf{x}), \dots, \phi_k(\mathbf{x})]^T \quad (10)$$

where $\phi_j(\mathbf{x})$ can be any scalar nonlinear basis function used in performing a basis function expansion of the final decision function of the classifier. If $\phi_j(\mathbf{x}) = x^{(j)}$ for all j where $x^{(j)}$ is the j^{th} component of \mathbf{x} , and $k = d$, then we recover the linear hyperplane classifier. However, if we used $\phi_j(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_j)$ where $K(\dots)$ is any Mercer kernel appropriate for use with the SVM, then we explicitly recover a classifier that has the same functional form as the SVM, i.e. our classification function becomes of the form:

$$f(\mathbf{x}) = \alpha_1 + \sum_{i=1}^M \alpha_{i+1} K(\mathbf{x}, \mathbf{x}_i) \quad (11)$$

3. THE RELEVANCE VECTOR MACHINE

While the logistic regression classifier discussed above performs well for our task, it has been improved to overcome certain limitations. The Relevance Vector Machine [6] discussed in this section replaces the Gaussian prior on \mathbf{a} with a Student-t prior distribution. This leads to interesting sparsity properties of the \mathbf{a} and hence, the classifier decision function as well. In kernel-machines that have a decision function of the form given in Eqn. 11, it is well known that sparsity of \mathbf{a} is an important indication of the margin of the classifier (since it prevents overfitting) [7]. In the SVM the sparsity is induced by the use of the L_2 norm of

the Reproducing Kernel Hilbert Space (RKHS) [7] of the kernel function $K(\dots)$ as the regularizer. In contrast, RVMs explicitly obtain sparsity by using sparsity-inducing hyperpriors in a principled Bayesian MAP estimation framework. Further, the RVM also provides the *a posteriori* class membership probabilities that we seek in this paper.

Since sparsity is an important indicator of the generalization ability, we use our intuition that most of α_i are exactly zero to guide our selection of the prior. However, we do not know *a priori* exactly which α_i should be set to zero because we do not know the support vectors beforehand. Therefore, we have to use our intuition in the form of a parametric prior whose parameters (further given their own hyperpriors) are estimated from the data. The basic idea is to make extensive use of hyperparameters in determining the priors $P(\alpha_i)$ on the individual expansion coefficients α_i . This is equivalent to letting the data select the most appropriate measure of smoothness/generalization to be used when finding the best decision function. In particular, we assume that the (hierarchical) parametric prior on α_i is a mean zero Gaussian with a variable parameter s_i (the inverse variance) that is estimated from the data. Following the notation and derivation of [7], we write:

$$P(\alpha_i / s_i) = \sqrt{\frac{s_i}{2\pi}} \exp\left(-\frac{s_i \alpha_i^2}{2}\right) \quad (12)$$

where $s_i > 0$ plays the role of a hyperparameter. In turn, s_i has its own corresponding gamma hyperprior:

$$P(s_i) = \Gamma(s_i / a, b) = \frac{s_i^{a-1} b^a \exp(-s_i b)}{\Gamma(a)}, s_i > 0 \quad (13)$$

For the RVM classifier, we usually choose $a, b = 10^{-4}$ or similar small values in the above, or even a non-informative (i.e. flat on a log-scale, with $P(\ln s_i) = \text{constant}$) but improper Jeffereys hyperprior with a, b set to zero in (13): $P(s_i) = (s_i)^{-1}$

The hierarchical priors used above can be integrated out to yield the effective prior on α_i . Here, this can be shown to be a Student-t distribution over α_i . The effective prior is given by:

$$P(\alpha_i) = \left(a + \frac{1}{2}\right) \ln\left(b + \frac{\alpha_i^2}{2}\right) \quad (14)$$

As seen from (14), $P(\alpha_i)$ is sharply peaked around $\alpha_i = 0$ suggesting a strong prior preference for sparsity.

Let us define the Gram matrix (also the design matrix) to be $G_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. Having prescribed a prior (here using a hierarchical prior), the next step of a Bayesian analysis is to specify the likelihood model. For this the RVM uses the Bernoulli distribution and a logistic transfer function σ (for the binary classification case) given by:

$$P(y = 1 | \mathbf{x}, t = f(\mathbf{x})) = \frac{1}{1 + \exp(-f(\mathbf{x}))} = \sigma(f(\mathbf{x})) \quad (15)$$

whereby we effectively introduce an intermediate latent variable $t = G\alpha = f(\mathbf{x})$ and perform a logistic regression using equation 15 as a model for the distribution of the labels y_i . As in regression, we use a kernel expansion for the latent variables. The negative log posterior is given by:

$$-\ln P(\alpha | \mathbf{Y}, s) = \sum_{i=1}^M -\ln P(y_i | t_i) - \sum_{i=1}^M \ln P(\alpha_i / s_i) + \text{const.} \quad (16)$$

However, we cannot maximize this with respect to α analytically and we have to resort to approximate methods. In the RVM, we use a Laplacian approximation where we approximate a PDF with a Gaussian centered on its mode, whose covariance is obtained directly from the Hessian. The gradient

and Hessian are found by computing the first and second derivatives of (16):

$$\text{Grad} = \partial_{\alpha} [-\ln P(\alpha | \mathbf{Y}, s)] = G\mathbf{c} + S\alpha \quad (17)$$

$$-\text{Hess} = \partial_{\alpha}^2 [-\ln P(\alpha | \mathbf{Y}, s)] = G^T Q G + S$$

where $S = \text{diag}(s_1, s_2, s_3, \dots, s_M)$, $\mathbf{c} = (\mathbf{t} + 1)/2 - \sigma(\mathbf{t})$, and $Q = \text{diag}\{\sigma(t_1)[1 - \sigma(t_1)], \sigma(t_2)[1 - \sigma(t_2)], \dots, \sigma(t_M)[1 - \sigma(t_M)]\}$.

Knowing the gradient and Hessian analytically, we use an iterative Newton-Raphson approach to find the mode of (16).

Thus, we can use an iterative Newton-Raphson maximization step to find the MAP estimate α_{MAP} that maximizes $P(\alpha | \mathbf{Y}, s)$ and thereafter, use a Laplacian approximation whereby $P(\alpha | \mathbf{Y}, s)$ is approximated by a Gaussian $N(\mu = \alpha_{\text{MAP}}, \Sigma = (-\text{Hess})^{-1})$. Thus, given any set of hyperparameters controlling the priors on the decision function, we can compute the optimal decision function and the evidence in favor of that set of hyperparameters. Iterating the above in a loop, we update the hyperparameters in each step in order to maximize the evidence:

$$s_i = \frac{1 - s_i \Sigma_{ii}}{\mu_i^2} \quad (18)$$

In other words, we perform a type II ML estimation, where we have made the simplifying assumption of approximating the posterior PDF over the hyperparameters by a delta function at its mode. The success of this approximation is verifiable experimentally from the performance of the RVM classifier, which is comparable to the SVM and other known state-of-the-art algorithms.

4. EXPERIMENTAL RESULTS AND DISCUSSION

We tested the performance of the logistic regression and RVM classifiers on the dataset originally analyzed by Golub, et al., in [4]. This dataset of samples derived from human patients with acute leukemia can be obtained from the web at <http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi>. Bone marrow or peripheral blood samples were collected from 72 patients with either acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). The dataset contains expression levels for 7129 human genes measured using Affymetrix oligonucleotide arrays. The numerical values in the dataset represent the level of gene expression after suitable normalization.

Following the methods of Golub, et al., we subtracted the mean of each profile, and then rescaled each profile to have unit variance. We selected a subset of the genes that were best suited for use as features of the classifier by using a Fisher Discriminant Ratio (FDR) score for each gene profile. In order to compare our performance with existing results in literature, we selected the top 500 genes based on the FDR. We note that all the classification methods proposed in this paper are quite robust to the number of genes used and achieved similar classification accuracy even using a much smaller number of genes; we selected 500 merely to compare with results in the existing literature.

Following the experimental setup of other researchers we performed leave-one-out-cross-validation (LOOCV) on the dataset in order to evaluate the classification accuracy of the proposed classifiers. The results of these tests (and the state-of-the-art results from other current techniques in the literature) are summarized in Table 1.

While it is commonly known that the SVM can find hyperplane classifiers (as in [2,3]), it is perhaps not as widely appreciated that not every possible hyperplane can be obtained as a classifier by the SVM. This is due to its inherent kernel basis function expansion on limited training samples and the extremely large dimensionality of the underlying feature space of the data. While the representer theorem [7] guarantees that the SVM will find the best decision surface under the specific regularizer induced by its kernel, it is by no means clear that better surfaces may not be induced by other regularizers in other classifiers such as the ones we propose. To test this, we ran both the RVM and the logistic regression classifiers using the linear inner product kernel as the design matrix, as well as directly on the space of the gene expression levels. Figure 1 depicts the posterior class (ALL) membership probabilities found by the logistic regression classifier and demonstrates the superiority (larger margin) of the direct logistic regression over linear-kernel based logistic regression. The results from the RVM are similar and have been omitted here due to space considerations.

In summary, the proposed classifiers successfully outperform the SVM on classification rates, while also providing posterior probabilities of class membership. However, we must note that these posterior probabilities are subject to interpretation and their validity is dependent on the underlying modeling assumptions. Identifying an appropriate biological interpretation for these probabilities is the subject of further investigation in our research and we hope to provide more substantive proof of the validity in our future work.

5. REFERENCES

- [1] S. Ramaswamy, et al. "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures", *PNAS*, Vol. 98, No. 26, Dec. 2001, pp. 15149-15154.
- [2] T. S. Furey, et al. "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Data", *Bioinformatics*, Vol. 16, No. 10, 2000, pp. 906-914.
- [3] A. Ben-Dor, et al. "Tissue Classification with Gene Expression Profiles", in *Proceedings of the Fourth International Conference on Computational Molecular Biology*, 2000.
- [4] T. Golub, et al. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science*, Vol. 286, 15 Oct. 1999, pp. 531-537.
- [5] T. P. Minka. "Algorithms for maximum-likelihood logistic regression", Technical Report, Dept. of Statistics, available at: <http://www.stat.cmu.edu/~minka/papers/learning.html>.
- [6] M. Tipping. "Sparse Bayesian Learning and the Relevance Vector Machine", *Journal of Machine Learning Research*, Vol. 1, Jun. 2001, pp. 211-244.
- [7] R. Herbrich. *Learning Kernel Classifiers*, MIT Press: Cambridge, MA, 2002.

Classifier Tested Using LOOCV	Number Correctly Classified (out of 72)
RVM (linear kernel)	68
RVM (no kernel, i.e. feature space)	70
Logistic Regression (linear kernel)	66
Logistic Regression (no kernel)	70
SVM (linear kernel) [3]	68
AdaBoosting (decision stumps) [3]	69

Table 1: Comparison of the number of correctly classified samples in a leave-one-out-cross-validation study. We trained the classifiers on $M-1=71$ samples and tested them on the other samples. By repeating this over all testing samples, we try to estimate the accuracy of the proposed classifiers on samples that have not been seen during training. The results for the SVM and Boosting (of the weak decision stump learner) are taken from [3]. These results represent the current state-of-the-art in the literature on this widely used benchmark dataset.

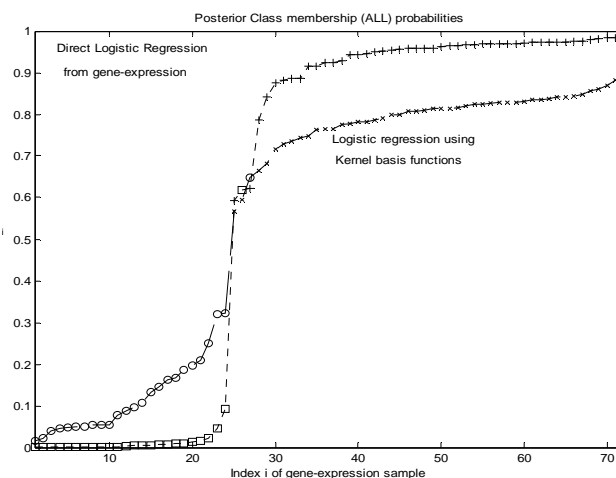


Figure 1: Comparison of posterior acute lymphoblastic leukemia (ALL) class membership probabilities from logistic regression directly on the gene expression levels, and from logistic regression using linear kernel basis functions. Direct logistic regression achieves a larger margin classifier. We trained and tested the classifier on the same set of 38 and 34 samples respectively used in [4]. We then arranged all the samples in order of their predicted ALL class membership probabilities. In the figure, \square and \circ denote samples from patients with acute myeloid leukemia (AML), and $+$ and \times denote samples from patients with acute lymphoblastic leukemia (ALL).