

Review av:

Efficient K-means clustering and the importance of seeding

Authors: Philip Eliasson, Niklas Rosén

Reviewer: Tobias Reinhammar

Frågor från oppositionsprotokollet

Var det lätt att förstå vad kexjobbet gick ut?

Ja, kexjobbets mål var lätt att förstå och ta till sig. Det förklaras både i referatet och i problem statement på ett utförligt vis.

Hur tycker du titeln avspeglar rapportens innehåll?

Titeln avspeglar innehållet mycket bra. Rapporten utvärderar en K-means algoritm och vilka algoritmer den kan förbättras med. Rapporten tar också upp hur seeding kan förbättra ett resultat dramatiskt vilket jag tycker titeln förmedlar på ett bra sätt.

Hur beskrev författaren bakgrunden till kexjobbet? Finns det en introduktion och översikt av området?

Ja det finns, den ger en generell introduktion till klustring och några av dess användningsområden.

Hur väl har författarna motiverat sitt val av metod att angripa problemet?

Metoden verkar vara väldigt logiskt upplagd och testerna är motiverade av ett behov av att utforska alla aspekter. Det enda jag saknar är kanske vad de använt för att köra sina tester.

Diskuterar författarna huruvida de förutsättningar som gäller för att metoden ska kunna användas är uppfyllda?

Ja. Eftersom författarna diskuterar när olika klustringsmetoder är lämpliga och när de presterar dåligt. De nämner också att datan är anpassad för k-means klustring eftersom de inte är intresserade av att undersöka värsta fallen.

Är metoden väl beskriven?

Metoden är mycket väl beskriven så när som på att den inte nämner något om vilken mjukvara som används för att utföra experimenten. I övrigt definieras metoder och parametrar ordentligt.

Har författarna presenterat sina resultat tydligt?

Ja resultaten är mycket tydligt presenterade med grafer och förklarande text samt en analys i samband med att resultaten presenteras.

Finner du författarnas slutsatser trovärdiga?

Resultaten ter sig vara tämligen ensidiga och jag kan inte annat än att hålla med författarna i deras slutsatser.

Vad tycker du om litteraturlistan? Vilken typ av litteratur finns med? Förefaller litteraturen relevant?

Jag tycker litteraturen verkar relevant för ämnet och den består av en handfull olika vetenskapliga rapporter. Jag saknar dock länkar till rapporterna och vid ett tillfälle verkar referensen lite överflödig.

Vilka avsnitt i rapporten var svåra att förstå?

2.2 Kd-tree var det avsnittet som var svårast att förstå, kanske går det att förtydliga på något vis med ytterligare en illustration eller likanande? Det avslutande stycket i 2.3 Weakness.... var också svårt att förstå.

Övriga kommentarer om rapporten och dess struktur

Rapporten är mycket bra skriven som helhet, det finns små språkfel, mestadels bara något s som saknas eller att en mening låter direkt översatt från svenska. Man kanske också borde tänka på att hålla sig till ett och samma tempus överallt i rapporten.

Vilka är rapportens starka sidor?

Rapporten presenterar resultaten på ett mycket tydligt sätt och metoden verkar mycket vetenskapligt upplagd. Rapporten är inte onödigt lång

Vilka är rapportens svaga sidor?

Rapporten kanske skulle dra fördel av en sektion för terminologi för att göra det lättare att förstå vissa delar.

Hur bedömer du arbetets nyhetsvärde?

Jag har svårt att bedöma nyhetsvärdet då jag inte är insatt i vad det forskas på inom området. De rapporter som refereras i rapporten har åtminstone ett par år på nacken så det kan tänkas .

Sammanfatta arbetet på ett par rader?

Arbetet gick ut på att jämföra hur bra 2 olika k-means algoritmer presterade med lite varierande parametrar. De utgick från Lloyd's algoritm och jämförde den mot en algoritm som utvecklats genom att använda sig av KD-tree. Vidare undersöker dem också hur stor påverkan valet av startvärden hade på resultaten. De kom fram till att man kan förbättra Lloyd's algoritm med bland annat KD-tree, men att det inte alltid ger bättre resultat för vissa typer av data. De kunde också se att man genom en grundligt val av startvärden kunde förbättra körtiden hos Lloyd's algoritm.

Frågor till författarna:

1. Vad använde ni för mjukvara i undersökningen?
2. Om ni skulle forska mer inom området, vad skulle ni då forska kring?
3. Random seeding, det står att man brukar göra det flera gånger, hur väljs den bästa?
4. Hur stor påverkan kan det ha att man inte inkluderar tids åtgången för partiell klustring i seeding undersökningen?

Vilket är ditt sammanfattande omdöme om rapporten?

Jag tycker det är en mycket bra rapport som helhet. Det finns ett par språkmässiga saker som man kan diskutera och snygga till och eventuellt någon liten extra förklaring att lägga till. På det stora hela tycker jag att rapporten tar upp det som är relevant, genomför en grundlig studie och presenterar resultat samt slutsatser på ett bra sätt.