Brief Overview: Pattern Computer

Discovering Patterns in Complex, High-Dimensional Datasets

Pattern Computer Inc.

August 2022

Abstract:

Pattern Computer Inc. (PCI) is a 6-year-old company addressing the simple question: "Can you create a computer system that is capable of *discovering* patterns in high-dimensional data?" The answer is Yes, and the details and results compelling. PCI has created a novel, powerful, scalable, high-performance computing system called the Pattern Discovery Engine™ (PDE), in which the data itself reveals the patterns within: the number of patterns, the key factors in each pattern, and how they contribute to the outcome. While significant credit has been given to the power of artificial intelligence (AI), in reality, the mathematical algorithms generally behind AI – neural networks – only *predict* a given outcome, e.g., whether a bank loan will perform or go into default. Neural networks do not tell you *why* a loan application was rated a high risk or identify the anomalous gene expressions associated with poor cancer survivability. Pattern Computer can and does do this. Pattern Computer will make the PDE available as Pattern Discovery as a Service (PDaaS).

Pattern Computer recently announced the development of its Explainable AI (XAI) engine – which decodes the decisions made by a neural network. The company has successfully engaged in projects in biology, aviation / aerospace, mining, microbiology, materials, and energy industries – discovering the patterns in their data, impacting their business success, and building accurate models to aid in understanding their data. Pattern Computer's mission is to create positive social impact in the world. Taking on diseases such as cancer is a key part of that mission. Using the PDE, Pattern Computer has identified anomalous gene-expression patterns associated with poor survivability in triple-negative breast cancer (TNBC) and ovarian cancer and is in the process of doing so with prostate cancer. With this information, Pattern Computer has developed a multidrug cocktail using off-the-shelf, approved drugs to target the patterns discovered. Pattern Computer is now in the fifth round of testing with TNBC and *in vitro* testing with ovarian cancer. It is also engaged with the Fred Hutch Cancer Center on colorectal cancer.

Using these pattern-discovery capabilities, Pattern Computer created the ProSpectral™ device, which uses broad-spectrum light patterns altered by the metabolomic response of two drops of saliva to detect a patient's response to a disease. At present, the PCI team is focused on detecting Covid-19. Currently in late-stage development, the ProSpectral device detects Covid-19 in 3 seconds (for $3-$5) as accurately as a PCR test, which takes at least 15 minutes for 10 times the cost. It is expected that ProSpectral will differentiate between Covid-19, influenza A&B, and RSV detection as required by the FDA. **ProSpectral could potentially detect any human disease**.

How is Pattern Computer different? Today's AI does not inform business decision-makers as to *why* a given predictive decision was made or *why* a manufacturing system failed. How can a company that has made itself dependent on AI-based decision-making be expected to resolve a problem without understanding which factors are impacting the outcome? Companies have pulled back on their investments in AI because they do not address their *business needs*. Pattern Computer has developed the Pattern Discovery Engine to solve that problem.

Introduction

Pattern Computer Inc. (PCI) is a six-year-old company based in Friday Harbor, Washington, with offices in Redmond and San Diego, and consists of 22 team members with interlocking talents. The team members have advanced degrees and/or successful careers in the areas of advanced mathematics, physics, biology / microbiology, bioinformatics, software development, hardware development, and/or electrical / electronic engineering. The PCI team is complemented by a small set of industry and academic experts to address specific areas of inquiry. While leveraging the technical and healthcare talent pools in the greater Seattle and greater San Diego areas, Pattern Computer has also established working relationships with the University of California schools (primarily UC San Diego and Berkeley), the Department of Energy's Los Alamos National Lab, and the Lawrence Berkeley National Laboratory. When the company was formed, it was agreed that the focus of the work would be to provide solutions that contribute to positive social impact, such as identifying the patterns behind cancers and other life-threatening diseases or reducing harmful emissions, pollution, etc. These complex areas involve large datasets with hundreds of thousands, or even tens of millions, of features (variables).

The Pattern Discovery Challenge

In the last decade, data scientists have been enamored with artificial intelligence (AI), which is effectively implemented by various realizations of neural networks. While neural networks can produce accurate predictions based on millions of examples of training data, they have two critical limitations in regard to pattern discovery. First, their computational requirements scale as the *square* of the number of covariates ($\mathcal{O}(n^2)$). This means that for reasonable computational speed, you need to either keep the number of covariates low or have a supercomputer on standby. Second, neural networks are somewhat arbitrarily designed, using a series of weighted decision nodes in a series of *hidden layers*. The decisions being made in the network's hidden layers are generally not known. Therefore, while many data scientists may be very competent in using neural networks to predict a specific outcome, based upon many training examples, they do not have data to indicate which of the factors (covariates) are associated with a specific outcome. This is the information that businesses need.

Given a simple manufacturing context, the neural network can predict whether a manufactured part will fail prior to the warranty expiration. But the challenge, of course, is that if that part happens to be a braking system on a train or a control for a flight surface on an airplane, you clearly want to know *why* this part fails early and how to prevent the defect in future parts. In this case, the engineer will need to know which test / validation measurements indicate the part failure and what steps can be taken to address the problem. These are not curious mathematical questions; they are critical business problems affecting both the company's reputation and its bottom line. The neural network is not able to provide the *why*.

Pattern Computer's approach is fundamentally different. It can identify and rank the key factors associated with a specific outcome. It builds mathematical models that describe the relationship between these key factors and can visually observe these relationships in up to eight dimensions using its Dimensional Navigator™. For data scientists, this means that they can use the Pattern Discovery Engine™ (PDE) to understand *why* failures are occurring. Then, working as a team with subject-matter experts (SMEs), they can review the patterns to reveal the scenarios or conditions under which these failures occur, as well as map out solutions to address the problem through additional test / validation of critical components related to the failure scenarios identified – or tighten the validation criteria for existing manufacturing methods. *The PDE becomes a tool enabling data scientists to move beyond simple AI prediction to understanding the nature of the data and the implications residing within.*

The critical difference between Pattern Computer's PDE and existing AI methods is that while neural networks can accurately predict an outcome, they are not particularly useful for solving an associated business problem by revealing the factors and mechanisms impacting the negative outcome(s). Pattern Computer's PDE does exactly that: it builds a mathematical model describing the relationship of those factors to one another and whether those relationships are directly, or inversely, related to the outcome. That is precisely the information a technical team or business decision-maker needs for addressing problems. Those results can positively impact the bottom line.

The Pattern Discovery Engine

The Pattern Discovery Engine is a set of algorithms working together in a complementary fashion to provide dimensional reduction in the size of the dataset without losing the information about the relationships that exist within the dataset. As these relationships are revealed and tagged with metadata, the discovery engine then calls other components within the PDE to further refine the data and reveal the patterns, in part by removing noise and biased information from the dataset. It then organizes the information in an acyclic digraph in a manner that maps out the pattern discoveries and then generates the mathematical models describing the structures discovered within the data.

The PDE also identifies important relationships between the covariates, whether they are collinear or have inverse relationships, etc. Hard outliers not having any relationship with the outcome are also identified. While these details may not seem particularly interesting, once the data scientists and subject-matter experts start investigating these results, the details fill in the "colors between the lines." It is important to note that the Pattern Discovery Engine also finds all of the significant patterns in the dataset – as there are typically multiple patterns resident therein. For example, in a dataset related to the patterns behind commercial flight departure delays, the first pattern was associated with flight departure delays initiated by the FAA's Operational Network – delays due to jet route closures because of intense regional storms. Secondary patterns included flight departure delays due to local inclement weather (fog, snow, and/or icing conditions) and delays due to late incoming flights.

It is common to see multiple patterns in the datasets. Particularly useful for businesses are not only the understanding of the patterns and the factors associated with these outcomes, but also a ranked list of the patterns themselves. When the researcher or business decision-maker wants to know the most principal factors that are impacting a specific desirable or undesirable outcome, they know where to start to have the greatest effect on the output. Also, by having this list of choices, one solution may emerge as being more cost-effective and easier to implement than the others. The business decision-maker can start to see some improvement as soon as possible, while the longer-term, perhaps more complex, solutions are being deployed.

As a practical matter, Pattern Computer ran the data from the public Breast Cancer Wisconsin Diagnostic Data Set[1] and determined that *with only three measurements*, breast cancer could be predicted with 98.2% accuracy[2]. When you think of such impactful findings, how valuable is it to get that information out to the medical community? To see examples of real datasets in which Pattern Computer

[1] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
[2] https://www.patterncomputer.com/wp-content/uploads/2022/02/Three-Results-Beyond-Published-Papers.pdf

not only exceeded the accuracy of the peer-reviewed, published results, but also identified the mathematical models, please review the short case study "Three-Results-Beyond-Published Papers."[3]

The Dimensional Navigator

As the PDE reveals the patterns and the ranked factors associated with that specific outcome, the next questions that often come up are: But what does that relationship *look* like? Are there relationships between the factors that are clear-cut, and is there a clear threshold between one or more of the variables when a specific outcome is known? Or, when looking at the key factors of the dataset, is there a mottled blend of factors associated with different outcomes?

When seeing these mottled relationships, one might ask whether all of the related variables have been included. Is there some other, additional factor that we haven't thought to include which may create a better delineation of the outcome? For example, in the Flight Operations identifications of the patterns associated with flight departure delays, the report out of the first pattern at the end of the first run was "Flight Date, Tail Number, and Route." Those patterns are not clear-cut; the SME looked at the dataset using those factors to identify the pattern. While the dataset included local weather at the departure and arrival airports at their respective takeoff and arrival times, the SME observed that when looking at the dataset knowing that Flight Date, Tail Number, and Route were principal factors (out of ~130 covariates), these patterns were associated with en-route weather events. On a few specific days, certain planes were multiply impacted by weather events on given routes that crossed over and back a large regional weather event (e.g., a line of thunderstorms from the Ohio Valley to North Texas). The SME determined that those types of delays would be issued by the FAA's Operational Network (OPSNET) group. When the OPSNET flight departure delay factors were included in the dataset, they popped to the top of the list, and a clearly defined set of factors appeared in the dataset.

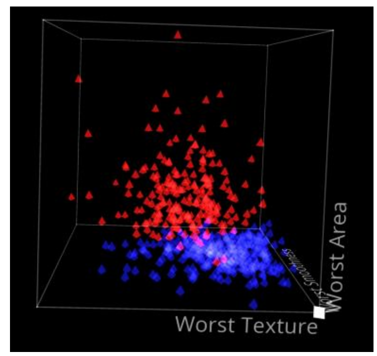

Figure 1: Dimensional Navigator visualization of Breast Cancer Wisconsin data in 2.5D

The Dimensional Navigator enables users to view up to eight dimensions in virtual reality, allowing them to "walk through" the dataset in a hypercube model, able to query specific datapoints as they move through the dataset. The user is able to re-assign covariates to specific axes to evaluate the relationships dynamically to better understand the nature of the relationships for general understanding and decision-making purposes.

The ProSpectral™ Device

Enabled by the power of the Pattern Discovery Engine, the ProSpectral device is a high-performance, accurate screening (and, ultimately, diagnostic) device for human diseases. Currently in late-stage development, the ProSpectral device is able to detect a person's metabolomic response to Covid-19 infection in 3 seconds as accurately as the gold-standard polymerase chain reaction (PCR) test, which takes at least 15 minutes.

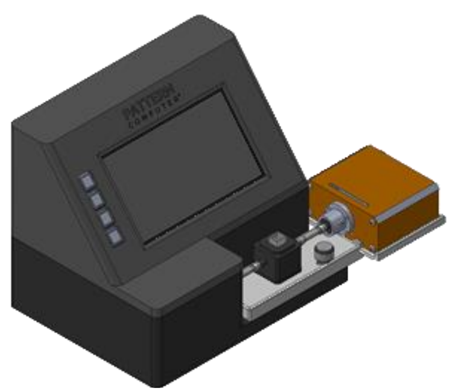
Figure 2: ProSpectral™ device

[3] Ibid.

Using broad-spectrum light with precision light sensors, ProSpectral captures the transmission of light through a polycarbonate cuvette holding two drops of a test candidate's saliva. No DNA is collected – just the light-absorption pattern of the contents of the cuvette. Notably, the collection method is not specific to Covid-19, or even to saliva. It is expected that ProSpectral will be able to detect monkeypox and other diseases, such as Zika, where early detection and treatment is important to the successful outcome. Other form factors of liquid samples could include urine and blood plasma.

Ultimately, *ProSpectral could detect any human disease.* Current costs for PCR tests in the US are approximately US$55; using the ProSpectral device, the individual test cost would be $3-$5, an order-of-magnitude cost differential per test. Another important use for ProSpectral would be in emerging countries where cost-effective diagnoses and access to medical facilities and doctors are limited. ProSpectral devices in remote villages, connected by cellular or satellite communications, could provide highly accurate disease diagnosis for remote treatment via drone delivery (via a third party). Instead of requiring a 20-mile walk to a hospital (which may further reduce a patient's immune response), the correct diagnosis and treatment is performed remotely. Notably, ProSpectral can be battery-powered and remotely updated via internet connectivity.

Explainable Artificial Intelligence (XAI)

The evolutionary development of artificial intelligence over the last 50 years has occurred in fits and starts, in many ways limited by the computational capabilities of the era. With rare exceptions, the progress of artificial intelligence has been through the development of machine-learning algorithms, which in the current generation of machine learning are variations of neural networks. The mathematical algorithms describe a network of layers of weighted nodes, interconnected with the nodes of neighboring layers, forming a web, layer by layer, through which decisions are made as to the predicted result of the network. The layer weights, and therefore the intermediate decisions, are determined through a training process whereby thousands, or even millions, of examples are fed through the neural network to "learn" the properties of the dataset on which it is being trained – for example, the last 10 years of loan applications processed by a regional bank.

In a loan scenario, the neural network learns by accessing all of the variable information received, along with the customer's account history data and the performance history of the loan, to determine whether that loan application was approved. Once trained on that data, the neural network can then process new loans, and, using its predictive power (based on all of the previously seen training data), recommend to the bank's loan officers whether to approve or deny a loan.

That is considered state-of-the-art today. Yet, these types of tools are subject to learning the mistakes of others as well. If bank managers demonstrated bias in their handling of loan applications – e.g., by providing preferential treatment based on the applicant's home location, age, or gender – then the neural network was trained to place more weight or consideration on those factors as well.

One of the key differences with Pattern Computer is its ability to understand the decisions being made within the neural network. Similar to how Pattern Computer can identify the patterns in given datasets, the ranked list of factors, and the accurate mathematical modeling of those relationships, using its explainable AI (XAI) engine can reveal the decisions being made within those neural networks. This ability is particularly valuable when companies want to know if their automated neural networks are biased in their decision-making process and they need to remove that biased information from their training datasets – or potentially be held liable by customers or governments that can also use Pattern Computer's XAI tool to reveal the decisions made within a neural network. Such decisions may or may not be compliant with the European Union's General Data Protection Regulation (GDPR) laws or the California Consumer Privacy Act (CCPA) regarding data misuse by commercial businesses operating in the state of California.

Future Work

Pattern Computer continues to expand the scope of its work, adding new capabilities to its Pattern Disovery Engine algorithms, as well as new areas of research. Pattern Discovery as a Service (PDaaS) is planned for early 2023, targeted toward industry, research organizations, and academic institutions. It is expected that some capabilities of the Dimensional Navigator will also be made available for users with compatible virtual reality systems to "walk through" and understand their data.

Pattern Computer will also continue its work in the field of biology, addressing the pattern discoveries behind the most critical forms of cancer. Work on triple-negative breast cancer, ovarian cancer, colorectal cancer, and prostate cancer is currently underway, with lung cancer on the horizon. Some of this work is currently limited by data availability.

As part of the FDA approval process, it is reasonable to assume that we will collect enough data to build accurate models for the ProSpectral device for detecting influenza A, influenza B, and RSV, along with Covid-19, by mid-2023. Pattern Computer is currently evaluating the need for monkeypox detection, but this is in the early stages, as data-gathering opportunities may be more limited.

Summary

Pattern Computer discovers pattern in complex, high-dimensional datasets. This is performed using a novel combination of mathematics and scalable information structures. This pattern-discovery method goes beyond traditional AI, which only make predictions. Using the Pattern Discovery Engine, Pattern Computer produces a ranked list of patterns discovered in the dataset. For each pattern, a ranked list of covariates and a mathematical model describing the relationships between the covariates can be viewed using the Dimensional Navigator. This is a breakthrough technology which incorporates an understanding of the patterns within these datasets. General machine learning and AI methods published today cannot do this. Additionally, Pattern Computer has delivered on the ability to perform XAI, which reveals the decisions being made within neural networks and, in fact, can go beyond the accuracy of the neural network itself by identifying additional information in subnetworks of a given dataset.

Using the Pattern Discovery Engine, Pattern Computer created the ProSpectral device, which uses high-resolution models created by observing changes in the light-transmission patterns through a cuvette containing a liquid (in this case, saliva) and has developed an accurate method of detecting whether a person is infected with Covid-19. This technique has been submitted for a patent and is expected to be useful for detecting most, if not all, human diseases; and it can perform the tests in seconds – in the case of Covid-19, 3 seconds, with the same accuracy as the gold-standard PCR test, which takes at least 15 minutes.

Pattern Computer comprises a small but exceptionally talented team of industry veterans and technical experts. They continue to push the frontiers of pattern discovery in different varieties of datasets while taking on the urgent issues facing society, including human diseases, pandemics, and environmental challenges such as pollution and energy management, as well as difficult industrial challenges.