Analýza dát o účastníkoch matematických súťaží

Jaroslav Pastorek

Viliam Sabol

Úvod

- Údaje o účastníkoch matematických súťaží a ich dosiahnutých výsledkoch
- Súťaže každoročne organizuje o.z. Strom pre žiakov základných a stredných škôl z celého Slovenska
- Pokúsime sa klasifikovať súťažiaceho do kategórií podľa dosiahnutého počtu bodov na základe poskytnutých atribútov

Organizácia súťaže

- Súťaž je rozdelená do troch kategórií:
  - Malynár – pre žiakov 4. - 6. ročníka ZŠ
  - Matik – pre žiakov 7. – 9. ročníka ZŠ
  - Strom – pre žiakov SŠ
- Každý súťažný rok je rozdelený na dva semestre a tie zas na niekoľko sérií príkladov
- Do dnešných dní sa z každej súťaže uskutočnilo asi 5 ročníkov, za ktoré sa zozbierali analyzované dáta

Dáta

- Získané z PostgreSQL databázy webovej aplikácie súťaže
- Select sme zostavili sami, avšak v danom čase ešte nebolo úplne jasné, ktoré dáta budú relevantné, neskôr už DB nebola dostupná
- Zhruba 20 000 riadkov
  - Každý riadok reprezentuje ohodnotenie riešenia jedného príkladu jedného účastníka
  - Atribúty:
    - ID účastníka, pohlavie, trieda(ročník), škola, mesto, číslo príkladu, semester, séria, dosiahnutý počet bodov

Úprava dát

- Dáta zagregujeme – vypočítame výsledný počet bodov daného účastníka za celý rok
- Problémy:
  - Nie všetci sa zúčastnili všetkých sérií
    - Riešenie – počítame priemer cez série v danom roku
  - Podľa pravidiel súťaže sa výsledný počet bodov upravuje podľa ročníka, ktorý účastník navštevuje
    - Napríklad najmladším sa najlepšie vyriešený príklad započíta dvakrát – tým by sa mali zotrieť rozdiely
    - Riešenie: počítame celkový počet bodov podľa pravidiel

Úprava dát II

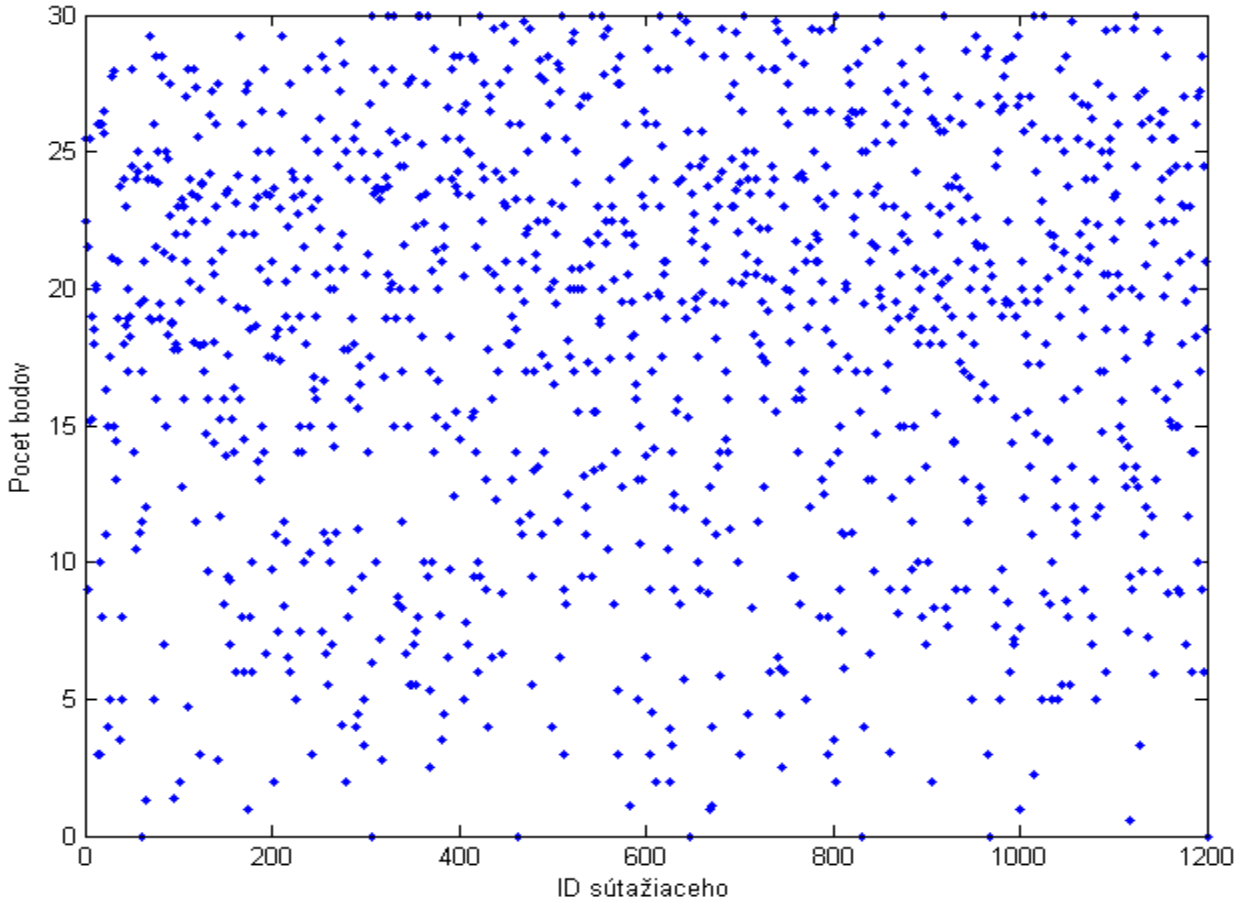
- Minulý rok sa zmenilo bodovanie – namiesto piatich bodov je možné za príklad dosiahnuť maximálne 9 bodov
  - Riešenie: Normalizujeme dáta z nového širšieho intervalu do pôvodného užšieho
- Vytvorenie nových atribútov
  - Zaviedli sme atribút „skúsenosť“, ktorý hovorí, koľkých predchádzajúcich ročníkov súťaže sa daný riešiteľ zúčastnil
  - Pre experimentovanie sme doplnili „kraj“, z ktorého účastník pochádza

Analýza

- Pre klasifikáciu účastníka súťaže sme sa rozhodli implementovať rozhodovací strom
  - Dôvody:
    - Jednoduché a efektívne klasifikačná metóda
    - Algoritmus nás vyslovene zaujal :-))
- Pre implementáciu sme sa rozhodli použiť platformu .NET a všetky potrebné programy sme písali v jazyku C#
- Na základe dostupných dát sme sa rozhodli vytvoriť samostatný rozhodovací strom pre každú zo súťaží
- Pre nedostatok miesta budeme prezentovať len výsledky zo súťaže Matik

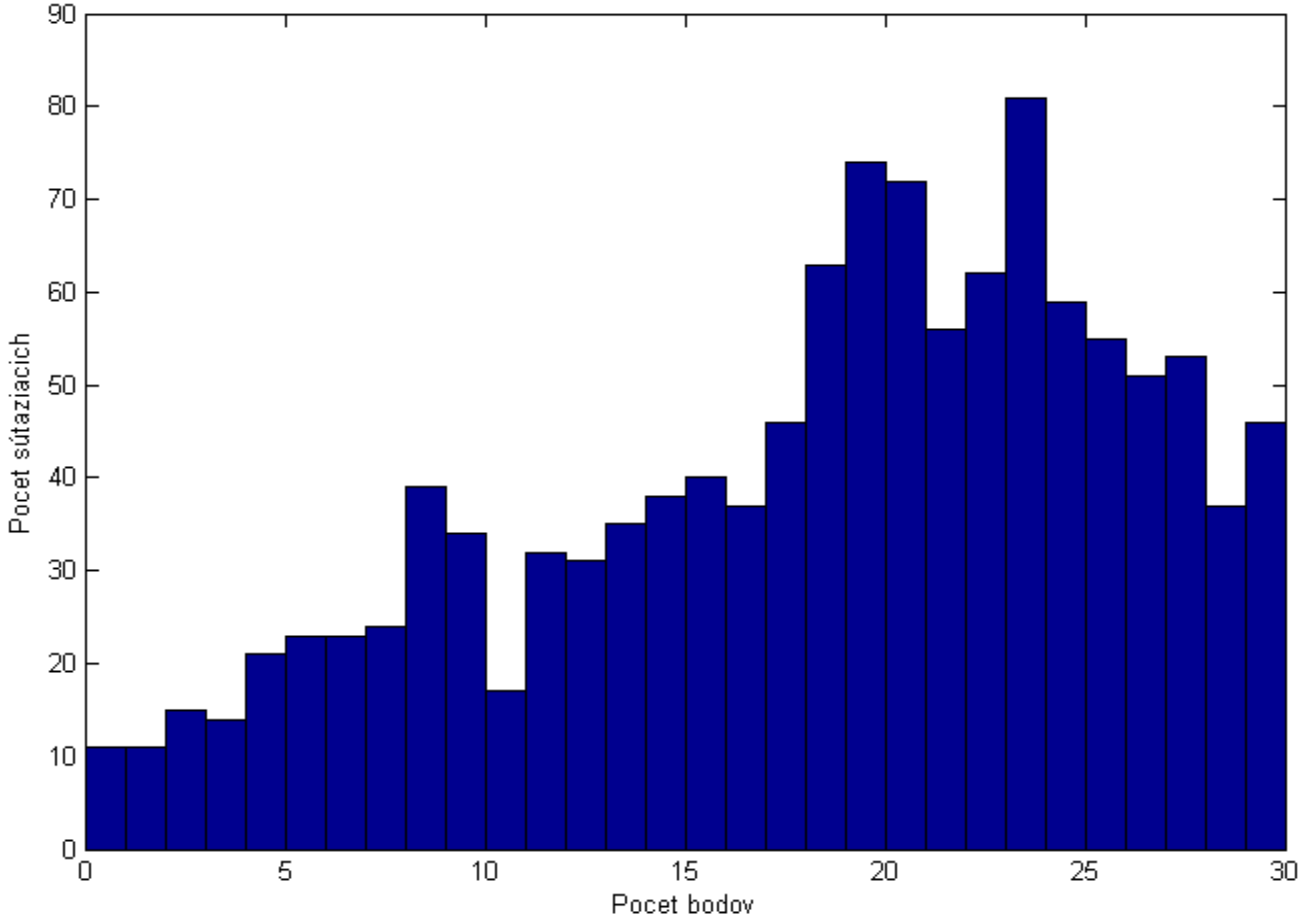
Prvý pohľad na dáta

- Rozloženie dosiahnutého počtu bodov



Druhý pohľad na dáta

- Histogram dosiahnutého počtu bodov



Klasifikačné skupiny

- Predošlé obrázky nás doviedli k presvedčeniu, že hranice, ktoré budú rozdeľovať klasifikačné skupiny, budú situované bližšie k hornej hranici intervalu
- Túto teóriu nakoniec potvrdili aj experimenty so samotným stromom
- Klasifikačné triedy:
  - 28 – 30 bodov
  - 24 – 27 bodov
  - 0 – 23 bodov

Implementácia

- Vytvorili sme aplikáciu, ktorá:
  - Načíta agregované dáta, rozdelí ich v pomere 2:1 na trénovacie a testovacie
  - Vytvorí rozhodovací strom metódou TDIDT, pričom deliaci atribút sa určuje podľa najmenšej entropie
  - Predloží stromu testovacie dáta a vyhodnotí výslednú klasifikáciu
- Implementácia je ľahko konfigurovateľná a veľmi univerzálna
  - Je možné meniť pohodlne klasifikačné kategórie, vstupné a cieľové atribúty...
  - V podstate je možné predložiť ľubovoľnú tabuľku vo forme CSV súboru

Vyhodnotenie

- Pre súťaž Matik vytvorený rozhodovací strom dosahuje na testovacej množine úspešnosť 75%
  - Trochu sme experimentovali a najlepšie výsledky sme dosiahli pri použití atribútov „pohlavie“, „kraj“, „ročník“ a „skúsenosť“
- Atribút, ktorý delí dáta najviac, je „ročník“
  - Mohlo by to naznačovať, že výpočet bodov nie je optimálny a nedostatočne zvýhodňuje mladších riešiteľov, prípadne na riešenie daných úloh je nutné mať znalosti získané vo vyšších ročníkoch

Výsledný strom