Oponentní posudek diplomové práce

Vypracoval: David Hauzar
Název práce: **A study of applying copulas in data mining**
Autor diplomové práce: Martin Ščavnický

Cílem diplomové práce bylo prozkoumat vhodnost jednotlivých typů kopulí v dobývání znalostí z dat. Práce popisuje základy teorie kopulí, použité rodiny kopulí, způsob prokládání kopulí daty a metriky kvality proložení. Práce shrnuje výsledky předchozího výzkumu vhodnosti kopulí pro modelování dat. Dále je v ní navržena modifikace algoritmu pro prokládání HAC kopulí daty tak, aby byly i tyto kopule schopné zachytit negativní závislost mezi atributy a porovnává kvalitu proložení pomocí původního a modifikovaného algoritmu. Dále práce porovnává kvalitu proložení u osmi rodin kopulí a čtyřech klasifikačních sadách dat a porovnává přesnost klasifikátoru založeném na těchto kopulích. Práce také popisuje a demonstruje způsob, jak pomocí hiearchických kopulí vizualizovat závislosti mezi daty. Součástí práce je implementace podpory pro d-dimenzionální archimédovské kopule a HAC kopule.

Práce v mnohém navazuje na studii o vhodnosti kopulí publikovanou Matthiasem Fischerem v roce 2007. Rozdíl obou prací je jednak v použití různých dat, kterými byly kopule prokládány. Fischerova studie zkoumala kvalitu proložení kopulí daty z finančnictví, tato práce používá data z dobývání znalostí z dat. V hodnocené práci bylo vyzkoušeno více způsobů prokládání kopulí daty a také více metrik kvality proložení kopulí daty. Na rozdíl od předchozí studie nebyla kvalita prokládání kopulí daty měřena pouze “umělými” metrikami, ale také výkonem proložených kopulí v reálné aplikaci - klasifikaci dat. Hodnocená práce potvrdila dobrý výkon eliptických kopulí z předchozí studie, na rozdíl od předchozí studie ukázala dobrý výkon také HAC kopulí.

Práce ukázala, že kopule je možné využít ke zlepšení klasifikace. Otázka, jak postupovat ve výběru kopulí zůstává ale víceméně otevřená.

Práce je psána v anglickém jazyce. Jazyková úroveň by mohla být lepší, text je místy hůře čitelný, podstatně to ale nesnižuje srozumitelnost textu. Rozsah implementace je spíše menší (necelé 3000 řádků kódu), vzhledem k povaze práce to ale nepokládám za nedostatek. Zdrojové kódy jsou dobře komentované a implementace je celkově na dobré úrovni.

Hodnocená práce má i několik nepříliš závažných nedostatků. Uvádím jejich výčet:

- Kopule nejsou porovnány z hlediska výkonu. Ve výsledcích chybí srovnání rychlosti proložení kopulí daty a srovnání rychlosti klasifikace za použití daných kopulí.
- Matematické vztahy by mohly být lépe popsány. Například na straně 34 chybí u popisu derivací informace, že parametr θ je parametr generátoru kopule. Příklad výpočtu derivace HAC kopule na straně 36 není dostatečně popsán apod.
- Pro snažší orientaci bych uvítal zahrnutí nezávislé kopule do tabulek s výsledky kvality proložení kopulí daty.

- Ačkoliv je struktura práce celkově dobrá, části o výpočtu distribuční funkce a hustoty pravděpodobnosti bych zařadil spíše do sekce Proposed methods and approaches než do sekce Implementation.

Doporučuji, aby práce byla přijata jako diplomová a připuštěna k obhajobě.

Otázky:
1) Jaký byste navrhl postup výběru vhodného typu kopulí v případě, že byste chtěl využít kopule v klasifikační úloze? Kdy byste použil hieararchické kopule spíše než eliptické kopule?

V Praze 7. 5. 2013

David Hauzar
Katedra distribuovaných a spolehlivých systémů MFF UK