# *Ансамблевые методы классификации*

Курс «Основы анализа текстовых данных»

Кафедра управления и информатики НИУ «МЭИ»

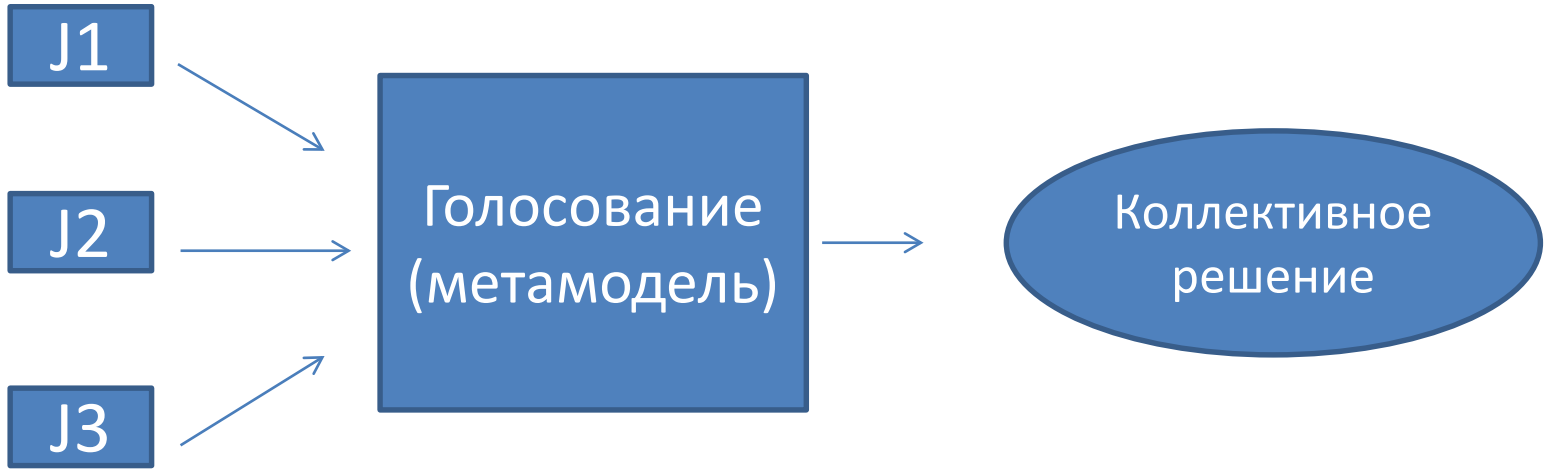
Весна 2023 г.

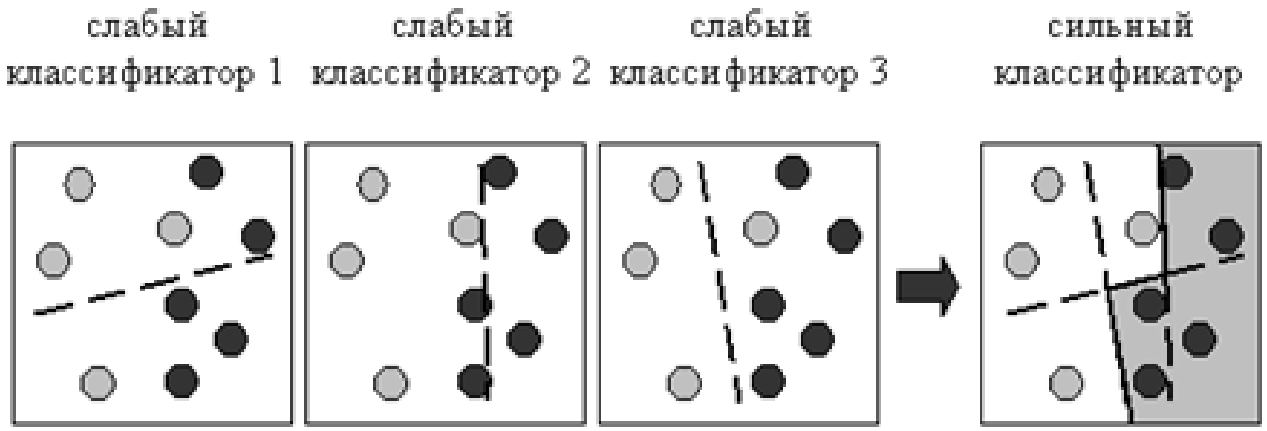
# *Ансамблевые методы классификации*

- Коллективы решающих правил (stacking)
- Bagging
- Boosting

# *Коллективная классификация*

*Коллективом решающих правил (КРП)* называется совокупность методов классификации, объединенных для выработки общего решения.



# *Коллективная классификация*

*Какие методы включать в коллектив?*

- Наиболее точные
- Наиболее разнородные – ошибающиеся на разных объектах

*Как померить разнородность?*

– Использовать меры сходства (см. методы выявления информативных терминов)

Способы обеспечения дополнительной разнородности:
- обучение *КРП* с помощью методов *bagging* и *boosting*;
- обучение комитета классификаторов на различных независимых обучающих выборках;

# *Коллективная классификация (2)*

*Сколько методов включать в коллектив?*

Вероятность правильной классификации в зависимости от количества и точности методов:

| | m = 3 | m = 5 | m = 7 | m = 9 |
|---|---|---|---|---|
| p = 0,6 | 0,648 | 0,682 | 0,710 | 0,733 |
| p = 0,7 | 0,784 | 0,837 | 0,874 | 0,901 |
| p = 0,8 | 0,896 | 0,942 | 0,966 | 0,980 |
| p = 0,9 | 0,972 | 0,991 | 0,997 | 0,999 |

# *Стратегии принятия решений*

- *Простое голосование* – каждый классификатор имеет равный вес при принятии решения. Новое наблюдение относится к тому классу, за который проголосовало большинство членов *КРП*
- *Взвешенное голосование* – каждому классификатору присваивается вес в зависимости от количества допускаемых ошибок $\Delta_p$ ( $\Delta_p$ –ошибка *p*–го классификатора, ). Решение об отнесении нового наблюдения к какому-либо из классов принимается по формуле:

$$C(\vec{X}_{N+1}) = \sum_{p=1}^{m} \left( \frac{\Delta_p}{\sum_{s=1}^{m} \Delta_s} J_p \right).$$

- *Определение областей компетенции* для классификаторов, включенных в комитет (например, в случае неоднородных *КРП* можно выявить для каждого *p*–го решающего правила «зону ответственности», в которой классификатор ошибается меньше других.

# *Что делать, если классификаторы не пришли к решению?*

Вводят понятие «Отказ от классификации» (метка «Джокер»)

Если все члены комитета присваивают полностью не совпадающие метки одному и тому же объекту, то это означает, что, скорее всего, данный объект является нехарактерным шумовым элементом и к нему целесообразно применить операцию “*Отказ от классификации*”.

При этом наблюдения, получившие метку “*Джокер*” не включаются в расчет общей ошибки, т.е. в этом случае общая ошибка вычисляется по формуле:

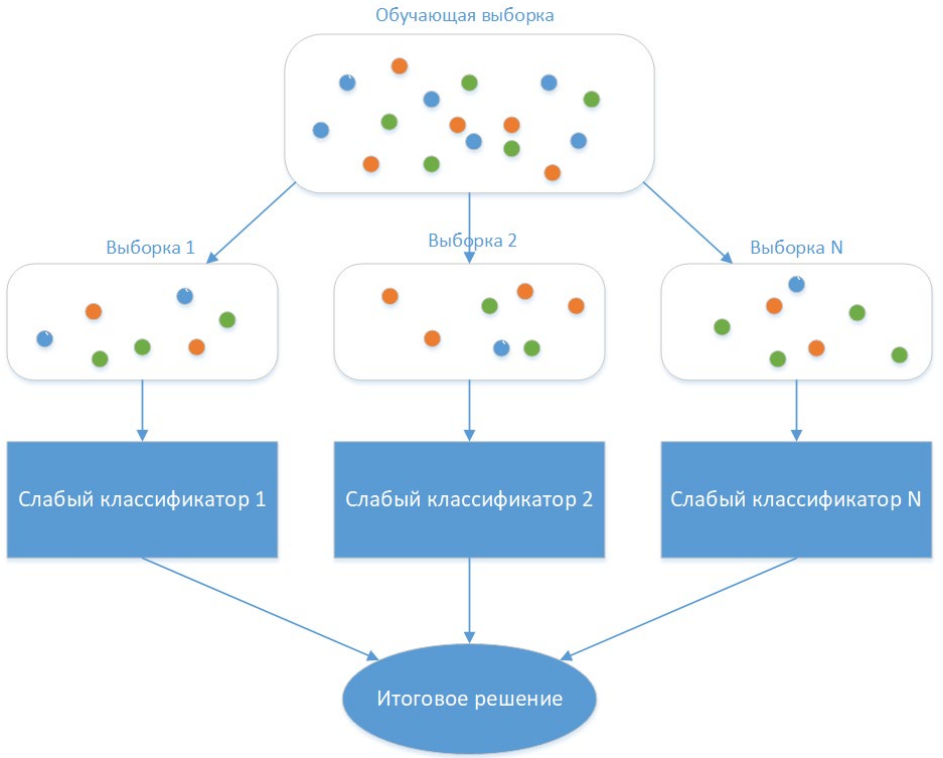
$$\Delta = \frac{(N^{-})^{*}}{N^{*}}$$

*$(N^{-})^{*}$* - число документов, отнесенных к неправильным классам, *$(N^{-})^{**}$* - число документов, получивших метку «Джокер», *N* – размер экзаменационной выборки, *$N^{*} = N - (N^{-})^{**}$*

# *Bagging*

Bagging – Bootstrap AGGregatING - для обучения «слабых» классификаторов используется случайная выборка с повторениями. Обучаем все классификаторы параллельно.
Результат классификации выдается на основе простого голосования каждой модели - обычно, вес каждого «слабого» классификатора одинаковый.

Классический пример – bagging над деревьями решений – «Случайный лес»



# *Случайный лес (Random Forest)*

Пусть обучающая выборка состоит из N примеров, размерность пространства признаков равна M, и задан параметр m

Все деревья комитета строятся независимо друг от друга по следующей процедуре:

1. Сгенерируем случайную подвыборку **с повторением** размером *N* из обучающей выборки. (Таким образом, некоторые примеры попадут в неё несколько раз, а в среднем $N(1 - 1/N)^N$, т.е. примерно *N/e* примеров не войдут в неё вообще)
2. Построим решающее дерево, классифицирующее примеры данной подвыборки, причём в ходе создания очередного узла дерева будем выбирать признак, на основе которого производится разбиение, не из всех M признаков, а лишь из m случайно выбранных.
3. Проводится построение дерева

Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев.

Часто используют «лес решающих пней» - деревья глубиной 1, т.е. каждое дерево проверяет только 1 признак

# *Boosting*

Хотим, чтобы каждый последующий классификатор улучшал качество классификации предыдущих => обучаем «слабые» классификаторы последовательно.

AdaBoost – адаптивный бустинг – Объектам, которые неправильно проклассифицированы текущим комитетом, назначается больший вес, и последующий классификатор строится с учетом веса этих объектов.

Градиентный бустинг – на каждом шаге оптимизируется функция потерь, зависящая от ответов комитетов на предыдущем шаге.