# Результаты, сгруппированные по методам машинного обучения

Для исследования было выбрано большое количество методов машинного обучения, которые были сгруппированы в следующие категории:

* Boosting
* Clustering
* Regression
* Decision trees
* Nearest neighbours
* Support vector machine
* Hyper pipes (heuristic)
* Voting Feature Intervals (heuristic)
* Meta classifiers (weak classifiers which used by another classifiers)

Классификация основана на классификации, предложенной в \*отсылка на кембриджскую книгу\* и трудах Воронцова.

Для получения обобщенной картины использовалось исследование 4 различных датасетов по одному форуму:

1. Выделение ключевых слов с помощью алгоритма RAKE, выделение цитирования на основе последнего сообщения автора
2. Выделение ключевых слов с помощью алгоритма RAKE, выделение цитирования на основе поиска процитированного текста
3. Выделение ключевых слов на основе частоты термов, наивное выделение цитирования на основе последнего сообщения автора
4. Выделение ключевых слов на основе частоты термов, наивное выделение цитирования на основе поиска процитированного текста

Усреднение производилось следующим образом:

1. Каждый метод машинного обучения внутри каждой группы подвергался обучению и проверке на контрольной выборке с помощью бутстрепа. Количество итераций бутстрепа – 100. На каждой итерации, после проверки на контрольной выборке осуществлялся расчет значений NCG для каждого количества отобранных постов [0; N], где N – размер контрольной выборки. После окончания всех итераций бутстрепа по рассматриваемому методу, результаты по всем итерациям усреднялись следующим образом: для каждого значения количества обобранных постов (ось аргументов графика NCG), в качестве нижней границы было взято 10% квантили, среднего – 50% квантили, верхнего – 90% квантили.
2. Процедура, описанная в п.1, проводилась для каждого из 4 датасетов. В результате обработки был получен большой список записей, где каждый метод машинного обучения был сопоставлен с усредненными результатами на NCG N раз, где N – количество входных датасетов:
3. Список с результатами, полученный в п.2, был сгруппирован в конечный результат. Для этого, по каждому числу k отобранных постов (ось аргументов графика NCG) из всех методов машинного обучения из соответствующей категории, было выбрано минимальное значение 10% квантили, среднее значение 50% квантили, максимальное значение 90% квантили. Тем самым, для каждой категории алгоритмов машинного обучения получилось «среднее» значение NCG по каждому количеству отобранных постов, а также минимальное и максимальное отклонение значение NCG (дисперсия).

Все результаты отображены в виде графиков на одной диаграмме:

Диаграмма 1

Как можно заметить, графики сильно накладываются друг на друга, ввиду чего далее представлены диаграммы, на которых изображены вместе отдельные графики из первой диаграммы.

Диаграмма 2

Диаграмма 2 хорошо отражает схожесть результатов обучения алгоритмов, основанных на деревьях решений, регрессии и бустинге – эти графики имеют очень похожую линию «среднего», а также большую площадь перекрытия. При этом, самой большой дисперсией обладает группа методов, основанных на деревьях решений, а самой меньшей – бустинг.

Диаграмма 3

Диаграмма 3 наглядно иллюстрирует превосходство малодисперсных методов SVM и метода ближайших соседей над эвристическими группами методов. Также заметно, что алгоритмы, основанные на методах ближайших соседей, проявляют себя лучше алгоритмов, рассмотренных на второй диаграмме – несмотря на то, что линии «среднего» имеют похожую форму, линия «среднего» для метода ближайших соседей сдвинута гораздо выше по оси ординат, а сам график NCG имеет гораздо меньшую дисперсию.

Диаграмма 4

На диаграмме 4 видно, что результаты, полученные для групп мета-классификаторов и кластеризации, значительно хуже по сравнению с ранее рассмотренными группами. Эти методы проявляют большую дисперсию и имеют достаточно низкое положение «средней» линии. Особый интерес вызывает график для кластеризации – заметно, что для небольшого количества постов (до 90) эта группа методов имеет крайне малую дисперсию.

Диаграмма 5

На последней диаграмме отражены группы методов, которые выдали наилучшие результаты по сравнению с остальными. Группа SVM методов выдает отличные результаты на небольшом количестве постов (до 50), после чего результативность несколько снижается, однако несмотря на это, имеет небольшую дисперсию.