Исследование алгоритма SVM SMO для задач классификации и сравнение его с другими методами машинного обучения

Курсовая работа
студента 344 группы
Кутькина Никиты

Научный руководитель:
Невоструев Константин

Постановка задачи классификации

Имеется множество объектов, разделенных на классы.

Дана обучающая выборка, для всех её элементов известно к каким классам они принадлежат.

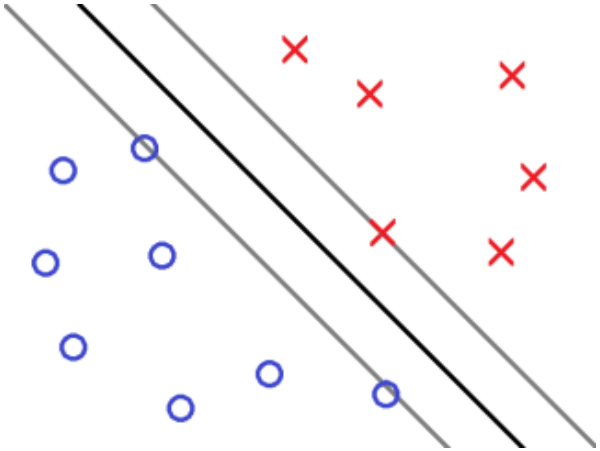
Необходимо построить алгоритм, определяющий класс произвольного объекта из исходного множества.

Цели

1. Реализация алгоритма SMO
2. Исследование и реализация улучшений и оптимизаций SMO
3. Исследование и реализация методов автоматического подбора параметров
4. Тестирование на реальных данных
5. Сравнение с другими алгоритмами машинного обучения

SVM

Основная идея - разделитель с максимальным зазором: если выборка линейно разделима, то потребуем чтобы разделяющая гиперплоскость максимально далеко отстояла от ближайших к ней точек обоих классов.



$$\begin{cases} \frac{1}{2}\lambda^T Q\lambda - e^T\lambda \to min_\lambda \\ 0 \le \lambda_i \le C \\ y^T\lambda = 0 \end{cases}$$

$$Q = y_i y_j K(x_{i,} x_i)$$

Итоговая задача в матричных обозначениях.

x – обучающая выборка

y – вектор ответов

K – функция ядра

λ – вектор переменных

e – вектор единиц

C – параметр алгоритма

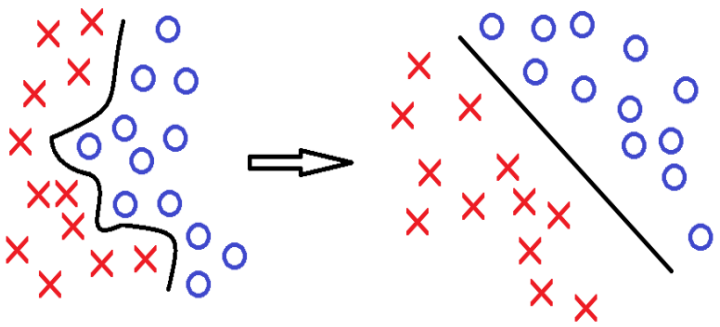
SMO

$$\begin{cases} \frac{1}{2}\lambda^T Q\lambda - e^T\lambda \to min_\lambda \\ 0 \le \lambda_i \le C \\ y^T\lambda = 0 \end{cases}$$

Общая схема SMO:

1. инициировать λ значениями, удовлетворяющими ограничениям
2. выбрать рабочее множество $\{\lambda_i, \lambda_j\}$
3. оптимизировать пару λ_i, λ_j
4. Пересчитать параметры, используемые в п.2
5. **если** параметры системы изменились, переход на п.2
6. Конец работы

Улучшения и оптимизации



1. Ядра и спрямляющие пространства
2. Мульти-классификация
3. Предварительное вычисление диагональных элементов матрицы Q и скалярных квадратов всех векторов
4. Кэширование
5. Автоматический подбор параметров

Автоматический подбор параметров

Необходимо выбрать константу C и параметры ядра.

Метод подбора параметров такой:

1. Выбираем граничные значения всех параметров и шаг
2. Строим соответствующую n-мерную сетку (n – число параметров)
3. Обучаем SVM с параметрами, соответствующими узлам сетки
4. Выбираем наилучшие из опробованных параметров

Тестирование

Необходимы специальные данные. Их нужно найти, привести к единому формату, провести предварительную обработку.

Используемые наборы данных (для наборов указаны количество классов и параметров, размер выборки):

1. MNIST (10, 784, 60000) – распознавание цифр
2. Poker (10, 85, 25010) – распознавание покерных рук
3. Adult (2, 123, 32561) – определить зарабатывает ли человек 50000$ в год
4. Titanic (2, 6, 1526) – определить утонул ли данный пассажир Титаника
5. Diabetes (2, 8, 576) – определить болен ли человек диабетом

Набор данных MNIST

Обучающая выборка: 60 000
Количество классов: 10
Количество параметров: 784

	SVM	ELM	Neural Network
Время обучения, сек	3578	836	900
Точность, %	98.57	96.1	96.39
Время на 1 предсказание, мсек	24	1.1	<1

Набор данных Adult

Обучающая выборка: 32 561
Количество классов: 2
Количество параметров: 123

	SVM	ELM	Neural Network
Время обучения, сек	197	30	75
Точность, %	85	85	85
Время на 1 предсказание, мсек	2.1	<1	<1

Результаты

1. Изучены стандартный и модифицированный алгоритмы SMO:
   - SMO (Platt, 1998)
   - SMO (Fan, Chen, Lin, 2005)
2. Реализован модифицированный алгоритм SMO
3. Исследованы возможные оптимизации и улучшения алгоритма. Реализованы некоторые из них
4. Проведено тестирование; выполнено сравнение с другими алгоритмами машинного обучения

Спасибо за внимание!