

## ОДИН ПОДХОД К КЛАССИФИКАЦИИ ТЕКСТОВ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ<sup>1</sup>

В.Д. Соловьев

Классификация текстов по смыслу является одной из ключевых проблем компьютерной лингвистики [1]. В связи с широчайшим использованием Интернета и электронной почты во многих приложениях задача классификации текстов должна решаться в режиме on-line. В настоящее время эта задача не только не решена, но и очень далека от удовлетворительного решения.

Наглядным примером этого является проблема спама, буквально захлестнувшего в последнее время электронные почтовые ящики. Пользователь легко отличает спам от полезных писем, и если бы существовали компьютерные программы, понимающие тексты на уровне, близком к человеческому, то этой проблемы попросту не существовало бы. Существующие антиспамовские программы действуют иначе. Например, одна из наиболее известных и достаточно хорошая антиспамовская программа SpamPal [2] даже не пытается анализировать содержание писем. Она составляет списки “плохих” электронных адресов и серверов, с которых ранее приходил спам, и “хороших” – с которых приходили нормальные письма. На основе этой информации и принимается решение о характере очередного пришедшего письма. Ясно, что такой лобовой подход имеет несколько слабых мест. Во-первых, не будет выловлен спам, пришедший с нового адреса, который раньше “не был замечен ни в чем плохом”. Во-вторых, существуют адреса, с которых приходит как спам, так и важные письма. Это возможно в следующей ситуации, которая имела место у автора статьи. Для организации международной научной конференции оргкомитет создал некий адрес для переписки с участниками. Все поступающие на этот адрес письма рассылаются ответственным членам оргкомитета и программного комитета. Опыт показал, что с этого адреса поступает и спам и заявки на участие в конференции от потенциальных участников. В рамках принятого в программе SpamPal подхода, совершенно не ясно как исключить поступающий с такого адреса спам, не выплеснув вместе с ним и полезные письма. Ограниченные возможности подхода, связанного с анализом только электронных адресов, заставляют разрабатывать подходы, связанные с анализом содержания писем.

<sup>1</sup> Работа выполнена при поддержке РФФИ, грант N 02-07-90230

## Основные подходы к классификации текстов

Выделяют следующие основные подходы к классификации текстов: экспертные системы, алгоритм ближайшего соседа и индуктивное обучение. Индуктивное обучение может быть основано на правиле Байеса, деревьях решений или линейных классификаторах.

Экспертные системы предусматривают предварительное составление вручную правил классификации с последующим их применением в компьютерной системе. Это наиболее трудоемкий подход. Примером его применения может служить система CONSTRUE [3], предназначенная для классификации новостных сообщений агентства Рейтер.

Алгоритм ближайшего соседа и индуктивное обучение реализуют совершенно иной подход. В этом случае способы классификации текстов не задаются заранее, а возникают в процессе самообучения системы. Для этого требуется иметь набор предварительно расклассифицированных текстов, на основе которого и проводится обучение.

В алгоритме ближайшего соседа [4] система запоминает принятые классификационные решения для набора исходных обучающих тестов и для каждого нового текста ищет наиболее близкий к нему (в смысле некоторой метрики) исходный текст. Новый текст относится к тому же классу, что и найденный.

Индуктивное обучение объединяет широкий класс методов, предусматривающих построение в процессе обучения образов классов, на которые предстоит разбивать тексты. Байесовский метод реализует вероятностный подход, в свою очередь имеющий ряд вариантов [5]. Деревья решений – хорошо известный метод, применяемый в искусственном интеллекте и теории принятия решений, который был адаптирован к данной задаче [6]. Он эффективен в том случае если разбиение на классы можно осуществить на основе небольшого числа признаков.

На наш взгляд наиболее интересным и многообещающим является применение линейных классификаторов. В этом подходе с каждым классом ассоциируется характеристический вектор  $C = (w_1, \dots, w_n)$ . Вектор такой же длины сопоставляется и с каждым классифицируемым документом  $D$ . Далее вычисляется скалярное произведение, определяющее степень близости документа  $D$  с классом  $C$

$$f_C(D) = \sum_{i=1}^n w_i d_i.$$

На основе этой характеристики и принимается классификационное решение. Как правило, документы и классы представляются векторами в  $n$ -мерном пространстве терминов данной предметной области ( $n$  - число терминов). Ключевым является вопрос о построении характеристического вектора класса. Здесь испытывалось много различных методов, в том чис-

ле, применялись и персептроны [7]. Хотя окончательного решения пока не найдено, все же почти общепринятым [1] является мнение, что применение линейных классификаторов является простым, но эффективным способом классификации.

## Система классификации текстов по лингвистике

### Общая структура и функции системы

Нами разрабатывается система классификации электронных писем на английском языке по лингвистической тематике. Классификация электронных писем относится к задаче сортировки (sorting) [1]. Система предназначена для классификации электронных писем, поступающих по рассылке LINGUIST LIST [8]. Список имеет более 10 000 подписчиков. Ежедневно приходит около 20 писем, иногда количество писем оказывается и значительно большим. Естественно, что далеко не все из них представляют интерес для получателя.

В свете вышесказанного было решено использовать линейные классификаторы в форме некоторого варианта персептронов. Как известно, персептроны являются частным случаем нейронной сети. Все необходимые для понимания данной работы понятия по нейронным сетям могут быть найдены в [9].

Общая структура представлена на рисунке 1.

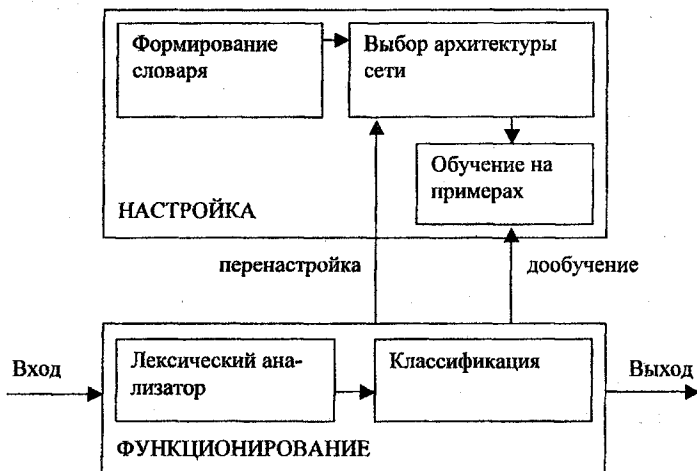


Рис. 1. Общая структура системы

## **Настройка**

### **1. Формирование словаря.**

Функцией блока “Формирование словаря” является выделение множества специальных для данной области терминов (слов и словосочетаний). Идеальным является вариант, когда такой словарь уже имеется в наличии. Если его нет, то его можно строить двумя способами.

Первый состоит в использовании списков слов, помещаемых в конце книг в данной области, и иных источников информации. Второй, наоборот, в использовании уже существующих списков общеупотребительных слов. Взяв возможно более представительный набор текстов в данной области (в частности, набор электронных писем) и удалив из них все общеупотребительные слова, в качестве “сухого остатка” получим набор терминов. Оба эти способа не лишены недостатков. Первый не гарантирует полноты покрытия. Второй – отсутствия в списке лишних слов. Оптимальной является комбинация этих способов. Из множества слов, полученных применением второго способа, выделим слова, которые не встречаются в списке, полученном применением первого способа, и предложим эксперту принять решение о включении их в окончательный список терминов данной предметной области.

В лингвистике существует электронный словарь основных терминов (более 800), входящий в набор словарей фирмы АВВУУ. Кроме того, в бумажном виде издан словарь [10]. Для формирования более представительного словаря предметной области осуществлен перевод в электронную форму указателей слов из нескольких наиболее широких по тематике лингвистических монографий [11].

В дальнейшем предполагается использовать более сложные алгоритмы обработки текстов с учетом некоторых семантических отношений. Для этого на основе полученного списка слов создается тезаурус, включающий следующие семантические отношения: синонимия, общее – частное, и обобщенное отношение “связан с”. В основу создаваемого тезауруса положен тезаурус по компьютерной лингвистике [12], но принятая в нем система отношений является чересчур подробной. Для данной прикладной задачи достаточно ограничиться вышеприведенными основными отношениями. Так как составление и отладка тезауруса – вещь очень трудоемкая, то в настоящее время построен фрагмент тезауруса, включающий лишь несколько сотен слов.

### **2. Выбор архитектуры сети.**

Нейронная сеть имеет столько входов, сколько терминов содержится в словаре данной предметной области. Количество выходов равно числу классов, на которые пользователь хочет разбить поступающие письма. С каждым выходом сопоставлен выходной нейрон. В простейшем варианте сеть двухслойная, каждый вход соединен с каждым выходным нейроном. Значения весов связей и порогов выходных нейронов в начальный момент

времени (до обучения) не столь существенны. Можно положить все веса равными 0.1, а пороги – равными 1. Сеть такой архитектуры по сути ни что иное, как простой персептрон. Конечно, стандартное определение персептрона предусматривает только один выход. Поэтому данная сеть является скорее совокупностью нескольких персептронов – по числу выходов.

Как известно, персептроны существенно ограничены в своих возможностях и способны решать только весьма узкий класс задач. В связи с этим мы введем некоторое усложнение структуры сети. В контексте данной задачи наибольший интерес представляют тормозящие связи. Под тормозящей связью понимается такая связь между входными и выходными нейронами, при подаче импульса по которой выходной нейрон не может перейти в возбужденное состояние, даже если суммарный импульс от возбужденных связей превосходит порог возбудимости.

Наблюдения над пользователями в процессе принятия ими решения об интересности письма показывают, что они руководствуются не только позитивными стимулами, но и негативными. Под этим имеется в виду следующее. В типичном листе рассылки, каким является, например, LINGUIST LIST, в заголовке каждого письма указывается его тип: conference, job, discussion и т. д., из которого понятен характер письма. Часто встречается ситуация, когда пользователя интересуют только письма определенного типа. Например, если он ищет работу или, наоборот, не собирается менять место работы. В этом случае сообщения, не интересующего пользователя типа, отклоняются им сразу, независимо от их содержания.

Слова в заголовке письма, обозначающие не интересующий пользователя тип, назовем негативными стимулами. Ясно, что в нейронной сети из соответствующих входов должны вести тормозящие связи. Нейронная сеть с тормозящими связями идеологически близка к комбинации персептрона и деревьев решений. В простейшем случае перечень негативных стимулов получается от пользователя. Это не составит сложности, т. к. такой перечень, обычно, крайне ограничен. Можно предложить и автоматическую процедуру выявления таких слов, однако, она достаточно сложна и здесь описываться не будет. Следует отметить, что вообще обучение нейронных сетей с тормозящими связями разработано слабо.

Интересной является идея использования многослойных персептронов. Вероятно, это приведет к некоторому повышению эффективности системы, однако, этот вопрос пока не изучался.

### 3. Обучение на примерах.

Для обучения системы используется информация, получаемая от пользователя скрытым образом. Предполагается, что пользователь выделяет в потоке поступающей корреспонденции заинтересовавшие его письма и помещает их в несколько директорий, классифицируя по темам. Остальные (неинтересные) письма удаляются, т. е. полагаем, что они помещаются в некоторую дополнительную директорию. Эта информация используется

для обучения. Подготовлена база из примерно 500 сообщений, разбитых на классы “когнитивная лингвистика”, “эволюция языка”, “эргативные языки”, “формальные модели языков”, “тюркские языки”, “все остальное”.

Применяется стандартный алгоритм обучения *back propagation* [13]. Все подготовленные примеры многократно прогоняются через сеть до тех пор, пока она не начнет давать безошибочные ответы. После этого сеть готова к работе.

### ***Функционирование сети***

#### **1. Лексический анализ.**

Прежде всего, все слова нового письма (не входящего в обучающую выборку) поступают на вход блока лексического анализа, где приводятся к словарной форме. Это осуществляется стандартными алгоритмами морфологического анализа, которые здесь не рассматриваются. При возникновении ситуации неоднозначной интерпретации слова генерируются все варианты. Далее, из получившегося списка слов письма выделяются термины путем сопоставления со словарем терминов.

#### **2. Собственно классификация.**

Для классификации письма выделенные слова подаются на вход сети, т. е. активируются нейроны, соответствующие этим словам. Возможны следующие варианты срабатывания сети.

А) Активирован ровно один выходной нейрон. Тогда поступившее на вход письмо относится к тому классу, которому соответствует этот нейрон.

Б) Активировано несколько выходных нейронов. Это означает, что анализируемый текст является “междисциплинарным”, относясь одновременно к нескольким областям. В этом случае письмо помещается во все соответствующие директории.

В) Ни один выходной нейрон не активирован. Тогда для каждого выходного нейрона подсчитывается сумма поданных на него импульсов, и письмо относится к классу, соответствующему нейрону с максимальной суммой.

### ***Усовершенствование сети***

В отличие от большинства систем на основе нейронных сетей, в которых после стадии обучения формируется окончательная структура, в создаваемой системе классификации текстов предусмотрено постоянное самообучение. Возможно самообучение двух типов: дообучение и перенастройка структуры.

#### **1. Дообучение.**

С этой целью предусматривается постоянный анализ правильности принимаемых решений. Если пользователь сочтет одно из принятых системой решений неправильным и переместит некоторое письмо из одной директории в другую, то это приводит в действие механизм обучения *back propagation*, соответствующим образом корректирующий веса связей.

В процессе дообучения применяется и еще один вид корректировки весов связей. Допустим, что имела место ситуация В, т. е. анализируемое письмо  $D$  отнесено к некоторому классу  $C$ , хотя  $f_C(D)$  и не превысило порог возбудимости соответствующего нейрона. Предположим также, что пользователь не внес корректив, т. е. признал это решение правильным. Тогда веса всех связей, ведущих от нейронов, соответствующих терминам документа  $D$ , к выходному нейрону класса  $C$ , увеличиваются на минимальную величину  $\alpha$ , выбранную так, чтобы  $f_C(D)$  превысило порог возбудимости нейрона класса  $C$ . Этот способ корректировки весов идейно близок к используемому в [14] алгоритму Widrow-Hoff, хотя там применены несколько иные формулы.

## 2. Перенастройка.

Несколько сложнее ситуация в случаях, когда пользователь изменяет множество директорий. Возможны следующие варианты:

- введение совершенно новой директории;
- деление одной из директорий на части;
- удаление директории;
- объединение директорий.

В этом случае адекватным образом изменяется число выходных нейронов и заново запускается механизм обучения сети.

Еще один вид перенастройки архитектуры связан с изменением словаря терминов. Пользователь имеет возможность редактировать словарь, добавляя новые термины и удаляя имеющиеся. В этом случае производится соответствующее изменение множества входных нейронов. Алгоритм обучения заново не запускается, т. к. вносимые изменения не меняют структуру сети радикально, и влияние изменений будет учтено в ходе постоянно функционирующей процедуры дообучения. Впрочем, за пользователем сохраняется возможность запустить процедуру обучения в любой момент.

## Заключение

В статье приведен краткий обзор методов автоматической классификации текстов и описан подход к построению системы классификации электронных писем в лингвистической области, включающий комбинацию методов. Выбранная архитектура нейронной сети может быть классифицирована как комбинация дерева решений с набором перцептронов. Алгоритм обучения также является комбинацией алгоритмов back propagation и Widrow-Hoff.

## Литература

1. Jackson P., Moulinier I. Natural language processing for online application. - Amsterdam: John Benjamins Com., 2002.
2. <http://www.SpamPal.org>
3. Hayes P. J., Weinstein S.P. CONSTRUE/TUE: A system for content-based indexing of a database of news stories // 2<sup>nd</sup> Annual conference on innovative applications of artificial intelligence. - 1990.
4. Larkey L., Croft W.B. Combining classifiers in Text Categorization // Proceedings of SIGIR'96. - Zurich, 1996.
5. McCallum A., Nigam K. A comparison of event models for Naive Bayes classification // Proceedings of AAAI-98 workshop on learning for text categorization. - 1998.
6. Cohen W. Fast effective rule induction // Proceedings of the 12<sup>th</sup> international conference on machine learning. - San Marino: Morgan Kaufmann, 1995.
7. Ng H., Goh W., Low K. Feature selection, perceptron learning, and a Usability Case study for text categorization // Proceedings of SIGIR. - 1997.
8. <http://www.linguistlist.org>
9. Осовский С. Нейронные сети для обработки информации. - М.: Финансы и статистика, 2002.
10. Демьянков В. З. Англо-русские термины по прикладной лингвистике и автоматической обработке текста. - М., 1982.
11. Nichols J. Linguistic Diversity in Space and Time. - Chicago: Univ. of Chicago Press, 1992.
12. Никитина С. Е. Тезаурус по теоретической и прикладной лингвистике. - М., 1978.
13. Короткий С. Нейронные сети: алгоритм обратного распространения. <http://www.orc.ru/~stasson/neurox.html>
14. Lewis D., Schapire R., Callan J., Parka R. Training algorithms for linear text classifiers // Proceeding of SIGIR'96. - Zurich, 1996.