РАЗДЕЛ I. ЭКОНОМИКО-МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

УДК 330.4

ОБ ОДНОРОДНЫХ АНСАМБЛЯХ ПРИ ИСПОЛЬЗОВАНИИ МЕТОДА БУСТИНГА В ПРИЛОЖЕНИИ К КЛАССИФИКАЦИИ НЕСБАЛАНСИРОВАННЫХ ДАННЫХ

В.Н. Никулин, к. физ.-мат. наук, доц. кафедры математических методов в экономике
Электронный адрес: vnikulin.uq@gmail.com
С.А. Палешева, соискатель кафедры математических методов в экономике
Электронный адрес: svetlanka-pluha@yandex.ru
Д.С. Зубарева, соискатель кафедры математических методов в экономике
Электронный адрес: ZubarevaDasha@yandex.ru
Вятский государственный университет, 610000, г. Киров, ул. Московская, 36

В работе приведены результаты международного соревнования по анализу данных в рамках конференции PAKDD-2007, подготовленных и предоставленных потребительской финансовой компанией с целью поиска лучших решений проблемы перекрестной продажи. Кроме того, рассмотрены два других соревнования на платформе Kaggle, которые также имеют финансовую интерпретацию.

Ключевые слова: ансамбль и элементарный классификатор; градиентные методы оптимизации; бустинг; случайный лес; дерево решений.

1. Введение

Ансамбли (включая методы голосования и усреднения) – это алгоритмы обучения, которые основаны на множестве элементарных классификаторов (base-learners) [6]. Как правило, ансамбли элементарных решений вычисляются по методу выборочного среднего. Хорошо известен тот факт, что ансамбли зачастую являются более точными, чем составляющие их элементарные классификаторы [3]. Например, ансамбль решающих деревьев, или так называемый «случайный лес» (random forest) [5], может быть использован в качестве иллюстрации эффективного классификатора. Другой пример – бэггинг (bagging) [4] с использованием в качестве элементарного классификатора метода опорных векторов (bagged support vector machines) – также очень важен, поскольку прямое применение этого метода ко всем данным не всегда является возможным. В случае применения метода опорных векторов мы заинтересованы в ограничении объема выборки, который совпадает с размерностью соответствующей ядерной матрицы. Бэггинг является очень перспективным методом, согласно которому каждый элементарный классификатор, используемый в ансамбле, основан на данных, которые были отобраны случайным образом из обучающей выборки без замещения.

Наш подход представляет компромисс между двумя основными интересами. С одной стороны, нам хотелось бы иметь дело со сбалансированными данными, с другой стороны, мы заинтересованы в использовании всей доступной информации. Соответственно мы принимаем во внимание большое число n относительно сбалансированных подмножеств, где каждое отдельное подмножество включает 2 части: 1) почти все «положительные» элементы (меньшинство) и 2) приблизительно такое же количество случайно отобранных «отрицательных» элементов. Метод сбалансированных случайных подмножеств является общим и может быть использован в сочетании с различными элементарными классификаторами.

В экспериментальной части статьи мы представили результаты, которые были получены на основе реальных данных, использованных в ходе международного соревнования по анализу данных PAKDD-2007[1]. Отметим, что данные

[1] http://lamda.nju.edu.cn/conf/pakdd07/dmc07/

являются сильно несбалансированными со значительно меньшим числом положительных случаев (1.49%), которые имеют следующую практическую интерпретацию: потребитель взял ипотеку в данной компании в течение 12 месяцев после открытия кредитной карты.

Регулируемая линейная регрессия (РЛР) представляет пример наиболее простой решающей функции [2]. В сочетании с квадратичной функцией потерь РЛР имеет существенное преимущество: используя метод градиентного спуска, мы можем оптимизировать размер шага. В результате процесс убывания целевой функции ускоряется.

По определению, регрессионные коэффициенты могут быть рассмотрены в качестве естественных показателей влияния соответствующих признаков [1]. В нашем случае мы имеем n векторов регрессионных коэффициентов и можем использовать их для изучения устойчивости влияния соответствующих признаков. Удачный отбор признаков может значительно снизить уровень перетренировки [10]. Мы исключаем признаки с неустойчивыми коэффициентами, затем пересчитываем регрессионные коэффициенты. Отметим, что устойчивость (степень влияния) коэффициентов может быть вычислена при помощи различных методов [9]. Например, мы можем применить t-критерий, который определен как отношение среднего к стандартному отклонению (так называемый метод среднее/дисперсия). Предлагаемый алгоритм является гибким в использовании. Мы не утверждаем, что алгоритм будет работать оптимально во всех возможных приложениях, поэтому возможность регулирования и подбора подходящих параметров может оказаться очень полезной.

Начальные результаты, полученные при использовании РЛР в ходе соревнования PAKDD-2007, представлены в [13]. В следующей работе [14], используя бустинг алгоритм (LogitBoost) [8] в качестве элементарного классификатора, мы улучшили все известные нам результаты. В настоящей работе мы предлагаем применить принципы бустинга к формированию семейств однородных ансамблей элементарных классификаторов. Более сложные наблюдения (согласно тренировочной ошибке) имеют большую вероятность быть отобранными [7]. С другой стороны, представляется целесообразным использовать принципы полунаправляемого обучения [11]. Согласно этим принципам, более простые наблюдения из тестирующей базы данных могут быть отобраны в тренировочное подмножество.

Работа структурирована следующим образом: в разделе 2 описан метод сбалансированных случайных подмножеств и метод среднее/дисперсия; в разделе 3.1 представлены новые методы; в разделе 4 поясняется экспериментальный процесс и наиболее существенные детали, относящиеся к рассматриваемой проблеме; наконец, раздел 5 подводит итог данной работы.

2. Техника моделирования

Пусть $X = (x_t, y_t)$, $t = 1, \ldots, m$, – выборка, основанная на результатах наблюдений, где x_t обозначает ℓ-мерный вектор признаков, y_t – целевая переменная: $y_t \in \{-1,1\}$. На практике значения целевой переменной y_t могут быть не доступны, и задача состоит в их оценке при помощи вектора признаков.

Рассмотрим простейший линейный классификатор

$$u_t = u(x_t) = \sum_{j=0}^{l} w_j x_{tj}, \qquad (1)$$

где x_{t0} – постоянный член.

Мы можем определить решающее правило как функцию классификатора (1) и параметра сдвига Δ:

$$f_t = f(u_t, \Delta) = \begin{cases} 1, \textit{если } u_t \geq \Delta; \\ 0, \textit{иначе.} \end{cases}$$

В данной работе мы использовали AUC (area under receiver operating curve) в качестве критерия оценки, где AUC – площадь под характеристической кривой (ROC). По определению, ROC-кривая показывает пропорцию верно классифицированных положительных наблюдений (True Positive Rates) как функцию пропорции неверно классифицированных положительных примеров (False Positive Rates).

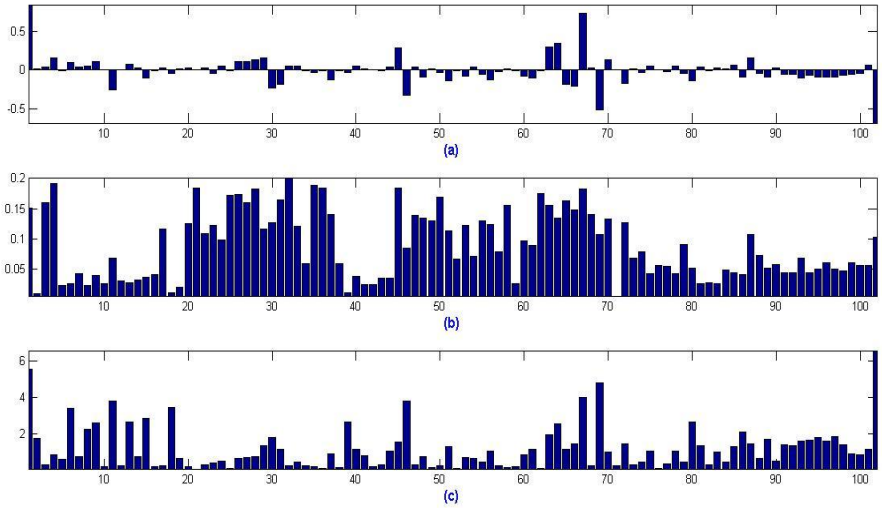
Согласно предложенному методу, мы принимаем во внимание большое число элементарных классификаторов, где каждый классификатор строится на основе относительно сбалансированного подмножества со всеми «положительными» и случайно отобранными (без замещения) «отрицательными» наблюдениями. Окончательная функция решения вычисляется как логистическое среднее элементарных классификаторов.

Определение. Обозначим упомянутые ранее случайно сформированные сбалансированные подмножества как $RS(\alpha, \beta, n)$, где α – число положительных случаев, β – число отрицательных случаев, n – общее число случайных множеств.

Модель сбалансированных случайных подмножеств включает два очень важных регулирующих параметра: 1) n и 2) $q = \frac{\alpha}{\beta} \leq 1$ - пропорция положительных случаев, при этом значение n должно быть достаточно большим, а значение q не может быть слишком маленьким. Мы рассматриваем n подмножеств из базы дан-

ных X, каждое из которых включает α положительных и $\beta = k \cdot \alpha$ отрицательных наблюдений, где $k \geq 1$, $q = 1/k$. Применяя метод градиентного спуска [12], мы можем вычислить матрицу коэффициентов линейной регрессии:

$$W = \{w_{ij},\ i = 1, \ldots, n,\ j = 0, \ldots, l\}.$$



Метод среднее/дисперсия: a – средние значения (μ); b – стандартные отклонения (s); c – отношения = |μ| / s (см. детальное описание в разделе 2)

Метод среднее/дисперсия (рисунок) [12] был предложен в качестве критерия для отбора наиболее важных признаков с целью уменьшения эффекта переобучения. Используя следующие отношения, мы можем оценить состоятельность влияния конкретных признаков:

$$r_j = \frac{|\mu_j|}{s_j},\ j = 1, \ldots, l, \qquad (2)$$

где μ_j и s_j – среднее значение и среднеквадратическое отклонение, соответствующие j-му столбцу матрицы W.

Низкий уровень отношения r_j означает, что влияние j-го признака неустойчиво. Отбор признаков осуществляется согласно условию

$$r_j \geq \gamma > 0.$$

Окончательная функция решения

$$f_t = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \exp\{-\tau \cdot u_{ti}\}},\ \tau > 0, \qquad (3)$$

вычислялась как логистическое среднее элементарных классификаторов

$$u_{ti} = \sum_{j=0}^{l} w_{ij} \cdot x_{tj},$$

где коэффициенты регрессии w были пересчитаны после отбора признаков на основании критерия (2).

Замечание. В разделе 4 продемонстрировано, что эффективность классификатора может быть существенно улучшена, если мы используем в уравнении (3) нелинейный элементарный классификатор. Наиболее популярным примером нелинейного классификатора могут служить решающие деревья.

3. Алгоритмы поддержки (бустинг алгоритмы)

Бустинг функционирует посредством последовательного применения классификационного алгоритма к обучающей выборке, в которой различные наблюдения имеют разные весовые коэффициенты, контролирующие важность наблюдений. Представляется естественным повысить внимание к «сложным» наблюдениям из тренировочной базы данных. Окончательное решение вычисляется по принципу взвешенного большинства голосов, взятого от последовательности элементарных классификаторов. Отметим, что для многих элементарных классификаторов такая несложная процедура существенно повышает эффективность.

3.1. Новые принципы «сложности» для бустинга и «простоты» для полунаправляемого обучения

В наших предыдущих публикациях мы использовали последовательность сбалансированных подмножеств, каждое из которых было отобрано независимо от предыдущих подмно-

жеств. Однако представляется логичным использовать в задаче отбора принципы бустинга на основе последнего элементарного классификатора. Так, мы предлагаем учитывать «сложность» наблюдения при принятии решения относительно включения этого наблюдения в тренировочное подмножество. Согласно основным принципам бустинга, уровень «сложности» наблюдения увеличивает вероятность отбора этого наблюдения.

3.1.1. Принципы бустинга в применении к формированию сбалансированных подмножеств

Предположим, что $s_t^{(\alpha)}$ – прогноз, соответствующий наблюдению *t*, *α* – индекс сбалансированного подмножества. Наблюдение *t* будет отобрано в следующее тренировочное подмножество *α+1* при условии

$$\xi \le \xi_1, \text{ если } y_t = 1;$$

$$\xi \le \xi_2, \text{ иначе,}$$

где

$$\xi_i = c_{i1} + (c_{i2} + c_{i3} \cdot \varphi) \cdot \cdot w(y_t, s_t^{(\alpha)}),\ i = 1,\ldots,2, \quad (4)$$

$$w(y, s) = | y - s |^{\beta}, \quad (5)$$

где ξ и φ – стандартные равномерные случайные величины,

β>0 и $c_{ij} > 0, i = 1,\ldots,2,\ j = 1,\ldots,3,$ – регулирующие параметры.

В качестве примера мы можем выбрать β = 1,75. Рекомендуемые значения коэффициентов *c* даны в табл. 1.

Таблица 1

Рекомендуемые значения коэффициентов $c_{ij} > 0, i = 1,\ldots,2,\ j = 1,\ldots,3$

0,75	0,35	0,15
0,02	0,003	0,002

Таблица 2

Некоторые новые (улучшенные) результаты, полученные при помощи функций *ada* и *gbm* (R); *neural* и *kridge* (CLOP, Matlab), (в столбце "RS" указано количество сбалансированных подмножеств)

Модель	AUC	RS	Параметры
ada	0,7052	20	*loss="l", iter=50,nu=0.25,type="discrete"*
gbm	0,7037	50	*distribution="adaboost",n.trees=1000,shrinkage=0.015*
-	-	-	*interaction.depth=14,n.minobsinnode=5*
neural	0,6874	200	*chain({standardize , neural({'units=6','maxiter=200'})})*
kridge	0,6867	200	*kridge({'shrinkage=0.01', 'balance=1'})*

3.1.2. О принципе «простота» в отношении тестовой базы данных

Формулы (4) и (5) могут быть применены только в отношении тренировочных данных, когда целевая переменная известна. Однако мы имеем значительное количество данных для тестирования (значение целевой переменной неизвестно). Эти данные также могут быть использованы в процессе машинного обучения, известного как полунаправляемое. Наблюдения из тестовой базы данных отбираются в соответствии с решающей функцией, исходя из имеющейся информации о том, что «негативные» наблюдения составляют подавляющее большинство. Таким образом, более «простые» наблюдения (из тестовой базы данных) имеют большую вероятность для отбора в случайное подмножество.

4. Результаты экспериментов

Исходные данные по ипотеке состоят из двух частей: 1) обучающая база данных с 700 положительными и 40000 отрицательными случаями/ наблюдениями; 2) тестовая база данных (целевая переменная отсутствует) с 8000 случаями. Каждое наблюдение, которое соответствует определенному клиенту, представляет собой вектор, состоящий из 40 непрерывных или категориальных объясняющих признаков.

4.1. Подготовка данных

Используя стандартный метод, мы привели категориальные признаки к численному (двоичному) формату. Также мы нормализовали непрерывные признаки, для того чтобы их значения лежали в интервале [0,...,1]. В результате описанных преобразований мы подготовили базу данных, содержащую *l=101* численных вторичных признаков. Затем мы применили фильтрацию согласно методу сред-

нее/дисперсия. Как следствие, число признаков было уменьшено со 101 до 44 (см. рисунок).

Таблица 3

Список 6 самых значимых признаков

№ п/п	Признак	μ	r
1	Кол-во контактов по ипотеке за последние 6 месяцев	0.729	4
2	Возраст	-0.683	6.6
3	Кол-во контактов по займам за последние 12 месяцев	-0.516	4.8
4	Кол-во контактов за последние 3 месяца	0.342	2.54
5	Кол-во членов семьи	-0.322	3.82
6	Кол-во контактов за последний месяц	0.299	1.92

4.2. Результаты

Сорок семь участников из различных представительств, в том числе научных сообществ, университетов, производственных объединений и консультационных компаний, представили на рассмотрение решения с результатами от 0.4778 до 0.701 в терминах AUC. Наш результат был 0.6888, что соответствует девятому месту в соревновании.

Заметим, что тренировочное значение критерия AUC, которое соответствует финальному решению, было 0.7253. Расхождение между тренировочным и тестовым результатами в случае модели с 44 признаками представляется довольно значительным. Начальное предположение относительно этого расхождения (после опубликования результатов) состояло в том, что это следствие перетренировки. Мы провели серию экспериментов с дополнительно отобранными признаками, однако это не привело ни к каким значительным улучшениям (см. табл. 4).

Далее мы решили применить в качестве элементарного классификатора ada-функцию в R. Лучший результат эксперимента AUC = 0.7023 был получен при использовании следующих параметров: *loss = e, v = 0.3, type = gentle.*

Таблица 4

Число использованных признаков

Число признаков	Тест AUC	Элементарный классификатор
44	**0.7023**	**LogitBoost**
44	0.688	РЛР
30	0.689	РЛР
17	0.688	РЛР
4	0.6558	РЛР

***** Некоторые интерпретации признаков (в случаях моделей с 4 и 17 признаками) могут быть найдены в табл. 3, 5 и 6.

Таблица 5

Четыре основных признака

№	Признак	μ	s	r
1	N1 (см. табл. 1)	1.1454	0.1011	11.3294
2	N3	-0.6015	0.0663	-9.0751
3	N5	-0.1587	0.0778	-2.0395
4	Возраст	-0.6831	0.0806	-8.4794

При проведении эксперимента мы использовали 100 случайных сбалансированных подмножеств. Отметим, что мы не смогли улучшить результаты, используя большее количество случайных подмножеств. Интересен тот факт, что мы не вносили никаких изменений в технику первичной обработки данных, которая применялась ранее [13], и использовали те же данные при проведении экспериментов в настоящей работе, а также в [14].

4.3. Обсуждение закономерностей, полезных для бизнеса, которые были выявлены как следствие анализа данных

Метод среднее/дисперсия предоставляет отличные возможности для оценки важности объясняющих признаков. Мы можем принять во внимание 2 показателя: 1) среднее значение μ; 2) t-статистику *r*, которые определены в формуле (2).

Таблица 6

17 основных вторичных признаков, отобранных методом среднее/дисперсия

№	Признак	μ	s	r
1	Семейное положение: женат (замужем)	0.0861	0.028	3.0723
2	Семейное положение: не женат (не замужем)	0.0419	0.0236	1.7786
3	Семейное положение: гражданский брак	0.09	0.0438	2.0572
4	Семейное положение: вдовец (вдова)	-0.2754	0.0766	3.594
5	Код заема: ипотека	0.0609	0.0191	3.1838
6	Код аренды: родители	-0.1285	0.0341	3.7692
7	Текущее жильё (в месяцах)	-0.0449	0.0101	4.4555
8	Текущая работа (в месяцах)	-0.0288	0.0111	2.586
9	Состав семьи (кол-во членов)	-0.3298	0.0807	4.085
10	Кол-во контактов за последний месяц	0.3245	0.183	1.7736
11	Кол-во контактов за последние 3 месяца	0.1296	0.1338	0.9691
12	Кол-во контактов по ипотеке за последние 6 месяцев	0.8696	0.1359	6.3982
13	Кол-во контактов по займам за последние 12 месяцев	-0.6672	0.0795	8.3905
14	Место жительства клиента: = 2	-0.1704	0.05	3.4067
15	Место жительства клиента: = 8	-0.1216	0.0397	3.063
16	Сектор клиента: = 9	-0.0236	0.0317	0.7453
17	Возраст	-0.654	0.0962	6.8015

В соответствии с табл. 3 и 6 мы делаем вывод, что более молодые люди (возраст: μ = -0.654) с меньшим составом семьи (μ = -0.3298), которые интересовались ипотекой в течение последних 6 месяцев, имеют большую вероятность подать заявление на ипотечный заем.

С другой стороны, обращения в бюро относительно займов представляют собой негативный фактор (μ = -0.672).

Принимая во внимание общие демографические характеристики, например семейное положение, мы можем утверждать, что одинокие люди менее заинтересованы (μ = -0.2754) в ипотечном заеме.

Кроме того, следует заметить, что постоянная работа (μ = -0.0288) или постоянное (длительное) местожительство (μ = -0.0449) могут быть рассмотрены как негативные факторы. По-видимому, эти люди уже имеют один дом или более, поэтому с большим нежеланием будут осуществлять дальнейшие инвестиции.

Замечание. Эксперименты с функцией “tree” в R подтвердили тот факт, что признак «кол-во заявок по ипотеке за последние 6 месяцев» является самым важным.

Мы полагаем, что при помощи нашей модели, разработанной в ходе соревнования, компания сможет существенно улучшить свою рекламную программу. Например, представляется целесообразной целевая кампания прямой рассылки писем с приглашениями относительно кредита тем клиентам, которых автоматическая система идентифицировала как наиболее перспективных. В независимой тестовой базе данных с 8000 случаев, из которых 350 случаев положительные, мы можем отсортировать клиентов в убывающем порядке согласно решающей функции, которая соответствует AUC = 0.7023 (см. табл. 4). В результате мы выявили интересный факт: 50, 60 и 70% всех «положительных» клиентов содержатся в областях из 1770, 2519 и 3446 клиентов, находящихся в классификационном списке на верхних позициях (согласно решающей функции).

4.4. Дополнительные эксперименты: финансовое моделирование при использовании методов вычислительной статистики

Используя вышеизложенный принцип сложность/простота (см. раздел 3.1) в популярном международном соревновании Credit на платформе Kaggle[2], мы продемонстрировали девятый результат из 970 активных участников. Кроме того, мы выиграли второй приз в аналогичном соревновании Carvana (582 участника, также на платформе Kaggle). Отметим, что оба соревнования относятся к области финансового моделирования.

В случае соревнования Credit необходимо было спрогнозировать платежеспособность клиента при предоставлении кредита на ближайшие два года на основе некоторых исторических и демографических данных. В случае соревнования Carvana необходимо было оценить качество автомобиля, продаваемого на вторичном рынке, при использовании различных цен общего характера на аналогичные автомобили.

[2] http://www.kaggle.com

5. Заключительные замечания

Предложенный метод основан на большом количестве сбалансированных случайных подмножеств и включает 2 основных этапа: 1) фильтрация признаков и 2) машинное обучение. В ходе соревнования PAKDD-2007 мы преодолели оба этапа, используя линейную регрессию. Заметим, что предложенный метод является общим и может быть реализован на основе различных элементарных классификаторов [13]. В последующей работе [14] мы представили улучшенные результаты, которые были получены при использовании пакета ADA в R.

В данной работе мы предложили и рассмотрели новые методы (см. раздел 3.1), которые могут быть использованы для отбора тренировочных подмножеств и формирования специальных семейств однородных ансамблей элементарных классификаторов. Опираясь на результаты, полученные при анализе данных PAKDD-2007 и двух других соревнований на платформе Kaggle, можем сделать вывод о перспективности предложенных методов.

Список литературы

1. *Шешукова Т.Г., Колесень Е.В.* Экономический потенциал предприятия: сущность, компоненты, структура // Вестн. Перм. ун-та. Сер. Экономика. 2011. Вып. 4. С. 118-127.

2. *Barinova O.* Incorporating posterior estimates into adaboost // Pattern Recognition and Image Analysis. 2009. Vol. 19 (3). P. 421-434.

3. *Biau G., Devroye L., Lugosi G.* Consistency of random forests and other averaging classifiers // Journal of Machine Learning Research. 2007. Vol. 9. P. 2015-2033.

4. *Breiman L.* Bagging predictors // Machine Learning. 1996. Vol. 24. P. 123-140.

5. *Breiman L.* Random forests // Machine Learning. 2001. Vol. 45(1). P. 5-32.

6. *Djukova E. V., Zhuravlev Yu. I., Sotnezov R. M.* Construction of an ensemble of logical correctors on the basis of elementary classifiers // Pattern Recognition and Image Analysis. 2011. Vol. 21(4). P. 599-605.

7. *Freund Y., Schapire R.* A decision-theoretic generalization of online learning and an application to boosting // J. Comput. System Sciences. 1997. Vol. 55. P. 119-139.

8. *Friedman J., Hastie T., Tibshirani R.* Additive logistic regression: a statistical view of boosting // Annals of Statistics. 2000. Vol. 28. P. 337-374.

9. *Frohlich B., Rodner E., Kemmler M., Denzler J.* Large scale gaussian process classification using random decision forests // Pattern Recognition and Image Analysis. 2012. Vol. 22 (1). P. 113-120.

10. *Guyon I., Weston J., Barnhill S., Vapnik V.* Gene selection for cancer classification using support vector machines // Machine Learning. 2002. Vol. 46. P. 389-422.

11. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer-Verlag, 2001.

12. *Nikulin V.* Learning with mean-variance filtering, SVM and gradient-based optimization // International Joint Conference on Neural Networks. Vancouver, BC, Canada. 2006. July 16-21. IEEE. P. 4195–4202.

13. *Nikulin V.* Classification of imbalanced data with random sets and mean-variance filtering // International Journal of Data Warehousing and Mining. 2008. Vol. 4. P. 63–78.

14. *Nikulin V., McLachlan G., Shu K.-N.* Ensemble Approach for the Classification of Imbalanced Data. Proceedings of the 22nd Australasian Joint Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence, Springer-Verlag, edited by A.Nicholson and X.Li. Melbourne, 2009. P. 291-300.