Метод недиагональной регуляризации в байесовском обучении

©2009г. Д. П. Ветров*, Д. А. Кропотов**, О. В. Курчин*
* 119992 Москва, Ленинские Горы, МГУ ф-т ВМиК;
** 119333 Москва, ул.Вавилова, д.40, ВЦ РАН
vetrovd@yandex.ru; dkropotov@yandex.ru; okurchin@gmail.com

Поступила в редакцию 01.06.2009

В данной статье предложен новый тип процедуры регуляризации в байесовских методах классификации. В основе рассматриваемого подхода лежит приведение матрицы гессиана логарифма функции правдоподобия к диагональному виду с последующей независимой регуляризацией вдоль ее собственных векторов. Подобное преобразование позволяет представить многомерный интеграл в выражении для правдоподобия модели (обоснованности) в виде произведения одномерных интегралов, каждый из которых зависит от своего коэффициента регуляризации (структурного параметра). Благодаря этому, процесс автоматического определения коэффициентов регуляризации сходится за одну итерацию. В качестве примеров показано использование данного подхода для случаев гауссовского и лапласовского априорных распределений над параметрами (весами) модели. В обоих случаях качество полученного решающего правила сравнимо с тем, которое дает метод релевантных векторов (MPB), при этом время обучения существенно меньше, а решающие правила получаются более разреженными.

Ключевые слова: распознавание образов, байесовское обучение, обоснованность модели, недиагональная регуляризация

1. Введение

В последние годы методы байесовского обучения получили достаточно широкое применение в области решения задач распознавания и классификации (см. [1, 2]). В рамках этой концепции параметры классификатора называются *весами*, а параметры, определяющие семейство возможных классификаторов, то есть задающие семейство априорных распределений вероятности над весами - *структурными параметрами* (или параметрами модели). На сегодняшний день существует два подхода к определению значений структурных параметров.

Один из подходов основан на автоматическом определении релевантности (AOP) и приводит к максимизации правдоподобия модели (обоснованности) (см. [3]). Типичным примером данного подхода является метод релевантных векторов (MPB) (см. [4]), в котором каждому весу соответствует свой собственный

структурный параметр, итерационно настраиваемый в процессе обучения. Данный метод является примером разреженного байесовского классификатора, большинство весов которого стремятся к нулю. Основным недостатком является тот факт, что на практике его можно использовать только при гауссовском априорном распределении весов, поскольку в этом случае интеграл в выражении для обоснованности может быть вычислен в явном виде. В то же время известно, что, например, распределение Лапласа сильнее поощряет получение разреженных решений, при этом часть весов полученного решения будет в точности равняться нулю (см. [5]), однако его использование в методе релевантных векторов невозможно в силу высокой вычислительной сложности численного интегрирования при подсчете обоснованности модели.

Альтернативным подходом является усреднение по структурному параметру (маргинализация), приводящее к выражению априорной вероятности для весов, не зависящему от структурных параметров, с последующей максимизацией произведения правдоподобия обучающей выборки на полученную при усреднении априорную вероятность.

Впервые данный подход был предложен для использования именно вместе с априорным распределением Лапласа (см. [5]) и впоследствии часто использовался для обработки медицинских данных большой размерности (см. [6]), а также для задач с несколькими классами (см. [7]). К сожалению, данный подход обладает существенным недостатком: часть полезных свойств может быть утеряна в процессе маргинализации исходного семейства классификаторов. Так, в случае линейных моделей, мы получаем классификатор, обладающий несколькими экстремумами (зачастую весьма большим количеством) (см. [8]).

В данной статье показано, что подход на основе правдоподобия модели (обоснованности) может применяться вне зависимости от типа априорного распределения вероятности над весами. Для того, чтобы это сделать приведем матрицу гессиана правдоподобия к диагональному виду, перейдя от естественного базиса в пространстве весов $\boldsymbol{w}$ к базису из собственных векторов полученной матрицы (пространство весов $\boldsymbol{u}$), и рассмотрим независимую регуляризацию вдоль новых координатных осей. Для каждой из осей, определяемой собственным вектором, зададим априорное распределение на u_i, а значения соответствующих гиперпараметров найдем с помощью АОР-подхода. После указанных выше преобразований выражение для обоснованности может быть представлено в виде произведения независимых одномерных интегралов, каждый из которых зависит от своего структурного параметра.

Этот подход является достаточно общим, поскольку не зависит от конкретного вида априорного распределения. Единственное требование, необходимое для его корректного применения, состоит в том, чтобы каждое априорное распределение регуляризовало свой вес u_i независимо.

Указанное преобразование факторизует обоснованность, то есть позволяет представить ее в виде произведения одномерных интегралов, каждый их которых может быть оптимизирован отдельно. Это в свою очередь позволяет существенно ускорить процесс обучения. При этом количество релевантных собственных векторов в большинстве случаев оказывается меньше, чем количество ненулевых весов в МРВ.

2. Обоснованность модели

2.1. Общая постановка задачи классификации в рамках байесовского подхода

Рассмотрим классическую постановку задачи классификации с учителем. Пусть $\widetilde{\mathcal{X}}$ - множество всевозможных описаний объектов, $\widetilde{\mathcal{T}}$ - множество меток (или наименований) классов (*генеральная совокупность в пространстве* $\widetilde{\mathcal{X}} \times \widetilde{\mathcal{T}}$). Существует неизвестное отображение вида $\mathcal{A} : \widetilde{\mathcal{X}} \to \widetilde{\mathcal{T}}$, значения которого известны только на объектах конечной *обучающей выборки* $\mathcal{D} = \{(\boldsymbol{x}_1, t_1), ..., (\boldsymbol{x}_n, t_n)\} = \mathcal{X} \times \mathcal{T}$, где $\mathcal{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, $\mathcal{T} = \{t_1, ..., t_n\}$. Требуется построить алгоритм $\mathcal{A}^* : \widetilde{\mathcal{X}} \to \widetilde{\mathcal{T}}$, для всех $\boldsymbol{x} \in \widetilde{\mathcal{X}}$.

В дальнейшем мы будем рассматривать один из наиболее распространенных типов задач классификации, а именно *задачу классификации на два класса с учителем*. В данной задаче считается, что $\widetilde{\mathcal{T}} = \{-1; 1\}$, $\widetilde{\mathcal{X}} = \widetilde{\mathcal{X}}_1 \cup \widetilde{\mathcal{X}}_2$, где $\widetilde{\mathcal{X}}_1$ – множество описаний объектов, соответствующих метке 1, а $\widetilde{\mathcal{X}}_2$ - множество описаний объектов, соответствующих метке -1. Для определенности будем считать, что описание объекта представляет из себя $d-$мерный вектор $\boldsymbol{x}_i = (x_i^1, ..., x_i^d)$ с вещественными координатами. Таким образом, необходимо построить алгоритм классификации (классификатор) $\mathcal{A}^*$, такой что $\mathcal{A}^*(\boldsymbol{x}) = 1$, при $\boldsymbol{x} \in \widetilde{\mathcal{X}}_1$ и $\mathcal{A}^*(\boldsymbol{x}) = -1$, при $\boldsymbol{x} \in \widetilde{\mathcal{X}}_2$.

Перейдем теперь к *вероятностной постановке задачи классификации*, которая формулируется следующим образом: пусть имеется множество объектов $\widetilde{\mathcal{X}}$ и множество имен классов $\widetilde{\mathcal{T}}$; пусть также имеется набор прецедентов $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^n = (\mathcal{X}, \mathcal{T})$, выбранных случайно и независимо из неизвестного распределения на генеральной совокупности $p(\boldsymbol{x}, t) = p(t|\boldsymbol{x})p(\boldsymbol{x})$. Другими словами, множество объектов обучения представляет собой совокупность d-мерных векторов с вещественными координатами - векторов признаков $\boldsymbol{x} \in \mathbb{R}^d$ и для каждого из которых известна метка класса - значение переменной, принимающей одно из двух значений: $t \in \{-1, +1\}$. Требуется восстановить неизвестную плотность распределения $p(t|\boldsymbol{x})$.

Для описания общей схемы поиска $p(t|\boldsymbol{x})$ введем понятие модели.

Определение 1. *Вероятностной моделью* алгоритмов восстановления плотностей назовем тройку

$$< \Omega, p(\widetilde{T}|\widetilde{X}, \boldsymbol{w}), p(\boldsymbol{w}) >,$$

где $\Omega = \{\boldsymbol{w}\}$ – множество допустимых значений параметров плотностей распределения, $p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}) = \prod_{i=1}^n p(t_i|\boldsymbol{x}_i, \boldsymbol{w})$ - функция правдоподобия выборки $(\mathcal{T}, \mathcal{X})$ при фиксированном значении $\boldsymbol{w}$ и $p(\boldsymbol{w})$ - априорное распределение на $\boldsymbol{w}$.

Определение 2. Назовем *вероятностным классификатором* некоторую функцию, определяемую значениями параметров $\boldsymbol{w}$, аргументом которой является вектор признаков $\boldsymbol{x}$, возвращающую значения апостериорных плотностей для каждого из классов: $p(-1|\boldsymbol{x}, \boldsymbol{w})$ и $p(+1|\boldsymbol{x}, \boldsymbol{w})$. Таким образом, при заданной модели, вектор $\boldsymbol{w}$ полностью определяет классификатор.

Зададим множество возможных классификаторов априорным распределением над $\boldsymbol{w}$: $p(\boldsymbol{w}|\boldsymbol{\alpha})$, где $\boldsymbol{\alpha} \in \mathcal{A}$ – вектор *структурных параметров* (параметров, определяющих семейство возможных классификаторов; параметров модели).

Определение 3. Назовем *байесовской оценкой вектора весов* классификатора $\boldsymbol{w}$ значение $\boldsymbol{w}_{MP}$, получаемое с помощью правила максимизации апостериорной вероятности

$$\boldsymbol{w}_{MP}(\boldsymbol{\alpha}) = \boldsymbol{w}_{MP} = \arg\max_{w} p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{\alpha})$$

при заданном значении $\boldsymbol{\alpha}$. Использование данного правила эквивалентно использованию аддитивного регуляризатора при оптимизации логарифма апостериорной вероятности.

Байесовский подход подразумевает, что решение принимается на основе результатов голосования, взвешенных по всему множеству возможных классификаторов в рамках одной модели и, в случае нескольких возможных моделей, по всему множеству моделей. После чего, апостериорная вероятность класса нового объекта $\boldsymbol{x}$ может быть записана в виде

$$p(t|\boldsymbol{x}, \mathcal{T}, \mathcal{X}) = \int_{\mathcal{A}} \int_{\mathcal{W}(\boldsymbol{\alpha})} p(t|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\alpha}) p(\boldsymbol{w}, \boldsymbol{\alpha}|\mathcal{T}, \mathcal{X}) d\boldsymbol{w} d\boldsymbol{\alpha} = \int_{\mathcal{A}} \int_{\mathcal{W}(\boldsymbol{\alpha})} p(t|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\alpha}) p(\boldsymbol{w}|\mathcal{T}, \mathcal{X}, \boldsymbol{\alpha}) p(\boldsymbol{\alpha}|\mathcal{T}, \mathcal{X}) d\boldsymbol{w} d\boldsymbol{\alpha}, \quad (2.1)$$

где $\mathcal{W}(\boldsymbol{\alpha})$ – множество допустимых векторов $\boldsymbol{w}$ при заданном значении вектора структурных параметров $\boldsymbol{\alpha}$, а $\mathcal{A}$ – множество допустимых векторов структурных параметров $\boldsymbol{\alpha}$.

Так как функцию правдоподобия считаем независимой от $\boldsymbol{\alpha}$, то выражение для апостериорной вероятности может быть записано в виде:

$$p(t|\boldsymbol{x}, \mathcal{T}, \mathcal{X}) = \int_{\mathcal{A}} \int_{\mathcal{W}(\boldsymbol{\alpha})} p(t|\boldsymbol{x}, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{\alpha}, \mathcal{T}, \mathcal{X}) p(\boldsymbol{\alpha}|\mathcal{T}, \mathcal{X}) d\boldsymbol{w} d\boldsymbol{\alpha}. \quad (2.2)$$

Ключевой проблемой при решении задач классификации в рамках байесовского подхода является вычисление многомерного интеграла (2.1).

2.2. Понятие обоснованности модели

Выражение (2.1) может быть упрощено с помощью аппроксимации $p(\boldsymbol{\alpha}|\mathcal{T}, \mathcal{X})$, предложенной МакКаем в работе [3]. Для приближения указанной апостериорной вероятности над $\boldsymbol{\alpha}$ он предложил использовать $\delta(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{ME})$, где $\boldsymbol{\alpha}_{ME} = \arg\max E(\boldsymbol{\alpha})$, а $E(\boldsymbol{\alpha})$ – правдоподобие модели (*обоснованность*).

Определение 4. Назовем *правдоподобием модели (обоснованностью)* величину, вычисляемую следующим образом:

$$E(\boldsymbol{\alpha}) = p(\mathcal{T}|\mathcal{X}, \boldsymbol{\alpha}) = \int_{\mathcal{W}(\boldsymbol{\alpha})} p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w}. \quad (2.3)$$

Использование данного способа аппроксимации, позволяет приблизить выражение (2.1) следующим образом:

$$p(t|\boldsymbol{x}, \mathcal{T}, \mathcal{X}) \approx \int_{\mathcal{W}(\boldsymbol{\alpha}_{\mathcal{ME}})} p(t|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\alpha}_{ME}) p(\boldsymbol{w}|\mathcal{T}, \mathcal{X}, \boldsymbol{\alpha}_{ME}) d\boldsymbol{w}. \quad (2.4)$$

2.3. Метод релевантных векторов

Одним из наиболее известных методов, допускающих использование принципа максимизации обоснованности для определения значений структурных параметров, является метод релевантных векторов. Для его описания нам понадобится понятие обобщенной линейной модели.

Определение 5. *Обобщенной линейной моделью*, будем называть модель вида

$$y(\boldsymbol{x}) = \sum_{i=1}^{M} w_i \phi_i(\boldsymbol{x}),$$

$\phi_i(\boldsymbol{x}) : R^d \to R$ - заранее фиксированный набор базисных функций, а $\boldsymbol{w} \in R^M$ - вектор весов.

В 2000 Типпинг в своей работе [4] применил метод на основе максимизации обоснованности для автоматической настройки отдельных структурных параметров в обобщенных линейных моделях. Функция правдоподобия при этом задавалась следующим образом:

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\alpha}) = \frac{1}{1 + \exp(-ty(\boldsymbol{x}, \boldsymbol{w}))} \tag{2.5}$$

с нормальным распределением для каждого веса $w_i \sim \mathcal{N}(0, \alpha_i^{-1})$. Для вычисления обоснованности Типпинг использовал приближение Лапласа подынтегральной функции $p(\mathcal{T}|\mathcal{X}, \boldsymbol{\alpha})$ в (2.3). Данное преобразование позволило ему реализовать автоматическое определение релевантности (АОР) с помощью итерационной процедуры настройки $\boldsymbol{\alpha}$ и получить разреженное решение с большой долей весов равных нулю. Альтернативный способ, описанный в работе [15], использует вариационный подход для получения альтернативной аппроксимации подынтегральной функции с помощью гауссианы.

Заметим, что веса нерелевантных базисных функций $\phi_i(\boldsymbol{x})$ только стремятся к нулю при α_i стремящихся к бесконечности. Использование распределения Лапласа, напротив, приводит к тому, что часть весов в точности становятся равными нулю, но при этом приближение обоснованности становится весьма трудоемкой задачей, поскольку подынтегральная функция перестает быть гладкой и для вычисления интеграла (2.3), его приходится разбивать на большое количество кусочно-гладких составляющих (до 2^M).

Важным свойством обобщенных линейных моделей, показанным в работе [8], является тот факт, что интеграл (2.4) может быть достаточно хорошо приближен с помощью значения в точке с наиболее вероятными значениями весов $\boldsymbol{w}_{MP}$, то есть с помощью байесовской оценки весов классификатора.

3. Предлагаемый подход

Без ограничения общности, здесь и далее мы будем в качестве функции правдоподобия для задач классификации рассматривать произведение сигмоид, аналогично тому, как это сделано в логистической регрессии. Тогда логарифм правдоподобия может быть записан в виде

$$L(\mathcal{T}|\mathcal{X}, \boldsymbol{w}, \boldsymbol{\alpha}) = -\sum_{i=1}^{n} \log(1 + \exp(-t_i y(\boldsymbol{x_i}, \boldsymbol{w}))). \tag{3.1}$$

Приблизим функцию правдоподобия $p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}, \boldsymbol{\alpha})$ гауссианой. Тогда выражение для обоснованности (2.3) может быть записано в следующем виде:

$$E(\boldsymbol{\alpha}) = \int\limits_{\mathcal{W}(\alpha)} p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}, \boldsymbol{\alpha}) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w} \approx$$

$$p(\mathcal{T}|\mathcal{X}, \boldsymbol{\mu}, \boldsymbol{\alpha}) \int\limits_{\mathcal{W}(\alpha)} \exp\left(\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^T H (\boldsymbol{w}-\boldsymbol{\mu})\right) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w},$$

где $\boldsymbol{\mu}$ и $(-H)^{-1}$ вектор средних и ковариационная матрица гауссианы, с помощью которой мы приблизили функцию правдоподобия. Это позволяет записать выражения для $\boldsymbol{\mu}$ и H в следующем виде:

$$\boldsymbol{\mu} = \boldsymbol{w}_{ML} = \arg\max_{\boldsymbol{w}} p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}, \boldsymbol{\alpha}), \tag{3.2}$$

$$H = \nabla_{\boldsymbol{w}} \nabla_{\boldsymbol{w}} \log p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}, \boldsymbol{\alpha})|_{\boldsymbol{w}=\boldsymbol{w}_{ML}}. \tag{3.3}$$

Основная идея предлагаемого подхода - ввести независимую регуляризацию относительно новых переменных. Для этого в качестве новых осей возьмем собственные векторы матрицы гессиана функции правдоподобия и проведем вдоль этих осей регуляризацию. Похожий метод диагонализации, используемый для построения разреженных ядровых методов в рамках байесовского подхода, описан в работе [9].

Для этого, приведя гессиан к диагональному виду $H = Q^T \Lambda Q$, где $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$, $\{\lambda_i\}_{i=1}^M$ – собственные значения гессиана и $Q^T = Q^{-1}$, перейдем к новым переменным $\boldsymbol{u} = Q\boldsymbol{w}$. Поскольку функция логарифма правдоподобия (3.1) является вогнутой, гессиан H неположительно определен и все собственные значения $\{\lambda_i\}_{i=1}^M$ неположительны. Обозначим $h_i = -\lambda_i \geq 0$. Введем независимую регуляризацию относительно новых переменных $\boldsymbol{u}$. Тогда, функция априорного распределения может быть записана как

$$p(\boldsymbol{u}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} P(u_i|\alpha_i),$$

а обоснованность может быть представлена в виде произведения одномерных интегралов

$$E(\boldsymbol{\alpha}) = p(\mathcal{T}|\mathcal{X}, \boldsymbol{u}_{ML}, \boldsymbol{\alpha}) \prod_{i=1}^{M} f_i(h_i, u_{ML,i}, \alpha_i) =$$

$$p(\mathcal{T}|\mathcal{X}, \boldsymbol{u}_{ML}, \boldsymbol{\alpha}) \prod_{i=1}^{M} \int \exp\left(-\frac{h_i}{2}(u_i - u_{ML,i})^2\right) p(u_i|\alpha_i) du_i, \tag{3.4}$$

где $\boldsymbol{u}_{ML} = Q\boldsymbol{w}_{ML}$.

После указанных преобразований, для определения значений структурных параметров $\boldsymbol{\alpha}$ применим процедуру автоматического определения релевантности (АОР). Описанный выше алгоритм получил название *метод релевантных собственных векторов* (МРСВ).

Заметим, что вместо приближения Лапласа (3.2)-(3.3) мы можем использовать любой другой метод, который позволяет приблизить правдоподобие или

Алгоритм 1 Гауссовский метод релевантных собственных векторов (ГМРСВ)

вход Обучающая выборка $(\mathcal{X}, \mathcal{T}) = \{\boldsymbol{x}_i, t_i\}_{i=1}^n$, $\boldsymbol{x}_i \in \mathbb{R}^d, t_i \in \{-1, 1\}$, набор базисных функций $\{\phi_i(\boldsymbol{x})\}_{i=1}^M$.

1: Найти максимум правдоподобия $\boldsymbol{w}_{ML} = \arg\max_{\boldsymbol{w}} \log p(\mathcal{T}|\mathcal{X}, \boldsymbol{w})$.

2: Вычислить гессиан в точке максимума правдоподобия $H = \nabla_{\boldsymbol{w}} \nabla_{\boldsymbol{w}} \log p(\mathcal{T}|\boldsymbol{w}, \mathcal{X})|_{\boldsymbol{w}=\boldsymbol{w}_{ML}}$.

3: Разложить гессиан на собственные значения: $H = Q^T \Lambda Q$, где $\Lambda = \mathrm{diag}(-h_1, \ldots, -h_M)$ и вычислить $\boldsymbol{u}_{ML} = Q\boldsymbol{w}_{ML}$.

4:

для $i = 1$ **до** M **цикл**

  если $h_i u_{ML,i}^2 > 1$ **тогда**

    $\alpha_i^* = h_i / (h_i u_{ML,i}^2 - 1)$

  иначе

    $\alpha_i^* = +\infty$

  конец если

конец для

5: Найти максимум регуляризованного правдоподобия $\boldsymbol{w}_{MP} = \arg\max_{\boldsymbol{w}} \log p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}) p(Q\boldsymbol{w}|\boldsymbol{\alpha}^*)$.

выход Решающее правило для классификации новых объектов $\boldsymbol{x}$: $f(\boldsymbol{x}) = \mathrm{sign}\left(\sum_{i=1}^M w_{MP,i} \phi_i(\boldsymbol{x})\right)$

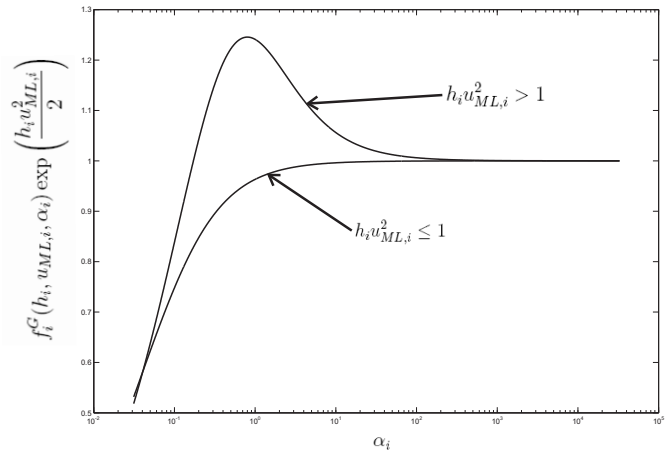
регуляризованное правдоподобие гауссианой, например метод вариационных границ (см. [10]) или метод прогонки ожидания (см. [11]). В остальном процедура регуляризации остается неизменной.

Применение данного метода позволяет в процессе поиска значения $\boldsymbol{\alpha}_{ME}$ вместо оптимизации многомерного интеграла осуществлять независимую оптимизацию одномерных интегралов. Это означает, что вместо экспоненциального, увеличение вычислительной сложности происходит линейно с ростом числа базисных функций M.

Поскольку в процессе преобразований в явном виде нигде не использовался тот факт, что в качестве модели выбрана обобщенная линейная модель, то данный метод применим для любых моделей, функция правдоподобия которых является дважды дифференцируемой относительно параметров $\boldsymbol{w}$ и допускает приближение гауссианой.

В предложенном методе изначально осуществлен переход от параметров классификатора $\boldsymbol{w}$ к величинам $\boldsymbol{u}$, которые характеризуют так называемые "степени свободы"классификатора. Регуляризация проводится вдоль степеней свободы (выраженных в терминах собственных векторов гессиана логарифма правдоподобия), поскольку они представляются более естественной мерой вариабельности классификатора, чем параметры $\boldsymbol{w}$.

В качестве примеров, рассмотрим два случая регуляризации: с помощью нормального распределения и распределения Лапласа.



Фиг. 1: Поведение одномерного интеграла $f_i^G(h_i, u_{ML,i}, \alpha_i)$

3.1. Гауссовское априорное распределение

Гауссовское априорное распределение задается следующим выражением

$$p(u_i|\alpha_i) = \sqrt{\frac{\alpha_i}{2\pi}} \exp\left(-\frac{\alpha_i {u_i}^2}{2}\right). \tag{3.5}$$

Рассмотрим одномерный интеграл $f_i(h_i, u_{ML,i}, \alpha_i)$ в выражении (3.4) с априорным распределением (3.5). Данный интеграл может быть вычислен аналитически:

$$f_i^G(h_i, u_{ML,i}, \alpha_i) = \sqrt{\frac{\alpha_i}{2\pi}} \int \exp\left(-\frac{h_i}{2}(u_i - u_{ML,i})^2 - \frac{\alpha_i}{2} u_i^2\right) du_i = \sqrt{\frac{\alpha_i}{h_i + \alpha_i}} \exp\left(-\frac{h_i \alpha_i u_{ML,i}^2}{2(h_i + \alpha_i)}\right). \tag{3.6}$$

В зависимости от h_i и $u_{ML,i}$ интеграл (3.6) либо имеет один максимум, либо растет по мере того, как α_i стремится к бесконечности. Поведение одномерного интеграла $f_i^G(h_i, u_{ML,i}, \alpha_i)$ в зависимости от h_i и $u_{ML,i}$ в случае гауссовского априорного распределения показано на фиг. 1. Функция f_i^G умножена на экспоненту с целью нормализации (обе кривые имеют общую асимптоту). Приравнивая производную (3.6) по α_i к нулю, получаем выражение для оптимального значения α_i:

$$\alpha_i^* = \begin{cases} \frac{h_i}{h_i u_{ML,i}^2 - 1} & \text{if } h_i u_{ML,i}^2 > 1 \\ +\infty & \text{иначе} \end{cases} \tag{3.7}$$

Похожие выражения для процедуры обучения МРВ с помощью метода покоординатного спуска получены в работе [12].

Алгоритм 1 представляет собой процедуру обучения разреженной байесовской модели с гауссовским априорным распределением. Заметим, что в отличие от МРВ, где для процедуры обучения необходим итерационный процесс, в гауссовском МРСВ (ГМРСВ) оптимальные значения $\boldsymbol{\alpha}$ вычисляются за одну итерацию. Результаты экспериментов (см. разд. 4.) показывают что ГМРСВ превосходит по скорости и позволяет получить более разреженное решающее правило по сравнению с МРВ.

Алгоритм 2 Лапласовский метод релевантных собственных векторов (ЛМРСВ)

вход Обучающая выборка $(\mathcal{X}, \mathcal{T}) = \{\boldsymbol{x}_i, t_i\}_{i=1}^n$, $\boldsymbol{x}_i \in \mathbb{R}^d, t_i \in \{-1, 1\}$, набор базисных функций $\{\phi_i(\boldsymbol{x})\}_{i=1}^M$.

1-3: Аналогично Алгоритму 1.

4:

для $i = 1$ **to** M **цикл**

Найти максимум (3.9) с помощью одномерной процедуры оптимизации:

$\alpha_i^* = \arg\max_{\alpha_i} f_i^L(h_i, u_{ML,i}, \alpha_i)$

конец для

5: Найти максимум регуляризованного правдоподобия вдоль векторов $\boldsymbol{u}$:

$\boldsymbol{u}_{MP} = \arg\max_{\boldsymbol{u}} \log p(\mathcal{T}|\mathcal{X}, Q^T\boldsymbol{u}) p(\boldsymbol{u}|\boldsymbol{\alpha}^*)$

при ограничениях $u_{ML,i} u_i \geq 0$ для всех i.

6: Найти веса $\boldsymbol{w}_{MP} = Q^T \boldsymbol{u}_{MP}$.

выход Решающее правило для классификации нового объекта $\boldsymbol{x}$: $f(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^M w_{MP,i}\phi_i(\boldsymbol{x})\right)$

3.2. Априорное распределение Лапласа

Априорное распределение Лапласа может быть записано как

$$p(u_i|\alpha_i) = \frac{\alpha_i}{4} \exp\left(-\frac{\alpha_i|u_i|}{2}\right). \tag{3.8}$$

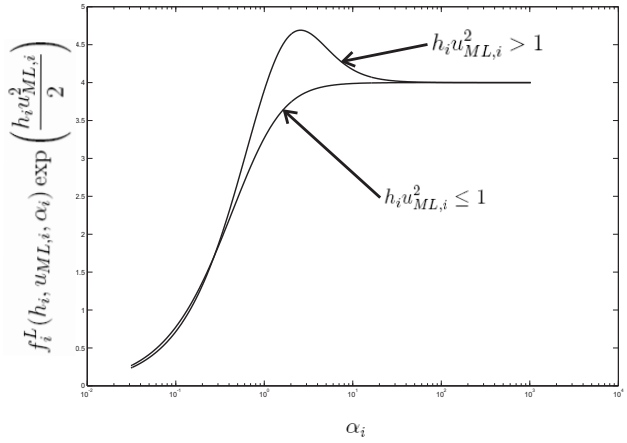
Подставляя выражение (3.8) в одномерный интеграл (3.4) получим

$$f_i^L(h_i, u_{ML,i}, \alpha_i) = \frac{\alpha_i}{4} \int \exp\left(-\frac{h_i}{2}(u_i - u_{ML,i})^2 - \frac{\alpha_i}{2}|u_i|\right) du_i = \frac{\alpha_i}{4}\sqrt{\frac{\pi}{2h_i}} \times \exp\left(-\frac{h_i u_{ML,i}^2}{2}\right) \left[\operatorname{erfcx}\left(\sqrt{\frac{h_i}{2}}\left(\frac{\alpha_i}{2h_i} - u_{ML,i}\right)\right) + \operatorname{erfcx}\left(\sqrt{\frac{h_i}{2}}\left(\frac{\alpha_i}{2h_i} + u_{ML,i}\right)\right)\right], \tag{3.9}$$

где $\operatorname{erfcx}(x) = \frac{2}{\sqrt{\pi}} \exp(x^2) \int_x^{+\infty} \exp\left(-t^2\right) dt$ – нормированное дополнение к функции ошибок[1]. Более детальное описание устойчивого вычисления выражения (3.9) см. в Приложении А. Последнее равенство задает унимодальную функцию относительно α_i, оптимальные значения которых могут быть эффективно найдены с помощью одномерных методов оптимизации. На фиг. 2 показано поведение одномерного интеграла $f_i^L(h_i, u_{ML,i}, \alpha_i)$ в зависимости от значений h_i и $u_{ML,i}$ в случае априорного распределения Лапласа. Функция f_i^L умножена на экспоненту с целью нормализации (обе кривые имеют общую асимптоту).

Алгоритм 2 описывает процедуру обучения для случая априорного распределения Лапласа. Аналогично ГМРСВ в лапласовском МРСВ (ЛМРСВ) оптимизация по $\boldsymbol{\alpha}$ может быть произведена за одну итерацию, что позволяет

[1] Функция реализована во многих математических пакетах, например в среде МАТЛАБ



Фиг. 2: Поведение одномерного интеграла $f_i^L(h_i, u_{ML,i}, \alpha_i)$

значительно сократить время обучения. Тем не менее, последний шаг алгоритма ЛМРСВ – оптимизация функции регуляризованного логарифма правдоподобия – нетривиален, поскольку данная функция не является гладкой в тех точках, где хотя бы один из весов равен нулю. Для такой оптимизации можно воспользоваться специальным методом ЛАССО (см. [13]), сводящим эту задачу к задаче квадратичного программирования, либо ограничиться рассмотрением одного гипероктанта, содержащего точку $\boldsymbol{u}_{ML}$. В этом случае мы можем свести задачу к оптимизации гладкой функции при наличии ограничений, которые задают границы гипероктанта, содержащего точку $\boldsymbol{u}_{ML}$. Эксперименты показали, что полученное таким образом приближенное решение близко (в смысле качества соответствующего решающего правила) к оптимальному или совпадает с ним в подавляющем большинстве случаев.

4. Эксперименты

В данной главе представлены результаты сравнения методов МРВ, ГМРСВ и ЛМРСВ. В качестве характеристик для сравнения были выбраны процент ошибок, время обучения и получаемая разреженность решения (для методов МРСВ разреженность означает количество ненулевых компонент в $\boldsymbol{u}_{MP}$) на наборах данных, полученных из репозитория UCI[1]. и сайта Исследовательской Группы Распознавания Образов университета Уэльса[2]. Для обучения МРВ использовался известный алгоритм, предложенный Типпингом и Фоль в работе [12]. Также были добавлены вариационные версии для МРВ и ГМРСВ (ВМРВ и ВГМРСВ соответственно). Для каждого набора данных номинальные признаки были преобразованы в набор бинарных, неизвестные значения были заменены на средние известных отдельно для каждого признака, после чего каждая выборка была нормализована таким образом, чтобы у каждого признака было нулевое математическое ожидание и единичная дисперсия. Все классификаторы были

[1] http://www.ics.uci.edu/∼mlearn/MLRepository.html

[2] http://www.informatics.bangor.ac.uk/∼kuncheva/activities/patrec1.html

Таблица 1: Доли ошибок со стандартными отклонениями (в процентах).

Набор данных	**ГМРСВ**	**ЛМРСВ**	**МРВ**	**ВМРВ**	**ВМРСВ**
Bupa	33.3±3.3	**30.7±1.1**	32.4±2.6	*30.8±2.1*	31.9±0.7
Heart	18.2±2.1	*17.4±0.8*	**17.3±1.7**	17.6±2.7	23.1±1.8
Hepatitis	*14.3±0.7*	19.4±0.8	20.3±1.3	**13.0±1.1**	17.8±1.5
Votes	*5.6±0.6*	5.9±0.3	6.4±2.0	5.6±0.6	**4.9±0.8**
WPBC	23.6±1.2	**23.1±1.4**	23.7±0.4	24.0±0.9	*23.4±0.3*
Contractions	19.2±2.2	18.0±1.4	18.9±1.9	*17.4±2.0*	**14.1±3.2**
Laryngeal1	**16.8±0.6**	17.5±1.1	17.5±0.6	*17.4±0.6*	18.3±2.8
Respiratory	*7.1±1.2*	7.5±1.6	9.4±2.5	8.2±1.9	**5.9±1.2**
Weaning	15.7±4.8	14.8±1.3	16.4±1.4	*13.6±1.2*	**13.3±2.1**
Рейтинг	**28.0**	**24.5**	**36.5**	**24.0**	**22.0**
Шрифтовая легенда	**Место 1**	*Место 2*	Место 3		

Таблица 2: Время обучения со стандартным отклонением (в секундах).

Набор данных	**ГМРСВ**	**ЛМРСВ**	**МРВ**	**ВМРВ**	**ВМРСВ**
Bupa	138.7 ± 30.5	19.8 ± 0.6	64.1 ± 0.8	62.9 ± 3.0	310.3 ± 22.2
Heart	39.9 ± 5.5	11.6 ± 0.4	46.1 ± 0.4	21.5 ± 0.5	115.6 ± 2.6
Hepatitis	8.9 ± 0.4	4.9 ± 0.1	25.2 ± 0.4	6.9 ± 0.3	37.2 ± 1.0
Votes	123.8 ± 10.0	31.5 ± 0.5	87.4 ± 0.8	82.1 ± 3.6	433.3 ± 14.6
WPBC	19.7 ± 0.8	7.5 ± 0.1	33.1 ± 0.2	16.3 ± 0.5	89.7 ± 3.3
Contractions	7.5 ± 0.7	3.0 ± 0.1	16.1 ± 0.1	3.8 ± 0.3	22.3 ± 0.6
Laryngeal1	24.2 ± 4.0	8.7 ± 0.4	36.3 ± 0.5	18.6 ± 0.9	100.5 ± 4.3
Respiratory	4.5 ± 0.2	2.4 ± 0.1	13.8 ± 0.2	2.4 ± 0.1	14.9 ± 0.1
Weaning	53.3 ± 1.3	14.1 ± 0.2	53.2 ± 0.2	32.5 ± 0.9	177.2 ± 4.9

построены на основе одинакового набора базисных функций $M = n + 1$, $\phi_i(\boldsymbol{x}) = \exp(-\|\boldsymbol{x} - \boldsymbol{x}_i\|/(2\sigma^2))$, $i = 1, \ldots, n$ и $\phi_{n+1}(\boldsymbol{x}) \equiv 1$. Оптимальное значение ширины σ выбиралось из множества $\{0.01, 0.1, 0.3, 0.6, 1, 2, 3, 5, 7, 10\}$ с помощью 5x2 кросс-валидации (см. [14]). Также, с помощью 5x2 кросс-валидации для каждого набора данных измерялись доля ошибок, время обучения и разреженность полученного решения. Результаты экспериментов приведены в таблицах 1, 2 и 3. Рейтинг рассчитывался следующим методом: для каждого набора данных победитель получает одно очко, следующий за ним - два очка и т.д., проигравший получает пять очков, после чего все полученные очки суммируются по всем задачам.

Полученные результаты позволяют сделать следующие выводы. Все алгоритмы показывают сравнимые результаты в терминах доли неверных ответов. Доля ненулевых параметров в ГМРСВ и особенно в ЛМРСВ значительно ниже, чем в МРВ. Заметим, что в МРСВ количество ненулевых параметров относится к собственным векторам (или степеням свободы) и потому полученная разреженность неявная. При этом все исходные веса $\boldsymbol{w}_{MP}$ необходимые для классификации

Таблица 3: Доля ненулевых параметров со стандартным отклонением

Набор данных	#Об./2	ГМРСВ	ЛМРСВ	МРВ	ВМРВ	ВГМРСВ
Bupa	172.5	23.1 ± 14.1	8.2 ± 1.4	5.8 ± 1.0	172.5 ± 0.0	17.2 ± 1.4
Heart	135	12.1 ± 4.0	9.0 ± 0.9	6.0 ± 1.1	135.0 ± 0.0	86.6 ± 13.5
Hepatitis	77.5	10.2 ± 2.7	4.6 ± 1.9	3.1 ± 1.1	77.5 ± 0.0	39.4 ± 17.3
Votes	217.5	14.6 ± 3.7	6.6 ± 1.1	4.7 ± 0.6	217.5 ± 0.0	58.7 ± 21.8
WPBC	99	20.8 ± 2.5	2.2 ± 1.0	2.4 ± 1.4	99.0 ± 0.0	27.0 ± 22.8
Contractions	49	14.5 ± 12.3	3.3 ± 1.4	2.3 ± 0.8	49.0 ± 0.0	34.2 ± 3.4
Laryngeal1	106.5	7.2 ± 1.8	3.9 ± 1.4	1.9 ± 0.6	106.4 ± 0.2	57.1 ± 22.6
Respiratory	42.5	5.7 ± 1.3	2.5 ± 0.6	1.6 ± 0.4	42.5 ± 0.0	22.4 ± 1.7
Weaning	151	12.9 ± 4.2	10.1 ± 2.1	6.8 ± 1.6	151.0 ± 0.0	120.2 ± 10.2

новых объектов, вообще говоря, отличны от нуля, что не позволяет использовать предложенный подход для определения множества релевантных признаков объектов обучения.

ГМРСВ оказался быстрее МРВ поскольку для оптимизации коэффициентов регуляризации $\boldsymbol{\alpha}$ в ГМРСВ необходима только одна итерация в отличие от МРВ, а соответствующие выражения для $\boldsymbol{\alpha}$ могут быть выписаны в явном виде. ЛМРСВ быстрее, чем МРВ на данных с большим количеством объектов и медленнее в остальных случаях. С одной стороны, ЛМРСВ выигрывает по скорости работы, поскольку вычисление коэффициентов регуляризации происходит за одну итерацию. С другой стороны, для вычисления $\boldsymbol{\alpha}$ необходимо M одномерных оптимизаций, так как в явном виде выражение для α_i выписать нельзя, а для нахождения $\boldsymbol{u}_{MP}$ необходимо применить условную оптимизацию.

Вариационная аппроксимация требует значительно больше времени для процедуры обучения. Сначала в ВГМРСВ находим гауссовское вариационное приближение нерегуляризованного правдоподобия, минимизируя дивергенцию Кульбака-Лейблера $KL(\mathcal{N}(\boldsymbol{\mu}, -H^{-1}) \| p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}, \boldsymbol{\alpha}))$. (подробноее см. [15]). Затем регуляризуем собственные значения H аналогично тому, как это сделано в случае приближения Лапласа (3.2)-(3.3). Заметим, что ВГМРСВ медленнее, чем ВМРВ поскольку приближение нерегуляризованного правдоподобия гауссианой при использовании вариационного подхода требует больше времени, чем приближение регуляризованного правдоподобия. Возможно, процесс может быть ускорен с помощью введения небольшого начального регуляризатора над $\|\boldsymbol{w}\|$. С другой стороны, хотя точность вариационных методов кажется более высокой, этот факт требует дальнейшего изучения для возможности делать обоснованные выводы. Доля ненулевых весов выше, чем в случае приближения Лапласа, но это общее свойство вариационного подхода в целом. Аналогично, вариационное приближение может быть использовано для ЛМРСВ.

5. Заключение

В данной работе представлен новый подход к регуляризации процедуры обучения классификаторов. В его основе лежит регуляризация по степеням

свободы, в качестве которых выбраны собственные векторы гессиана правдоподобия, вместо весов классификатора. В пространстве весов данному подходу соответствует использование априорного распределения с недиагональной матрицей, связывающей веса $\boldsymbol{w}$:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \frac{\sqrt{|A|}}{\sqrt{2\pi}^M} \exp\left(-\frac{1}{2}\boldsymbol{w}^T Q^T A Q \boldsymbol{w}\right)$$

для гауссовского априорного распределения, и

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \frac{|A|}{4^M} \exp\left(-\frac{1}{2}\sum_{i=1}^{M}\left|\sum_{j=1}^{M} q_{ij} w_j\right|\right)$$

для распределения Лапласа. Здесь $A = \operatorname{diag}(\alpha_1, \ldots, \alpha_M)$ и $Q = \{q_{ij}\}_{i,j=1}^{M}$. На наш взгляд количество степеней свободы является более естественной мерой вариабельности семейства классификаторов. Кроме того предлагаемый подход позволяет произвести декомпозицию выражения для обоснованности на произведение одномерных интегралов, которые могут быть оптимизированы независимо друг от друга. Это означает, что подход на основе обоснованности может быть эффективно использован для автоматического определения релевантности (АОР) с различными типами априорных распределений. Это было показано на примере распределения Лапласа, которое при использовании МРВ приводило к сложному для прямого вычисления интегралу.

Также важно отметить, что предложенный подход, в отличие от МРВ, является инвариантным относительно линейных преобразований, поскольку регуляризация производится не непосредственно над весами $\boldsymbol{w}$, а над степенями свободы $\boldsymbol{u}$.

Более разреженные решения, получаемые с помощью МРСВ, косвенно показывают, что рациональнее вводить независимые априорные распределения на степенях свободы $\boldsymbol{u}$, определяемых собственными векторами гессиана правдоподобия, чем на весах $\boldsymbol{w}$, которые могут одновременно относиться и к релевантным, и нерелевантным собственным векторам гессиана правдоподобия. Можно предположить, что собственные значения и направления собственных векторов являются характерными признаками функции правдоподобия и отвечают за обобщающую способность классификатора. В рамках данного предположения веса классификатора являются вторичными по отношению к степеням свободы и вместо них независимой регуляризации должны подвергаться непосредственно собственные векторы.

Для более общей постановки задачи регуляризации обобщенных линейных моделей может быть использована неотрицательно-определенная симметричная матрица R_0, коэффициенты которой могут быть найдены в процессе оптимизации выражения для обоснованности

$$R_0 = \arg\max_{R \in \mathcal{R}} p(\mathcal{T}|\mathcal{X}, \boldsymbol{w}) p(\boldsymbol{w}|R),$$

где $\mathcal{R} = \{R \in \mathbb{R}^{M \times M} \mid R^T = R,\ R \geq 0\}$ и $p(\boldsymbol{w}|R) = \sqrt{\frac{\det(R)}{2\pi^M}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T R \boldsymbol{w}\right)$. Данная матрица может быть найдена аналитически (см. [16]).

Список литературы

1. *Bishop C. M.* Pattern recognition and machine learning. New York: Springer, 2006.

2. *Neal R. M.* Bayesian learning for neural networks. New York: Springer, 1996.

3. *MacKay D. J. C.* The evidence framework applied to classification networks // Neural Comput.. 1992. V. 4. P. 720–736.

4. *Tipping M. E.* The relevance vector machine // Advances neural information processing systems. 2000. V. 12. P. 652–658.

5. *Williams P. M.* Bayesian regularization and pruning using a laplace prior //Neural Comput.. 1995. V. 7. P. 117–143.

6. *Cawley G. C., Talbot N. L. C.* Gene selection in cancer classification using sparse logistic regression with Bayesian regularization //Bioinformat.. 2006. V. 22. N. 19 P. 2348–2355.

7. *Cawley G. C., Talbot N. L. C., Girolami M.* Sparse multinomial logistic regression via bayesian l1 regularisation // Advances Neural Informat. Processing Systems. 2007. V. 19. P. 209–216

8. *Tipping M. E.* Sparse bayesian learning and the relevance vector machines // J. Mach. Learning Res. 2001. V. 1. P. 211–244.

9. *Cawley G. C., Talbot N. L. C.* A Simple Trick for Constructing Bayesian Formulations of Sparse Kernel Learning Methods // Proc. of IJCNN-2005. Montreal: IEEE Computer Society, 2005. P. 1425–1430.

10. *Jaakkola T., Jordan M. I.* Bayesian parameter estimation via variational methods //Stat. and Comput.. 2000. V. 10. P. 25–37.

11. *Minka T.* Expectation propagation for approximate Bayesian inference // Proc. of 17th Conf. on Uncertainity in Artificial Intelligence. San Francisco: Morgan Kaufmann, 2001. P. 362–369.

12. *Tipping M. E.,Faul A. C.* Fast marginal likelihood maximisation for sparse Bayesian models //Proc. of 9th Internat. Workshop on Artificial Intelligence and Stat. Key West: Society for AI & Stat., 2003. P. 3–6.

13. *Tibshirani R.* Regression shrinkage and selection via the lasso //J. Roy. Stat. Soc.. 1996. V. 58. P. 267–288.

14. *Dietterich T. G.*. Approximate statistical tests for comparing supervised classification learning algorithms //Neural Computat.. 1998. V. 10. P. 1895–1924.

15. *C. M. Bishop, M. E. Tipping* Variational relevance vector machines //Uncertainty in Artificial Intelligence. 2000. P. 46–53

16. *Kropotov D. A., Vetrov D.P.* On Equivalence of information-based and bayesian approaches to model selection for linear regression problems // Proc. of 9th Internat. Conf. Pattern Recognition and Image Analysis. Nizhni Novgorod: МАИК Наука/Interperiodica Publ., 2008. P. 419–422.

A Эффективное вычисление обоснованности для ЛМРСВ

Рассмотрим вычисление обоснованности (3.9) для случая априорного распределения Лапласа. Выражение (3.9) может быть записано как

$$C \exp\left(-\frac{h_i u_{ML,i}^2}{2}\right) \times \left[\exp\left(x_1^2\right) \operatorname{erfc}(x_1) + \exp\left(x_2^2\right) \operatorname{erfc}(x_2)\right], \tag{A1}$$

где C – некоторая положительная константа, а $\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty \exp(-t^2) dt$ – дополнительная функция ошибок и $x_{1,2} = \sqrt{\frac{h_i}{2}} \left(\frac{\alpha_i}{2h_i} \mp u_{ML,i}\right)$. При больших положительных значениях x произведение $\exp(x^2)\operatorname{erfc}(x)$ может быть представлено в виде нормированного дополнения к функции ошибок

$$\operatorname{erfcx}(x) = \exp(x^2)\operatorname{erfc}(x) \approx 1/(\sqrt{\pi}x).$$

Для достаточно больших отрицательных значений x целесообразно объединить $\exp(-h_i u_{ML,i}^2/2)$ и $\exp(x^2)$ в одно выражение

$$\exp(-h_i u_{ML,i}^2/2) \exp(x^2) = \exp(y),$$

где

$$y_{1,2} = \frac{\alpha_i^2}{8h_i} \mp \frac{\alpha_i u_{ML,i}}{2}.$$