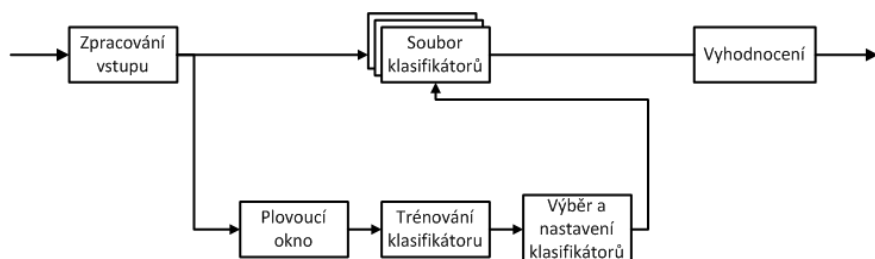


Klasifikace v datových proudech umožňuje klasifikovat nepřetržitě přicházející data v reálném čase. Data nemusí být nijak omezena, proto se musíme vypořádat s jejich objemem a dynamickými změnami.

Problémy zpracování datových proudů

- Nekonečnost
- Změna konceptu dat
- Evoluce konceptu dat
- Zapomínání

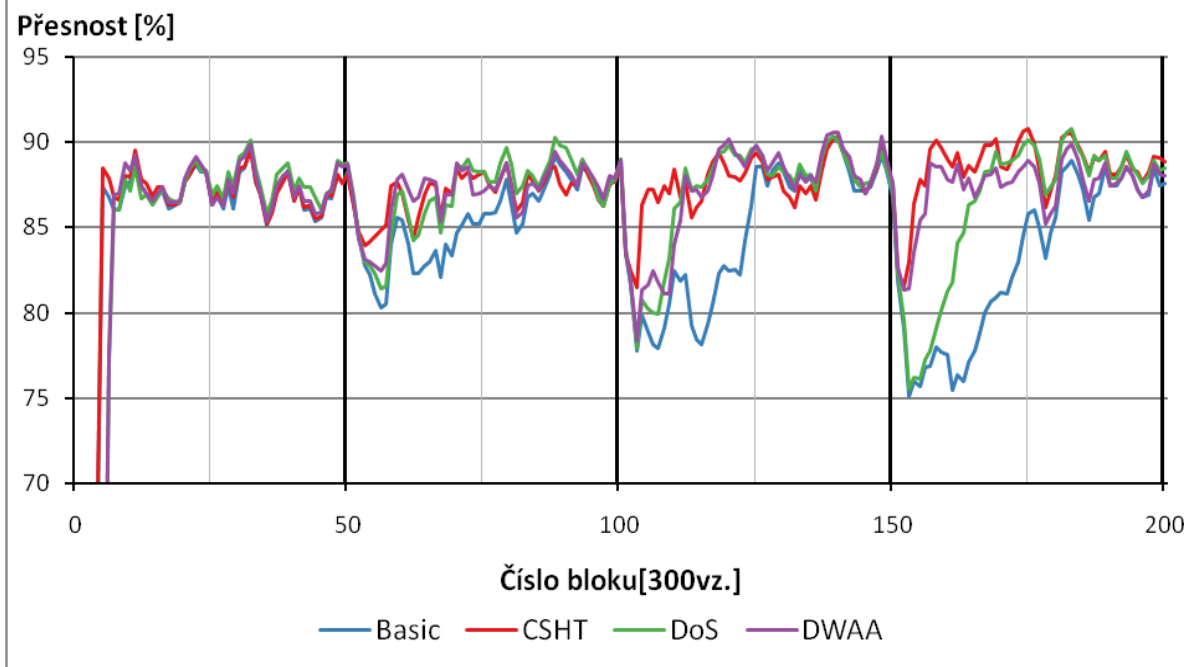
Proudový klasifikační systém



- Systém nepřetržitě přijímá data a **klasifikuje** je všemi klasifikátory v souboru. Výstup je pak získán váženým hlasováním použitých klasifikátorů.
- **Trénovací část** nasbírání dostatečné množství trénovacích dat, natrénuje nový klasifikátor a aktualizuje soubor.
- Systém je **paralelizován** rozdělením na úlohy pomocí *Task Paraller Library* obsažené v *.NET Frameworku 4*. Zpracování je pak škálováno dle dostupných výpočetních zdrojů.

Změna konceptu dat

Při změně konceptu dat dochází k propadům přesnosti klasifikace. Snažíme se minimalizovat množství potřebných trénovacích dat pro přizpůsobení. Byly implementovány a testovány 3 metody, jak se vyrovnat se změnou konceptu dat. Každá metoda reprezentuje jiný přístup.



Basic

- Referenční metoda bez úprav
- 25xNaive Bayes

CSHT

- Detekuje změnu konceptu testem hypotézy
- 25xNaive Bayes + 50 odložených

DoS

- Kombinuje SVM, C4.5 a Naive Bayes
- 25xklasifikátor + 51 odložených

DWAA

- Používá odměňovací strategii
- 25xNaive Bayes