Joint classifier and feature optimization for cancer diagnosis using gene expression data

Seminar Data Mining in Datenbanken

Matthias Hübenthal

Fachbereich Mathematik und Informatik
Martin-Luther-Universität Halle-Wittenberg

27. März 2006

1. Einleitung

2. Formulierung des Problems

3. BAYES'scher Ansatz

4. MAP-Schätzung der Parameter

5. Zusammenfassung

6. Literatur

Einleitung: Motivation

- Ziel: Konstruktion von Klassifikatoren zur Krebsdiagnose
- Ansatz: Vergleich von Genexpressionsprofilen von Geweben mit unbekanntem Krebsstatus mit Datenbanken, die Expressionsprofile von Gewebeproben mit bekanntem Krebsstatus enthalten
- klassisches Anwendungsgebiet überwachter und unüberwachter Methoden der Mustererkennung
- die wohl populärsten basieren auf der Technik der support vector machines (SVM)
- zeigen recht gute Klassifikationsergebnisse; unter der Voraussetzung, dass zuvor eine möglichst gute Merkmalsauswahl erfolgte
- meist sind unzählige Gene mit dem Krebsstatus eines Gewebes stark korreliert
- aktuelle Forschungen ergaben aber, dass zur korrekten Diagnose der meisten Krebsarten die Expressionslevel von weniger als 10 Genen ausreichen
- weiteres Problem: jeder Klassifikator sollte gut generalisieren; dies wird mit Zunahme der Dimensionalität der Trainingsdaten schwieriger (curse of dimensionality)
- Krishnapuram, Carin und Hartemink schlagen zur Lösung beider Probleme in [5] den JCFO-Algorithmus vor $\Rightarrow$ BAYES'sche Verallgemeinerung der SVMs

Einleitung: Idee

- JCFO = Joint classifier and feature optimization
- Algorithmus identifiert simultan sowohl den optimalen nichtlinearen Klassifikator, als auch die optimale Menge an Genen, auf deren Basis die Diagnose gestellt wird
- in bisherigen Ansätzen versuchte man, beide Probleme getrennt von einander zu lösen
- hier: feature selection integraler Bestandteil des Klassifikatordesigns
- zunächst wird das Expressionslevel eines jeden Gens mit einem positiven Skalierungsfaktor versehen
- diese, sowie die Parameter der Basisfunktionen des Klassifikators werden MAP-geschätzt
- durch geschickte Wahl der Prior wird bei der Estimierung mittels EM-Algorithmus der größte Teil der Parameter auf 0 gesetzt
- die Merkmalsauswahl erfolgt also implizit; Gene mit Expressionslevel 0 sind für die Diagnose irrelevant
- der Klassifikator bleibt möglichst einfach

Formulierung des Problems: Trainingsdaten

- gegeben seien N Genexpressionsprofile $x^{(i)} = \left[x_1^{(i)}, \ldots, x_d^{(i)}\right] \in \mathbb{R}^d,\ i \in \{1, \ldots, N\}$
- enthalten jeweils Expressionslevel von d Genen innerhalb der betrachteten Gewebeprobe
- sei des Weiteren $y^{(i)}$ die Klassenzugehörigkeit des Profils $x^{(i)}$
- entsprechend lässt sich die Trainingsmenge D wie folgt definieren:

$$D = \left\{\left\langle x^{(i)}, y^{(i)}\right\rangle : x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \{0,1\}\right\}_{i=1}^{N}$$

- unter der Annahme einer durch α parametrisierten, funktionalen Beziehung $y = f_\alpha(x)$ zwischen x und y sind bzgl. der Trainingsdaten die optimalen Parameter α zu bestimmen
- gesucht ist also eine binäre Funktion $y = f_\alpha(.) : \mathbb{R}^d \to \{0,1\}$, die Elemente aus D den Klassen 0=gesund bzw. 1=krebsartig zuordnet (vgl. SVM)

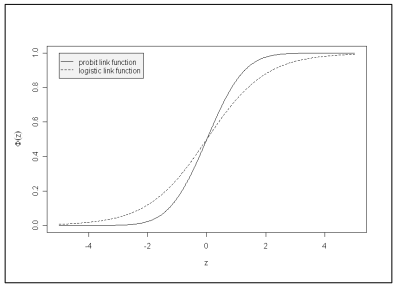
Formulierung des Problems: Klassenwahrscheinlichkeiten

- mit einer Funktion $y = g_\alpha(.) : \mathbb{R}^d \to [0,1]$ mit

$$P(y = 1|x) = g_\alpha(x) = \Phi\left(\beta^T h_\theta(x)\right)$$

  lässt sich nun einem Merkmalsvektor x die Wahrscheinlichkeit zuordnen, der Klasse 1 anzugehören
- dabei sind $\theta = [\theta_1, \ldots, \theta_d]^T$ die für die Merkmalsauswahl zuständigen Parameter der Unterscheidungsfunktionen $h_\theta(.)$
- β beeinflusst die Auswahl der Basisfunktionen des zu konstruierenden nichtlinearen Klassifikators
- die Parameter α der Klassifikationsfunktion $g_\alpha(.)$ haben entsprechend folgende Form: $\alpha = \left[\beta^T, \theta^T\right]^T$
- als Entscheidungsregel $\Phi(z)$ soll hier die kumulative Standardnormalverteilung verwendet werden

Formulierung des Problems: Kopplungsfunktionen als Entscheidungsregeln

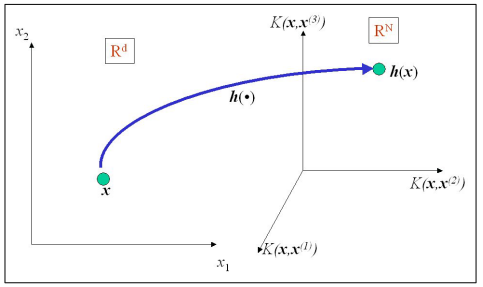


- engl.: standard gaussian cumulative distribution function oder standard normal CDF, auch: probit link function:

$$\Phi(z) = \int_{-\infty}^{z} N(x|0,1)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(\frac{-x^2}{2}\right) dx$$

- Alternative: logistische Kopplungsfunktion o.ä.
- um über die Klassenzugehörigkeit eines Merkmalsvektors entscheiden zu können, ist ein Schwellwert vorzugeben

Formulierung des Problems: Unterscheidungsfunktion I



- verwenden hier Kernklassifikatoren, d.h. betrachten N mit θ parametrisierte Kernoperatoren

$$h_\theta(x) = \left[1, K_\theta\left(x, x^{(1)}\right), \ldots, K_\theta\left(x, x^{(N)}\right)\right]^T$$

- definieren Abbildung zwischen dem Merkmalsraum $\subseteq \mathbb{R}^d$ und dem durch die Basisfunktionen aufgespannten Kernraum $\subseteq \mathbb{R}^N$

Formulierung des Problems: Unterscheidungsfunktion II

- für $g_\alpha(.)$ gilt demnach:

$$g_\alpha(x) = \Phi\left(\beta^T h_\theta(x)\right) = \Phi\left((\beta_0, \beta_1, \ldots, \beta_N) \begin{pmatrix} 1 \\ K_\theta\left(x, x^{(1)}\right) \\ \vdots \\ K_\theta\left(x, x^{(N)}\right) \end{pmatrix}\right)$$

- die Auswahl der Basisfunktionen erfolgt durch ein $\beta \in \mathbb{R}^{M=N+1}$
- H_θ sei die Designmatrix des Klassifikators mit

$$H_\theta = \left[h_\theta^T\left(x^{(1)}\right), \ldots, h_\theta^T\left(x^{(N)}\right)\right]^T = \begin{pmatrix} 1 & K_\theta\left(x^{(1)}, x^{(1)}\right) & \cdots & K_\theta\left(x^{(1)}, x^{(N)}\right) \\ \vdots & \vdots & & \vdots \\ 1 & K_\theta\left(x^{(N)}, x^{(1)}\right) & \cdots & K_\theta\left(x^{(N)}, x^{(N)}\right) \end{pmatrix}$$

- $K_\theta\left(x^{(i)}, x^{(j)}\right)$ ist dabei ein nichtlineares Maß für die Ähnlichkeit zwischen einer neuen, ungelabelten Probe $x^{(i)}$ und der gelabelten Trainingsprobe $x^{(j)}$

Formulierung des Problems: Unterscheidungsfunktion III

- es finden hier polynomielle Kernoperatoren n-ten Grades Anwendung, d.h.

$$K_\theta\left(x^{(i)}, x^{(j)}\right) = \left(1 + \sum_{l=1}^{d} \theta_l x_l^{(i)} x_l^{(j)}\right)^n$$

- die Auswahl bzw. Wichtung der Expressionslevel der für die Diagnose relevanten Gene erfolgt mittels θ
- H_θ hat entsprechend folgende Form:

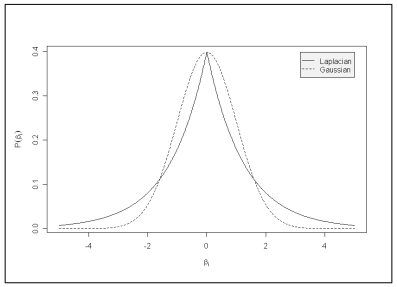
$$H_\theta = \begin{pmatrix} 1 & \left(1 + \sum_{l=1}^{d} \theta_l x_l^{(1)} x_l^{(1)}\right)^n & \cdots & \left(1 + \sum_{l=1}^{d} \theta_l x_l^{(1)} x_l^{(N)}\right)^n \\ \vdots & \vdots & & \vdots \\ 1 & \left(1 + \sum_{l=1}^{d} \theta_l x_l^{(N)} x_l^{(1)}\right)^n & \cdots & \left(1 + \sum_{l=1}^{d} \theta_l x_l^{(N)} x_l^{(N)}\right)^n \end{pmatrix}$$

- alternative Ähnlichkeitsmaße: radiale Basisfunktionen o.ä.

BAYES'scher Ansatz: sparse Priors

- Klassifikatordesign und Merkmalsauswahl können durch Estimierung der Parameter β und θ aus gegebenen Daten erfolgen
- Ziel: konstruiere Klassifikator mit möglichst wenigen $\beta_i \neq 0$ und $\theta_i \geq 0$
- Ansatz: verwende bei der MAP-Schätzung so genannte sparse Priors
- $P(\beta_i)$ bzw. $P(\theta_i)$ seien entsprechend groß, wenn β_i bzw. θ_i exakt 0
- werden sehr schnell kleiner, wenn $\beta_i \neq 0$ bzw. $\theta_i \geq 0$
- irrelevante oder redundante Komponenten werden aus den Parametervektoren entfernt
- im Falle von β führt dies zu einer Vereinfachung der Klassifikationsfunktion und dadurch zu einer Verbesserung der Generalisierbarkeit des Klassifikators
- ein analog geschätztes θ verringert deutlich die Dimensionalität des Merkmalsraumes
- $N(x|0, \sigma^2)$ als Prior weit verbreitet, aber ungeeignet
- verwenden Laplaceverteilung bzw. eine BAYES'sche Dekomposisition dieser Verteilung

BAYES'scher Ansatz: Laplaceverteilung als Prior



- Laplaceverteilung

$$P(\beta|\eta) = \prod_{i=1}^{M} \frac{\eta}{2} \exp\left(-\eta\,|\beta_i|\right) = \left(\frac{\eta}{2}\right)^M \exp\left(-\eta\,\|\beta_i\|\right)$$

- im Vergleich zur Normalverteilung ist Unterschied zwischen $P(0)$ und $P(\beta_i)$ für kleine β_i wesentlich großer
- d.h. bei der Maximierung des Posteriors werden $\beta_i = 0$ explizit favorisiert

BAYES'scher Ansatz: Dekomposition des Laplacepriors I

- zur Vermeidung numerischer Instabilitäten Verwendung einer BAYES'schen Interpretation der Laplaceverteilung
- Zerlegung des Modells in

$$P(\beta_i|\tau_i) = N(\beta_i|0,\tau_i) \quad \text{und} \quad P(\tau_i|\gamma_1) = \frac{\gamma_1}{2}\exp\left(-\frac{\gamma_1\tau_i}{2}\right)$$

  mit $\tau_i \geq 0$
- betrachten Gaußprior mit Mittelwert 0 und Varianz τ_i, die einer Exponentialverteilung (Hyperprior) mit Parameter γ_1 folgt
- durch BAYES und Marginalisierung erhält man:

$$\begin{aligned} P(\beta_i|\gamma_1) &= \int_0^\infty P(\beta_i,\tau_i|\gamma_1)d\tau_i \\ &= \int_0^\infty P(\beta_i|\tau_i,\gamma_1)P(\tau_i|\gamma_1)d\tau_i = \frac{\sqrt{\gamma_1}}{2}\exp\left(-\sqrt{\gamma_1}\,|\beta_i|\right) \end{aligned}$$

  also wieder eine Laplaceverteilung

BAYES'scher Ansatz: Dekomposition des Laplacepriors II

- prinzipiell analoges Vorgehen für θ, aber $\theta_i \geq 0$
- große Ähnlichkeit der Expressionslevel zweier Gene impliziert nicht unbedingt große Ähnlichkeit der zugehörigen Expressionsprofile
- sie kann aber keinesfalls eine Verringerung der Ähnlichkeit der Profile zur Folge haben
- daher gilt:

$$P(\theta_i|\rho_i) = \begin{cases} N(\theta_i|0,\rho_i) & \text{falls } \theta_i \geq 0 \\ 0 & \text{falls } \theta_i < 0 \end{cases} \quad \text{sowie} \quad P(\rho_i|\gamma_2) = \frac{\gamma_2}{2}\exp\left(-\frac{\gamma_2\rho_i}{2}\right)$$

  mit $\rho_i \geq 0$
- der Laplaceprior für θ_i lässt sich also auch in der Form

$$P(\theta_i|\gamma_2) = \begin{cases} \sqrt{\gamma_2}\exp\left(-\sqrt{\gamma_2}\theta_i\right) & \text{falls } \theta_i \geq 0 \\ 0 & \text{falls } \theta_i < 0 \end{cases}$$

  schreiben

MAP-Schätzung der Parameter: versteckte Variablen

- nach Vereinbarung der Prior lassen sich nun mit Hilfe des EM-Algorithmus die Parameter θ und β bestimmen, für die die a-posteriori-Verteilung (lokal) maximal wird
- sei zusätzlich zu den Hyperparametern ρ und τ durch

$$z(x,\theta,\beta) = \beta^T h_\theta(x) + \omega$$

  eine weitere versteckte Variable definiert (ω standardnormalverteilt)
- für den Klassifikator

$$y = \begin{cases} 1 & \text{falls } z(x,\theta,\beta) \geq 0 \\ 0 & \text{falls } z(x,\theta,\beta) < 0 \end{cases}$$

  erhält man wegen

$$P(y=1|x) = P(z(x,\theta,\beta) \geq 0) = \Phi\left(\beta^T h_\theta(x)\right)$$

  das zuvor definierte Modell
- zur Erinnerung: kumulative Standardnormalverteilung Φ, angewendet auf Kernklassifikatoren $h_\theta(x)$

MAP-Schätzung der Parameter: Posterior

- gegeben seien die Daten

$$D = \left\{\left\langle x^{(i)}, y^{(i)}\right\rangle, \ldots, \left\langle x^{(N)}, y^{(N)}\right\rangle\right\}$$

  sowie die versteckten Variablen z, τ und ρ
- dann hat die log-a-posteriori-Verteilung folgende Form:

$$\begin{aligned} \log P(\beta,\theta|z,\tau,\rho,y) &\propto \log P(z|\beta,\theta) + \log P(\beta|\tau) + \log P(\theta|\rho) \\ &\propto -\|H\beta - z\|^2 - \beta^T T\beta - \theta^T R\theta \\ &\propto -z^T z - \beta^T H^T(H\beta - 2z) - \beta^T T\beta - \theta^T R\theta \end{aligned}$$

- dabei sind

$$T = \begin{pmatrix} \tau_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \tau_M^{-1} \end{pmatrix} \quad \text{und} \quad R = \begin{pmatrix} \rho_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho_d^{-1} \end{pmatrix}$$

MAP-Schätzung der Parameter: E-Schritt I

- für gegebene y, $\hat{\beta}^{(t)}$ und $\hat{\theta}^{(t)}$ (hier als Λ notiert), ist der erwartete logarithmierte Posterior der Parameter β und θ zu berechnen

$$\begin{aligned} Q\left(\beta, \theta | \hat{\beta}^{(t)}, \hat{\theta}^{(t)}\right) &= \mathrm{E}\left[\log P(\beta, \theta | z, \tau, \rho, y) \mid \Lambda\right] \\ &\propto \mathrm{E}\left[-z^T z - \beta^T H^T (H\beta - 2z) - \beta^T T \beta - \theta^T R \theta \mid \Lambda\right] \end{aligned}$$

- da die Maximierung bzgl. β und θ erfolgt, lässt sich dies zu

$$-\beta^T H^T H \beta + 2\beta^T H^T \mathrm{E}[z|\Lambda] - \beta^T \mathrm{E}[T|\Lambda]\,\beta - \theta^T \mathrm{E}[R|\Lambda]\,\theta$$

  vereinfachen
- $\mathrm{E}[z|\Lambda]$, $\mathrm{E}[T|\Lambda]$ und $\mathrm{E}[R|\Lambda]$ sind dabei analytisch bestimmbar

MAP-Schätzung der Parameter: E-Schritt II

- sei $\upsilon_i = \mathrm{E}\left[z^{(i)} | \Lambda\right]$, dann ist

$$\upsilon_i = \begin{cases} h^T\left(x^{(i)}\right)\hat{\beta}^{(t)} + \frac{N\left(h^T\left(x^{(i)}\right)\hat{\beta}^{(t)} \mid 0,1\right)}{1-\Phi\left(-h^T\left(x^{(i)}\right)\hat{\beta}^{(t)}\right)} & \text{falls } y^{(i)} = 1 \\ h^T\left(x^{(i)}\right)\hat{\beta}^{(t)} - \frac{N\left(h^T\left(x^{(i)}\right)\hat{\beta}^{(t)} \mid 0,1\right)}{\Phi\left(-h^T\left(x^{(i)}\right)\hat{\beta}^{(t)}\right)} & \text{falls } y^{(i)} = 0 \end{cases}$$

- sei $\omega_i = \mathrm{E}\left[\tau_i^{-1} \mid y, \hat{\beta}_i^{(t)}, \gamma_1\right]$, dann ist

$$\omega_i = \frac{\int_0^\infty \tau_i^{-1} P\left(\tau_i | \gamma_1\right) P\left(\hat{\beta}_i^{(t)} | \tau_i\right) d\tau_i}{\int_0^\infty P\left(\tau_i | \gamma_1\right) P\left(\hat{\beta}_i^{(t)} | \tau_i\right) d\tau_i} = \gamma_1 \left|\hat{\beta}_i^{(t)}\right|^{-1}$$

- analog ist $\delta_i = \mathrm{E}\left[\rho_i^{-1} \mid y, \hat{\beta}^{(t)}, \hat{\theta}_i^{(t)}, \gamma_2\right] = \gamma_2 \left(\hat{\theta}_i^{(t)}\right)^{-1}$ für $\theta_i > 0$

MAP-Schätzung der Parameter: M-Schritt

- $\upsilon = [\upsilon_1, \ldots, \upsilon_N]^T$ entspricht den erwarteten Distanzen zwischen den Proben x und der durch den Klassifikator beschriebenen Hyperebene im Kernraum
- mit Hilfe der erwarteten Hyperparameter ω und δ definieren wir zudem in Analogie zu T und R

$$\Omega = \begin{pmatrix} \omega_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_M^{-1} \end{pmatrix} \quad \text{und} \quad \Delta = \begin{pmatrix} \delta_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \delta_d^{-1} \end{pmatrix}$$

- durch Einsetzen in die Funktion Q erhält man

$$Q\left(\beta, \theta | \hat{\beta}^{(t)}, \hat{\theta}^{(t)}\right) = -\beta^T H^T H \beta + 2\beta^T H^T \upsilon - \beta^T \Omega \beta - \theta^T \Delta \theta$$

- eine analytische Maximierung von Q bzgl. β und θ ist auf Grund des nichtlinearen Charakters der Designmatrix H nicht möglich
- der M-Schritt zur Estimierung der neuen Parameter $\hat{\beta}^{(t+1)}$ und $\hat{\theta}^{(t+1)}$ ist demnach numerisch durchzuführen

Zusammenfassung

- von Krishnapuram, Carin und Hartemink vorgeschlagener JCFO-Algorithmus: Versuch, existierende Algorithmen zur Krebsdiagnose in Vorhersagegenauigkeit zu übertreffen
- dabei Merkmalsauswahl integraler Bestandteil des Klassifikatordesigns
- es wird ein BAYES'scher Ansatz zur Parameterestimierung verwendet
- durch sparse Priors werden Koeffizienten zur Wichtung der Expressionslevel sowie der (polynomiellen) Kernbasisfunktionen bei diagnostischer Irrelevanz auf 0 gesetzt
- zur Parameterschätzung wird der iterative EM-Algorithmus verwendet
- die Erwartungswerte der versteckten Variablen sind analytisch bestimmbar
- die Maximierung der Q-Funktion efolgt numerisch
- betrachtete Testdatensätze: zur Unterscheidung zwischen akuter myeloischer (AML) und lymphatischer Leukämie (ALL) [4] bzw. zur Vorhersage von Dickdarmkrebs [1]
- LOOCV ergab: andere bekannte Algorithmen zeigen größere Klassifikationsfehler
- neben bekannten Markern zur Krebsdiagnose wurde Kandidatengene für weitere klinische Untersuchungen identifiziert

Literatur I

[1] Alon, U., N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack und A. J. Levine: *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*.
Proc. Natl. Acad. Sci. USA, 96:6745–6750, 1999.

[2] Figueiredo, M. A. T. und A. K. Jain: *Bayesian learning of sparse classifiers*.
In: *CVPR '01: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, S. 35–41, 2001.

[3] Furey, T. S., N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer und D. Haussler: *Support vector machine classification and validation of cancer tissue samples using microarray expression data*.
Bioinformatics, 16(10):906–914, 2000.

[4] Golub, T. R., D. K. Slonim, P. Tamayo, C. H. M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, und E. S. Lander: *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*.
Science, 286(5439):531–537, 1999.

Literatur II

[5] Krishnapuram, B., L. Carin und A. J. Hartemink: *Joint classifier and feature optimization for cancer diagnosis using gene expression data*.
In: *RECOMB '03: Proceedings of the seventh annual international conference on Research in computational molecular biology*, S. 167–175, 2003.

[6] Tipping, M. E.: *Sparse Bayesian Learning and the Relevance Vector Machine*.
Journal of Machine Learning Research, 1:211–244, 2001.