LUNDS UNIVERSITET

Hur maskininlärning kan ge oss ett effektivare moln

Albin Heimerson
Institutionen för Reglerteknik

Populärvetenskaplig sammanfattning av doktorsavhandlingen *Learning to Control the Cloud*, november 2023. Avhandlingen kan laddas ner från:
http://www.control.lth.se/publications

I dagens samhälle är det många viktiga funktioner som är beroende av att internet fungerar. Det kan vara allt från att kunna se på en film, till att kunna betala räkningar, eller arbeta. Om vi tar att strömma film som exempel, så finns det en dator någonstans som har filmen lagrad, och den skickar en liten del av filmen åt gången till din dator som då kan visa filmen. Om det är många som vill se filmen samtidigt så kan det bli för mycket för den datorn som har filmen lagrad, vilket resulterar i problem att visa filmen. Då behövs kanske två datorer som har filmen lagrad, och förfrågningarna måste fördelas mellan båda så att ingen av dem blir överbelastad. Detta kallas lastbalansering, och är en av många viktiga funktioner som är grunden till vad vi kallar *molnet*.

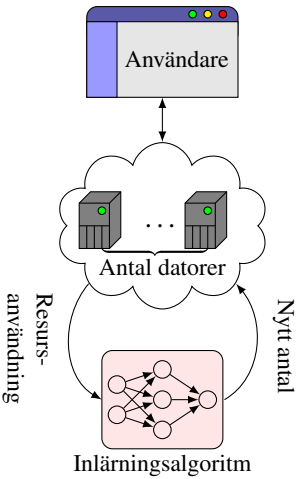
Molnet kan enkelt beskrivas som en stor samling datorer, där delar av dessa datorer kan hyras ut, snabbt och smidigt, till den som behöver. Folk som tillhandahåller tjänster på internet kan då hyra in sig på molnet istället för att köpa in och hantera egna datorer. Utöver att det kan vara både dyrt och krångligt att ha sin egen hårdvara så tillhandahåller molnet också en flexibilitet, där det är enkelt att hyra fler datorer och replikera sin tjänst på dem, eller flytta tjänsten till en ny världsdel om så skulle behövas. Enkelheten och flexibiliteten som molnet tillhandahåller är en drivande faktor till att internet har blivit så stort som det är idag.

Att ha så många tjänster som körs på molnet gör också att det finns mycket potential i att försöka effektivisera molnet. Att lastbalansera genom att skicka varannat besök till en dator och varannat till en annan är en enkel strategi, men kanske inte alltid den bästa. Egentligen skulle man vilja kolla på många olika faktorer, som hur mycket varje dator har att göra, kanske hur nära datorn är till den som vill se filmen, och sedan besluta vilken dator som är bäst lämpad utifrån en kombination av alla dessa faktorer. Att ta hänsyn till många små faktorer gör att vi har möjlighet hitta mer optimala sätt att styra molnet, men fler faktorer gör det också svårare att både definera vad som är en bra strategi och att hitta den bästa strategin.

Det är här maskininlärning kommer in i bilden, en teknik för att lära datorer att göra saker bättre genom att låta dem lära sig själva. Vi tittar på en speciell typ av maskininlärning som kallas för förstärkningsinlärning, där vi låter datorn prova olika saker och sedan ger den en poäng för hur bra det gick. Datorns uppgift är då att hitta den strategi som ger den högsta poängen. Detta är en väldigt generell metod,

och kan användas för att lära datorn att göra nästan vad som helst. Det är också en komplex metod som ofta kan komma fram till väldigt dåliga strategier, och det är inte alltid lätt att förstå varför datorn kom fram till en specific strategi.

Vårt mål har varit att undersöka hur vi kan använda förstärkningsinlärning för att styra olika delar av infrastrukturen i molnet på ett effektivare sätt. Ett populärt sätt att bygga applikationer i molnet är genom att dela upp funktionaliteten i små delar som körs på olika datorer, så kallade mikrotjänster. Som exempel kan det finnas en mikrotjänst som genererar hemsidan som visas, en som hanterar inloggning, en som hanterar rekommendation av filmer, och flera andra som hanterar strömning av filmerna. När en användare loggar in på sidan så behövs vissa av mikrotjänsterna, och senare när en film spelas upp behövs några andra. Att lista hur hur många datorer som behövs för varje mikrotjänst vid varje tidpunkt är svårt, och ofta används fler datorer än vad som egentligen behövs för att vara på den säkra sidan. Vi undersöker om förstärkningsinlärning kan lära sig att hitta mönster i användingen, så att när antalet användare som loggar in ökar så förstår den att snart kommer antalet som tittar på film också öka, och då kan den proaktivt skala upp antalet datorer innan vi får problem. Detta kan leda till att vi kan använda färre datorer, och därmed spara energi, utan att det påverkar användarupplevelsen.



Automatisk skalning av molnresurser baserat på användning.

Vidare har vi undersökt hur förstärkningsinlärning kan användas på andra områden, som att minska energin som används av datorernas kylsystem genom att styra hur lasten fördelas i samförstånd med hur kylsystemet styrs, eller hur lasten ska balanseras över geografiskt spridda datacenter för att optimera för både energikostnad och användarupplevelse. Vi visar att det är möjligt att hitta strategier som är bättre än de som är brett använda idag, men att det också finns nackdelar, som att vi inte alltid kan förstå varför datorn väljer en viss strategi, eller om strategin är bra i alla lägen eller bara i de lägen som vi har testat. Detta är något som vi måste ta hänsyn till när vi använder maskininlärning, och det är viktigt att vi är medvetna om både fördelar och nackdelar med tekniken.

Utöver att undersöka styrning av fler delar av molnet har vi också undersökt hur vi kan göra det enklare för datorn att hitta en bra strategi. När det kommer till förstärkningsinlärning finns det många små tricks som kan göra det enklare för datorn, som att baka in kunskap om problemets struktur in i själva inlärningsalgoritmen. På detta sätt kan vi återanvända mycket existerande kunskap, och endast låta maskininlärningen göra små förbättringar på toppen av det. Detta kan vara svårt att få till på ett bra sätt, men vi visar att det kan vara värt besväret.