ТАД для обработки геологических данных

Nidropakshin

Май 2023

Введение

Геологические данные помогают в обосновании технологических решений проектирования разработки, регулировании процесса разработки, регулировании и учете фонда скважин, в принятии решений о переводе скважины из одного состояния в другое, в контроле добычи нефти, газа и воды и их динамики по скважине. Для их обработки, анализа и интерпретации требуется специалист-геолог. Зачастую ему приходится выполнять рутинную работу, которую можно было бы автоматизировать с помощью алгоритмов машинного обучения.

Введение

Проблема

Специалисты-геологи тратят много времени на рутинную работу по интерпретации геофизических данных, собранных со скважин

Актуальность

Несмотря на то, что такие решения уже есть, мы хотим использовать другой, новый быстроразвивающийся метод анализа данных, Топологический Анализ Данных (ТАД), и проверить его эффективность на этой задаче

Цель

Изучить целесообразность применения ТАД для интерпретации результатов ГИС

Задачи

- Узнать какие геофизические данные собирают и обрабатывают геологи
- Найти базу данных для обучения модели
- Создать алгоритм, использующий методы ТАД, для автоматизации работы геологов
- Оценить эффективность созданного алгоритма

Методы геологических исследований скважин

Классификация методов ГИС может быть выполнена по виду изучаемых геофизических полей:

- Электрические методы - измеряются удельное сопротивление, электропроводность и естественные потенциалы горных пород
- Радиометрические методы - основаны на изучении естественного гамма-излучения и взаимодействия вещества горной породы с наведенным ионизирующим излучением
- Сейсмоакустические методы - основаны на изучении среды посредством пробных акустических волн, распространяющихся в земных породах
- и др.

Самой распространенной разновидностью ГИС является каротаж, который, в зависимости от задачи исследования, объединяет в себе некоторые из этих методов

База данных

Поскольку чаще всего применяют электрические и радиометрические методы каротажа, то баз данных, где собраны соответствующие измерения, больше и они полнее.
Одной из основных задач ГИС является определение геологического разреза, поэтому нас будут интересовать базы данных, в которых указаны еще и литотипы.
В силу названных выше факторов, а также ограниченности вычислительных мощностей, используемого нами оборудования, мы выбрали базу данных, которая использовалась в конкурсе Geophysical Tutorial Machine Learning Contest 2016.
Соответственно, сравнивать наши результаты мы будем с результатами победителей конкурса, которые использовали модели машинного обучения, отличные от нашей.

Кринж данных

База данных содержит в себе информацию о каротаже 10 скважин: литотипе, формации, названии скважины, глубине замера, измеренные значения ГК (GR), удельного сопротивления (ILD log10), разницы в пористости по плотности нейтронов (DeltaPHI), средней пористости нейтронной плотности (PHIND), фотоэлектрического эффекта (PE), относительного положения (RELPOS) и др.

	Facies	Formation	Well Name	Depth	GR	ILD_log10	DeltaPHI	PHIND	PE	NM_M	RELPOS	FaciesLabels	dist_mar_up	dist_mar_down	Formation_category	NM_M_label
0	3	A1 SH	SHRIMPLIN	2793.0	77.450	0.664	9.900	11.915	4.600	1	1.000	FSiS	-99999.0	21.5	1	non_marine
1	3	A1 SH	SHRIMPLIN	2793.5	78.260	0.661	14.200	12.565	4.100	1	0.979	FSiS	-99999.0	21.0	1	non_marine
2	3	A1 SH	SHRIMPLIN	2794.0	79.050	0.658	14.800	13.050	3.600	1	0.957	FSiS	-99999.0	20.5	1	non_marine
3	3	A1 SH	SHRIMPLIN	2794.5	86.100	0.655	13.900	13.115	3.500	1	0.936	FSiS	-99999.0	20.0	1	non_marine
4	3	A1 SH	SHRIMPLIN	2795.0	74.580	0.647	13.500	13.300	3.400	1	0.915	FSiS	-99999.0	19.5	1	non_marine
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4144	5	C LM	CHURCHMAN BIBLE	3120.5	46.719	0.947	1.828	7.254	3.617	2	0.685	MS	0.0	0.0	12	marine
4145	5	C LM	CHURCHMAN BIBLE	3121.0	44.563	0.953	2.241	8.013	3.344	2	0.677	MS	0.0	0.0	12	marine
4146	5	C LM	CHURCHMAN BIBLE	3121.5	49.719	0.964	2.925	8.013	3.190	2	0.669	MS	0.0	0.0	12	marine
4147	5	C LM	CHURCHMAN BIBLE	3122.0	51.469	0.965	3.083	7.708	3.152	2	0.661	MS	0.0	0.0	12	marine
4148	5	C LM	CHURCHMAN BIBLE	3122.5	50.031	0.970	2.609	6.668	3.295	2	0.653	MS	0.0	0.0	12	marine

База данных

Из всех характеристик мы оставим только те, где заполнены все колонки. В дальнейшем, для обучения алгоритма мы будем использовать 9 из 10 скважин, а предсказывать будем литологию оставшейся скважины.

	Well Name	Depth	GR	ILD_log10	DeltaPHI	PHIND	NM_M	RELPOS	Facies
0	SHRIMPLIN	2793.0	77.450	0.664	9.900	11.915	1	1.000	3
1	SHRIMPLIN	2793.5	78.260	0.661	14.200	12.565	1	0.979	3
2	SHRIMPLIN	2794.0	79.050	0.658	14.800	13.050	1	0.957	3
3	SHRIMPLIN	2794.5	86.100	0.655	13.900	13.115	1	0.936	3
4	SHRIMPLIN	2795.0	74.580	0.647	13.500	13.300	1	0.915	3
...	...	...	...	...	...	...	...	...	...
4144	CHURCHMAN BIBLE	3120.5	46.719	0.947	1.828	7.254	2	0.685	5
4145	CHURCHMAN BIBLE	3121.0	44.563	0.953	2.241	8.013	2	0.677	5
4146	CHURCHMAN BIBLE	3121.5	49.719	0.964	2.925	8.013	2	0.669	5
4147	CHURCHMAN BIBLE	3122.0	51.469	0.965	3.083	7.708	2	0.661	5
4148	CHURCHMAN BIBLE	3122.5	50.031	0.970	2.609	6.668	2	0.653	5

Алгоритм топологической классификации

Алгоритм топологической классификации был представлен в статье R.Kindelan, J.Frias, M.Cerda, N.Hitchfield, Classification based on Topological Data Analysis, Feb. 9 2021, где также сравнивался с другими моделями машинного обучения, такими как k-NN и wk-NN, на стандартном наборе баз данных, и местами превзошёл их. Алгоритм было решено реализовать на языке python с использованием библиотек pandas, numpy и gudhi.

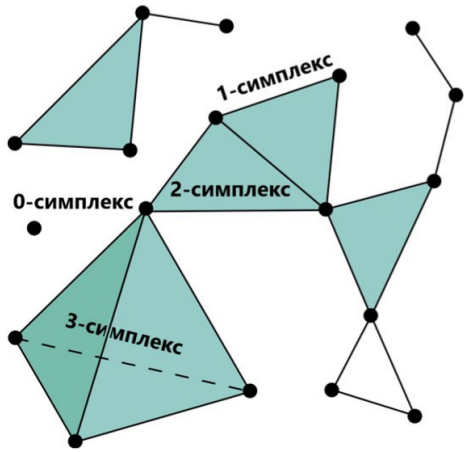
Пререквизиты

Определение

n-мерный Симплекс — геометрическая фигура, являющаяся n-мерным обобщением треугольника

Определение

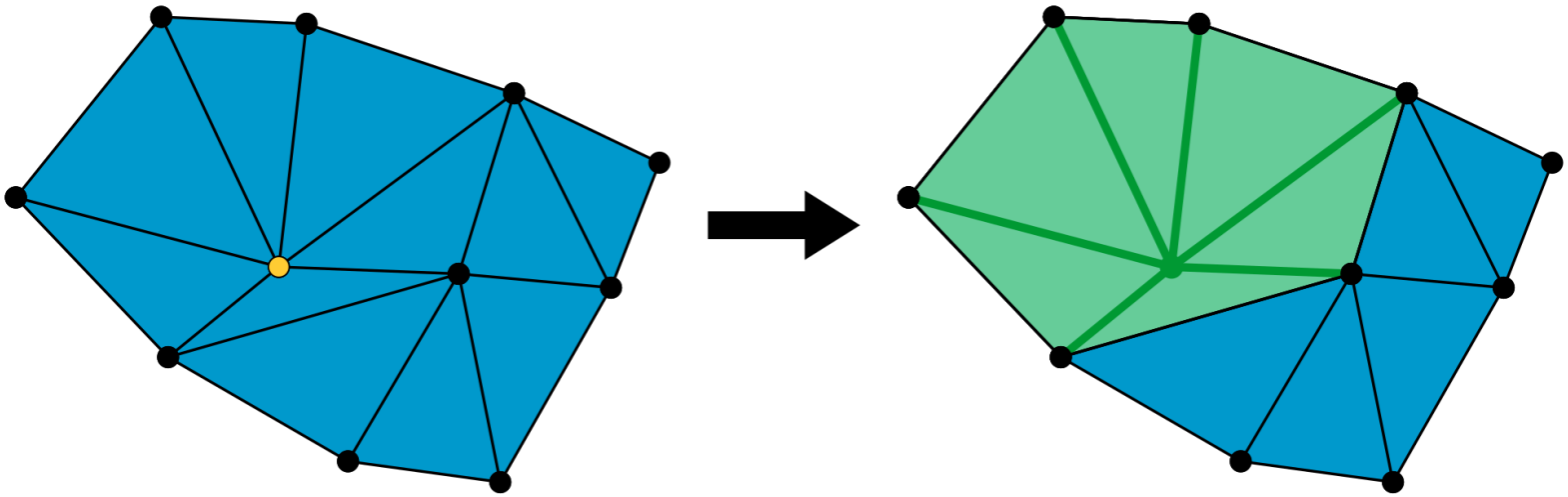
Симплициальный комплекс — топологическое пространство с заданной на нём триангуляцией, то есть, неформально говоря, склеенное из топологических симплексов по определённым правилам или обобщение графов на высшие размерности



Пререквизиты

Определение

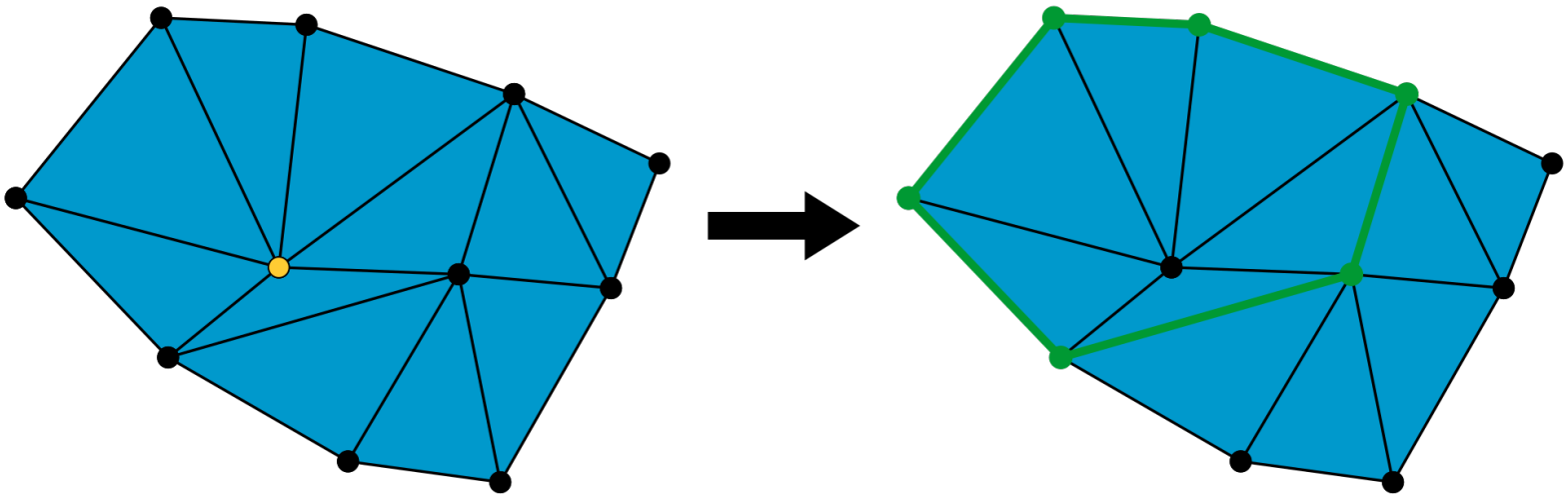
Звезда симплекса в симплициальном комплексе — это все симплексы в симплициальном комплексе, имеющие данный симплекс своей гранью:



Пререквизиты

Определение

Линк симплекса в симплициальном комплексе — обобщение окрестности вершины в графе. Линк вершины кодирует информацию о локальной структуре комплекса в её окрестности:

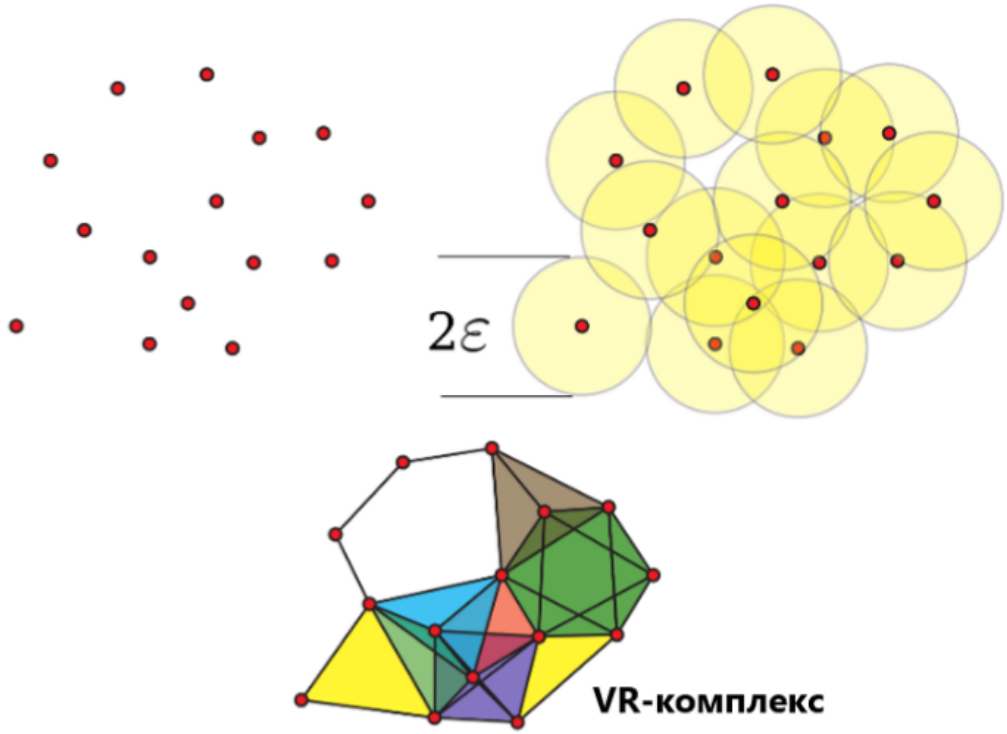


Зная звезду симплекса, можно легко вычислить и его линк

Пререкквизиты

Определение

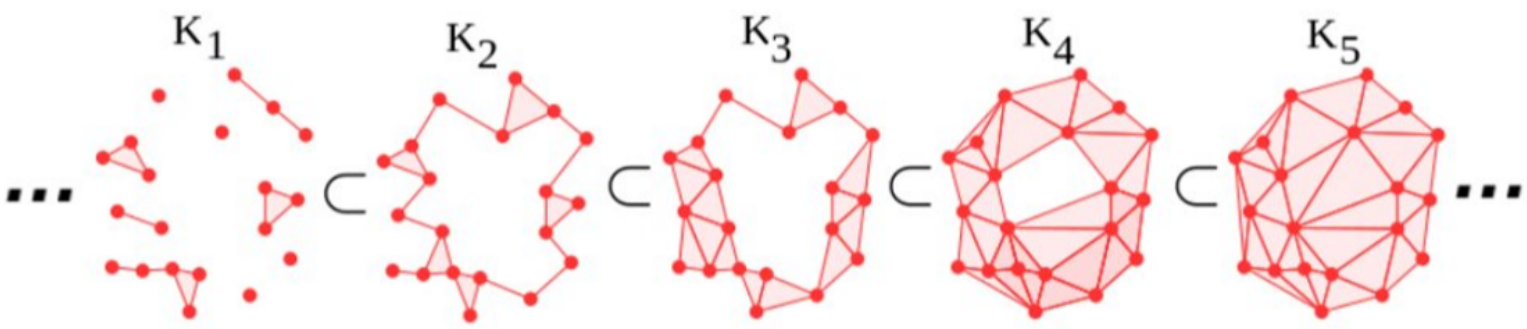
Комплекс Вьеториса-Рипса или **VR-комплекс** — симплициальный комплекс, полученный из облака точек путем объединения его подмножеств в симплексы при условии, что диаметр этого подмножества меньше 2ε, где ε — заданный наперед параметр



Прерeквизиты

Определение

Фильтрация симплициального комплекса — возрастающая последовательность его подкомплексов, т.е. каждый подкомплекс является подкомплексом следующего:

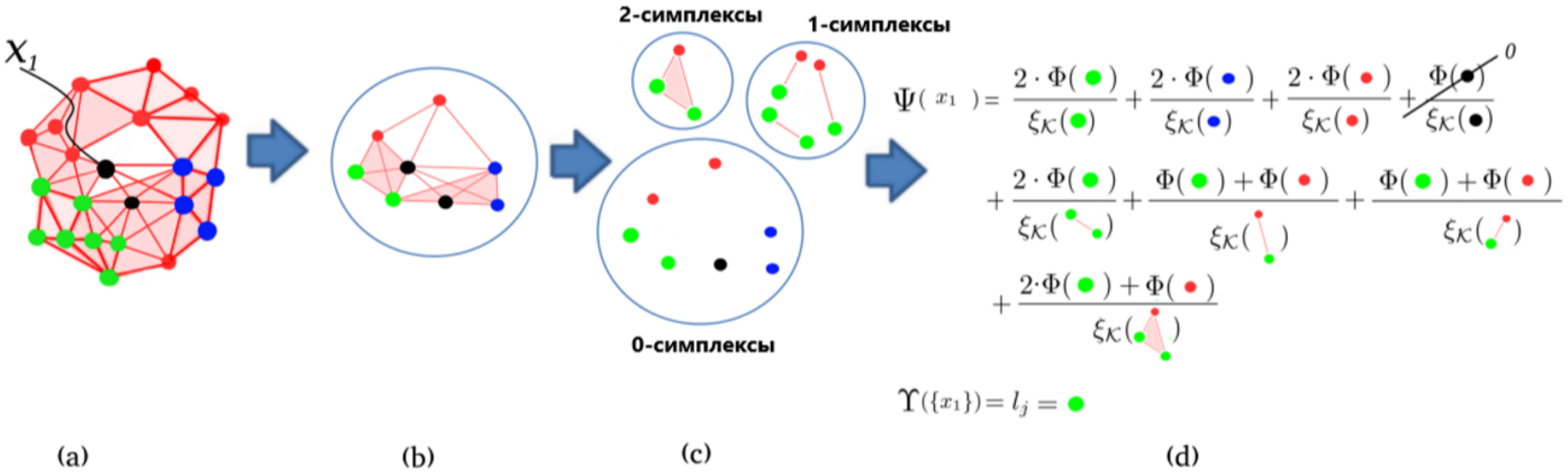


Фрагмент фильтрации симплициального комплекса

Таким образом, мы можем построить фильтрацию VR-комплекса данного облака точек, занумерованную значениями параметра ε и каждому симплексу приписать значение фильтрации, при котором он впервые появляется. Теперь перейдем непосредственно к описанию алгоритма

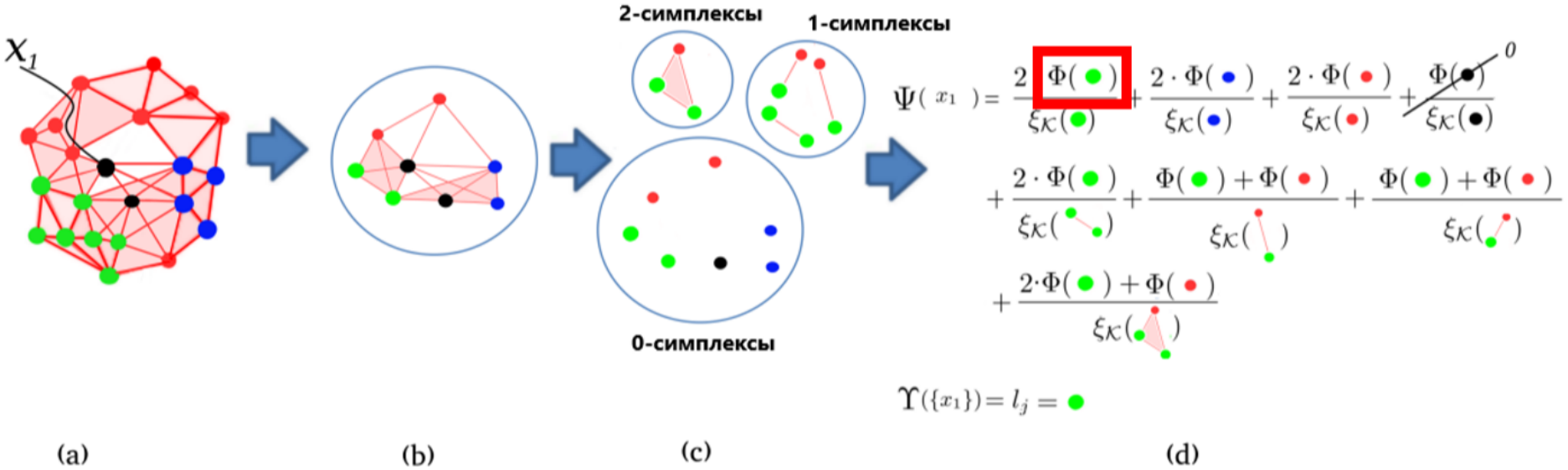
Алгоритм Топологической Классификации

Его работу на каждой отдельной вершине можно описать как последовательное применение трех функций:



Здесь, черные вершины - это те, которым мы присваиваем класс.

Алгоритм Топологической Классификации



Алгоритм топологической классификации

Первая функция Φ, **Ассоциирующая**, каждой вершине комплекса сопоставляет соответствующий её классу базисный вектор N-мерного пространства, где N — это кол-во классов. Т.е. мы ассоциировали базисные векторы N-мерного пространства и множество классов.

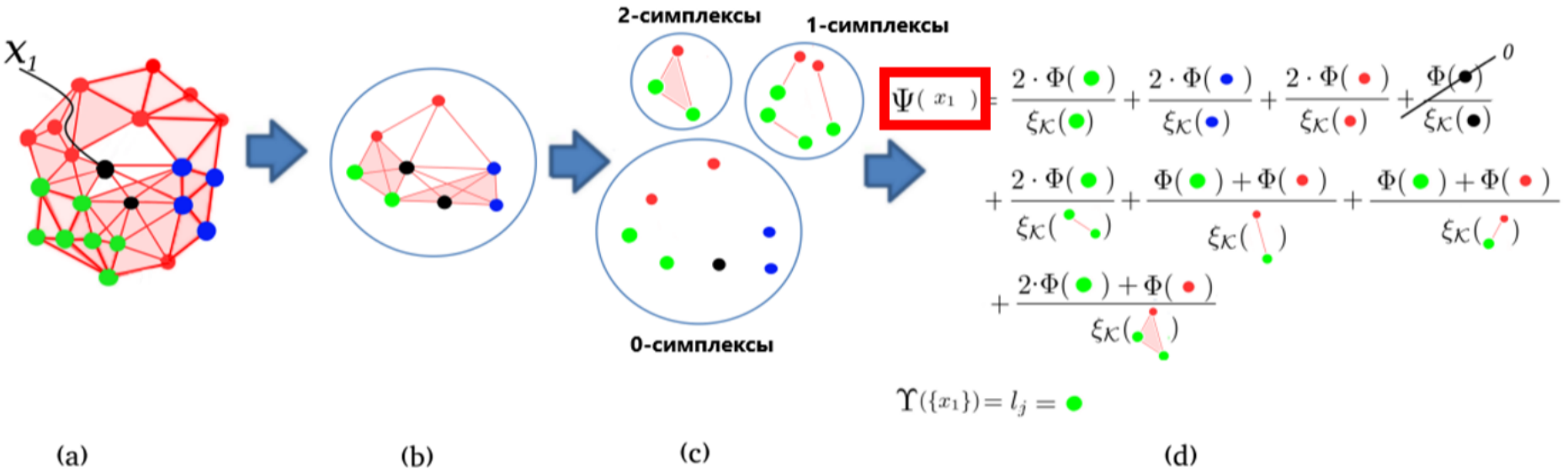
$$\Phi_\varepsilon(v) = \begin{cases} \hat{l}_i, \text{если } l_i \text{ - класс вершины } v \in K_\varepsilon \\ 0, \text{иначе} \end{cases}$$

где K_ε — VR-комплекс нашего облака точек со значением параметра ε. Эту функцию можно доопределить и для произвольных симплексов.

$$\Phi_\varepsilon(\sigma) = \sum_{v \in \sigma} \Phi_\varepsilon(v)$$

Таким образом, мы каждому симплексу при фиксированном значении ε сопоставили N-мерный вектор, компоненты которого — суть кол-во его вершин с данной маркой

Алгоритм Топологической Классификации



Алгоритм топологической классификации

Вторая функция Ψ, **функция Расширения**, берет симплексы близкие к нашей точке, вычисляет ассоциированные с ними векторы, "взвешивает" их и суммирует результат. В статье близкие симплексы берутся из линка вершины, мы же предлагаем брать их из звезды и посмотреть какой из способов будет лучше:

$$\Psi_\varepsilon(v) = \sum_{\sigma \in Lk_{K_\varepsilon(v)}} \frac{\Phi_\varepsilon(\sigma)}{\xi_K(\sigma)}$$

или

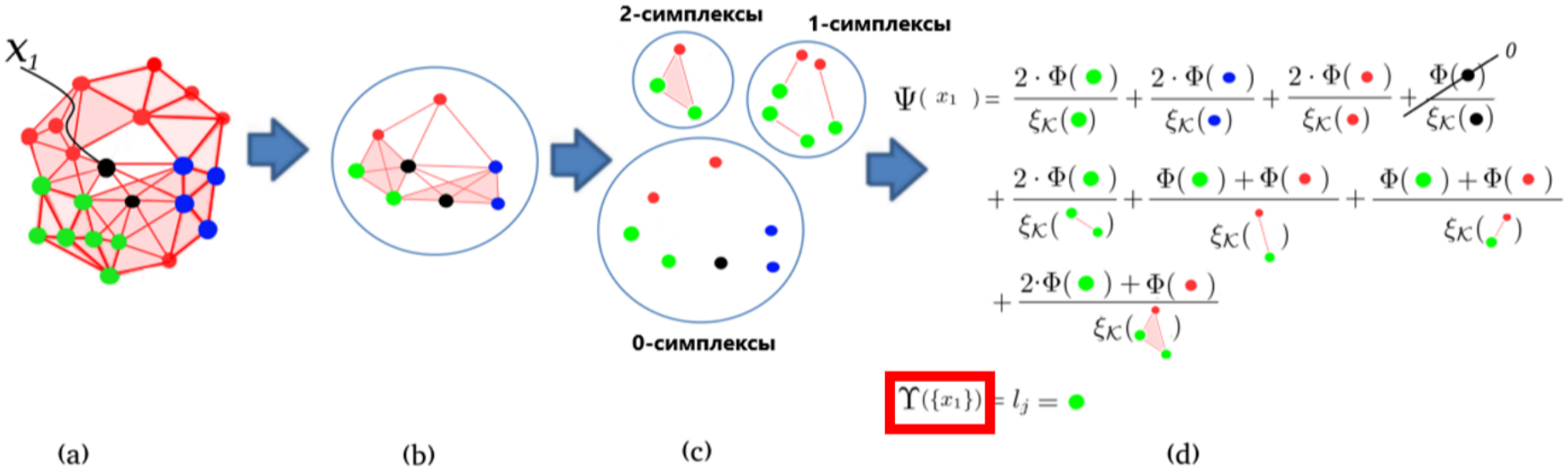
$$\Psi_\varepsilon(v) = \sum_{\sigma \in St_{K_\varepsilon(v)}} \frac{\Phi_\varepsilon(\sigma)}{\xi_K(\sigma)}$$

где $\xi_K(\sigma)$ — значение фильтрации, при котором появился симплекс сигма. Чем раньше наша вершина попала в комплекс сигма, тем меньше значение функции кси и тем больший вклад вносят вершины, которые он содержит.

Алгоритм Топологической Классификации

Теперь мы имеем N-мерный вектор, компоненты которого характеризуют сколько и как близко находятся вершины данного класса к классифицируемой. Заметим, что для вершин значение функции ξ_K по определению равно нулю, но на ноль мы делить не можем, поэтому в реализации предполагается, что значение ξ_K на вершинах равно одной миллионной (или любое другое достаточно малое число, это зависит от кол-ва точек)

Алгоритм Топологической Классификации

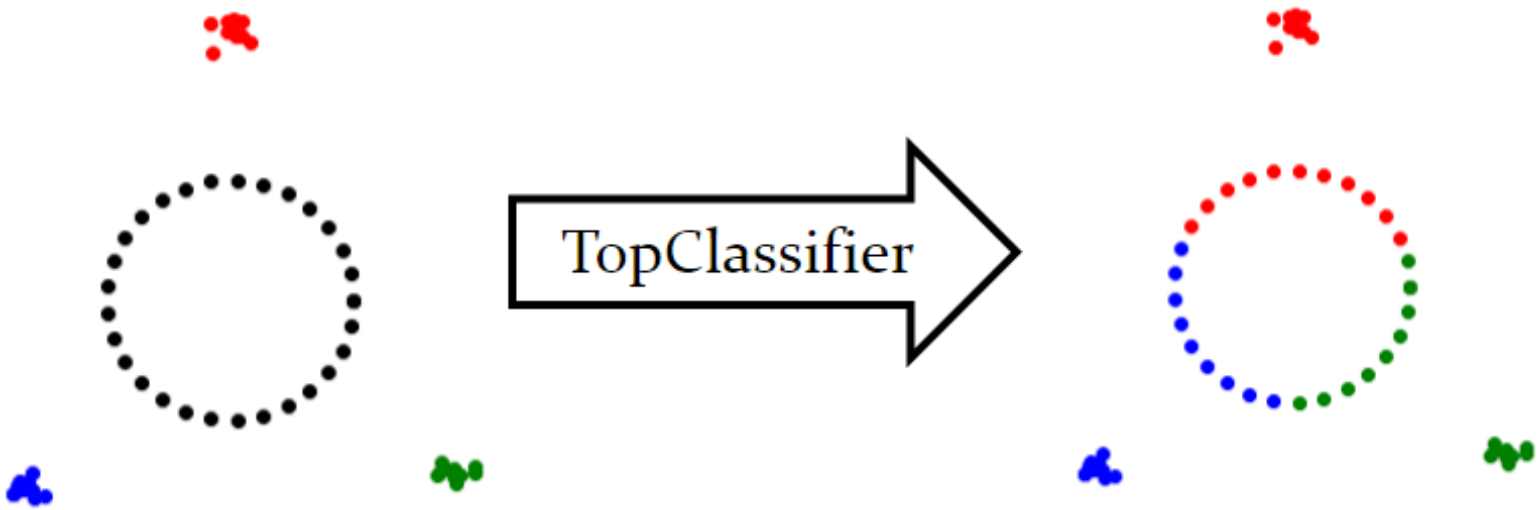


Алгоритм Топологической Классификации

Мы пришли к финальному этапу присвоения метки. По заданной точке мы построили вектор или, проще говоря, набор весов, которые определяют влияние положения этой точки на образование связей с точками соответствующего класса.
Последнее, что остается сделать — это выбрать класс с максимальным весом и присвоить его точке. За это и отвечает функция Υ. Если классов несколько, то можно всегда выбирать первый попавшийся, а можно выбрать случайно. Это, как и выбор линка или звезды для функции Расширения, влияет на точность предсказания. Выбор этих параметров заложен в реализацию нашего алгоритма.
Также мы оставили возможность изменять или не изменять облако точек, которое мы используем для определения метки, т.е. использовать только что помеченные точки для предсказания меток новых.

Алгоритм Топологической Классификации

Пример работы алгоритма.
Заданы три облака точек трех разных классов (красного, синего и зеленого).
На вход подается облако точек из круга и алгоритм классифицирует их:



Применение алгоритма

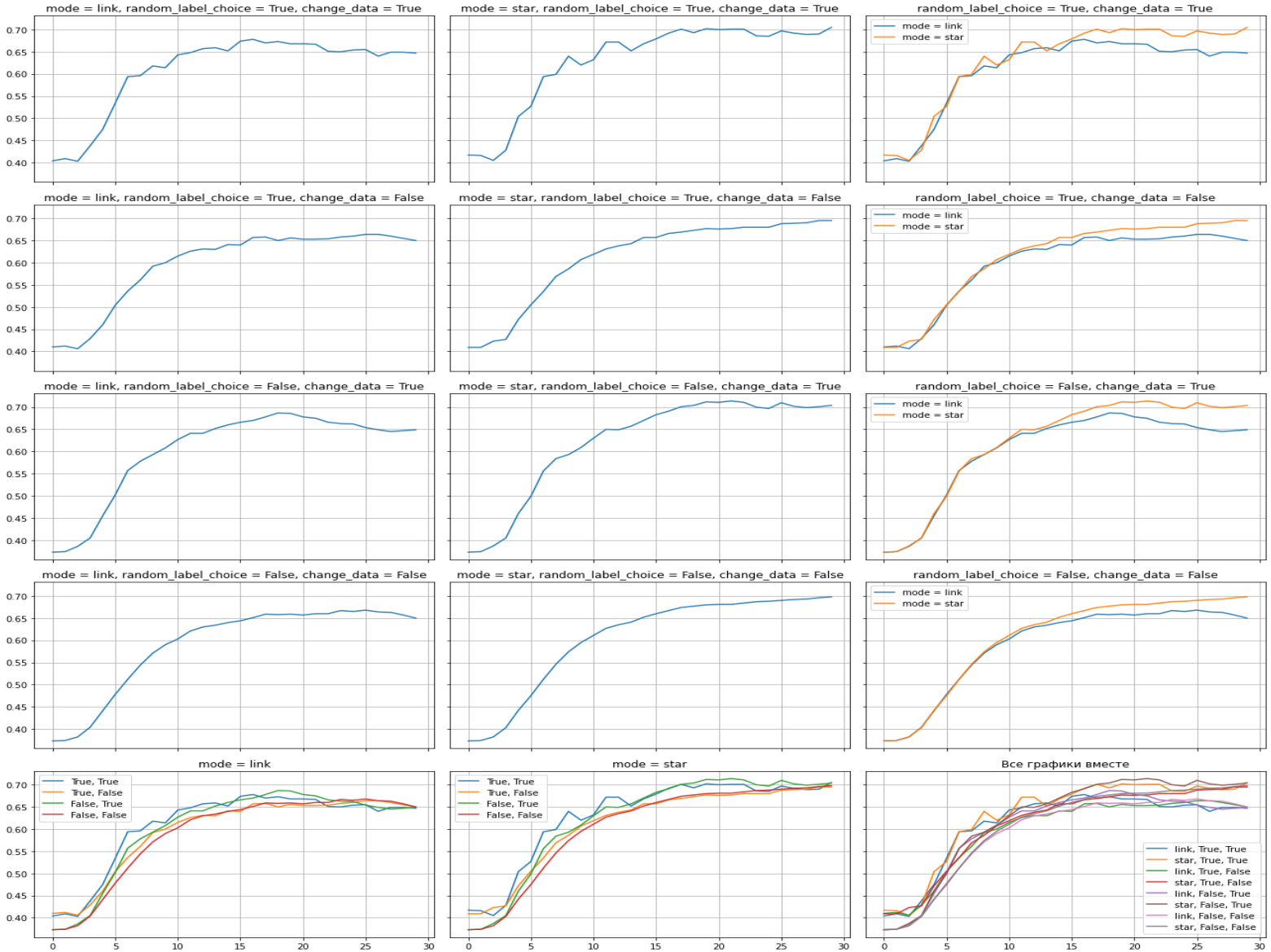
Прежде чем тестировать алгоритм на базе данных, упомянем способ оценки верности предсказания. Поскольку верно предсказать литотип сложно даже для специалиста и допускается некоторая погрешность в пределах схожих типов, то оценивать эффективность нашего алгоритма мы будем, руководствуясь таблицей, используемой в Geophysical Tutorial Machine Learning Contest 2016, в которой для каждого класса указаны схожие с ним.
В контесте для оценки использовалась метрика F1-micro, которая, как известно, в задаче многоклассовой классификации совпадает с метрикой точности, поэтому её мы и будем вычислять.

Facies	Label	Adjacent Facies
1	SS	2
2	CSiS	1,3
3	FSiS	2
4	SiSh	5
5	MS	4,6
6	WS	5,7
7	D	6,8
8	PS	6,7,9
9	BS	7,8

Применение алгоритма

Из всей базы данных для классификации мы использовали только четыре параметра: значения ГК (GR), удельного сопротивления (ILD log10), разницы в пористости по плотности нейтронов (DeltaPHI) и средней пористости нейтронной плотности (PHIND).
Мы изменяли параметры filt value, mode, random label choice и change data и получили следующие графики зависимости метрики точности от этих параметров.
Параметр maxdim был при этом всегда равен единице, поскольку эмпирически было установлено, что при maxdim > 1 точность при тех же параметрах уменьшается, к тому же кратно возрастает время вычислений, т.к. появляется много симплексов размерности 2.

Применение алгоритма



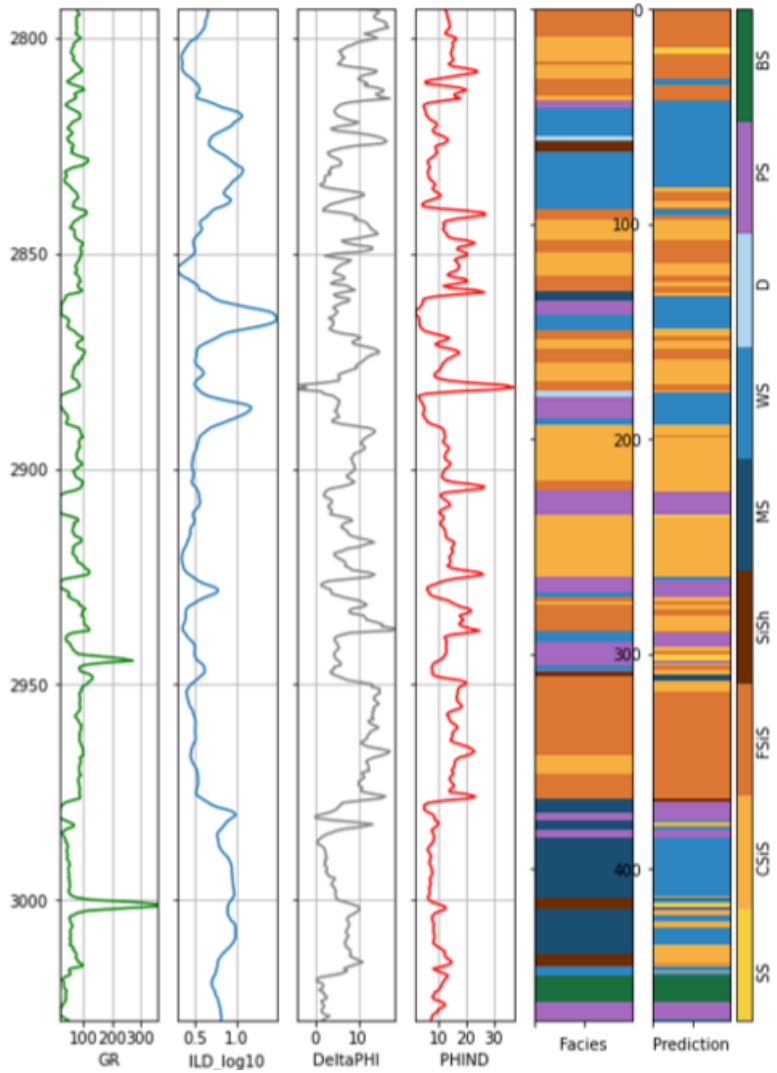
Применение алгоритма

Как видно из графиков, лучшим набором параметров является:

- filt value = 21
- maxdim = 1
- mode = star
- random label choice = False
- change data = True

Применение алгоритма

Пример предсказания литологии скважины STUART:



Обсуждение

Версия алгоритма	Accuracy	Adjacent Accuracy	F1-micro
TopClassifierLink	0.326	0.686	0.172
TopClassifierStar	0.376	0.713	0.236

Adjacent accuracy — модифицированная метрика accuracy, которая учитывает схожесть смежных классов.

Обсуждение

Позиция	Команда	F1	Алгоритм	Язык	Решение
1	LA Team	0.6388	Boosted trees	Python	Notebook
2	PA Team	0.6250	Boosted trees	Python	Notebook
3	ispl	0.6231	Boosted trees	Python	Notebook
4	esaTeam	0.6225	Boosted trees	Python	Notebook

По результатам контеста победил алгоритм, F1-micro метрика (то же, что и accuracy) которого приблизительно равна 0.64.
Учитывая тот факт, что Adjacent Accuracy чуть меньше чем вдвое больше Accuracy, то можно ожидать, что Accuracy этого алгоритма лежит в пределах 0.9 - 1.

Заключение

Мы увидели, что алгоритм TopClassifier плохо справился с классификацией сырых геофизических данных, поэтому, возможно, стоит провести feature engineering, но для этого нужны хорошие знания в геофизике.
Мы испробовали совсем немного методов ТАД, и, чтобы в полной мере раскрыть его потенциал, потребуется больше реальных данных, вычислительных мощностей и консультирования со стороны специалистов.
В заключение можно сказать, что на данном этапе использование алгоритма TopClassifier для предсказания геологического разреза не эффективно.