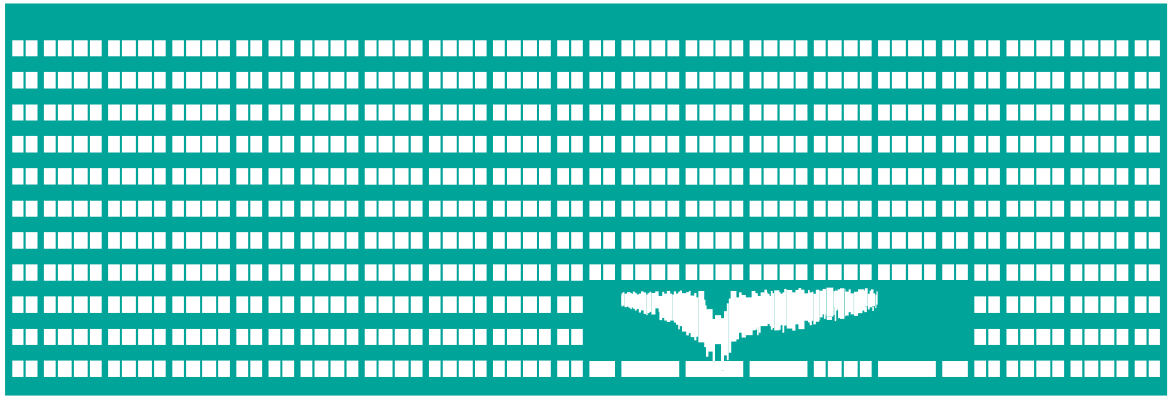
VŠB TECHNICKÁ UNIVERZITA OSTRAVA

VSB TECHNICAL UNIVERSITY OF OSTRAVA



www.vsb.cz

VŠB TECHNICKÁ UNIVERZITA OSTRAVA | FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Relief Feature Selection

Jakub Šenk

Vedoucí práce:

doc. Mgr. Miloš Kudělka, Ph.D.

Cíl práce

- Implementace algoritmů pro výběr atributů založených na reliéfu
- Webová aplikace pro provádění experimentů
- Samotné experimenty

Základní myšlenka

- Určení nejvýznamnějších atributů, na základě kterých je možné nejlépe určit typ objektu
- Porovnávání na základě nejbližších sousedů „nearest hit“ a „nearest miss“
- Přepočítání skóre pro každý atribut

	Atributy	Třída
Testovaná instance:	1 2 3 **4** 5 6	A
Nearest hit:	1 2 3 **6** 5 6	A

Skóre tohoto atributu se sníží.

	Atributy	Třída
Testovaná instance:	1 **2** 3 4 5 6	A
Nearest miss:	1 **5** 3 4 5 6	B

Skóre tohoto atributu se zvýší.

Základní verze reliéfu

- Neumí pracovat s chybějícími hodnotami
- Pouze 2 třídy objektů

$$Scores[j] = \sum_{j=0}^{a} -\frac{diff(A[j], i[j], h[j])}{n} + \frac{diff(A[j], i[j], m[j])}{n}$$

- Diff pro numerické hodnoty: $\frac{|i-j|}{max(a)-min(a)}$

Algorithm 2 Relief

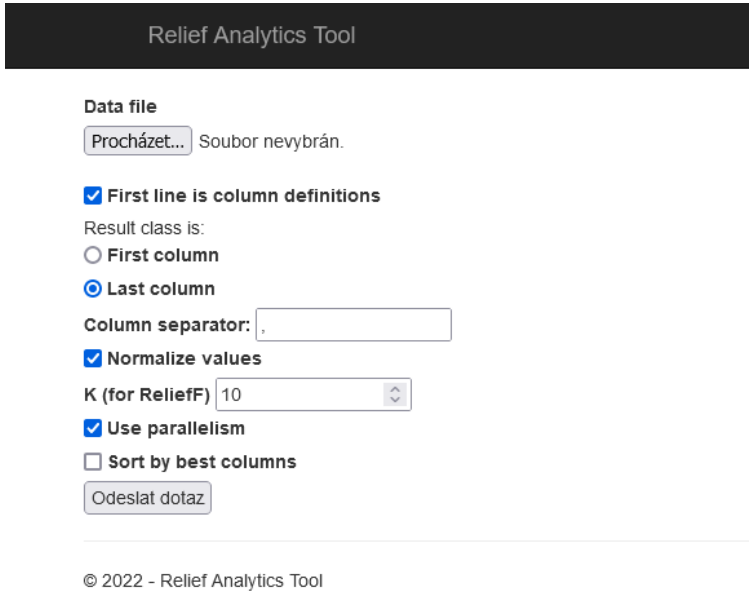
```
Scores ← List(a)
for i < n do
    ri ← data[random]
    hit ← NearestHit(i)
    miss ← NearestMiss(i)
    for j < a do
        Scores[j] ← Scores[j] − Diff(j, ri, hit) + Diff(j, ri, miss)
    end for
end for
for i < a do
    Scores[i] ← Scores[i]/n
end for

function DIFF(featureIndex, a, b)
    return Abs(a[featureIndex] − b[featureIndex])/Max(data[featureIndex]) −
Min(data[featureIndex])
end function

function NEARESTMISS(sampleIndex)
    shortestDistance ← MAX
    currentDistance ← NULL
    index ← NULL
    for i < n do
        if class(data[sampleIndex]) ≠ class(data[i]) then
            currentDistance ← 0
            for j < a do
                currentDistance ← currentDistance + data[sampleIndex][j] − data[i][j]
                if currentDistance² < shortestDistance² then
                    shortestDistance ← currentDistance
                    index ← i
                end if
            end for
        end if
    end for
    return data[index]
end function
function NEARESTHIT(sampleIndex)
    ...
    Pseudokód funkce NearestHit je téměř totožný, liší se pouze v
podmínce
    if class(data[sampleIndex]) = class(data[i])ANDsampleIndex! = i then
    end if
    ...
end function
```

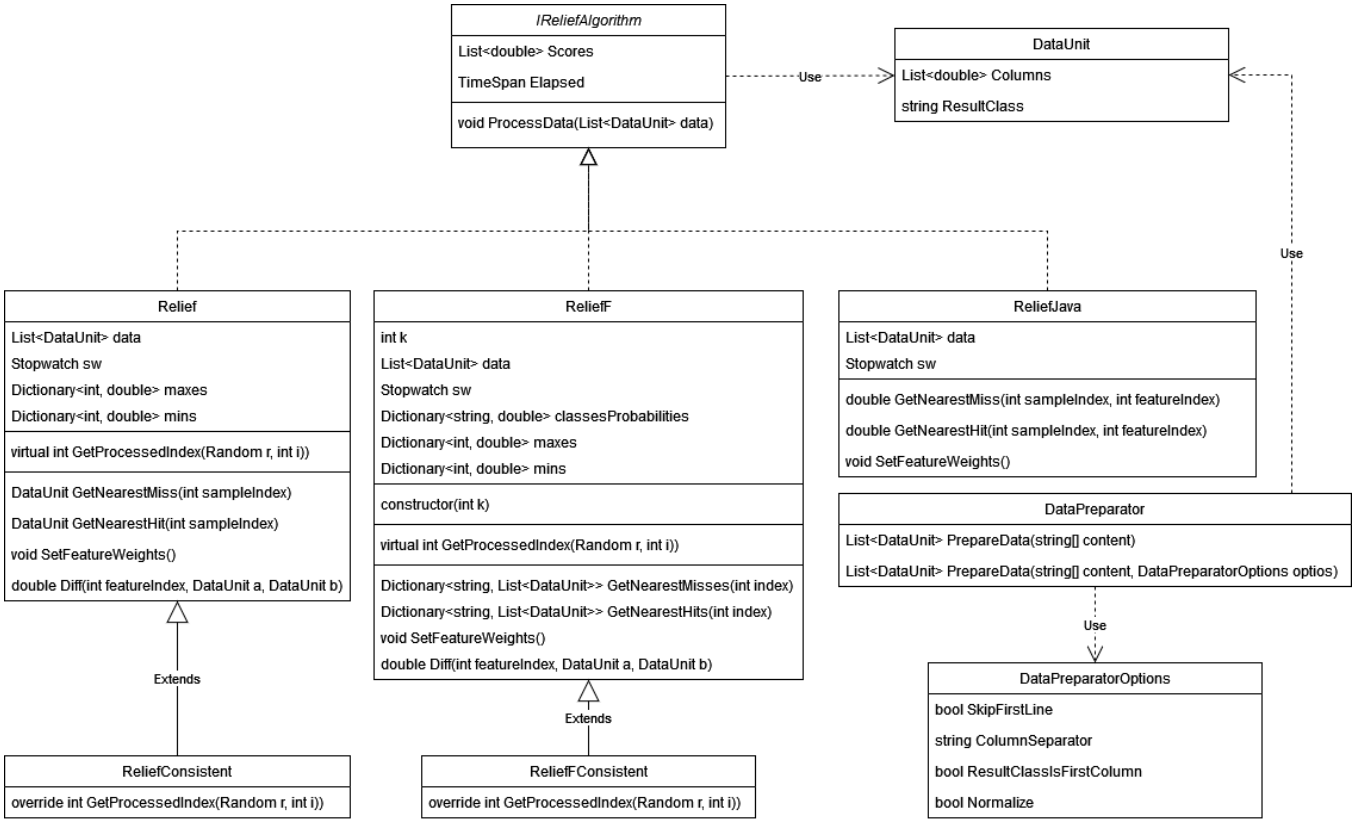
Experimentální aplikace

- Webová aplikace umožňující zpracování datasetů a určení nejvýznamnějších atributů
- Implementace 5 relief algoritmů
- Nezávislost jádra s algoritmy na GUI
- Přehledné setřídění všech atributů



Relief algoritmy

- ReliefJava
- Relief
- ReliefF
- ReliefConsistent
- ReliefFConsistent



Výsledek

Dataset WineQT.csv (https://www.kaggle.com/datasets/yasserh/wine-quality-dataset)

Algorithm	Best column	Calculation time
ReliefJava	total sulfur dioxide	00:00:02.6469415
Relief	alcohol	00:00:01.2319185
ReliefConsistent	alcohol	00:00:01.1816663
ReliefF	alcohol	00:00:02.1121193
ReliefFConsistent	alcohol	00:00:02.1089962

Column	ReliefJava	Relief	ReliefConsistent	ReliefF	ReliefFConsistent
fixed acidity	-0,00006	0,03171	0,03340	0,01002	0,00933
volatile acidity	-0,00012	0,04376	0,04256	0,01540	0,01794
citric acid	-0,00029	0,05353	0,05393	0,01211	0,01418
residual sugar	-0,00007	0,02069	0,01636	0,00228	0,00216
chlorides	-0,00071	0,01614	0,01643	0,00388	0,00346
free sulfur dioxide	-0,00008	0,03823	0,03802	0,00888	0,00953
total sulfur dioxide	**0,00029**	0,02868	0,03376	0,01829	0,01804
density	-0,00007	0,03059	0,03577	0,01505	0,01354
pH	-0,00013	0,03391	0,03142	0,00674	0,00449
sulphates	-0,00031	0,03166	0,03249	0,01557	0,01311
alcohol	-0,00015	**0,06915**	**0,07883**	**0,04735**	**0,04702**

Výsledek

Dataset iris.csv (https://www.kaggle.com/datasets/arshid/iris-flower-dataset)

Algorithm	Best column	Calculation time
ReliefJava	petal_width	00:00:00.0189065
Relief	sepal_width	00:00:00.0107623
ReliefConsistent	sepal_width	00:00:00.0088134
ReliefF	petal_width	00:00:00.0301261
ReliefFConsistent	petal_width	00:00:00.0275352

Column	ReliefJava	Relief	ReliefConsistent	ReliefF	ReliefFConsistent
sepal_length	0,00236	0,01407	0,02722	0,08583	0,10248
sepal_width	0,00023	**0,20972**	**0,23417**	0,23247	0,23124
petal_length	0,02974	0,15469	0,16847	0,34377	0,35432
petal_width	**0,04222**	0,15000	0,17056	**0,34779**	**0,35928**

Výsledek

Dataset Date_Fruit.csv (https://www.kaggle.com/datasets/muratkokludataset/date-fruit-datasets)

Algorithm	Best column	Calculation time
ReliefJava	SHAPEFACTOR_2	00:00:04.9395066
Relief	MeanRR	00:00:01.5451867
ReliefConsistent	PERIMETER	00:00:01.5349026
ReliefF	MeanRR	00:00:02.1362201
ReliefFConsistent	MeanRR	00:00:02.1398484

Column	ReliefJava	Relief	ReliefConsistent	ReliefF	ReliefFConsistent
AREA	0,00024	0,14610	0,15455	0,15912	0,16232
PERIMETER	0,00124	0,17974	**0,18761**	0,17594	0,18106
MAJOR_AXIS	0,00091	0,12966	0,13482	0,12349	0,12702
MINOR_AXIS	0,00003	0,09722	0,10091	0,12218	0,12351
ECCENTRICITY	0,00008	0,04662	0,03875	0,06678	0,06497
EQDIASQ	0,00028	0,11759	0,12289	0,12324	0,12584
SOLIDITY	-0,00044	0,03151	0,02782	0,02847	0,02789
CONVEX_AREA	0,00023	0,14900	0,15814	0,16007	0,16346
EXTENT	0,00019	0,02916	0,02844	0,02571	0,02737
ASPECT_RATIO	0,00000	0,00125	0,00012	-0,00068	-0,00077
ROUNDNESS	0,00019	0,03044	0,02866	0,03326	0,03356
COMPACTNESS	0,00004	0,02246	0,01786	0,02849	0,02757
SHAPEFACTOR_1	-0,00001	0,00190	0,00081	0,00001	-0,00005
SHAPEFACTOR_2	**0,00185**	0,09967	0,10029	0,09227	0,09405
SHAPEFACTOR_3	0,00004	0,03540	0,02874	0,04674	0,04530
SHAPEFACTOR_4	-0,00030	0,01902	0,01459	0,01512	0,01512
MeanRR	0,00102	**0,18253**	0,18439	**0,20425**	**0,20796**
MeanRG	0,00031	0,15174	0,15003	0,17092	0,17377
MeanRB	0,00002	0,13704	0,13780	0,14580	0,14858
StdDevRR	-0,00001	0,07480	0,07864	0,07952	0,07916
StdDevRG	0,00004	0,05603	0,06055	0,06029	0,05961

Výsledek

Dataset Pumpkin_Seeds.csv (https://www.kaggle.com/datasets/muratkokludataset/pumpkin-seeds-dataset)

Algorithm	Best column	Calculation time
ReliefJava	Area	00:00:14.5997739
Relief	Aspect_Ration	00:00:07.1607979
ReliefConsistent	Aspect_Ration	00:00:07.1095925
ReliefF	Aspect_Ration	00:00:12.5362310
ReliefFConsistent	Aspect_Ration	00:00:12.4418499

Column	ReliefJava	Relief	ReliefConsistent	ReliefF	ReliefFConsistent
Area	**0,00043**	0,00541	0,00482	0,00476	0,00498
Perimeter	0,00008	0,03917	0,03756	0,03743	0,03776
Major_Axis_Length	0,00013	0,08375	0,07933	0,08028	0,07990
Minor_Axis_Length	0,00027	0,08208	0,07839	0,08219	0,08040
Convex_Area	0,00042	0,00521	0,00468	0,00461	0,00485
Equiv_Diameter	0,00034	0,00548	0,00499	0,00484	0,00513
Eccentricity	0,00017	0,07423	0,07034	0,07453	0,07340
Solidity	-0,00002	0,00310	0,00293	0,00293	0,00292
Extent	0,00004	0,02797	0,02414	0,01910	0,02019
Roundness	0,00005	0,09615	0,09262	0,09410	0,09286
Aspect_Ration	0,00036	**0,12638**	**0,12065**	**0,12272**	**0,12103**
Compactness	0,00020	0,12338	0,11763	0,12106	0,11922

Výsledek

Dataset thrombin.data (https://pages.cs.wisc.edu/~dpage/kddcup2001)

Calculation time: 00:01:02.4826765
Algorithm: ReliefJava

Column	Score
16403	**0,27273**
86855	0,27273
37133	0,25455
10942	0,23636
38649	0,23636
38668	0,23636
59640	0,23636
62275	0,23636
79416	0,23636
79420	0,23636
9674	0,21818
10081	0,21818
20973	0,21818
25283	0,21818
35549	0,21818
38653	0,21818
38858	0,21818
38872	0,21818
38877	0,21818
55130	0,21818
62272	0,21818
79464	0,21818
79697	0,21818
86166	0,21818

Calculation time: 00:00:20.4549088
Algorithm: Relief

Column	Score
244	**0,45455**
2583	0,45455
104654	0,45455
42841	0,43636
66795	0,43636
41074	0,40000
41096	0,40000
102459	0,40000
40405	0,38182
40553	0,38182
40971	0,38182
41320	0,38182
49861	0,38182
133619	0,38182
40396	0,36364
40400	0,36364
40540	0,36364
41321	0,36364
73451	0,36364
74377	0,36364
74573	0,36364
8574	0,34545
38640	0,34545
40418	0,34545

Calculation time: 00:00:20.3786104
Algorithm: ReliefConsistent

Column	Score
244	**0,38182**
2583	0,38182
104654	0,38182
40405	0,36364
41074	0,36364
41096	0,36364
74573	0,36364
102459	0,36364
46836	0,34545
49861	0,34545
86855	0,34545
133619	0,34545
40400	0,32727
40540	0,32727
41037	0,32727
40971	0,30909
42841	0,30909
61398	0,30909
5449	0,29091
6483	0,29091
23951	0,29091
29337	0,29091
37133	0,29091
38657	0,29091

Calculation time: 00:00:33.9523228
Algorithm: ReliefF

Column	Score
86855	**0,50000**
38668	0,47818
104654	0,47455
79416	0,44545
79662	0,43455
10694	0,42909
13509	0,42727
10942	0,42545
38649	0,41636
16409	0,41091
16886	0,41091
79420	0,40909
16403	0,40909
73693	0,39636
40405	0,39091
16365	0,38727
28558	0,38545
59640	0,38545
38877	0,38182
13784	0,38182
25194	0,38000
9680	0,38000
9250	0,37818
10799	0,37455

Calculation time: 00:00:34.2628136
Algorithm: ReliefFConsistent

Column	Score
86855	**0,45455**
16403	0,41273
104654	0,40182
79662	0,38727
38668	0,38000
10694	0,37091
10799	0,36727
38543	0,36364
16886	0,36000
16664	0,35636
38649	0,35455
13509	0,35455
10942	0,34909
10909	0,34545
16409	0,34545
13618	0,34182
79416	0,34000
25194	0,34000
79420	0,33273
28558	0,32909
9250	0,31818
37133	0,31636
13784	0,31455
16365	0,30545

Výsledek

Dataset car.csv

Algorithm	Best column	Calculation time
ReliefJava	Capacity	00:00:00.0001742
Relief	Capacity	00:00:00.0001348
ReliefConsistent	Capacity	00:00:00.0001349
ReliefF	Capacity	00:00:00.0002995
ReliefFConsistent	Capacity	00:00:00.0002627

Column	ReliefJava	Relief	ReliefConsistent	ReliefF	ReliefFConsistent
ID	0,02160	0,25735	0,18382	0,06912	0,02794
Capacity	**0,33832**	**0,58047**	**0,56049**	**0,56160**	**0,54562**
Doors	0,00000	0,29412	0,29412	0,03529	0,00000

Sumarizace

- Pochopení jak fungují algoritmy založené na reliéfu
- Implementovány 3 verze relief algoritmů + 2 konzistentní odvozeniny
- Implementována experimentální aplikace
- Výsledky ověřeny s výsledky programu WEKA