МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБОУ ВПО «Благовещенский государственный педагогический университет»

Физика математический факультет

КУРСОВАЯ РАБОТА

на тему: Марковские процессы

по дисциплине: Математика

|  |  |  |  |
| --- | --- | --- | --- |
| Исполнитель: студент группы 2 «А» | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ дата | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_подпись | И. А. Магда |
| Руководитель: к.п.н, доцент | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ дата | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_подпись | О.А. Семочкина |

Благовещенск 2015

СОДЕРЖАНИЕ

|  |
| --- |
| ВВЕДЕНИЕ……………………………………………………………………… 3 |
| 1. Основные понятия теории Марковских процессов………………………. 5 |
| 1.1 Марковский процесс……………………………………………………….. 5 |
| 1.2 Система состояний………………………………………………………….. 6 |
| 1.3 Граф состояний…………………………………………………………….. 7 |
| 1.4 Случайная последовательность…………………………………………… 8 |
| 1.5 Цепи Маркова……………………………………………………………… 10 |
| 2. Использование теории Марковских процессов в работе с матричными объектами……………………………………………………………………… 15 |
| 2.1 Матрица частот…………………………………………………………… 15 |
| 2.2 Матрица переходных вероятностей…………………………………….. 17 |
| 2.3 Матрица признаков………………………………………………………. 18 |
| 2.4 Приведенные матрицы признаков………………………………………. 19 |
| 3. Расстояния Хемминга на множестве приведенных матриц признаков... 21 |
| 3.1 Мера расстояния. Расстояние Хэмминга……………………………….. 21 |
| Заключение…………………………………………………………………….. 23 |
| Список использованных источников………………………………………… 24 |

**Введение**

Марковские процессы названы в честь великого русского математика Андрея Андреевича Маркова. Он впервые начал изучение вероятностной связи случайных величин, а также создал теорию марковских процессов. В дальнейшем основы этой теории явились исходной базой общей теории случайных процессов, а также таких важных прикладных наук, как теория диффузионных процессов, теория надежности, теория массового обслуживания и т.д.

Благодаря сравнительной простоте и наглядности математического аппарата, высокой достоверности и точности получаемых решений особое внимание аппарат теории марковских процессов с дискретными состояниями и цепей Маркова получил в различных областях науки, в исследовании операций и теории систем. Это обусловлено так же:

* Многие реальные технические системы имеют конечные множества возможных состояний, а их поведение в процессе работы можно смоделировать используя марковские процессы.
* Теория марковских процессов с дискретными состояниями и цепей Маркова изучена настолько глубоко, что позволяет решать широкий спектр прикладных задач.

На сегодняшний день существует проблема определения авторства текста. Например, периодически находят письма, произведения литературного искусства без указания авторства. Можно проводить экспертизу почерка для идентификации автора или еще существующие специфические методы. Мы же анализируем стиль письма, используя статистические методы. Исходя из изученной теории нами была выдвинута гипотеза о том, что аппарат марковских процессов применим к задаче авторства текста.

Цель данной работы: разработать программное обеспечение реализующее теорию марковских процессов для получения приведенной матрицы признаков в задаче определения авторства текста.

Цель работы определила задачи исследования:

* Познакомиться с теорией марковских процессов.
* Реализовать процедуры и функции для работы с матричными объектами в рамках аппарата марковских процессов.

Реализовать алгоритмы:

* Построения переходной матрицы вероятностей для слов из заданного текста.
* Построения матрицы признаков для заданного текста.
* Построения приведенных матриц признаков для двух заданных текстов.
* Определение расстояния Хемминга на множестве приведенных матриц признаков.

**1. Основные понятия теории марковских процессов**

**1.1 Марковский процесс**

Основополагающим понятием в теории марковских процессов является понятие марковского процесса. Марковский случайный дискретный процесс, протекающий в системе S, характеризуется не только возможными состояниями, в которых система может находиться случайным образом, но и теми моментами времени, в которые могут происходить ее переходы из состояния в состояние. Эти моменты времени могут быть заранее известны или случайны. Случайный процесс называется марковским, если для любого момента времени состояние системы в будущем зависит только от состояния в настоящем и не зависит от того, как и сколько времени развивался этот процесс в прошлом.

Случайный процесс называется процессом с дискретными состояниями, если множество его возможных состояний конечно или счетно. Случайные величины меняющиеся в процессе опыта, называются случайными процессами. Случайным процессом называется семейство случайных величин X(t,ω).

В нашей работе используется дискретный марковский процесс.

**1.2 Система состояний**

Пусть S это некоторая физическая система с возможными дискретными состояниями, которая случайным образом с течением времени переходит из одного состояния в другое. Если этот процесс является марковским, то тогда это марковский процесс с дискретными состояниями. Под системой S будем понимать всякое целостное множество взаимосвязанных элементов, которое нельзя разделить на независимые подмножества. Связи между элементами системы в одну или обе стороны могут быть как непосредственными, так и опосредованными. Элементы системы и связи между ними изменяются.

В системе S протекает случайный процесс, если система S с течением времени t изменяет свои состояния S(t) случайным образом. В любой момент времени система может находиться только в одном из состояний, то есть для любого момента времени t найдется единственное состояние такое, что . Если множество состояний не более чем счетно, то оно дискретно. Если множество состояний более чем счетно, то оно непрерывно. В зависимости от времени нахождения системы в каждом состоянии различают процессы с дискретным временем. Системы с непрерывным временем предполагают, что переход системы из одного состояния в другое может осуществляться в любой момент времени, т.е. время нахождения системы в каждом состоянии представляет непрерывную случайную величину.

**1.3 Граф состояний**

Для удобного визуального представления системы можно использовать граф состояний, который изображает возможные состояния системы и возможные переходы этой системы из одного состояния в другое, указываемые стрелками. Граф состояний системы представляет собой совокупность вершин и совокупность ветвей.

Пример:

Система S состоит из двух узлов с номерами 1 и 2, каждый из которых в процессе функционирования может выйти из строя. Система представлена на рисунке 1. Возможные состояния системы:

* - оба узла работают.
* - первый узел отказал, а второй работает.
* - первый узел работает, а второй отказал.
* - оба узла отказали.

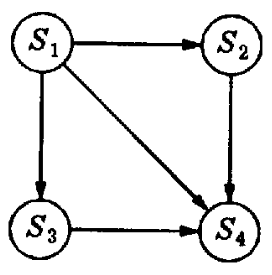


Рис 1

**1.4 Случайная последовательность**

Если для системы S переход из состояния в состояние возможен лишь в фиксированные моменты времени то эти моменты времени принято называть шагами или этапами марковского процесса. В промежутках между соседними шагами система сохраняет свои состояния. Не исключается возможность, что на некоторых шагах система не изменит своего состояния. Следовательно случайный процесс с дискретным временем можно представить случайной последовательностью этих событий которую называют также цепью Маркова, если для каждого шага вероятность перехода системы S из любого состояния в любое другое состояние не зависит от того, когда и как она попала в это состояние.

Во многих прикладных дисциплинах в большинстве случаев вместо термина “случайная последовательность” употребляют термин случайный процесс с дискретным временем.

Если ввести случайное событие состоящее в том, что после j этапов исходная система S находится в состоянии то для каждого фиксированного j ≥ 1 имеем группу событий т.е

Пример:

Цель – самолет обстреляна из зенитного автомата очередью в четыре снаряда. Если – интервал между последовательными выстрелами, а - время первого выстрела, то

Возможные состояния системы:

1. цель невредима.
2. цель получила незначительные повреждения.
3. цель получила значительные повреждения, но еще продолжает функционировать.
4. цель поражена, т.е. не может дальше функционировать.

Если в начальный момент времени система S находилась в состоянии то граф ее состояний принимает вид, как на рисунке 2:

Рис 2

**1.5 Цепи Маркова**

Для формализованного описания цепи Маркова удобно употреблять понятия вероятностей состояний и переходных вероятностей. Поэтому введем необходимые в дальнейшем обозначения.

Пусть множество состояний системы S. Вероятность исполнения случайного события состоящее в том, что после j этапов система находится в состоянии обозначают и называют вероятностью состояния. Вектор вероятностей состояний системы S после j этапов обозначим

а вектор вероятностей начальных состояний

Если ввести матрицу-строку

то равенство

можно представить в виде

Если множество возможных состояний системы S, а случайное событие, состоящее в том, что после j этапов система находится в состоянии то условную вероятность события при условии обозначают:

и называют переходной вероятностью.

Матрицу для каждого фиксированного называют матрицей переходных вероятностей.

Цепь Маркова называют однородной, если матрица P переходных вероятностей системы не зависит от номера этапа Иначе цепь Маркова называют неоднородной.

Рассмотрим свойства матриц переходных вероятностей.

Сумма элементов любой строки матрицы переходных вероятностей равна 1, т.е.

Действительно, полная группа событий для любого Следовательно,

Вектор вероятностей состояний после j этапов равен произведению транспонированной матрицы переходных вероятностей на вектор вероятностей состояний после (j – 1) этапов,

Так как полная группа событий при любом натуральном j, то по формуле полной вероятности имеем

или то же самое,

Вектор вероятностей после j этапов однозначно определяется матрицами переходных вероятностей и вектором вероятностей начального состояния При этом, если цепь Маркова является однородной, то и

где я степень матрицы P.

С учетом свойства о том, что вектор вероятностей равен произведению транспонированной матрицы переходных вероятностей на вектор вероятностей имеем

При рассмотрении однородных цепей Маркова зачастую бывает удобно пользоваться графом состояний, на котором у стрелок выписаны соответствующие переходные вероятности. Граф такого типа принято называть размерным графом состояний.

Вероятности состояний цели определяются из предыдущего примера об обстреле самолета, если в начальный момент времени она находилась в состоянии а размеченный граф состояний изображен на рисунке 3.

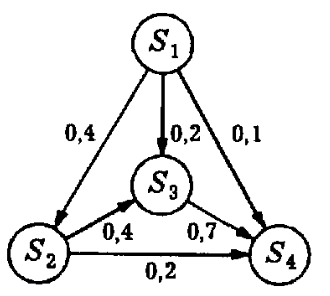


Рис 3

Из размеченного графа состояний имеем:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

где вероятности найдены из соотношений

Таким образом, матрица переходных вероятностей принимает вид:

А так как по условию то, согласно свойству матрицы переходных вероятностей, находим:

Таким образом, найдены вероятности всех исходов обстрела цели одной очередью из четырех выстрелов:

* Цель не повреждена:
* Цель получила незначительные повреждения:
* Цель получила значительные повреждения:
* Цель поражена:

Пример:

Цель – самолет обстреляна из зенитного орудия тремя выстрелами с корректировкой наводки после каждого выстрела. Цель может находиться в одном из четырех возможных состояний, определённых в предыдущих примерах, а матрицы переходных вероятностей равны:

Определим вероятности состояний цели после каждого выстрела, если в начальный момент времени она находилась в состоянии т.е.

Согласно свойствам матрицы переходных вероятностей имеем:

**2. Использование теории Марковских процессов в работе с матричными объектами**

**2.1 Матрица частот**

Матрица частот – это одна из основных структур данных, необходимых для дальнейшей работы с матричными объектами в рамках аппарата марковских процессов. Она имеет свойства:

* Содержит в себе частотную информацию о нахождение слов в тексте – количестве слов стоящих до текущего слова.
* Слова, упорядоченные в алфавитном порядке и строчном регистре.
* Каждое слово имеет свой порядковый номер, начиная с нуля.
* Таблица состоит из n столбцов и n строк.
* Все элементы таблицы – целые числа.

Состоянием системы в нашем случае является любое слово текста. Частотой будет количество слов стоящих до этого слова.

Для построения матрицы частот практически не имеет значение какой длины задан текст. Ее можно построить и из маленького текста, так и из большого, пример большого текста: «Война и мир» – роман-эпопея Льва Николаевича Толстого. Чем больше текст, тем больше размер матрицы частот, потому что она квадратичная, а ее размер зависит от количества уникальных слов в заданном тексте. Процедура построения матрицы занимает непродолжительное время, не более 1-2 секунд. Алгоритм последовательно отделяет каждое предложение текста. Затем смотрится какие слова стоят до текущего рассматриваемого слова и их счетчик присутствия увеличивается на единицу, т.е. каждое слово будет иметь вектор, состоящий из слов, где каждое слово будет отображено с частотой появления до рассматриваемого слова. Основные элементы таблицы: слова и их частота появлений.

Пример:

Построение матрицы частот, исходя из заданного текста.

Исходный текст: Сегодня завтра было вчера, а еще сегодня было завтра.

Таблица 1 – Матрица частот

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | а | было | вчера | еще | завтра | сегодня |
| а | 0 | 1 | 0 | 1 | 1 | 1 |
| было | 1 | 1 | 1 | 1 | 2 | 1 |
| вчера | 1 | 1 | 0 | 1 | 1 | 1 |
| еще | 0 | 1 | 0 | 0 | 1 | 1 |
| завтра | 1 | 2 | 1 | 1 | 1 | 1 |
| сегодня | 1 | 3 | 1 | 1 | 3 | 1 |

**2.2 Матрица переходных вероятностей**

Матрица переходных вероятностей или таблица относительных частот получается из матрицы частот путем проведения операции деления каждого значения частоты на сумму частот в строке. Без матрицы переходных вероятностей мы бы не имели возможность получения матрицы признаков, поэтому матрица переходных вероятностей играет важную связующую роль. Она необходима для дальнейшей успешной работы алгоритма.

Таблица 2 – Матрица переходных вероятностей

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | а | было | вчера | еще | завтра | сегодня |
| а | 0 | 0.25 | 0 | 0.25 | 0.25 | 0.25 |
| было | 0.142 | 0.142 | 0.142 | 0.142 | 0.285 | 0.142 |
| вчера | 0.2 | 0.2 | 0 | 0.2 | 0.2 | 0.2 |
| еще | 0 | 0.333 | 0 | 0 | 0.333 | 0.333 |
| завтра | 0.142 | 0.285 | 0.142 | 0.142 | 0.142 | 0.142 |
| сегодня | 0.1 | 0.3 | 0.1 | 0.1 | 0.3 | 0.1 |

**2.3 Матрица признаков**

Матрица признаков является нашим собственным изобретением. Ни в каких других источниках упоминание о ней не встретить. Матрица признаков несет в себе практически основную роль исследования. Ведь имея ее можно получить приведенные матрицы признаков, которые будут иметь одинаковые размерности, следовательно в дальнейшем можно провести операцию сравнения, с целью определения авторства текста.

Основным внутренним свойством данной таблицы является двумерная таблица, где каждый i – тый столбец возведен в i – ю степень двумерной таблицы переходных вероятностей. Столбцов в таблице столько же, сколько максимальное количество слов в предложениях исходного текста. Начальным значением для возведения в степень, является i = 2.

Используя свойство матрицы переходных вероятностей составляется матрица признаков.

Таблица 3 – Матрица признаков

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0.09642 | 0.10519 | 0.106193 | 0.10619 | 0.10619 | 0.10619 | 0.10619 | 0.10619 |
| 0.26647 | 0.24783 | 0.247791 | 0.24778 | 0.24778 | 0.24778 | 0.24778 | 0.24778 |
| 0.09642 | 0.08643 | 0.08850 | 0.08849 | 0.08849 | 0.08849 | 0.08849 | 0.08849 |
| 0.09642 | 0.13347 | 0.132737 | 0.13274 | 0.13274 | 0.13274 | 0.13274 | 0.13274 |
| 0.26547 | 0.24783 | 0.247791 | 0.24778 | 0.24778 | 0.24778 | 0.24778 | 0.24778 |
| 0.17976 | 0.17921 | 0.176988 | 0.17699 | 0.17699 | 0.17699 | 0.17699 | 0.17699 |

**2.4 Приведенные матрицы признаков**

Для сравнения текстов используются матрицы признаков. Из матриц признаков выбираются одинаковые состояния систем, что приводит к получению приведенных матриц признаков. Из матриц признаков выбираются необходимые столбцы и строки, соответствующие словам, которые встречаются в сравниваемых матрицах. Таким образом, приведенные матрицы признаков для разных текстов будут иметь идентичные размерности, что в дальнейшем нам позволит произвести задачу определения авторства текста.

Таблица 4 – Матрица признаков текст 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | … |  |
| 0 |  |  |  |  |  |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| … |  |  |  |  |  |
|  |  |  |  |  |  |

Таблица 5 – Матрица признаков текст 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | … |  |
| 0 |  |  |  |  |  |
| 1 |  |  |  |  |  |
| 2 |  |  |  |  |  |
| … |  |  |  |  |  |
|  |  |  |  |  |  |

Таблица 6 – Приведенная матрица признаков текст 1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | … |  | … |  |
| 0 |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Таблица 7 – Приведенная матрица признаков текст 2

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | … |  | … |  |
| 0 |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Пример:

Построение матрицы признаков, исходя из заданного текста.

Исходный текст: Сегодня завтра было вчера, а еще сегодня было завтра.

**3. Расстояния Хемминга на множестве приведенных матриц признаков**

**3.1 Мера расстояния. Расстояние Хэмминга**

На множестве приведенных матриц признаков вводится мера расстояния. И решается задача идентификации автора текста. Мерой расстояния является расстояние Хэмминга.

Расстояние Хэмминга – метрика различия объектов одинаковой размерности. Расстояние Хэмминга применяется для строк одинаковой длины любых q–ичных алфавитов и служит метрикой различия (функцией, определяющей расстояние в метрическом пространстве) объектов одинаковой размерности. Например, В нашем случае это метрика для приведенных матриц признаков.

Расстояние Хэмминга уже довольно широко используется для различных задач, таких как поиск близких дубликатов, распознавание образов, классификация документов, исправление ошибок, обнаружения вирусов и т.д.

Расстояние Хэмминга обладает свойствами метрики, так как удовлетворяет ее [определению](http://neerc.ifmo.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D1%80%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D0%BF%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D1%81%D1%82%D0%B2%D0%BE#def1).

* ~d(x, y) = 0 \iff x = y (Если расстояние от x до y равно нулю, то x и y совпадают (x = y))
* ~d(x,y)=d(y,x) (Объект x удален от объекта y так же, как объект y удален от объекта x)
* ~d(x,y) \le d(x,z) + d(z,y) (Расстояние от x до y всегда меньше или равно расстоянию от x до y через точку z. Это свойство обычно называют неравенством треугольника за его естественную геометрическую аналогию: сумма двух сторон треугольника всегда больше третьей стороны.)

Таблица 8 – Промежуточная матрица разностей

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2 | 3 | 4 | … |  | … |  |
| 0 |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Таблица 9 – Детерминированная матрица разностей

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2 | 3 | 4 | … |  | … |  |
| 0 |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

**Заключение**

В ходе нашего исследования были изучены основные понятия марковских процессов. Был найден способ приложить полученные знания на практике – использование аппарата марковских процессов в задаче распознавания авторства. Поставленные перед нами задачи были реализованы. Мы успешно использовали теорию аппарата марковских процессов для получения матричных объектов, приведенной матрицы признаков в задаче определения авторства текста. Анализ письма, используя статистические методы оказался не опровергнутым, из этого следует, что аппарат марковских процессов применим в задаче авторства текста.

Таким образом нами разработана библиотека, реализующая теорию марковских процессов для получения приведенной матрицы признаков в задаче определения авторства текста.

Она была апробирована на текстах:

* Толстой Л.Н. «Анна Каренина».
* Чехов А.П. «Попрыгунья», «Жалобная книга».
* Русские народные сказки.
* Научные статьи.

**Список использованных источников**

1. Волков, И.К. Случайные процессы: Учебник для вузов / И.К. Волков, С.М. Зуев, Г.М. Цветков – издательство МГТУ им. Н.Э. Баумана, 1999. – 448 с.
2. Вентцель, Е.С. Теория вероятностей: Учебник для вузов / Е.С. Вентцель – Москва: 6-е издание – «Высшая школа», 1999. – 576 с.
3. Волков, И.К., Загоруйко Е.А. Исследование операций: Учебник для вузов / Под ред. В.С. Зарубина, А.П. Крищенко. – М.: издательство МГТУ им. Н.Э. Баумана, 2000. – С. 239.
4. <http://habrahabr.ru> Статья. Что такое скрытые модели Маркова.
5. <http://habrahabr.ru> Статья. Поиск лиц на основе скрытых марковских моделей из песочницы.
6. <http://habrahabr.ru> Статья. Скрытые марковские модели в распознавании речи из песочницы.
7. <https://ru.wikipedia.org> Статья. Скрытые марковские модели.