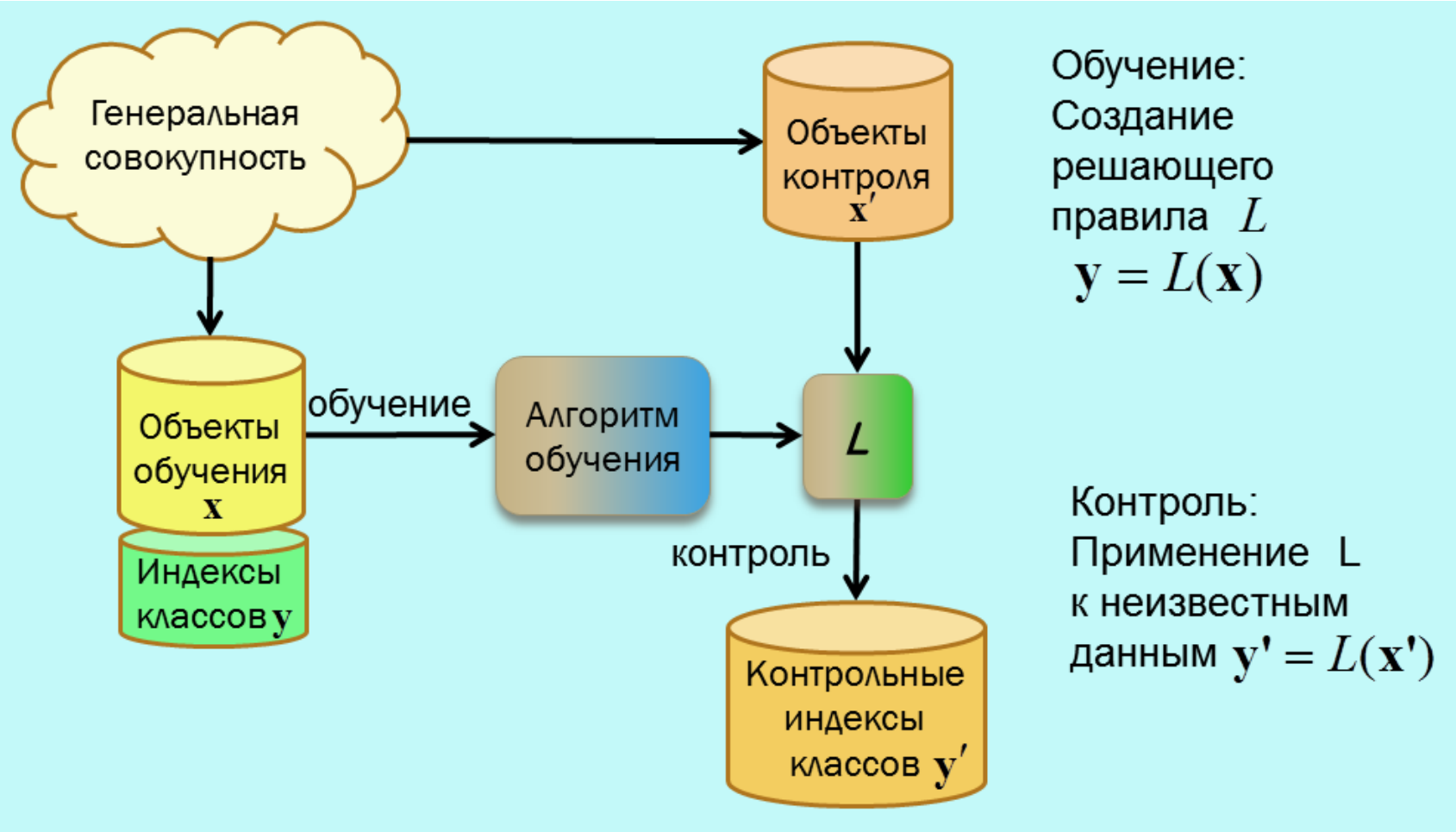
Отбор признаков в задаче классификации потоков данных при смещении решающего правила

О.В. Красоткина, П.А. Турков

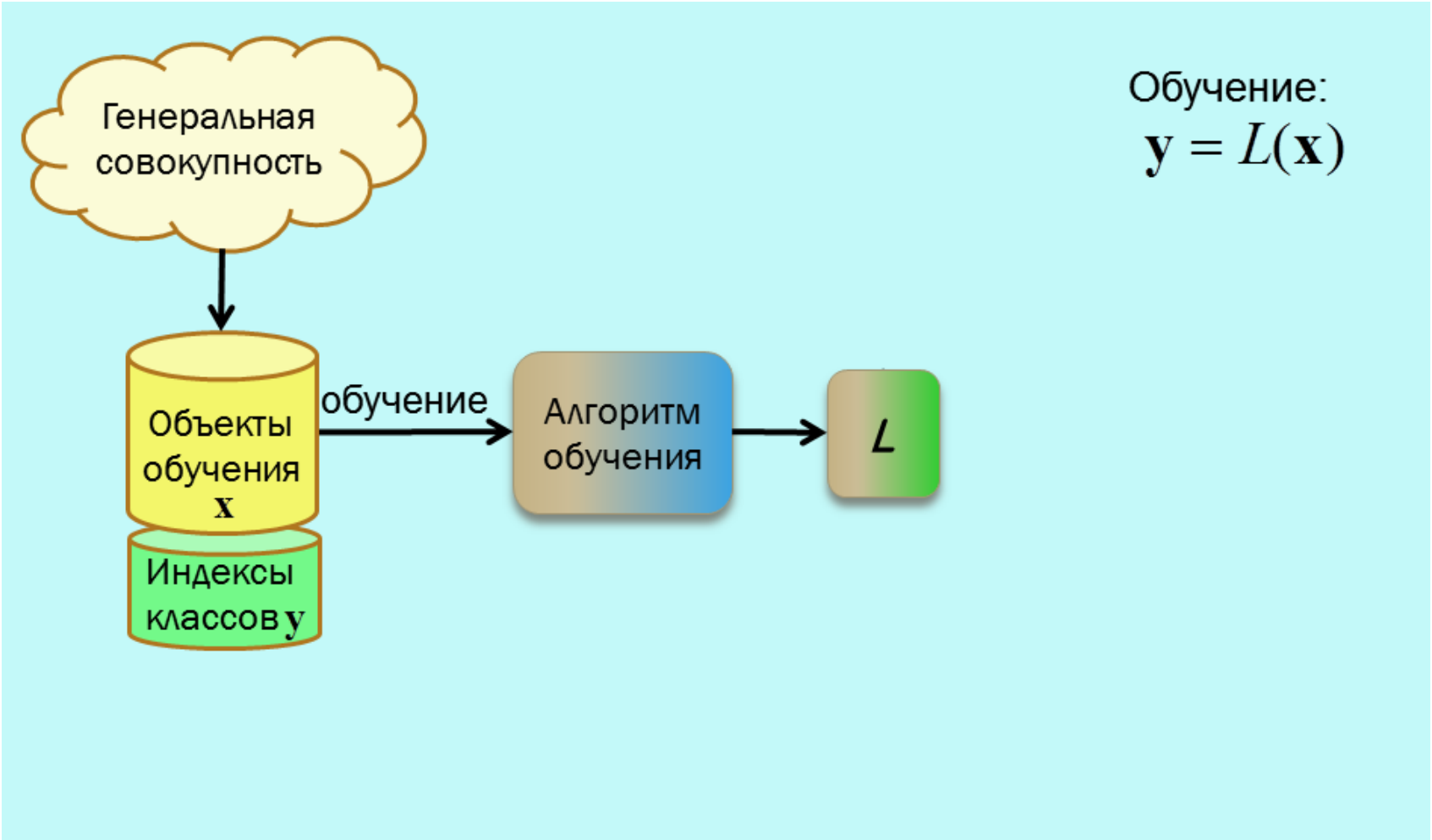
ФГБОУ ВО
Тульский государственный университет

17-я Всероссийская конференция
Математические методы распознавания образов
Светлогорск, 19 - 24 сентября 2015

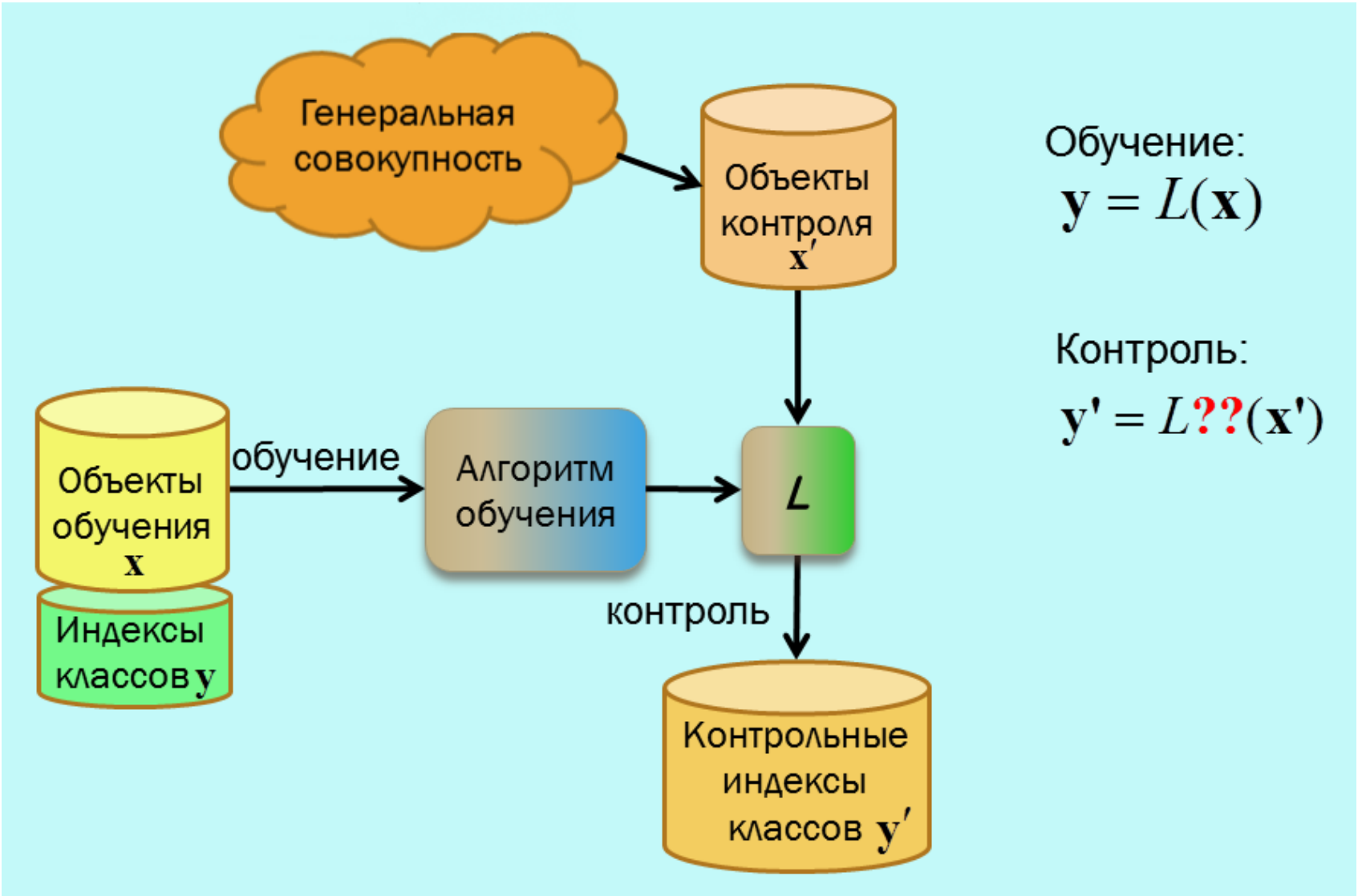
Классическая процедура обучения распознаванию образов



Процедура обучения распознаванию образов в потоках данных



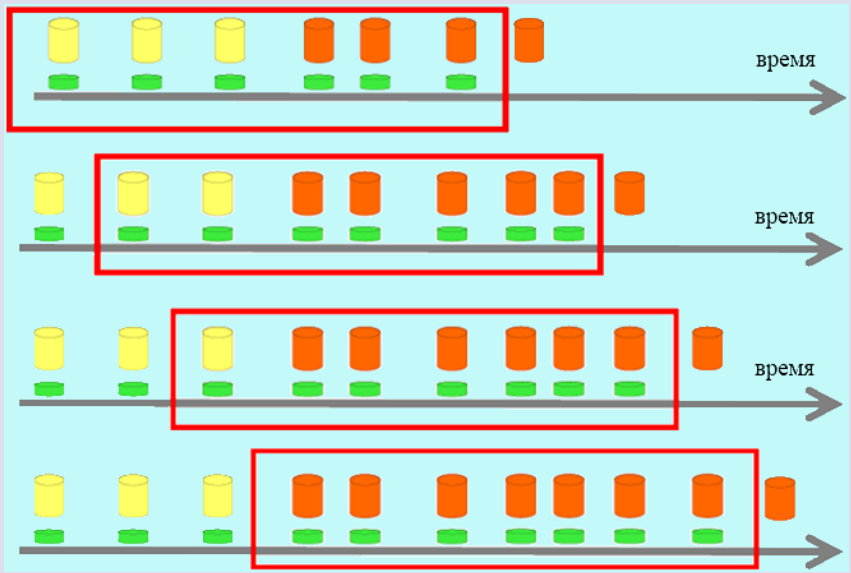
Процедура обучения распознаванию образов в потоках данных



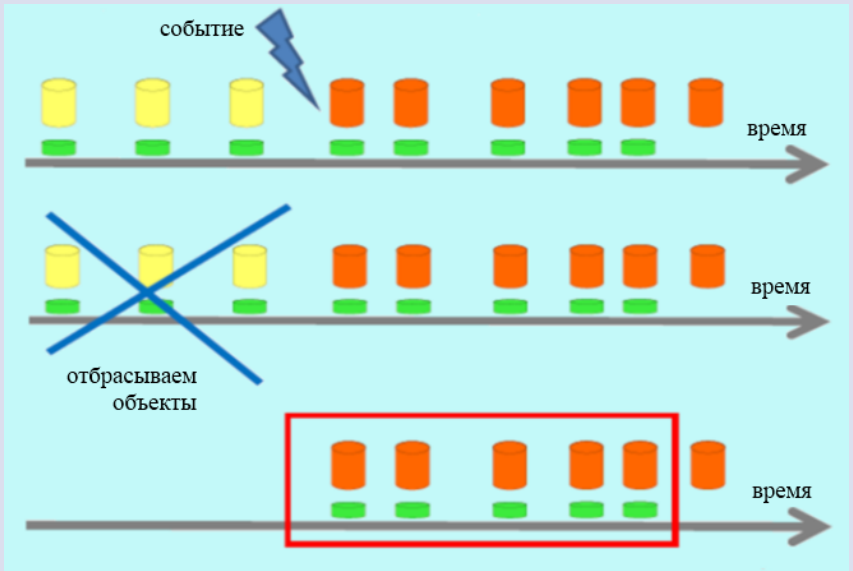
Обзор существующих методов обучения распознаванию в потоках данных

Обучение с использованием одного классификатора

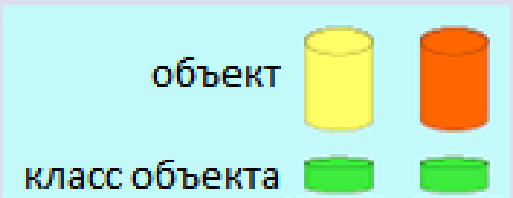
Окно постоянной длины (Widmer, Kubat, 1996)



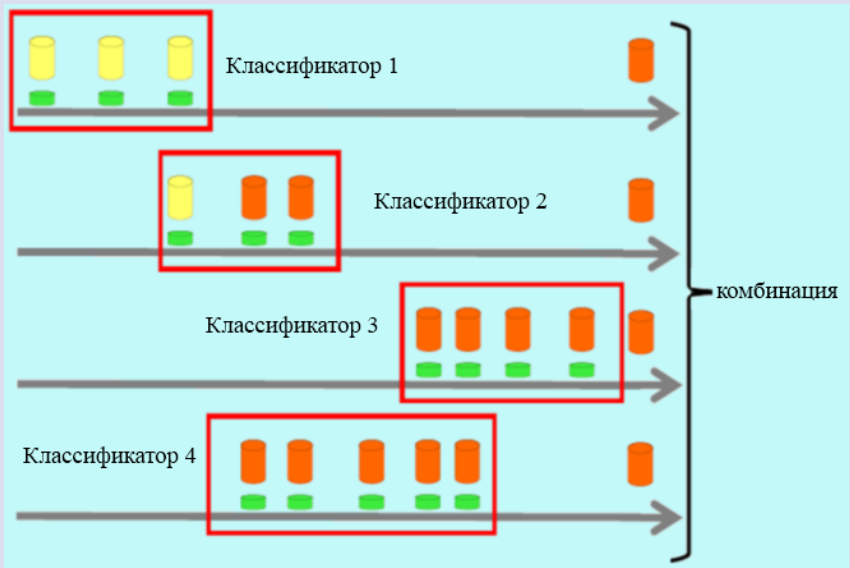
Окно переменной длины (Patist, 2007)



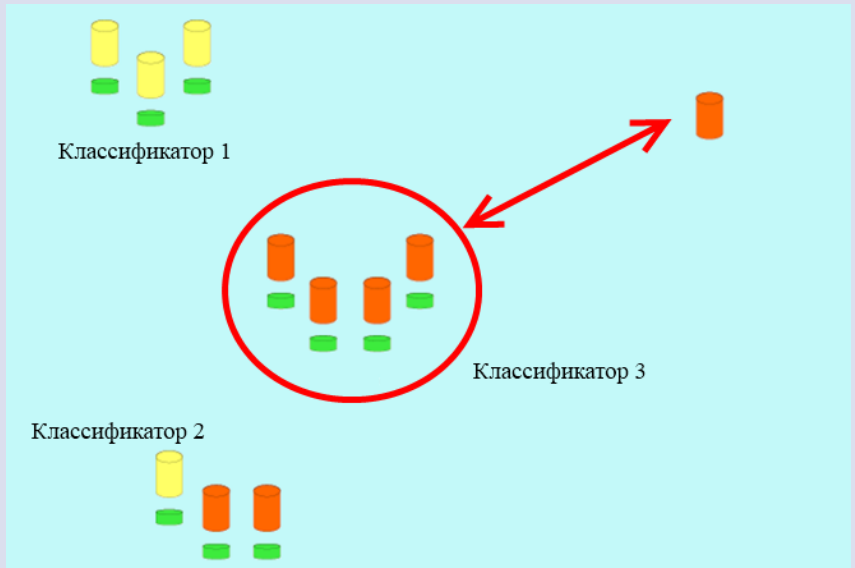
где



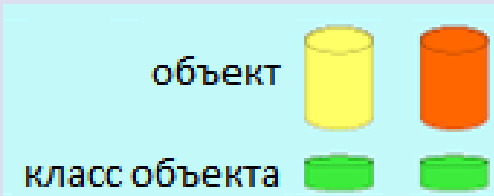
Bagging & Boosting (Kolter, 2007)



Stacking (Street, 2001)



где



Постановка задачи обучения распознаванию образов в нестационарной генеральной совокупности

Обучающее множество

- Пусть каждый объект генеральной совокупности $\omega \in \Omega$ представлен точкой в линейном признаковом пространстве $\mathbf{x}(\omega) = \big(x^1(\omega), \ldots, x^n(\omega)\big) \in \mathbb{R}^n$, а его скрытая фактическая принадлежность к одному из двух классов описывается значением индекса класса $y(\omega) \in \{1, -1\}$.
- Наличие нестационарности в исследуемых данных требует рассмотрения дополнительно к признаковому описанию объекта момента времени его получения (ω, t). В результате обучающее множество приобретает вид $\big\{(\mathbf{X}_t \in \mathbb{R}^n,\ \mathbf{Y}_t,\ t)\big\}_{t=1}^{T}$, $(\mathbf{X}_t, \mathbf{Y}_t) = \big\{(\mathbf{x}_{k,t}, y_{k,t})\big\}_{k=1}^{N_t}$ - подмножество объектов, поступивших в момент времени t.

Постановка задачи обучения распознаванию образов в нестационарной генеральной совокупности

Решающее правило

- Модель генеральной совокупности понимается как априори существующая дискриминантная функция, описываемая как гиперплоскость с направляющим (нормальным) вектором $\mathbf{a}$ и параметром положения b:
  $f\big(\mathbf{x}(\omega)\big) = \mathbf{a}^T\mathbf{x} + b$ преимущественно > 0 если $y(\omega) = 1$, и < 0 если $y(\omega) = -1$.
- Наличие нестационарности в исследуемых данных требует рассмотрения параметров $\mathbf{a}$ и b как функций времени:
  $\mathbf{a}_t : T \to \mathbb{R}^N$

Иерархическая вероятнотстная модель для оценивания параметров нестационарного решающего правила

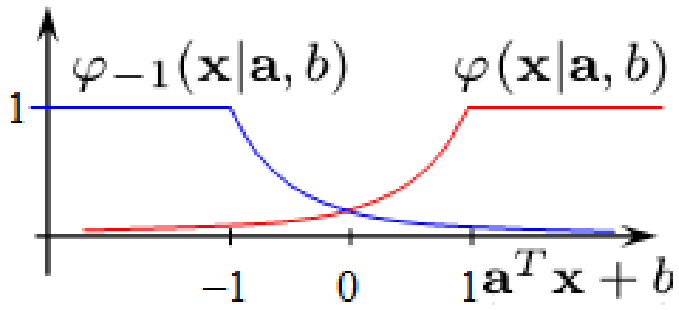
Апостериорное распределение параметров нестационарного решающего правила:

$$P\left((\mathbf{a}_t, b_t)_{t=1}^T | (\mathbf{X}_t, \mathbf{Y}_t)_{t=1}^T\right) = \\ = \frac{\Psi(\mathbf{a}_t, b_t, t = \overline{2,T}) \prod_{t=1}^T \Phi(\mathbf{X}_t | \mathbf{Y}_t, \mathbf{a}_t, b_t)}{\int\limits_{\mathbb{R}} \int\limits_{\mathbb{R}^n} \Psi(\mathbf{a}'_t, b'_t, t = \overline{2,T}) \prod_{t=1}^T \Phi(\mathbf{X}_t | \mathbf{Y}_t, \mathbf{a}'_t, b'_t) d\mathbf{a}'_t db'_t}$$

Оценка параметров по методу максимального правдоподобия приводит к следующему критерию обучения:

$$(\hat{\mathbf{a}}_t, \hat{b}_t)_{t=1}^T = \arg \max_{\mathbf{a}_t, b_t, t=1,\cdot,T} P\left((\mathbf{a}_t, b_t)_{t=1}^T | (\mathbf{X}_t, \mathbf{Y}_t)_{t=1}^T\right)$$

В качестве условных плотностей распределения объектов обучающей выборки возьмем следующие несобственные распределение:



$$\varphi_1(\mathbf{x}|\mathbf{a_t},b_t)=\begin{cases}1, \mathbf{a_t}^T\mathbf{x}+b_t>1,\\ \exp\left[-c\left(1-(\mathbf{a_t}^T\mathbf{x}+b_t)\right)\right], \mathbf{a_t}^T\mathbf{x}+b_t<1,\end{cases}$$

$$\varphi_{-1}(\mathbf{x}|\mathbf{a_t},b_t)=\begin{cases}1, \mathbf{a_t}^T\mathbf{x}+b_t<-1,\\ \exp\left[-c\left(1+(\mathbf{a_t}^T\mathbf{x}+b_t)\right)\right], (\mathbf{a_t}^T\mathbf{x}+b_t)>-1.\end{cases}$$

- Несобственные плотности распределений объектов генеральной совокупности $y_{j,t} = \pm 1$:

$$f(\mathbf{x}_{j,t}|y_{j,t}, \mathbf{a}_t, b_t) =$$
$$= \begin{cases} 1, y_{j,t}(\mathbf{a}_t^T \mathbf{x}_{j,t} + b_t) > 1, \\ \exp\left[-c\left(1 - y_{j,t}(\mathbf{a}_t^T \mathbf{x}_{j,t} + b_t)\right)\right], y_{j,t}(\mathbf{a}_t^T \mathbf{x}_{j,t} + b_t) < 1, \end{cases}$$

- Совместная плотность распределения объектов обучающей выборки $\mathbf{X}_t, \mathbf{Y}_t$ в момент времени t :

$$\Phi(\mathbf{X}_t|\mathbf{Y}_t, \mathbf{a}_t, b_t) = \prod_{j=1}^{N_t} f(y_{j,t}|\mathbf{x}_{j,t}, \mathbf{a}_t, b_t).$$

- Ключевым элементом является понимание зависящих от времени параметров гиперплоскости $(\mathbf{a}_t, b_t)$ как скрытых случайных процессов с марковскими свойствами:

$$\mathbf{a}_t = q\mathbf{a}_{t-1} + \boldsymbol{\xi}_t, M\left(\boldsymbol{\xi}_t\right) = \mathbf{0}, M(\boldsymbol{\xi}_t\boldsymbol{\xi}_t^T) = d\mathbf{I},$$
$$b_t = b_{t-1} + \nu_t, M\left(\nu_t\right) = 0, M(\nu_t^2) = d', q = \sqrt{1-d}, 0 \leq q < 1$$

Дисперсии d и d' описывают скрытую динамику смещения решающего правила.

- Априорная плотность распределения последовательности параметров решающего правила:

$$\Psi(\mathbf{a}_t, b_t, t = 2, \ldots, T) = \prod_{t=2}^{T} \left[\mathcal{N}\left(\mathbf{a}_t | \sqrt{1-d}\,\mathbf{a}_{t-1}, d\mathbf{I}\right)\mathcal{N}(b_t | b_{t-1}, d')\right]$$

Способы отбора признаковой информации

Методы-фильтры

Фильтры применяются на множестве всех признаков до восстановления зависимости, независимо от используемого метода восстановления. Как правило используются различные переборные стратегии. Метод отбора признаков не учитывает особенности искомой зависимости и используемых для ее восстановления алгоритмов.

Встроенные методы

Встроенные методы отбора признаков непосредственно инкорпорируются в метод решения задачи и, следовательно, существенно зависят от его специфики. Окончание процесса обучения одновременно является окончанием процесса отбора признаков.

Для наделения нашей нестационарной модели способностью к отбору признаков, введем в модель нестационарного решающего правила еще один уровень иерархии, рассматривая дисперсии параметров направляющего вектора как случайные

$$\psi(a_{t,j}|r_j, a_{t-1,j}) = \frac{1}{\sqrt{2\pi r_j d}} \exp(-(a_{t,j} - \sqrt{1-d}a_{t,j})^2/2r_j d)$$

В качестве распределения для величин, обратных дисперсиям, используем гамма-распределение:

$$\gamma(1/r_i|\alpha, 1/\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot (\frac{1}{r_i})^{\alpha-1} \cdot \exp\left(-\frac{\beta}{r_i}\right)$$

Для упрощения процедуры подбора параметров используем вместо двух параметров α и β один - μ. Выбор значения этого параметра обусловлен следующими условиями:

$$\mu \to 0 \Rightarrow \begin{cases} E(1/r_i) \to 1 \\ Var(1/r_i) \to 0 \\ \left(\sqrt{Var(1/r_i)}/E(1/r_i)\right) \to 0 \end{cases}$$

$$\mu \to \infty \Rightarrow \begin{cases} E(1/r_i) \to \infty \\ Var(1/r_i) \to \infty \\ \left(\sqrt{Var(1/r_i)}/E(1/r_i)\right) \to 1 \end{cases}$$

Одним из вариантов, удовлетворяющим этим условиям, является следующий: $\alpha = (1+\mu)^2/2\mu, \beta = 1/2\mu$

Критерий обучения

$$\left(\hat{\mathbf{a}}_t, \hat{b}_t, \hat{\mathbf{r}}, \hat{\boldsymbol{\delta}}_t | (\mathbf{x}_j, y_j), \mu, j = 1, ..., N_t, t = 1, ..., T\right) =$$

$$\arg\max_{\hat{\mathbf{a}}_t, \hat{b}_t, \mathbf{r}} \left[- \sum_{t=1}^{T} \sum_{i=1}^{n} a_{t,i}^2 / r_i - C \sum_{t=1}^{T} \sum_{i=1}^{N_t} \delta_{t,j} - \right.$$

$$- \frac{1}{2} \sum_{i=1}^{n} \log r_i - (\frac{(1+\mu)^2}{2\mu} - 1) \sum_{i=1}^{n} \log r_i - \frac{1}{2\mu} \sum_{i=1}^{n} \frac{1}{r_i} -$$

$$\left. - \frac{1}{2d} \sum_{t=2}^{T} (\mathbf{a}_t - \sqrt{1-d}\mathbf{a}_{t-1})^T (\mathbf{a}_t - \sqrt{1-d}\mathbf{a}_{t-1}) - \frac{1}{2d'} \sum_{t=2}^{T} (b_t - b_{t-1})^2 \right]$$

при ограничениях

$$1 - y_{t,j}(\mathbf{a}_t^T \mathbf{x}_{t,j} + b_t) < \delta_{t,j},$$

$$\delta_{t,j} > 0, j = 1, \ldots, N_t, t = 1, ..., T$$

Оценка параметров решающего правила

Для оптимизации критерия воспользуемся методом покоординатного спуска по двум группам переменных $\mathbf{r}$ и $\mathbf{a_t}, b_t, t = 1...N$
При зафиксированных дисперсиях $\mathbf{r}$ на каждом шаге получаем следующую задачу оптимизации для оценивания нестационарного направляющего вектора и порога

$$\frac{1}{2d}\sum_{t=2}^{T}(\mathbf{a}_t - \sqrt{1-d}\mathbf{a}_{t-1})^T(\mathbf{a}_t - \sqrt{1-d}\mathbf{a}_{t-1}) + \frac{1}{2d'}\sum_{t=2}^{T}(b_t - b_{t-1})^2 + \\ + \sum_{t=1}^{T}\sum_{i=1}^{n}\frac{a_{t,i}^2}{r_i} + C\sum_{t=1}^{T}\sum_{j=1}^{N_t}\delta_{t,j} \to \min_{\mathbf{a}_t, b_t, \boldsymbol{\delta}_t, t=1,...T}$$

с ограничениями

$$1 - y_{t,j}(\mathbf{a}_t^T\mathbf{x}_{t,j} + b_t) < \delta_{t,j},$$
$$\delta_{t,j} > 0, j = 1, \ldots, N_t, t = 1, ..., T$$

Оценка параметра селективности

При зафиксированных параметрах решающего правила на каждом шаге дисперсии коэффициентов можно найти по формуле

$$r_i = \frac{\sum_{t=1}^{T} a_{t,i}^2 + \frac{1}{2\mu}}{(1+\mu)^2/2\mu - 1/2}$$

- Использовалось множество данных, описывающее электронные письма, из репозитория UCI. https://archive.ics.uci.edu/ml/datasets/Spambase
- Всего представлено 4601 письмо, характеризуемых 58 признаками, упорядоченные по времени их получения.
- Данные классифицированы как "спам"и "не-спам". "Спам"составляет 39.4% (1813 объектов) всего множества.
- Для обучения использовались первые 3600 объектов. Контрольное множество составляли оставшийся 1001 объект.
- Для сравнения получаемых результатов использовалось несколько алгоритмов из состава программного пакета Massive Online Analysis (A. Bifet, G. Holmes, R. Kirkby, B. Pfahringer MOA: Massive Online Analysis http://sourceforge.net/projects/moa-datastream/ //Journal of Machine Learning Research (JMLR), 2010). Параметры алгоритмов подбирались на контрольном множестве.

Экспериментальное исследование на реальных данных
Результат

Алгоритм	Доля ошибочно классифицированных объектов к их общему числу, %
OzaBagASHT	22,278
OzaBagAdwin	20,879
SingleClassifierDrift	39.361
AdaHoeffdingOptionTree	23.876
LimAttClassifier	29,271
ApproxSVM	14,785

- Каждый момент времени двумя нормальными двумерными распределениями с математическими ожиданиями 1 и -1, соответственно, генерируется по 10 объектов каждого класса
- К каждому объекту добавляется еще 98 шумовых признаков;
- В каждый момент времени центры генерирующих распределений поворачиваются на 0.0314 радиана.

Алгоритм	θ
OzaBagASHT	11.16
OzaBagAdwin	12.45
SingleClassifierDrift	17.81
AdaHoeffdingOptionTree	7.22
LimAttClassifier	8.75
ApproxSVM	5.7

$$\theta = \frac{min[|\hat{a}|_1, |\hat{a}|_2]}{max[|\hat{a}|_3, |\hat{a}|_{100}]}$$

Алгоритм	θ
OzaBagASHT	3.82
OzaBagAdwin	2.42
SingleClassifierDrift	1.32
AdaHoeffdingOptionTree	1.12
LimAttClassifier	1.81
ApproxSVM	455.56