ADVANCED SYMBOLICS

Beyond knowledge

Creating perfect samples online using Conditional Independence Coupling

Executive Summary

The online world is rich in market data – what brands people like, which political parties they support and which government programs they agree with. The problem is that without a scientific methodology that is as rigorous as what we've done with traditional polling, the conclusions drawn from this data are without merit. To put online research on the same footing as traditional polling, we need a more rigorous sampling methodology.

Conditional Independence Coupling (CIC), is a new algorithm which is better for sampling large, online networks than older methods such as Random Walk and Metropolis-Hastings. Current methods crawl the network to collect as many people as they can in order to create a sample. For instance, let's say we start with John, and then add all of John's friends to the sample. The problem is that John's friends are likely to be like him, and therefore, our sample is not diversified. So how do we decide who to add to our sample and who not to add? That is called the "stopping problem." CIC adopts a new stopping condition, called Conditional Independence (CI). This algorithm is scalable for sampling large networks without any bias as compared to previous algorithms. CIC is mathematically proven to provide a random sample that is identical to the stationary distribution.

Introduction

Online Networks (ONs) are large graphs consisting of nodes and links. Mining these large networks has become a big challenge. The most popular algorithms use Markov Chain Monte Carlo (MCMC) techniques.

Metropolis-Hastings and Random Walk are typical MCMC algorithms for producing samples. In theory, they are very simple. They start with one sample, and then travel the network, looking for a node that is not related to the first one. When it is found, it is added to the sample. This is called "stopping" as it tells the algorithm when to stop looking for a new person to add to the sample – that person has been found. Then it starts again, until it is told to stop, each time adding incrementally to the sample. In this way, it ensures that we have a sample of people that are not related to each other, but are still representative of the population we wish to measure.

In practice, these algorithms have difficulty determining if a node is related to the first node. The question is how far from the initial choice do you need to be before a node is un-correlated to the original? Detecting how far away is good enough is called the *stopping problem.* What if I go100 nodes away, but now I've looped back again to one of your close friends, or worse, to you? Then I'm either double counting or over-representing a certain group of people.

A common solution to the stopping problem is to employ ***convergence diagnostics***. These techniques monitor the chains and use heuristic tests to detect if the chain is sufficiently far away from the starting point. There are 3 problems with these techniques:

1. they have no guarantee of stopping automatically, which requires manually monitoring and stopping the chain if it goes too long. Then the process has to be started again, which makes it time consuming and expensive
2. they are very expensive to run, as the algorithm has to produce very long chains, with lots of data being collected and discarded before it can be sure the nodes are not correlated

3. they are at best approximate estimates, as they are heuristics and not mathematical proofs and never fully guarantee that the selected node is really uncorrelated with the initial choice

An alternative to convergence diagnostics is to use perfect sampling techniques such as Coupling From The Past (CFTP). These algorithms give an exact stopping condition that is mathematically proven.

There are two main obstacles to overcome with CFTP algorithms. Firstly, unlike Markov Chains, CFTP requires that the algorithm **consider all the nodes in the network at once** in order to prove mathematically that they are unrelated. Secondly, the whole network must be stored in memory, an impossible task for even a medium-sized network. Clearly not suitable for sampling social media, which is too large.

The new work undertaken by ASI takes the advantages of CFTP and overcomes the two difficulties – the stopping condition and the in-memory condition. CIC introduces a new stopping condition, called Conditional Independence (CI), for efficient and effective stopping. This new condition works without the need of storing — or even knowing the details about — the entire online network. The key innovation is a new criterion for determining how far away a node is from the starting node.

Conditional Independence Coupling

Storing the whole network is prohibitive at the cost of time (crawling the entire network) and memory (storing the entire network). It is also impossible in ONs because the whole network is unavailable!

The first step in CIC is to make a more concrete definition of ***near nodes*** and ***far nodes***. We do this by creating an initial first network called the ***Small State Space (SSS).*** The SSS is a randomly generated starting network that is contained within the larger network. The SSS is not a copy or sub-sampled version of the entire network. It is simply a very small, local network that is used as the starting point for the CIC algorithm.

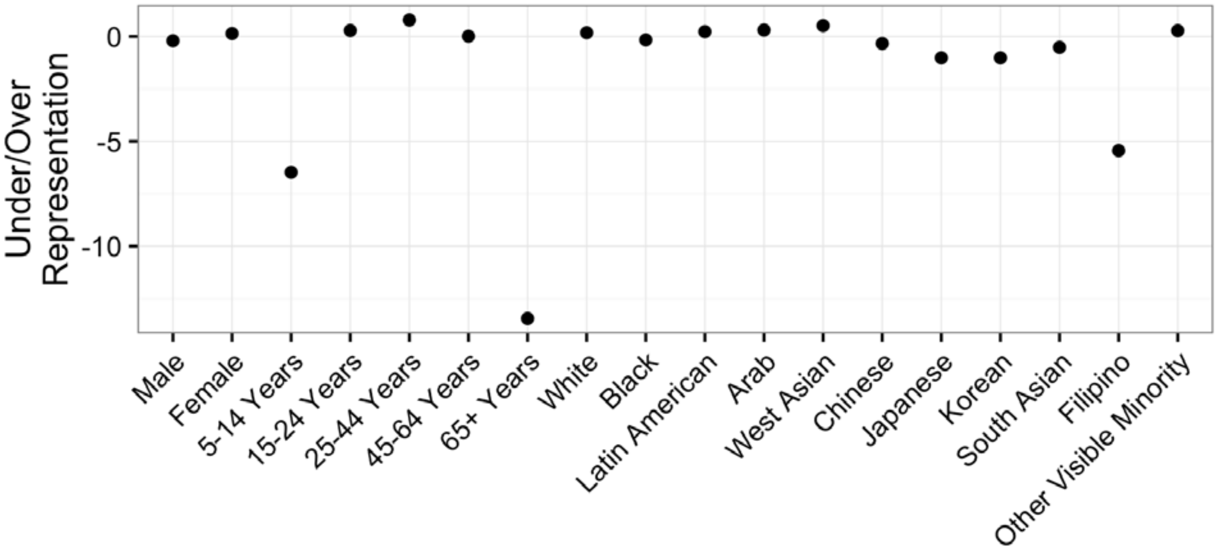
Using the SSS, we make a clear distinction between a *near* node and a *far* node. Pick a node in the larger network at random that is not part of the SSS. If we start a Markov Chain on that node, and it has a high probability of visiting a node inside the SSS, it is a **near node**.

Conversely, a **far node** is one that, if we started a Markov Chain on that node, has a low probability of visiting the SSS .This idea of near and far nodes is the basis behind Conditional Independence.

With Conditional Independence, we can mathematically prove that the probability of picking this node is the same as if we had access to the entire network. In other words, this is a ***perfectly randomized sample*** regardless of where we started on the network.

We are now ready to run multiple coupled Markov Chains starting from the initial SSS.
The technical core of CIC is to create two sets: W, called working set and V, called Visited Set. Initially, V is the SSS with the starting nodes and W is empty. We then run a one-step Markhov

Chain on every node in "V". The new nodes that we hit go into "W". We now set V=V+W, in other words, all the nodes that we've visited. Now, we run a Markhov Chain of length "2" (i.e. two hops) and add those nodes to "W" and set V=V+W. We continue doing this, incrementing the length of the chain each step, until all but one node in W is present in V. This leftover node is, by mathematical proof, a **far node**, and that is what we put in the sample. At this point the algorithm stops, a new initial SSS is created around the sampled node, and a new forward coupling is started.



Experimental Study

We have also run experiments to measure how representative the sample is of the population (see below). Using CIC, we collected a random sample of 5,000 Canadians. We then provided the social profiles of these 5,000 people to a committee of university researchers who were tasked with identifying the age, gender and ethnicity from the social profile. If 2 out of 3 researchers agreed on the labeling of a particular trait, then it was assumed to be true. This is a standard practice for labeling data sets in science.

We then compared this labeled data set to the Canadian Population (2011 Census), to see how representative it was across gender, age, and ethnicity. We found it was representative with 3 notable exceptions:

1) Children 14 and younger are under represented. This is because many online sites require their members to be at least 14 years of age and parents often will curtail online activity of their children.
2) Adults over 65 years are also under represented. Online is a new form of media and as has happened in the past, the oldest members of our population are slower to adopt this new medium.
3) Filipinos are under represented. This is most likely due to researchers mis-labeling Filipinos as Asian or Hispanic (the Filipino ethnicity is a unique blend of Asian and Hispanic traits).

Conclusion

Conditional Independence Coupling (CIC) is an algorithm that allows researchers to create random samples of the population as is done for traditional polling.   It is important to note that, because all but 3 demographics are well represented online, this sample tells us what the population as a whole thinks, not just what people online think. In a similar vein, people who conduct telephone polling are not telling us what people with telephones think, they are designing a research study that tells us what the demographic thinks, not just the people who answered the survey.

Once the sample is created, it is fed into ASI’s Artificial Intelligence (AI) where it is used in conjunction with other technologies to enable accurate market research forecasting.