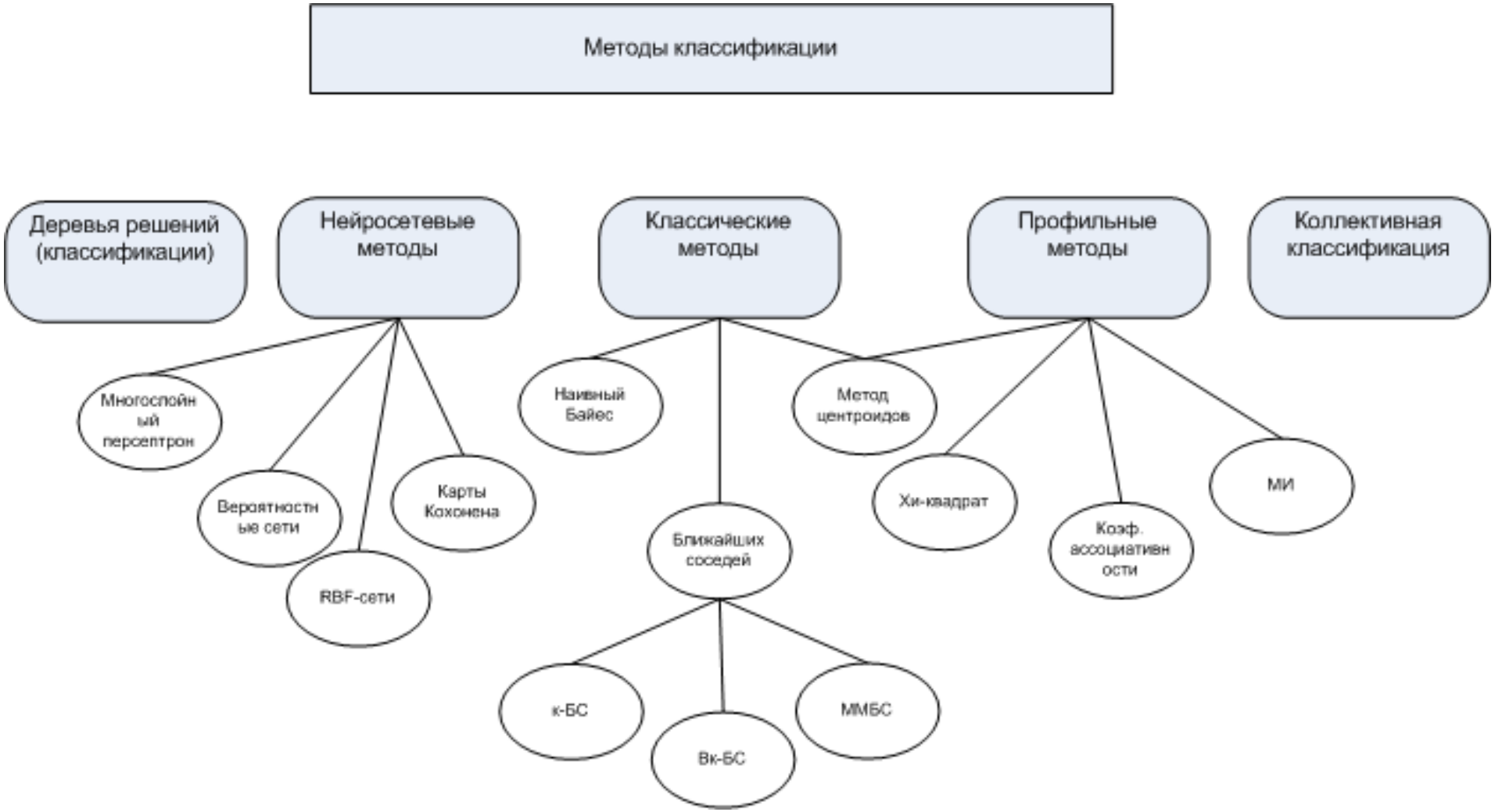
Методы классификации текстовых документов

Курс «Интеллектуальные информационные системы»

Кафедра управления и информатики НИУ «МЭИ»

Осень 2017 г.

Систематизация методов классификации

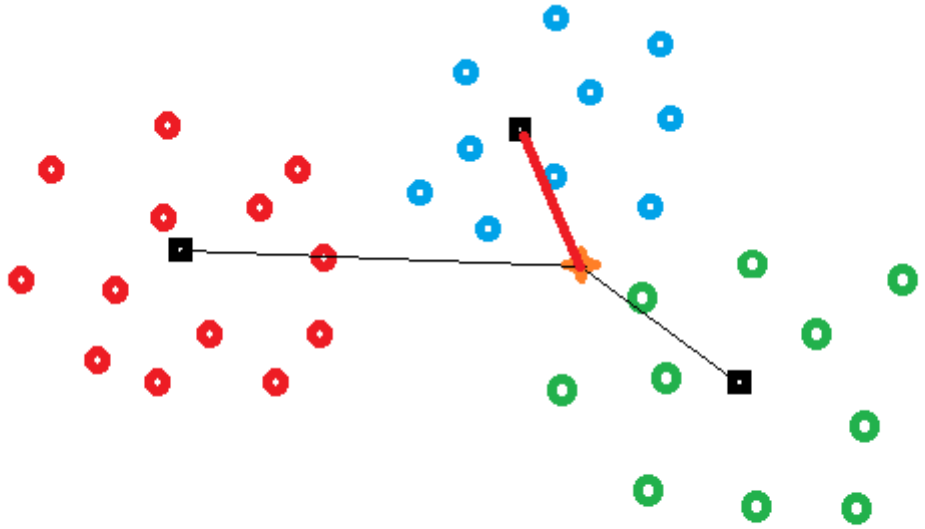


Центроидный метод

Центроид – вектор со средними значениями весов терминов документов данного класса. «Центр тяжести».
Классифицируемый объект относится к классу с наиболее близким центроидом.

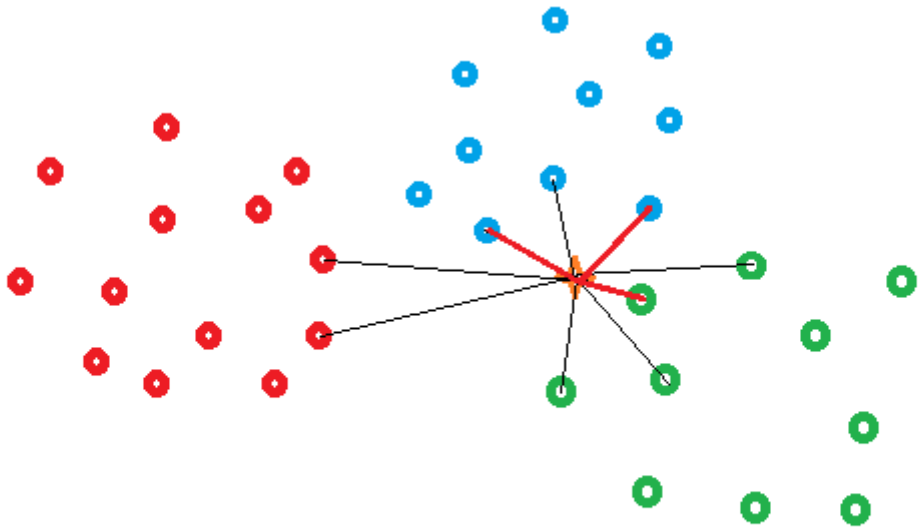
$$\vec{\mathbf{C}}_k = \frac{1}{N_k}\sum_{j=1}^{N_k}\vec{X}_j$$

Роккио: $$\vec{C}_k = \alpha\frac{1}{N_k}\sum_{j=1}^{N_k}\cos(\vec{C}_k,\vec{X}_j) - \frac{\beta}{N-N_k}\sum_{l=1}^{N-N_k}\cos(\vec{C}_k,\vec{X}_l)$$



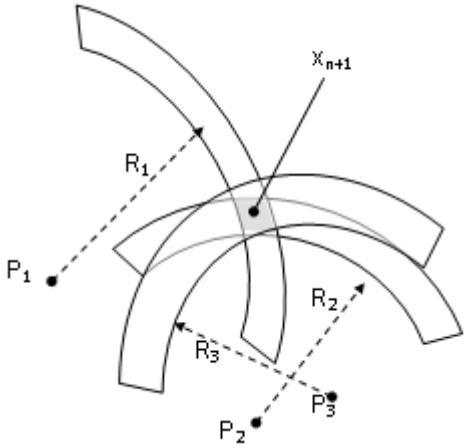
Правило ближайшего соседа (БС)

Классифицируемый объект относится к тому классу, к которому относится ближайший к нему сосед.



Семейство методов БС

- кБС – Решение принимается на основании анализа к ближайших соседей. Обычно к - нечетное число [5;25]
- Взвешенный кБС – наиболее близкие соседи имеют больший вес при голосовании.
- Модифицированный МБС – поиск соседей только определенной области признакового пространства, с целью сокращения вычислительных операций.



Наивный байесовский метод (НБ)

теорема Байеса: $$P(Q_k \mid \vec{X}) = \frac{P(\vec{X} \mid Q_k)P(Q_k)}{P(\vec{X})}$$

позволяет определить вероятность какого-либо события при условии, что произошло другое статистически взаимозависимое с ним событие.

- $P(\vec{X})$ - одинакова для различных классов и может быть исключена из дальнейшего рассмотрения
- Допущение: поскольку мы используем модель «мешок слов», условная вероятность документа аппроксимируется произведением условных вероятностей всех слов входящих в документ

$$P(\vec{X} \mid Q_k) = \prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$$

$$P(Q_k \mid \vec{X}) = P(Q_k) \prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$$

Наивный байесовский метод (2)

$$P(Q_k \mid \vec{X}) = P(Q_k)\prod_{i=1}^{M} P(x^{(i)} \mid Q_k)$$

- $\widehat{P}(Q_k) = \frac{N_k}{N}$ - оценка для P(Qk) – вероятность встретить документ класса Qk в корпусе документов
- $\widehat{P}(x^{(i)} \mid Q_k) = \frac{N_{ik}}{N_k}$ - вероятность встретить термин x(i) в классе Qk
- Часто используется уточненная формула: $\widehat{P}(x^{(i)} \mid Q_k) = \frac{1+N_{ik}}{M+N_k}$, где M – общее количество терминов во всех документах выборки

$$P(Q_k \mid \vec{X}) = \arg\max_{k \in K} \frac{N_k}{N} \prod_{i=1}^{M} \frac{1+N_{ik}}{M+N_k}$$