Результати роботи

Програма розпізнавання мови і класифікації текстів написана на Java. Система розпізнає такі мови: російську, китайську, українську, англійську, німецьку, французьку. У нашій роботі для класифікації використаний SVM метод. Було протестовано три групи з 5000 документів. Результати: середня повнота: 86,3%. середня точність: 89,2%. Основними недоліками описаного алгоритму є нестійкість визначення мови малих текстових документів: окремі пропозиції на схожих мовах розпізнаються невпевнено і з помилками. Крім того, системи автоматичної класифікації текстів показують кращі часові результати на 64-бітовій апаратній платформі (ширина регістрів сучасних процесорів), ніж на 32-бітовій апаратній платформі, яка працює повільніше.

Висновки

Пропонований метод і результати роботи підтверджують можливість створення ефективної системи автоматичної класифікації документів для кінцевої множини мов за критерієм приналежності до певної галузі знання, використовуючи сучасні засоби обчислювальної техніки.

Література

[1] Лифшиц Юрий. Курс «Алгоритмы для Интернета». — М. : РАН 2006.

[2] Salton G., Buckley C. Term-Weighting Approaches in Automatic Text Retrieval // Information Processing and Management. —1988. — P. 513–523.

[3] Burges C. J. C. A tutorial on support vector machines for pattern recognition. // Data Mining and Knowledge Discovery, 1998. — P. 955–974.

КЛАССИФИКАЦИЯ МЕТОДОМ k-БЛИЖАЙШЕГО СОСЕДА

Юрий Юрьевич Игнатко
НТУУ «КПИ», г. Киев, просп. Победы, 37
E-mail: yura. ignatko@gmail. com

Методы Data Mining помогают решить многие задачи, с которыми сталкиваются аналитики. Одной из основных задач, которые приходится решать, является задача классификации. Методы классификации позволяют отнести неизвестные ранее объекты к тому или иному раннее известному классу. Например, когда человек обращается в банк за предоставлением ему кредита, банковский служащий должен принять решение: кредитоспособен ли потенциальный клиент или нет. Такое решение принимается на основе данных об исследуемом объекте (в данном случае — человеке): его место работы, размере заработной платы, возрасте, составе семьи и т. п. В результате анализа этой информации, служащий банка должен отнести человека к одному из двух известных классов: «кредитоспособен» и «некредитоспособен».

Другим примером задачи классификации является фильтрация электронной почты. В этом случае программа фильтрации должна классифицировать входящее сообщение как спам или как письмо. Данное предложение принимается на основании частоты появления в сообщениях определенных слов (например, имени получателя, безличного обращения, слов и словосочетаний: «приобрести», «заработать», «выгодное предложение» и т. п.). [1]

Одним из наиболее простых и популярных методов, решающих данный класс задач, является метод k-ближайшего соседа (*k*-nearest neighbor). Данный метод относится к классу метрических классификаторов. Основная идейная составляющая состоит в том, что похожие объекты очень часто находятся «рядом», т. е. на близком расстоянии друг от друга в метрическом пространстве.

Следует сразу отметить, что метод k-ближайшего соседа относится к классу методов, работа которых основывается на хранении данных в памяти для сравнения с новыми элементами. При появлении новой записи для прогнозирования находятся отклонения между этой записью и подобными наборами данных, и наиболее подобная (или ближний сосед) идентифицируется. Например, при рассмотрении нового клиента банка, его атрибуты сравниваются со всеми существующими клиентами данного банка (доход, возраст и т. д.). Множество «ближайших соседей» потенциального клиента банка выбирается на основании ближайшего значения дохода, возраста и т. д. [2]

Число k — это количество соседних объектов в пространстве признаков, которое сравнивается с классифицируемым объектом. Иными словами, если k = 10, то идентифицируемый объект сравнивается с 10-ю соседями.

Выбор параметра k противоречив. С одной стороны, увеличение его значения повышает достоверность классификации,

но при этом границы между классами становятся менее четкими. На практике хорошие результаты дают эвристические методы выбора параметра k, например, перекрестная проверка [3].

Рассмотрим вышесказанное на абстрактном примере, который покажет работу данного метода. У нас есть 3 класса объектов. Количество объектов в каждом классе равно 5. На рис. 1 изображено размещение этих точек в двумерном пространстве координат. Крестиком отмечена точка, которую нужно отнести к одному из классов. Для этого нужно рассчитать расстояния между заданной точкой и точками всех объектов и найти минимальные. Число k в данном случае выберем равным 4.

Рис. 1 — Размещение объектов классов в пространстве

Из 4 соседей 3 принадлежат объекту изображенному треугольниками, а одна точка принадлежит объекту в виде квадратов. Степень схожести неопределённого объекта с объектами в виде треугольников равна 0,75(3/4), следовательно, точка, скорее всего, принадлежит классу «треугольники» [4].

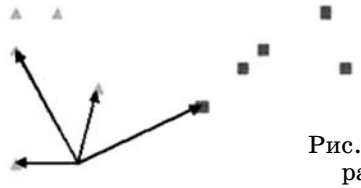

Рис. 2 — Определение минимального расстояния до объектов классов

Метод k-ближайшего соседа широко используется в распознавании образов, классификации текстов, экспертных системах, а также в обнаружении мошенничества (новые случаи могут быть похожи на те, что происходили в прошлом) и медицине (классификация пациентов по разным показателям, основываясь на данных прошедших периодов). Главными достоинствами данного метода являются его простота реализации

и интерпретации результатов, возможность модификации под конкретную задачу и устойчивость к аномальным выбросам (вероятность попадания такого объекта в k-ближайших соседей мала). Среди недостатков можно выделить не эффективный расход памяти (приходиться хранить всю выборку целиком) и большое количество операций при классификации, в ледствии чего — вычислительная трудоемкость [5].

Література

[1] А. А. Барсегян, М. С. Куприянов «Методы и модели анализа данных OLAP и Data Mining» — БХВ-Петербург, 2004

[2] Метод «ближайшего соседа» или системы рассуждений на основе аналогичных случаев[Электронный ресурс]/ Интернет университет информационных технологий. — Режим доступа: www. URL:http://www. intuit. ru/department/database/datamining/10/3. html- 5. 09. 2010

[3] Алгоритм ближайшего соседа [Электронный ресурс]/ BaseGroup Labs. — Режим доступа: www. URL:http://www. basegroup. ru/glossary/definitions/nearest_neighbor/-5. 09. 2010

[4] Математические основы kNN [Электронный ресурс]/ TradeExperts. — Режим доступа: www. URL:http://forex-tradexperts-to. narod. ru/kNN_Osnovi. htm-5. 09. 2010

[5] Метод k-ближайших соседей [Электронный ресурс] / BaseGroup Labs. — Режим доступа: www. URL:http://www. basegroup. ru/library/analysis/regression/knn/-5. 09. 2010

РОЗРОБКА МЕТОДИКИ ВІЗУАЛІЗАЦІЇ СТРУКТУРИ БАГАТОФАЗНОГО ПОТОКУ

Олена Кабанова

Івано-Франківський національний технічний університет нафти і газу, 76019, м. Івано-Франківськ, вул. Карпатська, 15

В період експлуатації газових і нафтових родовищ характеристики продуктивних пластів змінюються так, як і характеристики родовищ вцілому, — не в кращу сторону. Ці зміни контролюються технологічними і геологічними службами, які, згідно регламенту, встановлюють оптимальний режим експлуатації свердловин.

Оперативною інформацією для прийняття рішень є наземні параметри — тиск і температура газу на виході свердловини,