Московский государственный университет имени
М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра системного программирования

Курсовая работа

«Определение языка сообщений социальной сети Twitter»

Выполнил студент 327 группы

М. А. Кольцов

Научный руководитель

В. Д. Майоров

Москва, 2014

Содержание

1 Введение

В настоящее время человечество имеет доступ к огромному запасу знаний, накопленному за тысячи лет. Немалая часть этих знаний представляется в виде текстов на различных языках. В связи с этим активно разрабатываются методы, предназначенные для автоматического извлечения и преобразования информации, данной в символьном представлении. Возникло научное направление «обработка естественных языков». Одной из его фундаментальных задач является определение языка текста.

Стандартным подходом к этой задаче является применение машинного обучения. А именно, если у нас есть база из сотен документов на нескольких языках, то можно «предсказать» язык поступившего на рассмотрение документа, сравнив его с имеющимися. В общем случае, нужно по имеющимся данным построить так называемую модель, а затем все действия с текстами проводить в терминах этой модели. Классическим примером является метод, описанный в 1994 году: каждому документу сопоставим «профиль документа» — упорядоченный по числу встречаний список n-грамм — последовательностей длины n подряд идущих символов. «Профиль языка» — это совокупность профилей документов, которые имеются на этом языке. Теперь, если нужно для какого-то документа определить язык, то последовательность действий такова:

1. Составляется профиль этого документа
2. Сравнивается с имеющимися профилями языков
3. Тот язык, чей профиль наиболее похож на профиль документа, объявляется результатом

Вышеописанный метод плохо работает для коротких сообщений. В то же время, поток информации в виде коротких шумных сообщений нельзя игнорировать — в социальной сети Twitter среднее количество сообщений в день составляет примерно 58 000 000, а длина каждого ограничена 140 символами. Такой формат текстов обусловил появление новых алгоритмов. В данной работе рассматриваются современные методы решения задачи определения языка, а также предлагается улучшение для одного из них. Проводится тестирование, показывающее превосходство по полноте распознавания языка полученного результата над существующими.

2 Постановка задачи

2.1 Формальное описание задачи автоматического определения языка

Пусть L - множество меток, сопоставленных естественным языкам. По заданному тренировочному корпусу

$$T = \{(msg_1, lang_1), (msg_2, lang_2), \ldots, (msg_N, lang_N)\}$$

(здесь $msg_i, i \in \overline{1..N}$, - текст на естественном языке, $lang_i \in L$ - метка этого языка) нужно построить классификатор, который произвольному входному сообщению new_msg на языке $some\_lang$ сопоставит метку $l \in L$, соответствующую этому языку, или же сообщит, что язык текста невозможно достоверно распознать.

2.2 Цели и задачи курсовой работы

В данной работе в качестве текстов выступают так называемые «твиты» - сообщения из социальной сети Twitter[1], а множество L соответствует 18 языкам, которые можно разделить на три группы по типу алфавита:

- Кириллические: болгарский, чувашский, русский, татарский, украинский
- Арабские: арабский, персидский (фарси), хинди, маратхи, непальский, урду
- Латинские: нидерландский, французский, английский, немецкий, итальянский, испанский, турецкий

Цели работы:

1. Исследовать современные решения задачи автоматического определения языка коротких сообщений
2. Провести совместное сравнительное тестирование некоторых методов решения задачи и выяснить, действительно ли они показывают заявленное авторами качество классификации
3. Исследовать возможность улучшения какого-либо алгоритма решения задачи автоматического определения языка коротких сообщений

[1] https://twitter.com/

3 Метод решения задачи

Определение 1. Множеством достижимости достижимости в момент времени t называется множество $\mathcal{X}[t]$ всех точек x, в которые можно попасть из начального множества $\mathcal{X}_0$ в момент времени t при выборе какого-либо допустимого управления u:

$$\mathcal{X} = \{x | \exists u(s) : t_0 \leqslant s \leqslant t \Rightarrow x(t, t_0, x_0) = x\}.$$

Определение 2. Трубкой достижимости называется множество $X[\cdot] = \mathcal{X}[\cdot, t_0, \mathcal{X}_0]$.

Определение 3. Множеством достижимости при фазовых ограничениях в момент времени t $Y(t)$ и начальном положении $(t_0, \mathcal{X}_0)$ называется множество

$$\mathcal{X}[t] = \{x | \exists u(s) : t_0 \leqslant s \leqslant t \Rightarrow x(t, t_0, x_0) = x \in Y(t)\}.$$

Аналогично для трубки достижимости при фазовых ограничениях.
Для решения задачи воспользуемся эволюционным уравнением:

$$\lim_{\sigma \leftarrow 0} \frac{1}{\sigma} h\left\{\mathcal{X}[t+\sigma], (\mathcal{X}[t] + \sigma B(t)\mathcal{P}[t]) \cap \mathcal{Y}[t+\sigma]\right\} = 0.$$

Предполагая непрерывность по Хаусдорфу множеств $\mathcal{X}[t]$ и $\mathcal{Y}[t]$, перепишем выражения для этих множеств для момента времени $t+\sigma$ в следующем виде:

$$\mathcal{X}[t+\sigma] = \mathcal{X}[t] + \sigma A(t)\mathcal{X}[t] + \sigma B(t)\mathcal{P}[t],$$

Таким образом, исходное эволюционное уравнение эквивалентно следующему уравнению:

$$\mathcal{X}[t+\sigma] = ((I + \sigma A(t))\mathcal{X}[t] + \sigma B(t)\mathcal{P}[t]) \cap \mathcal{Y}[t+\sigma] + o(\sigma).$$

Будем строить внутренние эллипсоидальные оценки множества достижимости. Пусть эллипсоид $\mathcal{E}_-(q_-[t], Q_-[t])$ — внутренняя эллипсоиадальная аппроксимация множества достижимости в момент времени t без фазовых ограничений. Тогда для момента времени $t+\sigma$ справедливо:

$$\mathcal{E}_-(q_-[t+\sigma], Q_-[t+\sigma]) = ((I+\sigma A(t))\mathcal{E}_-(q_-[t], Q_-[t]) + \sigma B(t)\mathcal{E}_-(p(t), P(t)) =$$

$$= \mathcal{E}_-((I+\sigma A(t))q_-[t], (I+\sigma A(t))Q_-[t](I+\sigma A(t))^T) + \mathcal{E}_-(\sigma B(t)p(t), \sigma B(t)P(t)\sigma B^T(t)$$

$$= \mathcal{E}_-\left(q_1 + q_2, S_1 Q_1^{\frac{1}{2}} + S_2 Q_2^{\frac{1}{2}}\right),$$

где

$$q_1 + q_2 = (I + \sigma A(t))q_-[t] + \sigma B(t)p(t),$$
$$Q_1 = I + \sigma A(t))Q_-[t](I + \sigma A(t))^T,$$
$$Q_2 = \sigma B(t)P(t)\sigma B^T(t),$$

а матрицы S_1 и S_2 удовлетворяют следующим свойствам:

$$S_i S_i^T = S_i^T S_i = I, i = 1, 2.$$

Данная формула дает возможность итерационного построения внутренней эллипсоидальной оценки множества достижимости — с некоторым шагом σ будем строить множество достижимости до тех пор, пока не достигнем момента времени t_1, а за начальное значение $\mathcal{X}[t]$ возьмем эллипсоид $\mathcal{E}(x_0, X_0)$.

Для того, чтобы выполнялись фазовые ограничения $\mathcal{Y}(t)$ на множество достижимости, будем на каждом шаге t полученную оценку пересекать с множеством $\mathcal{Y}(t)$ и строить эллипсоидальную оценку пересечения двух множеств средствами Ellipsoidal Toolbox, а именно с помощью функции intersection_ia.

Для того, чтобы касание эллипсоидальной оценки происходило по направлению $l, (l \in \mathbb{R}^n, ||l|| = 1)$, нужно чтобы выполнялось следующее соотношение:

$$S_1 Q_1^{\frac{1}{2}} l = \lambda S_2 Q_2^{\frac{1}{2}} l, \lambda > 0.$$

Поэтому, будем брать матрицу S_1 равную единичной, а матрицу S_2 находить из этого соотношения. Для этого воспользуемся функцией ell_valign, входящей в состав Ellipsoidal Toolbox.

Для более точного построения множества достижимости будем проводить перебор направлений, вдоль которых происходит касание эллипсоидальной оценки и множества. В силу того, что по условию задачи необходимо построить проекции трубки достижимости и множества достижимости на некоторую плоскость, порожденную векторами l_1 и l_2, то перебор направлений будем производить по единичной сфере, принадлежащей этой плоскости, а в соотношении для матриц S_1 и S_2 в качестве матриц Q_1 и Q_2 будем использовать проекции конфигурационных матриц эллипсоиадальных оценок, полученных из эволюционного уравнения. После этого полученные оценки будем объединять.

Проекция матрицы Q и вектора q на плоскость (l_1, l_2) вычисляются по следующим формулам:

$$P = (l_1, l_2),$$
$$\widehat{Q} = P'QP,$$
$$\widehat{q} = P'q.$$

4 Вывод уравнений движений системы из примера

Построим уравнения движения маятника. Для этого возьмем за обобщенные координаты углы между вертикалью и положением первого и второго стержня и обозначим их за φ_1 и φ_2. Потенциальная энергия системы выражается как сумма потенциальных энергий первого и второго шариков:

$$\Pi = -mgl_1 \cos\varphi_1 - mg(l_1 \cos\varphi_1 + l_2 \cos\varphi_2).$$

Если выразить линейные скорости шариков через их угловые скорости, то получим, что

$$v_1 = l_1\dot{\varphi}_1,$$
$$v_2 = v_1 + l_2\dot{\varphi}_2 = l_1\dot{\varphi}_1 + l_2\dot{\varphi}_2.$$

Используя эти соотношения, построим выражение для кинетической энергии:

$$K = \frac{mv_1^2}{2} + \frac{mv_2^2}{2} = \frac{m}{2}\left(2l_1^2{\dot{\varphi}_1}^2 + 2l_1l_2\dot{\varphi}_1\dot{\varphi}_2 + l_2^2{\dot{\varphi}_2}^2\right).$$

Выпишем лагранжиан системы и с помощью уравнений Лагранжа второго рода получим уравнения движения системы:

$$L = K - \Pi = \frac{m}{2}\left(2l_1^2{\dot{\varphi}_1}^2 + 2l_1l_2\dot{\varphi}_1\dot{\varphi}_2 + l_2^2{\dot{\varphi}_2}^2\right) + mg(2l_1 \cos\varphi_1 + l_2 \cos\varphi_2).$$

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i} - \frac{\partial L}{\partial q_i}\right) = 0,$$

где за q_i обозначены обобщенные координаты (в нашем случае — φ_1 и φ_2).

$$\frac{\partial L}{\partial \dot{\varphi}_1} = 2ml_1^2\dot{\varphi}_1 + ml_1l_2\dot{\varphi}_2,$$

$$\frac{\partial L}{\partial \dot{\varphi}_2} = ml_2^2\dot{\varphi}_1 + ml_1l_2\dot{\varphi}_1,$$

$$\frac{\partial L}{\partial \varphi_1} = -2mgl_1 \sin\varphi_1,$$

$$\frac{\partial L}{\partial \varphi_2} = -mgl_2 \sin\varphi_2.$$

Окончательно получим:

$$\begin{cases} 2ml_1^2\ddot{\varphi}_1 + ml_1l_2\ddot{\varphi}_2 + 2mgl_1 \sin\varphi_1 = 0, \\ ml_2\ddot{\varphi}_2 + ml_1l_2\ddot{\varphi}_1 + mgl_2 \sin\varphi_2 = 0. \end{cases}$$

Сократим на m оба уравнения и на l_1 и l_2 первое и второе уравнения соответственно, затем вычитая одно уравнение из другого получим два уравнения, в которые входят только $\ddot{\varphi}_1$ и $\ddot{\varphi}$:

$$\begin{cases} l_1\ddot{\varphi}_1 + g(2\sin\varphi_1 - \sin\varphi_2) = 0, \\ l_2\ddot{\varphi}_2 - g(2\sin\varphi_1 - 2\sin\varphi_2) = 0. \end{cases}$$

Приведем систему к нормальной форме:

$$\begin{cases} \dot{\varphi}_1 = \psi_1, \\ \dot{\varphi}_2 = \psi_2, \\ \dot{\psi}_1 = -g\dfrac{2\sin\varphi_1 - \sin\varphi_2}{l_1}, \\ \dot{\psi}_2 = g\dfrac{2\sin\varphi_1 - 2\sin\varphi_2}{l_2}. \end{cases}$$

В силу того, что по условию задачи маятник совершает малые колебания, можно заменить $sin\varphi$ на φ. Сделав это преобразование, получим окончательный вид системы, которая будет являться линейной и стационарной:

$$\begin{cases} \dot{\varphi}_1 = \psi_1, \\ \dot{\varphi}_2 = \psi_2, \\ \dot{\psi}_1 = -g\dfrac{2\varphi_1 - \varphi_2}{l_1}, \\ \dot{\psi}_2 = g\dfrac{2\varphi_1 - 2\varphi_2}{l_2}. \end{cases}$$

4.1 Некоторые сведения из теории управления

Для начала запишем уравнения системы с использованием управления. Так как управляющее устройство прикреплено только ко второму шарику, то уравнения движения системы с управлением примут вид:

$$\begin{cases} \dot{\varphi}_1 = \psi_1, \\ \dot{\varphi}_2 = \psi_2, \\ \dot{\psi}_1 = -g\dfrac{2\varphi_1 - \varphi_2}{l_1}, \\ \dot{\psi}_2 = g\dfrac{2\varphi_1 - 2\varphi_2}{l_2} + u. \end{cases}$$

5 Примеры работы программы

5.1 Пример 1

В данной системе матрицы A, B, P, X_0 являются единичными матрицам в $\mathbb{R}^3$, векторы p, x_0 — нулевыми. Диапазон времени: $t_0 = 0$, $t_1 = 3$, фазовые ограничения отсутствуют. За статичные направления l_1 и l_2 взяты векторы $[1, 0, 0]$ и $[0, 1, 0]$, за динамичные — $l_1(t) = [\sin(t); \cos(t); t], l_2(t) = [\cos(t); \sin(t); t]$. %endcenter

5.2 Пример 2

В данной системе:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1, \end{pmatrix}, B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1, \end{pmatrix} P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1, \end{pmatrix}$$

$$x_0 = [0, 0, 0]^T, p = [0, 0, 0]^T, t_0 = 0, t_1 = 3.$$

В данной системе есть фазовые ограничения $x_1 \leqslant 3$.

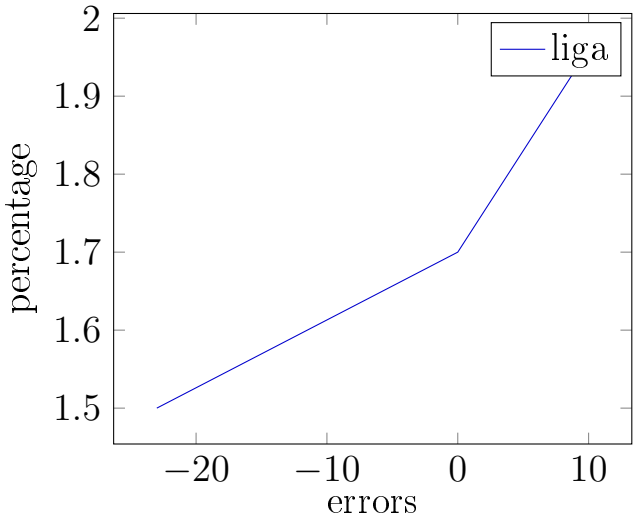
5.3 Пример 3

Рассмотрим колебательную систему из задания прошлого семестра. В этой системе:

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2\frac{g}{l_1} & \frac{g}{l_1} & 0 & 0 \\ 2\frac{g}{l_2} & -2\frac{g}{l_2} & 0 & 0, \end{pmatrix}, B = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1, \end{pmatrix} P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1, \end{pmatrix}$$

$$x_0 = [0.1, 0.3, 1, 1]^T, p = [0, 0, 0, 0]^T, t_0 = 0, t_1 = 3, l_1 = 2, l_2 = 1.$$

На систему наложены фазовые ограничения $|x_1| \leqslant y_1$, $y_1 = 1$.



Проекция множества достижимости на статическую плоскость (l_1, l_2).

Проекция трубки достижимости на статическую плоскость (l_1, l_2).

Проекция трубки достижимости на статическую плоскость (l_1, l_2).

Из рисунков трубки достижимости видно, что фазовые ограничения выполняются.

6 Библиография

Список литературы

[1] Голубев Ю. Ф. Основы теоретической механики: Учебник. 2-е изд., перераб. и дополн. — М.:Изд-во МГУ, 2000.

[2] P. Gagarinov, Alex A. Kurzhanskiy Ellipsoial toolbox: ver. 2.0 beta 1, 2013.