NATIONAL RESEARCH UNIVERSITY

Can FCA-based Recommender System Suggest a Proper Classifier?

Yury Kashnitsky, Dmitry Ignatov

Higher School of Economics,

School of Applied Mathematics and Information Science

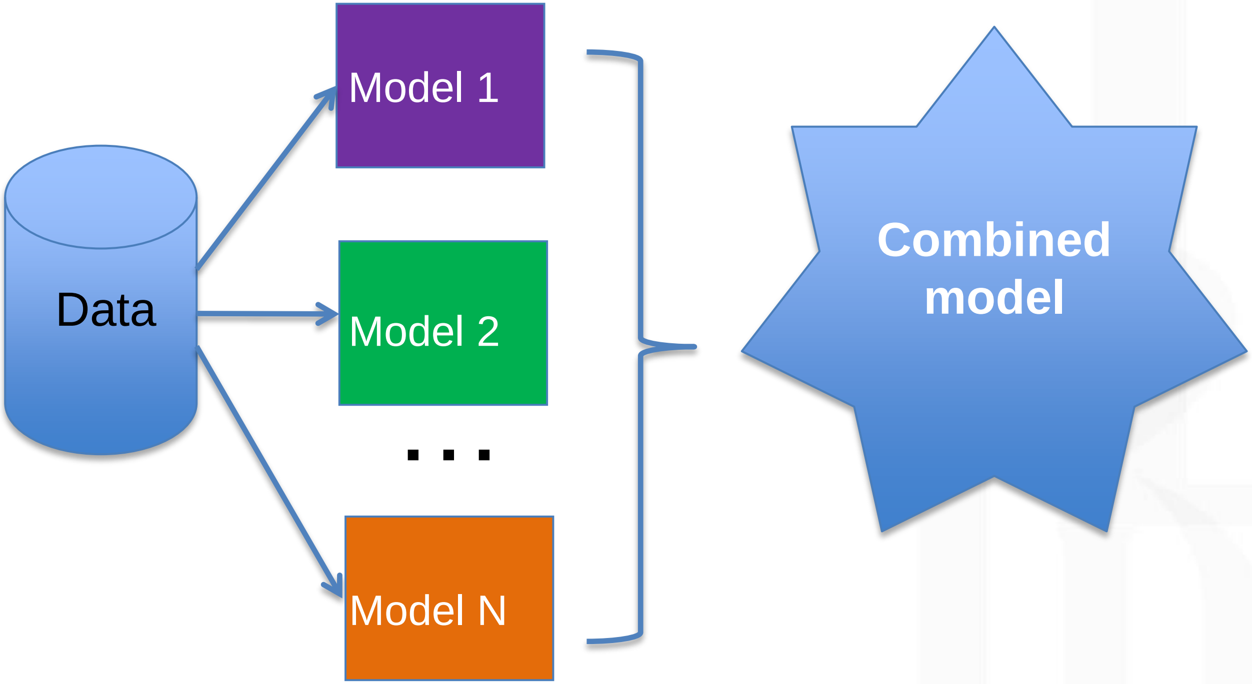
3rd Workshop "What can FCA do for Artificial Intelligence?"

21st European Conference on Artificial Intelligence

Background

Definition

Ensemble classification - aggregation of predictions of multiple classifiers in order to obtain better predictive performance



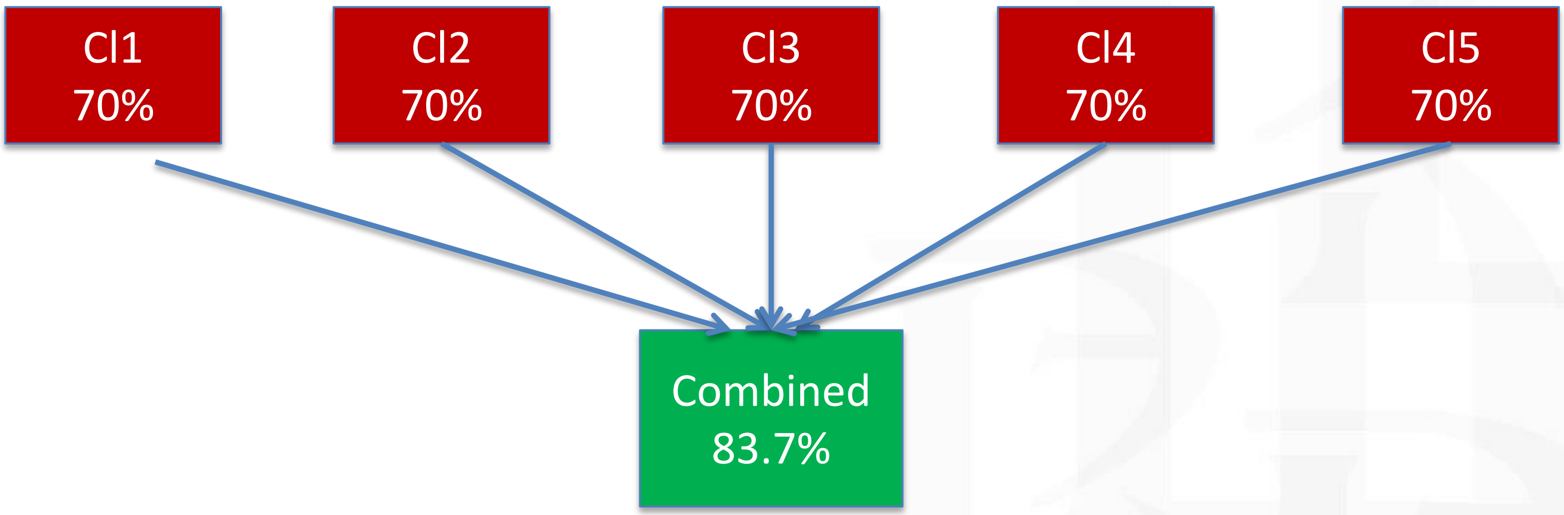
Notion variation

- Ensemble classification algorithm
- Multiple classifier system
- Committee machine
- Meta-algorithm

Condorcet jury theorem

The Condorcet jury theorem («the wisdom of the crowd», 1785)

If a population makes a group decision and each voter most likely votes correctly, then adding more voters increases the probability that the majority decision is correct.



$$Acc_{comb} = C_5^3(0.7)^3(0.3)^2 + C_5^4(0.7)^4 0.3 + C_5^5(0.7)^5 = 0.837$$

In case of 100 independent classifiers $Acc_{comb} = 0.999$

Popular ensemble learning algorithms

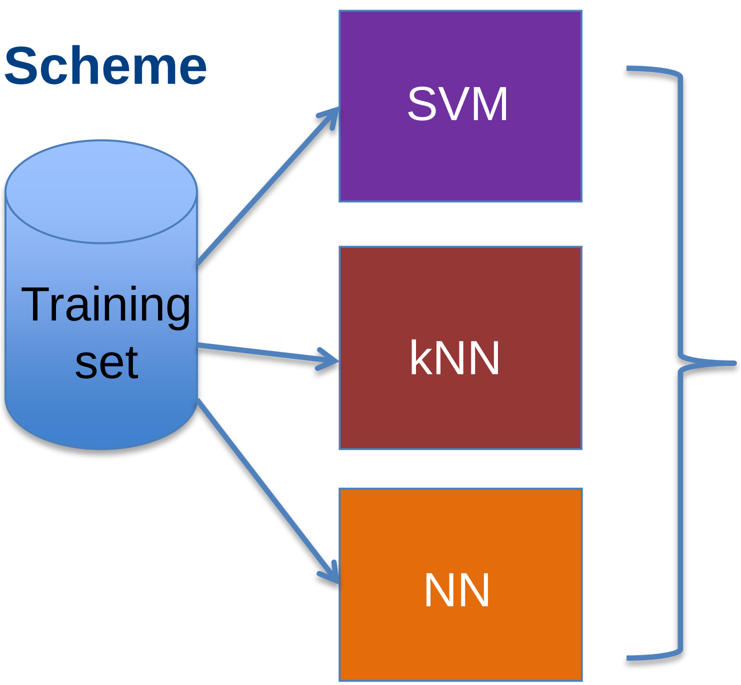
- **Bagging** (bootstrap aggregation) – training different classifiers on multiple random samples of the initial training set.
- **Boosting** – iterative building of a succession of models, each one being trained on a data set in which points misclassified by the previous model are given more weight.
- **Stack generalization** (stacking) – combining (maybe non-linearly) base classifiers of different types according to their cross-validation performance.
- **Random forests** – bagging trees + random split selection.
- **Random subspace method (RSM)** - training learning machines on randomly chosen subspaces of the original input space (i.e. the training set is sampled in the feature space).

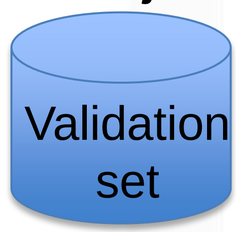
RMCS basics

Intuition

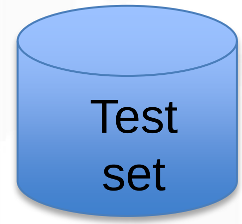
A classifier is likely to classify an object correctly if it predicts the labels of similar objects correctly.
Then this classifier is “recommended” to the object.

Scheme





	SVM	kNN	NN
1	X	X	
2		X	X
…	…	…	…
N_{val}	X		X



Object	Recommended classifier
1	NN
2	SVM
…	
N_{test}	kNN

A toy example

A toy dataset

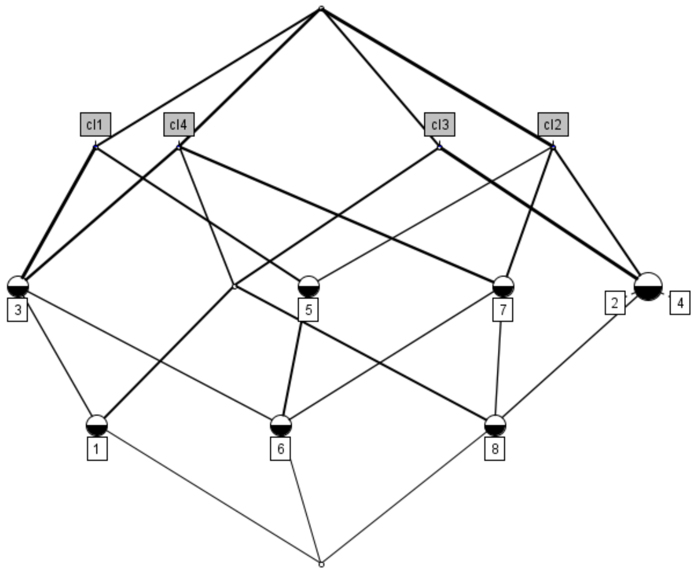
G\M	m_1	m_2	m_3	m_4	Label
1	X	X		X	1
2	X			X	1
3		X	X		0
4	X		X	X	1
5	X	X	X		1
6		X	X	X	0
7	X	X	X		1
8			X	X	0
9	X	X	X	X	?
10		X		X	?

Classification context

G\Cl	Cl1	cl2	cl3	cl4
1	X		X	X
2		X	X	
3	X			X
4		X	X	
5	X	X		
6	X	X		X
7		X		X
8		X	X	X

A toy example (2)

Concept lattice of the classification context



"Top" concepts

(G, {})
({1,3,5,6},{cl1})
({2,4,5,6,7,8},{cl2})
({1,2,4,8},{cl3})
({1,3,6,7,8},{cl4})

$G_{test} \setminus G_{val}$	1^{st} neighbor	2^{nd} neighbor	3^{rd} neighbor	Recommended classifier
9	4	5	7	cl2
10	1	6	8	cl4

Distance metrics

Distance metrics:

- **Minkowski distance:** $d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}$
- for p = 2: Euclidean, for p = 1: Manhattan.
- **Hamming distance:** $d(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i == y_i)$
- **Others:** Jaccard, Chebyshev, cosine distance and so on.
- **Weighted Hamming distance**: $d(x,y) = \frac{1}{n}\sum_{i=1}^{n} w_i(x_i == y_i)$,

  w_i - weight of i^{th} attribute.

  Here we use IG-based (Information Gain) attribute weight.

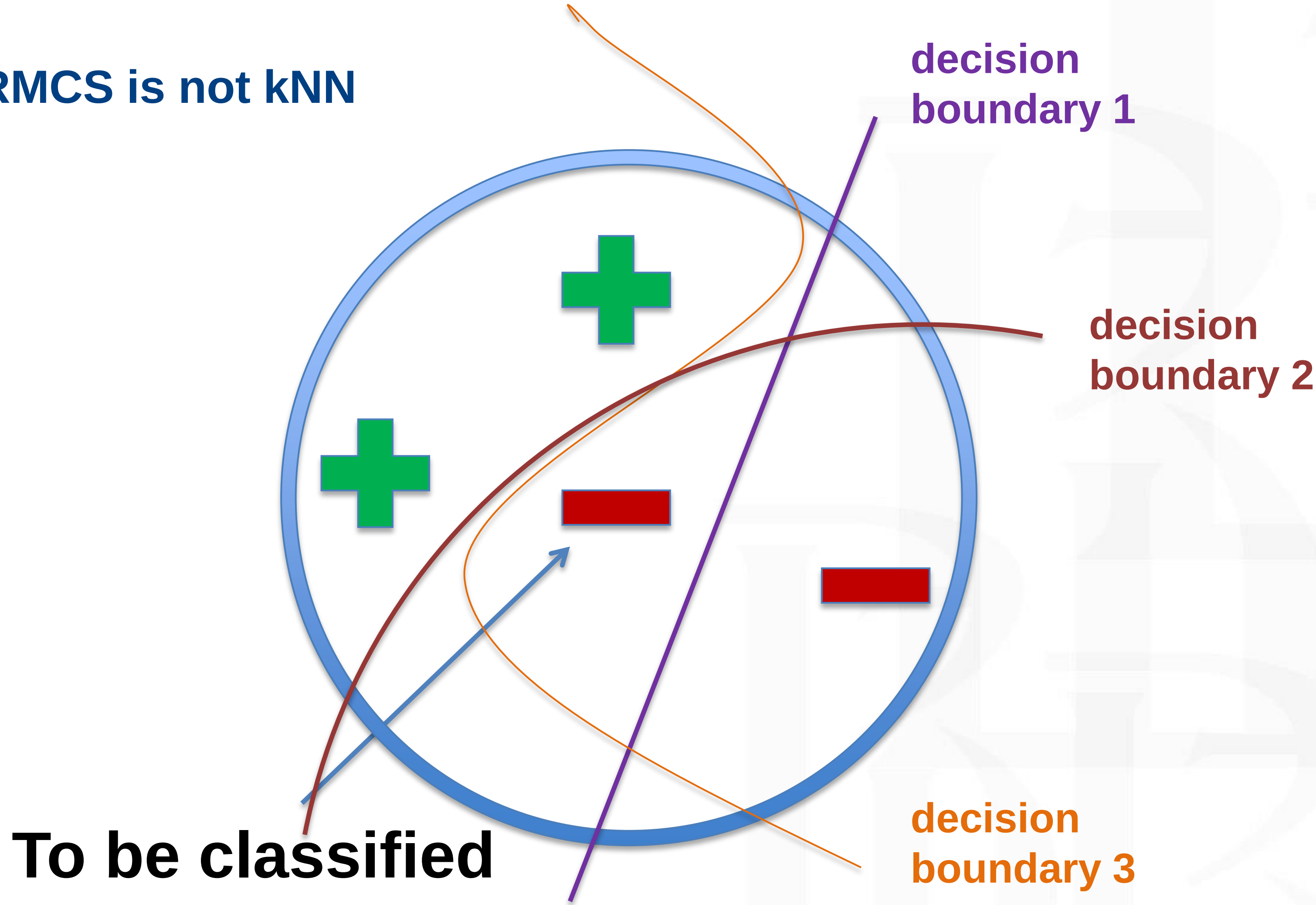
Information gain: $IG(O,a,w) = H(O,w) - \sum_{i=1}^{v}\frac{|O_i|}{|O|}H(O_i,w)$

O is a set of objects
w is the target property
a is an attribute

O_i is a set of objects which have the i^{th} value of the attribute a
v is the number of different values of the attribute a
H is entropy

What it is not

RMCS is not kNN



Experiments

UCI datasets – mushrooms, ionosphere, and hand-written digits (archive.ics.uci.edu/ml/datasets),
Scikit implementation of base learners (www.scikit-learn.org)

Dataset	**SVM** with RBF kernel ($C=10^3$, $\gamma=0.02$)	**Logit** ($C=10^3$)	**kNN** (minkowski, p=1, k=5)	**RMCS** (k=5, n_folds=10)	**Bagging SVM** 50 estimators ($C=10^3$, $\gamma=0.02$)	**AdaBoost** on decision stumps 100 iterations
Mushrooms	0.998 0.16 sec.	0.999 0.16 sec.	0.989 0.012 s.	0.999 29.45 sec.	0.999 3.54 sec.	0.998 49.56 sec.
Ionosphere	0.906 $4.3*10^{-3}$ sec.	0.868 10^{-2} sec.	0.887 $8*10^{-4}$ s	0.9 3.63 sec.	0.925 0.23 sec.	0.934 31.97 sec.
Digits	0.937 2.4 sec.	0.87 0.3 sec.	0.847 0.03 sec.	0.951 580.4 sec.	0.927 85.17 sec.	0.921 2484 sec.

Further work

The directions for further work on RMCS:

- exploring the impact of different distance metrics (such as the one based on attribute importance or information gain) on the algorithm's performance
- experimenting with various types of base classifiers
- investigating the conditions preferable for RMCS (in particular, when it outperforms bagging and boosting),
- improving execution time of the algorithm
- analyzing RMCS's overfitting

Thank you for your attention!

20, Myasnitskaya str., Moscow, Russia, 101000
Tel.: +7 (495) 628-8829, Fax: +7 (495) 628-7931
www.hse.ru