УДК 004.822

Классификация узлов ассоциативно-вербальной сети по когнитивным областям: этап построения классификаторов[1].

Ю.Н. Филиппович, А.В. Сиренко

Ю.Н. Филиппович – к.т.н., профессор каф. Медиасистем и технологий, Московского государственного университета печати им. Ивана Фёдорова

А.В. Сиренко – соискатель каф. Медиасистем и технологий, Московского государственного университета печати им. Ивана Фёдорова

y_philippovich@mail.ru, alexander.sirenko@gmail.com

В статье рассматривается построение классификаторов узлов ассоциативно-вербальной сети по когнитивным областям при наличии обучающей выборки. Классификация производится комбинацией линейных и локальных методов регрессионного анализа, с построением классификатора для каждой когнитивной области.

The paper is devoted to classification of associative-verbal network's nodes with cognitive areas, with provided educational set of nodes. Classification is carried out by combination of linear and local methods of regression analysis. For each cognitive area own classifier is constructed.

Ключевые слова: ассоциативная сеть, классификация, когнитивная область, регрессионный анализ.

Key words: Associative network, classification, cognitive domain, regression analysis.

[1] В данной статье приводятся результаты исследований, выполненных при поддержке грантов РГНФ №12-04-12039в, №12-04-12059в и гранта Президента РФ №НШ-3661.2012.6

Одной из центральных идей моделирования языкового сознания, представленных в работах авторов (на них имеются ссылки в литературе), и получивших теоретическое и практическое воплощение, является «идея построения когнайзера» — семиотической машины (автомата), как инфокогнитивной компьютерной системы/технологии, реализующей возможные модели операциональных отношений, существующих в сознании носителя языка культуры, между языковыми единицами (ЯЕ), которые зафиксированы в различных ассоциативных и когнитивных экспериментах. Результатами этих экспериментов являются две сетевые конструкции — ассоциативно-вербальная и когнемная сети (соответственно: АВС и КС).

Теоретическое и практическое воплощение идеи когнайзера актуально как для решения общей проблемы когниции, так и для решения частных проблем многих наук, междисциплинарных исследований и разработок, для интегрированных и/или конвергентных технологий, в частности NBIC-технологий.

***

1. Постановка задачи.

Свойство принадлежности когнитивной области является типичным для сетей, объединяющих элементы множества предметных областей.

Построение подобных сетей требует автоматизации назначения данных

атрибутов. При построении семантической сети за основу, как правило, берутся лексикографические объекты, некоторые из которых содержат признак области (например, тематические толковые и идеографические словари, тезаурусы), другие нет (словари синонимичных, антонимичных отношений, частотно-дистрибутивных связей). В автоматизированной системе научных исследований психолингвистических экспериментов (далее АСНИ) [1] используются ассоциативно-вербальная сеть (далее AVN), а также таблица когнитивных единиц (когнем). AVN включает узлы-понятия, объединенные при помощи направленных взвешенных ассоциативных связей:

$$AVN = (V, E),$$

где V– множество узлов сети;

E– множество связей сети, представленных кортежами

$$E = \{e_{ij}\}$$

$$e_{ij} = (v_i, v_j, p_{ij}),$$

где p_{ij}– вероятность реакции v_jна стимул v_iв ассоциативном эксперименте.

Под когнемой подразумеваем пятикомпонентную структуру Ю.Н. Караулова [2], включающую в себя: ЗНАК, ФОРМУЛУ_СМЫСЛА, ОБЛАСТЬ, СПОСОБ и ФУНКЦИЮ.

ЗНАК представляет собой понятие, зафиксированное в когнеме. По условиям эксперимента, знак представлен единичным словом.

ФОРМУЛА_СМЫСЛА представляет собой естественно-языковое суждение, относящееся к ЗНАКу некоторым СПОСОБом, например: дефиниция, загадка, метафора.

ОБЛАСТЬ содержит неупорядоченное множество когнитивных областей, к которым принадлежит концепт знака.

ФУНКЦИЯ отражает значимость когнемы с точки зрения носителя языка, определяющуюся ее знанием.

Основу семантической сети АСНИ составляет AVN, узлы которой не имеют атрибутов принадлежности когнитивной области. Множество когнем такими атрибутами обладает.

В [3] задача прогнозирования с количественной (quantitative) выходной величиной именована задачей регрессионного анализа, а задача с качественной (qualitative) обозначена как задача классификации. Мы будем рассматривать задачу нечеткой классификации, в которой:

- узел принадлежит подмножеству множества классов, оно может быть пустым;
- принадлежность элемента классу имеет вероятностную характеристику.

Задача определения когнитивной области узлов будет нами сформулирована в следующем виде: для каждого узла v_iопределить неупорядоченное множество пар:

$KAr_{ij} = (area_j, K_{ij})$,

где $area_j$– когнитивная область;

K_{ij}- коэффициент принадлежности узла v_iк области $area_j$. Данная задача будет именоваться нами далее классификацией узлов сети по когнитивным областям.

2. Исходные данные для классификации.

Обучающую выборку узлов будем строить на основе пересечения множеств ЗНАКОВ когнем и узлов ассоциативной сети по признаку совпадения написания, присвоив им когнитивные области с единичным коэффициентом принадлежности. При данном принципе переноса областей ухудшению результата способствует омонимия и многозначность. Необходимо производить перенос в лемматизированной сети, поскольку ЗНАКи когнем в большинстве представлены в основной форме.

Число узлов сети: 28 288, множество когнем: 18 281. При пересечении множеств формируется обучающая выборка из 4 076 узлов сети. Подлежит классификации: 24 212 узлов. Заметим, что среди узлов есть местоимения, служебные части речи и другие элементы, не относящиеся к некоторой предметной области, что должно найти выражение в результатах классификации.

Определим признаки, на основе которых будет производиться классификация. Эти признаки в обучающей выборке служат расчету параметров классификатора, а в классифицируемой – непосредственно классификации. Будем исходить из предположения, что узлы сети одной когнитивной области располагаются в виде компактной группы с позиции достижимости в сети. Данный подход близок к идее латентно-семантического анализа, использующего параметры встречаемости слов для индексирования текстовых документов [4].

В работе [5] предложена методика кластеризации ассоциативной сети, метрика близости узлов-стимулов в которой определяется через общность множеств их реакций. С позиции специалистов психолингвистики это согласовано с представлением семантики слова через семантические множители, а именно, оценку сходства через анализ множеств [6, гл. 7].

3. Классификационные признаки.

Классификатор будет строиться отдельно для каждой области $area_{cur}$. Это вызвано предположением, что свойства классификатора будут зависеть от числа элементов обучающей выборки по $area_{cur}$, а также, области могут иметь более общий или ограниченный характер.

Обозначим как «→» непосредственную достижимость узла v_iиз v_j. Тогда узлы контекста v_iдля области $area_{cur}$есть совокупность множеств узлов:

1. $Vout_i = v_j : v_i \to v_j, v_j \in area_{cur}, i \neq j$- непосредственно достижимых из v_i для текущей области;
2. $Vin_i = v_j : v_j \to v_i, v_j \in area_{cur}, i \neq j$– из которых v_iнепосредственно достижим для текущей области;
3. $Vcom_i = v_j : Vout_i \cap Vin_j \neq \emptyset$– узлы текущей области, имеющие общие реакции с v_i.

Представим контекст узла v_iв виде кортежа параметров:

$$Context_{i,area} = \left(Kin_{i,area_{cur}}, Kout_{i,area_{cur}}, Kcom_{i,area_{cur}}\right), \quad (1)$$

$Kin_{i,area}$– сумма весов входящих связей от узлов Vin_i.

$Kout_{i,area}$– сумма весов входящих связей от элементов $Vout_i$.

$Kcom_{i,area} = \sum_{\forall k} min(p_{ik}, p_{kj}) : v_j \in Vcom_i$ – сумма минимумов пар связей к общему узлу реакции v_k от узлов V_i и V_j, $j \neq i$.

Методика расчета весов связей приведена в [7]. Напомним, что вес связи между узлами v_i и v_j ассоциативной сети соответствует вероятности фиксации в эксперименте реакции v_j на стимул v_i. Рисунок 1 иллюстрирует, какие связи мы учитываем при расчете параметров контекста.

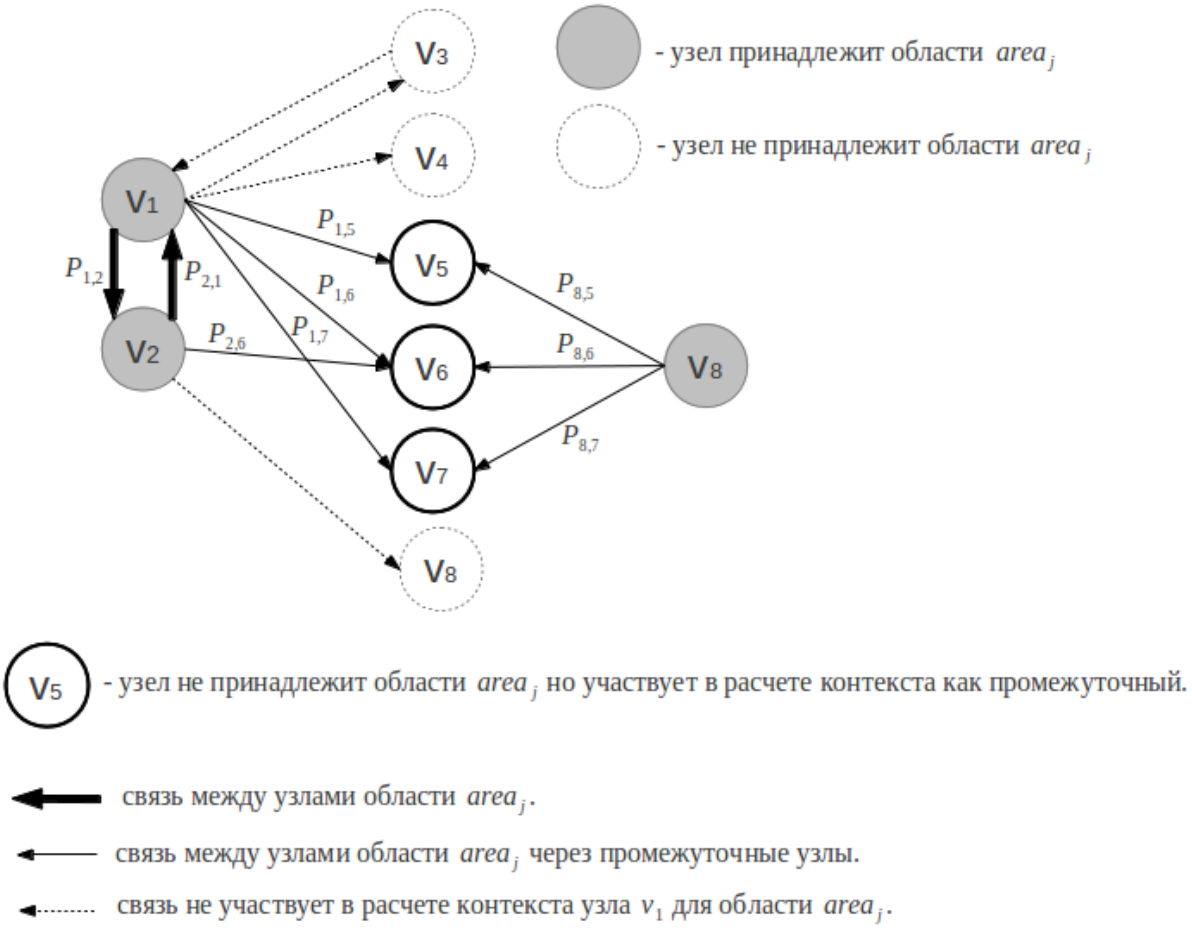

Рисунок 1. Контекст узла v_1

Для узла v_1 и области $area_j$ параметры контекста:

$Kin_{1,area_j} = P_{2,1}$;

$Kout_{1,area_j} = P_{1,2}$;

$Kcom_{1,area_j} = \min(P_{1,5}, P_{8,5}) + \min(P_{1,6}, P_{2,6}, P_{8,6}) + \min(P_{1,7}, P_{8,7})$;

Обучающие выборки для отдельных областей могут пересекаться на множестве узлов сети выборки, но будут иметь собственные значения параметров контекста, в общем случае не совпадающие. Алгоритм классификации узлов ассоциативной сети изображен ниже.

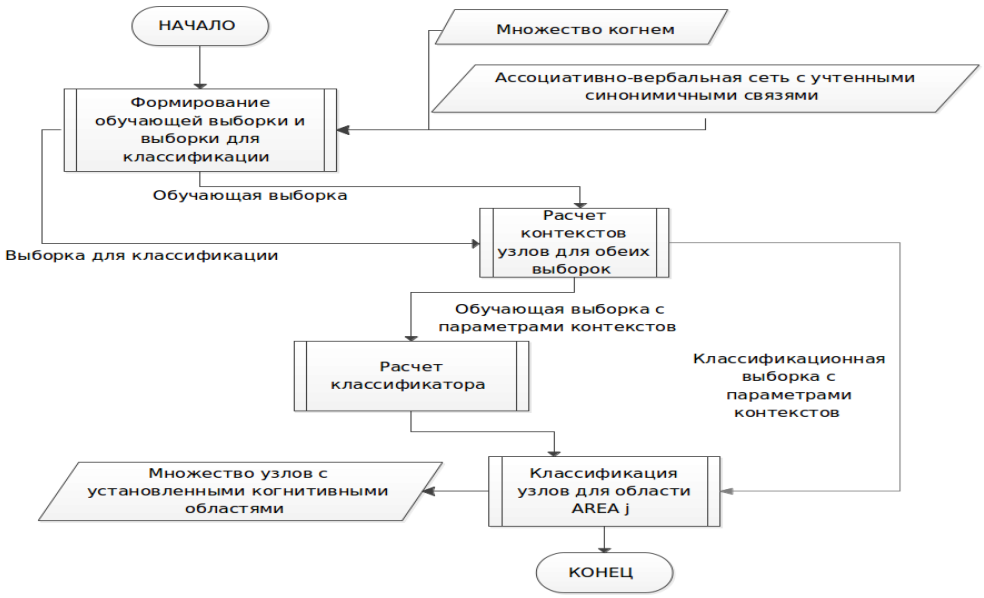

Рисунок 2. Алгоритм классификации узлов на примере области $\boldsymbol{Area_j}$.

Фрагмент таблицы узлов сети после расчета контекстов на примере области АВИАЦИЯ:

Узел	Kin	Kout	Kcom	Принадлежит области	Принадлежит обучающей выборке
парашют	0	0,172	1,008	TRUE	TRUE
пантера	0	0	0	FALSE	TRUE
Венера	0	0	0	FALSE	FALSE
Пар	0	0	0,186	FALSE	TRUE

Таблица 1: Контексты узлов.

По признаку «Принадлежит обучающей выборке» можем разделить таблицу на обучающую выборку и выборку, подлежащую классификации, после чего перейти к построению модели классификатора.

4. Построение классификатора.

Данные для обучения представляют собой выборку, каждый элемент которой включает три вещественных предикторных (прогностических) величины и одну результирующую вещественную величину «принадлежит области» / «не принадлежит области» с диапазоном значений от [0..1].

Выбор классификационной модели должен находиться в соответствии распределению объектов, подлежащих классификации. Например, использование локальной классификационной модели (K-ближайших) при линейной зависимости распределения по классам от предикторных параметров потребует большей обучающей выборки, чем для линейной модели для достижения равного качества классификации. И наоборот, линейная модель при отсутствии на то оснований приводит к низкому качеству классификации. Закономерно, что локальная регрессионная модель LOWESS, а также ядерные непараметрические модели, не накладывая жестких ограничений на модель, требуют большой и «плотной» обучающей выборки [8,9], которой мы не располагаем.

В работе предлагается совместить подходы локальных классификаторов (K-ближайших) и структурированных моделей (линейных). Это позволит

учитывать окружение каждого классифицируемого узла и настроить классификатор при небольших размерах обучающей выборки (меньших, чем требуется методу K-ближайших). Параметры окружения узлов сети были определены нами при расчете контекста (1), что соответствует локальному подходу.

Рассмотрим две группы линейных классификационных моделей, являющихся базовыми и подходящими для поставленной задачи: модель логистической регрессии и наивный классификатор Байеса.

В основе сети Байеса лежит формула вычисления апостериорной вероятности некоторого события при наступлении некоторых условий. В нашем случае этими условиями могут быть значения коэффициентов $Kin_{i,area}, Kout_{i,area}, Kcom_{i,area}$, а наступающим событием – принадлежность к области. Чтобы снизить требования к размеру обучающей выборки, будем рассматривать наивный дискретный классификатор Байеса, для чего необходимо перейти от непрерывных значений коэффициентов к некоторым интервалам, проведя дискретизацию. При нормальном распределении предикторных параметров для фиксированного результата классификации и увеличении обучающей выборки, наивный классификатор Байеса становится эквивалентным классификатору логистической регрессии [10]. Так как наивный классификатор Байеса более требователен к исходным данным

классификации (не с точки зрения объема обучающей выборки, а с позиций независимости условий и характера их распределения), выбор делается в пользу логистической регрессии.

5. Расчет параметров модели.

Множественная линейная регрессия позволяет прогнозировать значение некоторой величины на основе значений предикторных величин:

$$Y = c_{inter} + c_{in} * X_{in} + c_{out} * X_{out} + c_{com} * X_{com}, \qquad (2)$$

где $c_{inter}, c_{in}, c_{out}, c_{com}$– коэффициенты регрессии;

X_{in}, X_{out}, X_{com}– предикторные переменные;

Y- абстрактная величина.

Подписи в нижнем регистре выражают соответствие переменных и коэффициентов параметрам контекста классифицируемого узла сети.

Согласно свойствам логистической регрессии значение функции

$$r(Y) = \frac{1}{1 + e^{-Y}} \qquad (3)$$

при соблюдении требований к плотностям распределений классов (должны принадлежать к экспоненциальному семейству), интерпретируется как вероятность принадлежности классу,

закодированному большим значением прогнозируемой величины [11]. Так как мы не проверяем данное свойство плотности вероятности прогнозируемой величины, то будем использовать величинуr совместно с порогом отсечения (optimal cut-off value, cutoff).

Типовые стратегии определения порога отсечения связаны с требованиями к чувствительности (Se) и специфичности (Sp) модели [12]:

- превышение Se или Sp некоторого значения;
- $max(Se + Sp)$;
- $min(Se - Sp)$;

Здесь $Se = \frac{TP}{TP+FN}$, $Sp = \frac{TN}{TN+FP}$;

TP – верно классифицированные положительные примеры;

FN – неверно классифицированные отрицательные примеры;

TN – верно классифицированные отрицательные примеры;

FP – неверно классифицированные положительные примеры;

6. Выходные данные расчета параметров модели.

Расчет параметров модели осуществляется в программном пакете статистических расчетов R [13] с помощью встроенной функции glm, осуществляющей подбор коэффициентов регрессии методом наименьших

квадратов.

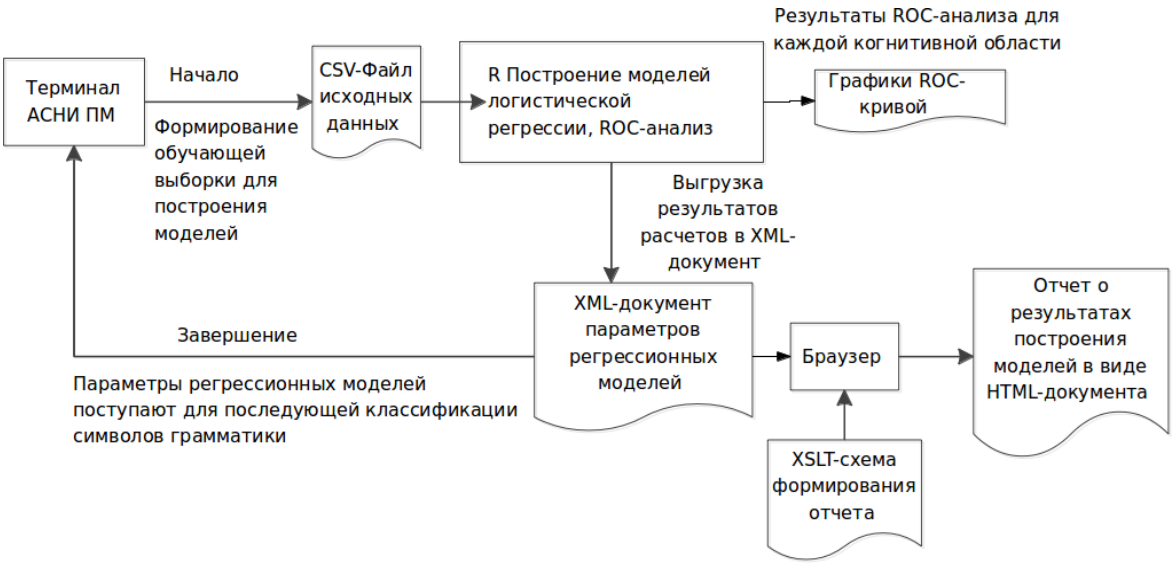

Рисунок 3. Информационные потоки построения регрессионных моделей.

CSV-файл исходных данных содержит совокупность кортежей:

$$learnEntry = (key, name, X_{in}, X_{out}, X_{com}, inArea, areaKey, areaName),$$

где $key, name$– ключ и наименование узла сети;

$areaKey, areaName$– ключ и наименование области;

$inArea$– логическое значение принадлежности узла области согласно множеству когнем;

X_{in}, X_{out}, X_{com}– параметры контекста (1) узла key для области $areaKey$.

Таким образом, общее число записей CSV-файла определяется как $AreasCount * SignSymsCount$, где $SignSymsCount$– число символов в ЗНАКах когнем.

Результаты расчетов компонуются в XML-документ следующей структуры:

```
<?xml version='1.0' encoding='UTF-8'?>
<?xml-stylesheet type='text/xsl' href={файл XSLT-преобразований}?>
<doc>
   <area name           = {Наименование области}
   fitted         = {Логический признак построения модели, TRUE/FALSE}
   in_val         = {Математическое ожидание Cin}
   in_std_err     = {Стандартное отклонение параметра Cin}
   out_val        = {Математическое ожидание Cout}
   out_std_err    = {Стандартное отклонение параметра Cout}
   com_val        = {Математическое ожидание Cout}
   com_std_err    = {Стандартное отклонение параметра Ccom}
   inter_val      = {Математическое ожидание Cinter}
   inter_std_err = {Стандартное отклонение параметра Cinter}
   auc            = {Площадь под ROC-кривой}
   cutoff_nearest_lc = {Значение выбранного уровня отсечки}
   roc_graph      = {Имя файла ROC-графика}
   />
</doc>
```

Для визуализации результатов в браузере префикс документа содержит ссылку на XSLT-схему (eXtensible Stylesheet Language Transformations).

Библиографический список.

1. *Филиппович Ю.Н., Сиренко А.В.* Программный комплекс исследований психолингвистической модели вербального сознания на основе когнитивного и ассоциативного экспериментов // Вопросы психолингвистики. – 2011. – № 1. – С. 126-139.

2. Караулов Ю.Н. Концептография языковой картины мира. Статья 1. Первый этап «восхождения» к образу мира: от элементарных фигур знания к предметно-референтным областям культуры // Проблемы прикладной лингвистики. Выпуск 2. Сборник статей. / под ред. Н.В. Васильевой. Москва: Азбуковник, 2004.- С. 7-17.

3. Hastie T., Tibshirani R., Friedman J.H. The elements of statistical learning: data mining, inference and prediction. 2-nd ed. New York: Springer, 2009.

4. Deerwester. United States Patent: 4839853. Режим доступа: http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=4,839,853.PN.&OS=PN/4,839,853&RS=PN/4,839,853 (дата обращения: 25.01.2011).

5. Филиппович Ю.Н., Черкасова Г.А., Дельфт Д. Ассоциации информационных технологий: эксперимент на русском и французском языках. Москва: Издательство “МГУП,” 2002. 304 с.

6. Караулов Ю.Н. Общая и русская идеография. М.: “Либроком” (УРСС), 1976. 360 с.

7. А.В. Сиренко. Алгоритмы поиска в ассоциативно-вербальных сетях психолингвистических экспериментов // Научная школа для молодых ученых "Компьютерная графика и математическое моделирование

(Visual Computing)": тезисы и доклады. М.-2009 г.-204 с. ISBN 978-5-902948-53-7.

8. Расин Д. Непараметрическая эконометрика: вводный курс. 2008. № 4. С. 7–56.

9. Cleveland W.S., Devlin S.J. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. 1988. Vol. 83, № 403. Pp. 596–610.

10. Mitchell T.M. Generative and discriminative classifiers: naive bayes and logistic regression // Machine Learning. 1997. McGraw Hill. ISBN: 9787111115021.

11. Авторский коллектив machinelearning.ru. Логистическая регрессия. Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Логистическая_регрессия (дата обращения: 25.01.2011).

12. Паклин Н. Логистическая регрессия и ROC-анализ - математический аппарат. Режим доступа: http://www.basegroup.ru/library/analysis/regression/logistic/ (дата обращения: 25.01.2011).

13. The R Project for Statistical Computing. Режим доступа: http://www.r-project.org/ (дата обращения: 25.01.2011).