МОДЕЛЬ РАСПОЗНАВАНИЯ НА ОСНОВАНИИ АНАМНЕЗА И ЕЕ ПРИЛОЖЕНИЯ

О. В. Герман, А. В. Сироткин
Кафедра информационных технологий автоматизированных систем
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: salexvlad@gmail.com, ovgerman@tut.by

В статье представлен подход, позволяющий определить предположительные заболевания на основании анамнеза в свободной форме.

Введение

Рассматривается задача постановки диагноза на основании представленного текстового файла с анамнезом. Основой распознавания служит текст, для которого выполняется трехступенчатая процедура обработки: общая классификация, классификация в пределах группы, классификация в пределах релевантных представителей подмножества группы. Описанная методика может служить основанием для применения в различных областях, где требуется text mining [1,2] – on-line консультации, обучение без учителя, системы интерактивной диагностики в интернет и т.п. Подход использует комплексное применение моделей распознавания, его цель состоит как в повышении производительности, так и в обеспечении более релевантных ответов.

Математические предпосылки

В поисковых системах типа Google [3] используются поисковые паттерны на базе ключевых слов. В основе лежит первый закон Ципфа [4] – чем больше частота слова, тем более точно (весомо) оно идентифицирует текст. Однако это не всегда так. В медицинских анамнезах слово «боль» (и его производные) может использоваться очень часто, но его диагностическая «сила» не высока. Это касается и других слов, например, «самочувствие», «поведение». Вместе с тем, словосочетание «падает зрение» может быть использовано один-два раза, а его диагностическая способность достаточно высока. Таким образом, закон Ципфа и поисковые системы, которые его активно применяют, ориентированы на первый общий этап классификации и выдают, как правило, огромное число найденных документов для заданного поискового паттерна. Напрашивается мысль применить следующий этап распознавания – классификацию в пределах группы (кластера) документов. Итак, система распознавания получает на входе текст, содержащий анамнез заболевания. На первом этапе выполняется определение ключевых слов и текст классифицируется как относящийся к медицинскому профилю (кластеру). На следующем этапе выполняется выделение из кластера документов наиболее релевантных тексту с анамнезом. Здесь используем процедуру опорных векторов [5], которая задействует формулу косинуса угла между двумя векторами $x = (x_1, x_2, ..., x_n)$, $y = (y_1, y_2, ..., y_n)$

$$cos(\phi) = \frac{x_1 \cdot y_1 + x_2 \cdot y_2 + ... + x_n \cdot y_n}{\sqrt{x_1^2 + x_2^2 + ... + x_n^2} \cdot \sqrt{y_1^2 + y_2^2 + ... + y_n^2}}.$$

В качестве разрядов векторов используем нечеткие аналоги взвешенных частот встречаемости ключевых слов в тексте анамнеза и тексте документа. Этот важный момент требует пояснения. Координаты векторов $x = (x_1, x_2, ..., x_n)$, $y = (y_1, y_2, ..., y_n)$ суть члены вида $\lambda_i \cdot \phi(f_i)$, где λ_i есть «семантический вес» ключевого слова и $\phi(f_i)$ - функция полезности от его частоты f_i. Семантическая веса назначаются экспертами. Для их оцифровки используем числовую шкалу Харрингтона [6], представленную в таблице 1.

Таблица 1 – Шкала Харрингтона

Числовое значение (шкала отношений)	Содержательное описание (шкала наименований)
0,8-1,0	Очень высокая (семантическая значимость)
0,64-0,8	Высокая
0,37-0,64	Средняя
0,2-0,37	Низкая
0,0-0,2	Очень низкая

Таким образом, например, если речь идет о диагностике сахарного диабета, то для ключевого словосочетания «плохое зрение» («ухудшается зрение», «падает зрение») семантический вес оценивается как высокий и очень высокий. Задача эксперта – указать значение из второго столбца приведенного числовой шкалы. Аналогично, функция полезности от частоты имеет тот же смысл. Предположим, частоты слов распределились в документе следующим (общим образом)

Таблица 2 – Предпололожительное распределение частот

Ключевое слово	x_1	x_2	...	x_n
Частота	f_1	f_2	...	f_n

Строим упорядоченную по убыванию частот последовательность и определяем в ней максимальную и минимальную частоту f_{max}, f_{min} .

В соответствии со шкалой Харрингтона разбиваем диапазон $[f_{min}, f_{max}]$ на пять равных поддиапазонов с длиной $\frac{f_{max}-f_{min}}{5}$ каждый. Самый левый поддиапазон получает оценку как «очень высокая частота», следующий за ним – как «высокая» и т.д. Таким образом, получаем нечеткий аналог частоты в рассматриваемой модели распознавания. Итак, все данные определены и рассматриваемы этап распознавания выбирает документы с наибольшей релевантностью (опять же по Харрингтону). Пример. На первом этапе общей классификации было определено, что текст с анамнезом относится к категории медицинских текстов. На рассмотренном втором этапе этот текст классифицирован как класс (кластер) эндокринологических заболеваний. По-прежнему, требуется последующая детализация диагноза. В силу вступает третий этап модели распознавания. Он базируется на использовании нечетких правил вида (1).

Такого вида правила используются в экспертных нечетких заключениях типа Мамдани [7]. В левой части правила записывают значения ключевых слов. Например, зрение=«падает» и т.д. Заключением правила является, например, «сахарный диабет» (0.6). Опять же требуется некоторая модификация системы вывода с учетом того, что важен весь ансамбль признаков и их нечетких значений. Таким образом, правила типа Мамдани позволяют получить окончательную диагностическую информацию (в примере из группы документов по эндокринологии будут отобраны те, которые относятся к сахарному диабету). В соответствии с методикой Мамдани для каждого диагноза вычисляется мера истинности по формуле

$$\mu(answer_k) = \frac{\sum\limits_t \mu(rule_t) \cdot \mu(answer_k)}{\sum\limits_t \mu(rule_t)},$$

где $\mu(rule_t)$ - мера сходства (близости) посылочной части правила вывода с номером t входному вектору x. В этом месте имеется сложность, поскольку методика Мамдани не указывает, как измерять нечеткую меру сходства между двумя векторами. Например, как оценить меру сходства двух векторов $< 1,\ 2 >$ и $< 2,\ 1 >$? Требуется дополнительно механизм нечеткой кластеризации и обучающая нечеткая выборка, на которой такую кластеризацию можно было бы провести. С другой стороны, можно полагать, что сами нечеткие правила устанавливают такие кластера (этот пункт дискуссионен). В качестве иллюстрации рассмотрим два правила:

$$< x_1 = 1,\ x_2 = 2 > \rightarrow\ answer(0.8),$$

$$< x_1 = 2,\ x_2 = 2 > \rightarrow\ answer(0.6).$$

Пусть входной вектор есть $< x_1 = 3,\ x_2 = 2 >$. Измерим расстояние по Евклиду данного входного вектора до каждого из векторов-предусловий указанных правил: $\rho_1 = \sqrt{(3-1)^2 + (2-2)^2} = 2$; $\rho_2 = \sqrt{(3-2)^2 + (2-2)^2} = 1$. Тогда, полагая, что мера близости обратно пропорциональна расстоянию, можно заключить, что $\mu_1/\mu_2 = 1/2$ и $\mu_2 = 2 \cdot \mu_1$. Применяя формулу Мамдани для заключения правила, получаем

$$\mu(answer) = \frac{\mu_1 \cdot 0.8 + 2\mu_1 \cdot 0.6}{\mu_1 + 2\mu_1} = \frac{2}{3} = 0.67.$$

Итак, мы последовательно рассмотрели предпосылки предложенной модели распознавания и указали математические принципы, которые лежат в ее основе.

Обобщения

Предложенная модель может быть использована в системах обработки естественного языка [8]. Сначала выполняется общая классификация текста, представленного естественноязыковой фразой. Затем производится «сужение» до релевантного кластера текстовых знаний и, наконец, на третьей фазе из этого релевантного кластера отбирается окончательный вариант ответа. Остается еще реализовать фазу выборки ответа из найденного текста. Очевидно, эта фаза требует реализовать каким-то образом алгоритм грамматического разбора, но это уже отдельный вопрос.

1. Барсегян, А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров / СПб.: БХВ-Петербург, 2009. – 512 с.
2. Паклин Н. Б., Орешков В. И. Бизнес-аналитика: от данных к знаниям. –СПб.:Питер, 2013. – 704 с.
3. Википедия-Google [Электронный ресурс] / Википедия-Google. – Режим доступа: https://ru.wikipedia.org/wiki/Google. – Дата доступа: 07.09.2015.
4. Википедия-Закон Ципфа [Электронный ресурс] / Википедия-Закон Ципфа. – Режим доступа: https://ru.wikipedia.org/wiki/Google. – Дата доступа: 07.09.2015.
5. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979. – 448 с.
6. Антонов А. В. Системный анализ. – М.: Высшая школа, 2004. –454 с.
7. Леоненков А. В. Нечеткое моделирование в среде MATLAB и fuzzyTECH / А. Леоненков. – СПб: БХВ-Петербург, 2003. – 736 с.
8. Люгер Д. Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. 4-е издание. – М.: Вильямс, 2003. – 864 с.

$$if(x_1 = (\geq, \leq)\alpha_1\ \&\ x_2 = (\geq, \leq)\alpha_2\ \&\ ...\ \&\ x_n = (\geq, \leq)\alpha_n)\ then\ answer(\mu_t) \qquad (1)$$