Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики

Факультет информационных технологий и программирования
Кафедра компьютерных технологий

Определение метрики качества для известного типа задач

Шведов Денис Владимирович
Группа M3436

Научный руководитель: к.ф.-м.н. доцент кафедры КТ
А. А. Фильченков

Решаемая проблема

Цель исследования

Компания VeeRoute занимается разработкой специальных алгоритмов, которые позволяют выстраивать маршруты в реальном времени. По данным, предоставленным компанией, построить классификатор, который определит тип новых наборов данных.

- Грамотно подобрать признаки для классификатора
- Выбрать лучший алгоритм классификации для решения поставленной задачи
- Протестировать на разных наборах данных и добиться хорошего результата

Решаемая проблема

Актуальность

- В компанию приходит заказчик с определенной бизнес-задачей и своими датасетами.
- При этом в компании не знают, корректную ли вообще задачу ставит заказчик.
- Вполне возможно, что он ошибается.

Решаемая проблема

Актуальность

- Проверяется похожесть нового датасета, к одной из групп, ранее отобранных.
- Можно посмотреть, какая бизнес-задача решалась на уже известных данных и сравнить с тем, что предоставил заказчик.
- Таким образом, можно еще на первом шаге устранить ошибку и переформулировать задачу.

Решение

Описание исходных данных

Каждый набор данных представляет собой информацию о

- Заказах
- Грузах
- Водителях и их передвижениях
- Траспортных средствах и их передвижениях
- Локациях

Решение

Выбор признаков

Было выбрано примерно более 50 признаков для построения классификатора.
Их можно разбить на следующие категории.

- Количественные признаки (количество заказов, транспортных средств, водителей, локаций в одном наборе данных)
- Матрицы совместимости (например, между исполнителем-транспортом, транспортом-локацией, исполнителем-заказом, грузом-отсеком траспнортного средства и т.д.)
- Статистические признаки (например, среднее количество рабочих смен для исполнителей, среднее количество грузов в заказах, средняя длительность временного отрезка в минутах (для исполнителя) длительность пути в метрах для транспорта и т.д.)
- Геокоординаты

Решение

Типы многоклассовой классификации

При решении использовались следующие типы классификаций

- Дерево решений (C4.5)
- Многоклассовый метод опорных векторов
- Многоклассовая логиститическая регрессия
- Random Forest

Решение

Описание решения

- Для исследования брались 4 разных класса наборов данных, не связанных друг с другом.В каждом классе выбиралось одинаковое числов наборов для рассмотрения
- В качестве обучающей выборки — 50 % от каждого класса наборов данных
- В качестве тестовой выборки — 50 % оставшихся данных
- При каждом запуске данные перемешивались.
- Далее будут рассмотрены результаты нескольких категорий тестирования. В каждой категории какое-то количество признаков зашумлено и не участвует в построении классификатора.

Результаты

Все признаки используются

Таблица: Средняя F_1 мера

Дерево Решений	0.949
SVM	0.949
LogReg	0.949
RandomForest	0.949

Результаты

Не используется признак количества заказов

Таблица: Средняя F_1 мера

Дерево Решений	1.0
SVM	0.949
LogReg	0.778
RandomForest	0.845

Результаты

Не используются признаки количества заказов,транспорта и водителей

Таблица: Средняя F_1 мера

Дерево Решений	0.949
SVM	0.949
LogReg	0.896
RandomForest	0.899

Результаты

Не используется половина рандомных признаков

Таблица: Средняя F_1 мера

Дерево Решений	0.949
SVM	0.949
LogReg	0.949
RandomForest	0.899

Результат

Выводы

- Наиболее лучший результат по итогам тестирования был показан при использовании метода "Дерево Решений". F_1 мера составила примерно 0.95
- В целом, выбранные признаки хорошо классифицируют наборы данных

Спасибо за внимание!