Постановка задачи двуклассового распознавания

1. Описание объекта. Пространство признаков.
2. Обучающее множество. Truth информация.
3. Решающее правило. Выход решающего правила. Примеры решающих правил: линейное правило, ближайший сосед.
4. Ошибка решающего правила. Веса ошибок.
5. Формальная постановка задачи распознавания. Распознаватель – это множество решающих правил + критерий ошибки. Обучение – выбор правила с наилучшим (на обучающем множестве) значением критерия.

I. Зачем нужно обучение?

1. Ошибка на обучающем множестве. Ошибка на тестовом множестве.
2. Цель распознавания – уменьшить ошибку на тестовом множестве. Обобщение. Вероятностные утверждения об ошибке.
3. Основная гипотеза индуктивного обучения: если сложность множества решающих правил не велика, то с высокой вероятностью ошибка на обучающем множестве будет мало отличаться от ошибки на тестовом множестве.
4. Оказывается, что можно определять меры сложности C множества решающих правил так, чтобы доказывать неравенства типа $P(|Err_{test}-Err_{train}| > d) < f(C,n,d)$, где $f \to 0$ при n, стремящемся к бесконечности.

II. Зачем нужно обучение?

`

1. С заданной вероятностью можно написать, что

$Err_{test} < Err_{train} + f(C, n)$. К сожалению, уменьшив Err_{train} с помощью построения более сложных правил, мы увеличиваем C и f(C, n).

2. Чем больше мы знаем об истинном правиле, тем более простое множество правил, обеспечивающее малую ошибку, можно построить.

Распознаватель «Кора».

1. Пространство признаков – логические утверждения. Симптомы. 3 значения синдрома.
2. Множество решающих правил – конъюнкции – синдромы.
3. Отбор синдромов по частотам. Экзамен – голосование. Возможное усложнение – веса.
4. Естественная мера сложности – количество оцениваемых синдромов + количество отобранных синдромов.

I. Что можно надежно утверждать об экспрессии генов?

1. Резко выраженная дифференциальная экспрессия.

Мы видели, что после нормализации и сложной обработки можно достаточно надежно заметить, что экспрессия изменилась в 2 и более раза. Это значит, что можно строить синдромы типа: 1, если $E_g > a$, 0, если $E_g < b$, не определено, в остальных случаях, при условии, что $a > 2b$.

2. Утверждения об экспрессии, не требующие нормализации.

Монотонно возрастающие функции.

А) Модель, не учитывающая неспецифической гибридизации

Интенсивность j-ого зонда гена g на k-том чипе

$I(g, j, k) = C_k(f(j)E(g))$, где $C_k()$ – монотонное нелинейное влияние k-ого чипа,

$f(j)$ – эффективность j-ого зонда, $E(g)$ – экспрессия гена g.

Из монотонности следует, что $I(g_1, j_1, k) > I(g_2, j_2, k) \Leftrightarrow E(g_1)/E(g_2) > f(j_2)/f(j_1)$

Важно, что $f(j_2)$ и $f(j_1)$ не меняются от чипа к чипу. Поэтому, если $I(g_1, j_1, k) > I(g_2, j_2, k)$ выполняется часто на одном классе и редко на другом, то это хороший симптом.

II. Что можно надежно утверждать об экспрессии генов?

Б) Модель, учитывающая неспецифическую гибридизацию.

$I(G, j, k) = C_k(\Sigma_g f(j,g)E(g))$,

Здесь I(G, j, k) – интенсивность для зонда j гена G, а f(j, g) – эффективность этого зонда для гена g.

Аналогично предыдущему $I(g_1, j_1, k) > I(g_2, j_2, k) \Leftrightarrow \Sigma_g f(j_1,g)E(g) > \Sigma_g f(j_2,g)E(g)$

Последнее неравенство формально зависит от экспрессий всех генов и поэтому может быть очень неустойчивым. Однако, поскольку все f по прежнему не зависят от чипа, если оно выполняется достаточно часто на одном классе и достаточно редко на другом, это хороший симптом. Поскольку Affymetrix специально выбирал олигонуклеотиды так, чтобы снизить влияние неспецифической гибридизации, то есть надежда, что в достаточно большой части случаев f таковы, что эта модель сводится к предыдущей, и, значит выполняется достаточно часто.

III. Что можно надежно утверждать об экспрессии генов?

B) Как выразить утверждение “высокая экспрессия гена” ?

Мы поняли, что утверждения о соотношений экспрессий двух генов могут быть выражены способом, не требующим нормализации. Но естественно предполагать, что не менее, а может и более важными являются утверждения об экспрессии конкретного гена типа “при раке данный ген сильно экспрессирован”. Прямое сравнение экспрессии с порогом невозможно без нормализации. Однако мы можем заменить сравнение с порогом на сравнение с квантилем. То есть вместо утверждения “данный ген сильно экспрессирован” можно использовать утверждение типа “данный ген больше ¾ генов на этом чипе”.

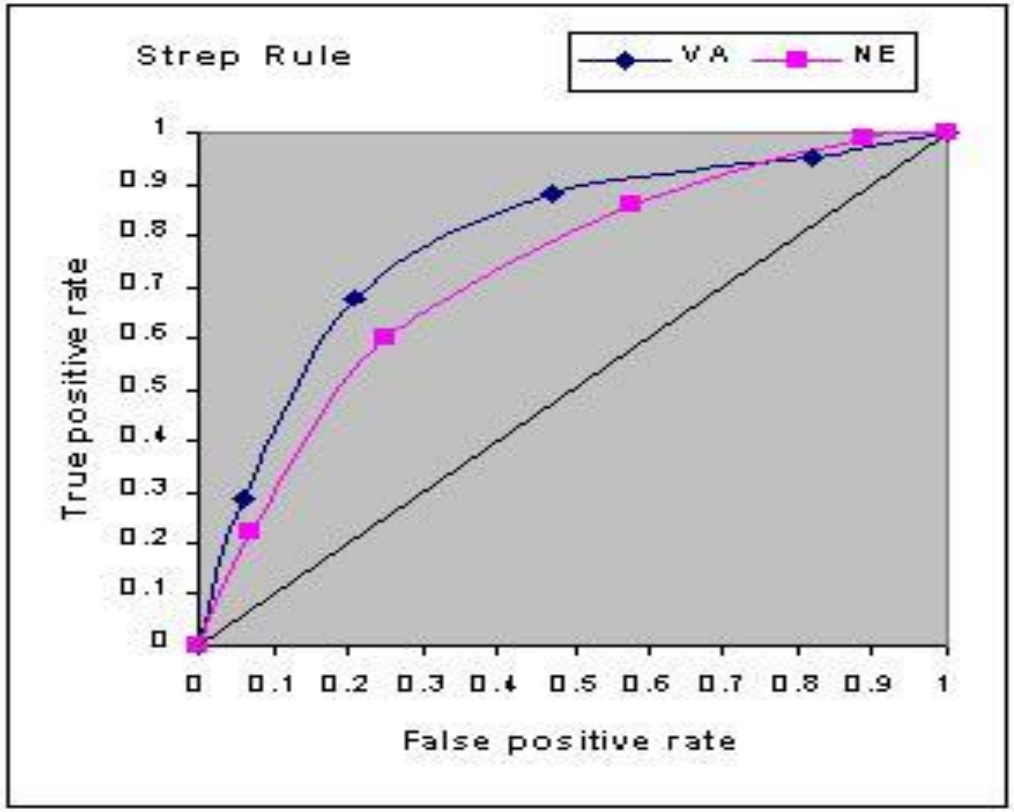
Как измерять ошибку распознавания?

1. Ошибка на обучающем множестве всегда занижена.
2. Лучший способ – разбиение на обучающее и тестовое множество. Еще лучше – разбиение на обучающее, верификационное и тестовое множества. На верификационном подбирают параметры обучения, а само обучение проводят на обучающем.
3. Скользящее обучение (leave-one-out)
4. Уверенность (конфиденс) ответа. Реджектная кривая.

ROC curve

FPR=1-specificity

TPR = sensitivity



Медицина, основанная на симптомах и медицина, основанная на примерах

1. Мера похожести и метод ближайшего соседа.
2. Автоматический выбор типичных представителей.
3. SVM как обобщение метода ближайшего соседа.