IMPLEMENTATION OF GENETIC ALGORITHM FOR DATA MINING SYSTEM

Rostislav Stríž
Bachelor Degree Programme (3), FIT BUT
E-mail: xstriz03@stud.fit.vutbr.cz

Supervised by: Michal Šebek
E-mail: isebek@fit.vutbr.cz

ABSTRACT

Data collecting plays important role in many aspects of today's businesses. Via process called Knowledge Discovery in Databases we are able to extract hidden usable information from stored data. This paper is presenting possible use of Genetic Algorithms in complex Data Mining System.

1 ÚVOD

Vzhledem k současné rozšířenosti informačních technologií se většina získaných dat ukládá v elektronické podobě do souborů, databází či datových skladů. S narůstající kvantitou takovýchto dat narážíme na problém, jak se v nich smysluplně orientovat, případně jaké zajímavé informace z nich můžeme vyzískat. Tyto úvahy stály za vznikem procesu **získávání znalostí z databází** (někdy zkráceně označovaném jako *dolování z dat*). Cílem procesu získávání znalostí z databází je nalézt *netriviální* a potencionálně *zajímavé* informace primárně ve velkých objemech uložených dat. Pojmem netriviální označujeme informace, které nejsou přímo uloženy v datových záznamech a které nejsme schopni získat pomocí jednoduchého dotazu (např. SQL). Co se týče zajímavosti vydolovaných znalostí, požadujeme uživatelskou přínosnost, ideálně sloužící ke komerečnímu využití.
Samotný proces získávání znalostí z databází je poměrně komplikovaný, skládá se z několika částí – *předzpracování dat* (čištění, integrace, výběr a transformace dat), jehož cílem je připravit surová data pro dolování, následuje samotné *dolování z dat*, kdy se pomocí specifických algoritmů provádí operace nad daty za vzniku modelů nebo vzorů a na závěr proběhne *vyhodnocení* a *prezentace výsledků* – veškeré detaily lze nalézt v [1]. Pro jeho praktickou prezentaci studentům FIT VUT v Brně byl zahájen vývoj softwarového nástroje a právě návrh implementace genetického algoritmu pro dolování do tohoto systému bude náplní následujících odstavců.

2 GENETICKÉ ALGORITMY

Genetický algoritmus je stochastická optimalizační metoda založená na principech evoluční biologie. Jednotlivé datové záznamy jsou chápany jako jedinci s vlastním genetickým kódem – genomem. **Genom** (jindy *chromozom*) je souborem **genů**, které popisují jednotlivé vlastnosti jedince, **populací** potom označujeme skupinu takovýchto jedinců. Stěžejní částí algoritmu je tzv.

fitness funkce, jejímž úkolem je hodnotit jednotlivé genomy. Na základě ohodnocení populace poté pokračuje její evoluce – silnější jedinci přežívají (genomy s vysokou fitness hodnotou), slabší z populace mizí.
Na počátku tedy vhodně zakódujeme náhodné hodnoty genů (z jejich domény) do čitelných genomů, ze kterých vytvoříme *inicializační populaci*. Jednotlivé genomy ohodnotíme fitness funkcí a vybereme skupinu nejlepších, nad kterou provedeme množinu operací popsaných v následující sekci 2.1. Pomocí fitness funkce zhodnotíme výsledky evoluce a z nejsilnějších jedinců vytvoříme novou populaci, následně cyklus opakujeme. Ukončení algoritmu je individuální pro jednotlivé implementace (např. maximální počet populací, minimální hodnota „vylepšení“ populace od jejího předchůdce, . . .).

2.1 METODY VÝVOJE POPULACE

Evoluce probíhá pomocí speciálních operací prováděných nad populací.

- **Výběr** (*selekce*, *reprodukce*) – vybereme již existujícího jedince a v nezměněné formě jej přesuneme do nové populace.
- **Křížení** (z angl. *cross-over*) – probíhá mezi nejméně dvěma jedinci, podle binární masky vyměníme jednotlivé geny mezi chromozomy navzájem (nejčastěji se používá půlení).
- **Mutace** – náhodný gen změníme na libovolnou hodnotu z jeho domény, lze nastavit i na prázdnou hodnotu (NULL), čímž lze zjistit důležitost konkrétního genu.

Pro obecné genetické algoritmy platí, že je nutné nastavit správné pravděpodobnosti mutace a křížení, které mají zásadní vliv na úspěšnost a časový průběh algoritmu.

2.2 MOŽNÉ PROBLÉMY

Obecně vychází úspěšnost algoritmu z dobrého návrhu fitness funkce a z kvalitního zakódování, které ma velký vliv především na výkon celého procesu. Z popisu algoritmu lze logicky odvodit, že při práci s velkým objemem dat nedosahuje příliš velké výpočetní rychlosti, vzhledem k poměrně složitým manipulacím s jednotlivými datovými záznamy. Dnešní doba tak přeje rozvoji metod pro *paralelní zpracování* (angl. parallel processing) a *vzorkování* (angl. sampling).

3 GENETICKÉ ALGORITMY PŘI DOLOVÁNÍ Z DAT

Genetický algoritmus lze použít i při procesu získávání znalostí z databází – spadá do kategorie *klasifikačních* algoritmů. Na základě klasifikační dolovací úlohy jsme schopni vytvořit *model*, podle něhož lze následně zařadit nově příchozích jedince.

KLASIFIKAČNÍ ALGORITMY

Cílem klasifikace je rozdělit data na **třídy** a vytvořit jejich popis, který bude použit ve výše zmíněném modelu. Na počátku dodáme algoritmu data již správně roztříděná a započne fáze **trénování** – na základě dodané *trénovací množiny dat* vytvoříme klasifikační model. Důležitá je fáze **testování**, kdy vzniklý model testujeme na datech, u nichž stále známe příslušnou třídu

a tím určujeme správnost aktuální verze modelu. V případě úspěšného ukončení testování můžeme přistoupit k **aplikaci** modelu na data neznámé třídy, což při správnosti modelu povede k jejich zařazení. U genetických algoritmů se tento koncept může změnit, jelikož vycházíme s množiny náhodných prvků, které porovnáváme s množinou známých, rozdělení na trénovací a testovací data je tak irelevantní.

4 NÁVRH POUŽITÍ GENETICKÉHO ALGORITMU V DOLOVACÍM PROCESU

Naším cílem bude vytvořit samostatný modul, který bude pomocí genetického algoritmu schopen vytvářet klasifikační model pro zadaná data. Přes jádro systému bude komunikovat s databázovým serverem Oracle a podle zadaných parametrů bude provádět dolovací úlohu, o jejímž výsledku bude uživatele informovat výsledným *reportem*. V následujících odstavcích budou nastíněny hlavní kroky návrhu daného modulu – vychází z [2].

- **Reprezentace genomu** – každý datový záznam bude reprezentován víceprvkovou relací hodnot jednotlivých atributů. Všechny tyto relace jsou podmnožinou kartézského součinu kompletních domén všech atributů, navíc bude každý genom označen unikátním identifikátorem. Spojité atributy budou diskretizovány a binárně zakódovány.

- **Formát popisu záznamů** – je nutné definovat způsob pro popis, resp. výběr jednotlivých záznamů. V našem případě se budeme snažit o matematický popis asociace – **konjunkce predikátů**, pomocí kterých dokážeme popsat určitou skupinu záznamů se stejnými vlastnostmi. Cílem algoritmu je popsat všechny třídy právě těmito pravidly – hledáme užitečný a dříve neznámý výraz pro popis třídy pomocí co nejmenšího počtu operací. Na základě skupin neužitečných výrazů (konjunkce stejných atributů,. . . , další v [2]) jsme schopni vytvořit základní pravidla pro relevantní výrazy – nebudeme používat disjunkci, výraz musí obsahovat min. jednu konjunkci a každý atribut ve výrazu bude unikátní. Tímto urychlíme práci algoritmu a ušetříme ho procházení nezajímavými prostory.

- **Fitness funkce** – fitness funkce bude navržena tak, aby zohledňovala poměr správně a špatně ohodnocených jedinců populace za předpokladu, že by testovaný jedinec byl výrazem pro jejich popis.

5 ZÁVĚR

Článek představuje základní informace o genetických algoritmech a o jejich možném použití pro dolování z dat. Dále pak seznamuje čtenáře se základem návrhu implementace takového algoritmu pro dolovací systém.

REFERENCE

[1] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Elsevier Inc., second edition, 2006, 770s., ISBN 978-1-55860-3

[2] Ghosh, A., Tsutsui, S.: Advances in evolutionary computing: theory and applications, Springer-Verlag New York, Inc., 2003, 1004 s., ISBN 3-540-43330-9