ESSAY

IMMUNOLOGICAL INSPIRED NOVELTY DETECTION

Technical Report
No. 4-01

Jason Brownlee
Master of Information Technology, Swinburne University of Technology, 2004
Bachelor of Applied Science, Computing, Swinburne University of Technology, 2002
(jbrownlee@ict.swin.edu.au)

Centre for Intelligent Systems and Complex Processes
Faculty of Information & Communication Technologies
Swinburne University of Technology

April 2005

Table of Contents

List of Figures

List of Tables

1. Introduction

This document is interesting for a number of reasons. Ultimately, this work provides a description of the space I am working in, without defining the specifics of what that research will consist of. A landscape is painted that provides a broader perspective for my proposed research project indicating where the cogs that will be the focus of the work fit into the larger system and how that larger system is connected to the real-world.

The selected research area is that of using inspirations from the biological immune system of vertebrate organisms as the basis for a system that addresses the constraints and requirements of novelty detection in dynamical and high data load environments. The general research question as to how this could be achieved, and the effectiveness of the resulting solution is framed within clear definitions. A concise mean of biological inspired computation is provided and a powerful yet simple framework is defined. This framework permits both existing related work to be naturally integrated and related, as well as providing a guide for identifying relationships and ultimately scope for research questions posed in the area of biologically inspired computation. Further, very clear roles are defined for the selected biological metaphor and its relation to the selected problem domain.

A problem centric view is taken, where all characteristics of the proposed solution are defined in the context of the domain, in this case novelty detection. The problem is explored from both a reductionist standpoint, so as that it can be easily related to existing work, and form a specialist standpoint so that it can be related to practical engineering problem domains. Problem constraints are clearly laid out, and from those constraints, useful requirements are defined of which a solution must demonstrate meeting to satisfy the selected problem domain. The two primary problem constraints of interested are high volume input data, and a dynamical underlying models both obviously in the context of novelty detection.

The metaphor is briefly discussed, and though it is the inspiration for a proposed general solution, its usefulness ends after the formulation of a solution. A general thesis is proposed as already mentioned, and a general broader system is described that represents an ideal or vision for a system. The system is intentionally wide reaching and vague. It consists of a number of specific processes, the aggregation of which defines the capability of the proposed solution. It is these processes, which are expected to be the focus of my work, though with the knowledge that each algorithm addresses a specific need within the described larger system.

Finally, some useful analytical and empirical techniques are identified and discusses as being potentially useful as evidence in constructed arguments defending the proposed solution. Ultimately, this work is intended to provide an induction and broad encompassing understanding of the nature of the proposed research problem. It is indented to demonstrate the usefulness of the selected problem and indicate where the fruit of such work fits into the broader scope of addressing practical engineering problems using inspirations from nature.

2. Methodology

Artificial intelligence or intelligent systems are a misnomer. Perhaps the only useful terms used are artificial (the systems are not natural) and system (collection of related elements) – though the term "artificial system" does not describe the intent of the field either. I choose to refer to what I do as biologically inspired computation. That is, the constructing of systems using computers that is in some way inspired by a biological processes. I choose to define the goal of this field in broad practical terms: to address engineering problems. Specifically I choose to define the field as follows:

Biologically Inspired Computation: The ideal of providing alternative – biologically inspired solutions to difficult engineering problems that add value over conventional (read non-biologically inspired) approaches

The definition implies three aphorisms:

1. Solutions are based on biological metaphor in some way
2. Problem domains are measurably hard
3. Solutions offer measurable benefits over existing approaches

It should be pointed out that this is an ideal and not requirements of such in the field. This ideal can be used as a tool for putting existing work into perspective, as well as for defining the broad elements required for work in this field.

2.1 Perspective and Requirements

Work in the field of biologically insured computation can be viewed in terms of two key elements:

1. The biological metaphor used
2. The addressed problem domain

It is proposed that work in the field observes both elements, though with specific constraints (defined further on). The realistic focus of work is expected to address sub-elements of some broader solution, yet still observe that some broader solution or "system" exists in which the sub-elements are applicable. This so-called larger system is a solution to an equally larger engineering problem, where the sub-problems addressed relate to data or processes (algorithms) within that the larger system.

Thus, the two-element view proposed above can be extended to the following:

1. The inspirational biological metaphor used
2. The addressed problem domain
3. The proposed solution (broader system)

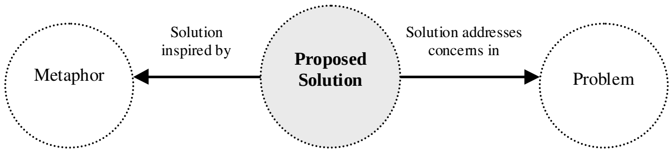

Figure 1 - Shows the three main elements for work in biologically inspired computation

This perspective is general and broad sweeping, yet highlights that larger scale systems define the works existence in the context of the inspiration and problem. The danger of not understanding or even acknowledging the broader system and its context is that work may be defunct or trivial; it may miss the point so to speak. The benefit of such a broad model is that it provides clear purpose and perspective. If research in this field is abstracted to such concepts then the relationships to related fields and existing approaches become obvious.

2.2 Research Model

The clear question raised is to what detail should the biological metaphor, problem domain, and broader solution be described? A concern is too much attention to the metaphor in the solution and the work can be considered biological modelling. Too much effort on the problem domain and the work can be considered engineering. I choose to define a continuum between the two poles that permits existing work in the field to be placed on a scale, and provide a guide for tailoring proposed research endeavours.

I argue that system components inspired by metaphor must complement or address problem features, and thus the argument "*because biology has it / does it*" have no validity. The metaphor provides a path or inspiration for a solution, though the specific characteristics of the solution must be defended in the context of the selected problem domain. The following figure highlights these concerns.

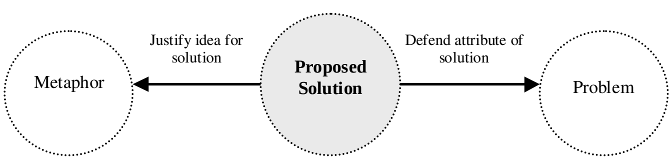

Figure 2 - Shows a relationship between solution, metaphor and problem in the context of constructing arguments

Arguments can be constructed relating features of the solution to complement features of the problem, though the arguments require evidence. Evidence is provided in either analytical or empirical forms. A mathematical model can be used to show system behaviour in an unambiguous way; where as implementation-based testing on problem sets can pseudo-practically demonstrate system behaviour. Analytical analysis is used where appropriate for theoretical performance, and empirical evidence is confirmatory of proposed theories. Further, given the complexity of solutions, empirical evidence can be used to show behaviours too difficult to demonstrate with theoretic methods.

Work is expected to take the form of a thesis that requires defending arguments and supporting evidence. It is proposed that a simple model can be constructed that qualitatively represents the desires of such a work. This proposed model is a basic two-dimensional graph where the x-axis covers the metaphorical and practical concerns and the y-axis covers the analytical and empirical concerns of the work. Simple as it is, it covers the thesis-based treatment of work described thus far. The result that a work can be plotted as some discontinuous area over the graphs surface.

All work in some ways covers some portion of the positive and negative extents of the two axes. A metaphor is described, as is a problem, the combination of the two being the focus of the work. The work exists within the context of a broader system, which exists somewhere on the x-axis. Arguments are used to justify the inspiration for the solution and defend the characteristics of the solution, evidence of which is taken from areas on the y-axis. The following figure represents the proposed simple biological computational research model.

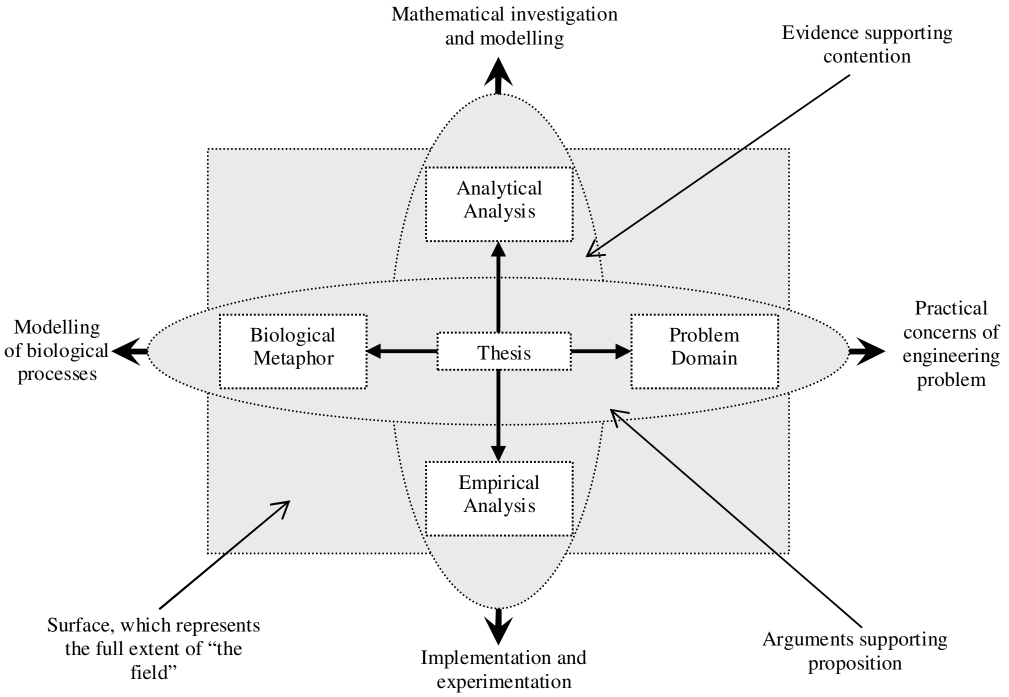
Mathematical investigation and modelling
Evidence supporting contention
Analytical Analysis
Modelling of biological processes
Biological Metaphor
Thesis
Problem Domain
Practical concerns of engineering problem
Empirical Analysis
Surface, which represents the full extent of “the field”
Implementation and experimentation
Arguments supporting proposition

Figure 3 - Simple model for work in biologically inspired computation

The presented model represents the extent of a “field” which is really a subfield within biologically inspired computation as fields are common termed by their underlying biological metaphor. Here, a field is taken as the combination of a metaphor and a problem. This field realistically has no bounds and intersects uncountable other areas of study. It is reasonable to conjecture that a field has bound on relevance, and it is reasonable that a single work (a research project) has area coverage within those bounds. When a research project is evaluated or analysed in context of the model some interesting things are observed.

Field Trajectory Biologists model phenomena, which provide the kernel for biologically inspired computation. This has been observed with some of the well-known areas such as evolutionary computation, artificial neural networks, and artificial immune systems. This occurs gradually as the connections between the problems addressed by nature are made to engineering problems and the abstractions of nature grow either in maturity or popularity. Eventually once the biologically inspired systems become mature enough at solving the problem, they are implemented in practical application by engineers.

On the graph this could be observed as a trajectory that would start at the top left corner, move down below x intercept in a soft L shape and move across passed the y-intercept over into the lower right of the plot. It is not until the plot comes back up somewhere towards the top right of the screen that practical adoption occurs. This intuitive progression occurs mostly in the on the empirical lower side of the graph because of the computer-simulation focus in the area. This ultimately results in slow adoption times that would be expected for techniques that do not undergo scientific rigor.

Project Trajectory A researcher new to a field (metaphor and problem) in biologically inspired computation will have a unique yet generally common trajectory though the plot. Entry point is commonly close to the centre guided by intuition on behalf of the researcher. Movement initially consist of jumps then slowly becomes oscillations along the x-axis close to the zero point. As an internal model is constructed of the field, the horizontal oscillations dampen slightly and there are oscillations in the vertical axis as various empirical and analytical techniques are identified and evaluated in terms of applicability.

Ultimately, the area carved out by the oscillations represents the mastery of the field and should approximate the reasonable bound within the general scope of the project. The summed knowledge in a field is simply the union of all approximate areas. The final work produced will be some small discontinuous concave area within the larger mastery area. As mentioned, there is a common bias towards the negative portion of the x-axis towards empirical computer-simulation and this has a self-reinforcing effect for researches new to the field.

The model is a useful tool, though is clearly limited. The axis are too few and the labels are too simplistic. It is interesting to consider further generalisations of the model. As a device for identifying concerns in a new research project or analysing existing work in the field of biologically inspired computation, it does provide a useful initial guide for knowledge acquisition and organisation. It demonstrates that in such a work the elements are not completely separable, and that some kind of balance is required, given a natural skew.

3. Problem Domain

Nailing down the general problem is difficult without a specific implementation. I choose to refer to the selected problem as the general problem of novelty detection. The term has connotations, specifically of identification or discovery of new and unusual things. Taking the problem of novelty detection alone is insufficient; there is simply not enough information about the problem for there to be a specific problem to be solved. The problem needs to be reduced into its basic elements and thus made easier to relate to existing work. Constraints are needed on the problem to limit its scope to something relevant to the field of biologically inspired computation, specifically to make it hard. Finally, real-world examples or case studies are required to show that ultimately the problem is practical or real world. It is only after these steps have been addressed (at the very least), that it would be possible to begin considering a solution, let alone a solution inspired by a biological metaphor.

3.1 Description

It is common for the problem of novelty detection to be described interchangeably with the problems of anomaly detection and change detection. These terms are broad and too ambiguous to be used as problem definitions without explanation. I choose to define novelty detection as the identification of anything new or unusual in input data. This means anything that is new and or interesting relative to an internal model of historic input data. Anomaly detection and change detection can be taken as two specific cases of novelty detection. Change detection is identification of any ***difference*** in input data from an internal model whether the change is unusual or not. Anomaly detection is the identification of input that ***deviates*** in some way from an internal model. The differences are subtle. Perhaps a continuum exists between crisp and fuzzy for novelty detection where the sensitivity or configuration of what is new or unusual determines the position a system takes on the axis. A crisp configuration would be considered a change detection system, whereas a fuzzy tolerance based system would be considered an anomaly detection system.

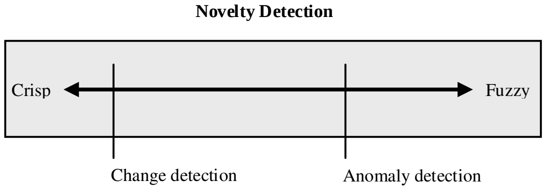

Figure 4 - Novelty detection continuum showing change detection and anomaly detection

The positions of the two specialisations on the novelty detection graph are ad hoc but provide some indication of their relationship and distinctiveness. To recapitulate, there are intuitive assumptions about the problem domain. Input data flows into the system in discrete units, and there is an internal model of normal input activity. For a given system with such a model, all data that flows into the system is new and unknown and must be classified. Data that is novel is new or unusual in that it is not the same as data in the known model (is a change) or it varies from the expected model (is an anomaly).

Given the assumption that there is an internal model of normal or expected input, the problem of identification can be reduced to that of a two-class classification problem. Unknown input data enters the system, it is compared in someway to an internal model and classified as either normal or new and unusual (abnormal). Taking anomaly detection as the example, unknown input data can be either normal or an anomaly. The problem is that real world problems are less clear-cut. It is common for a problem to have a class label of anomalies that can be further broken down. There are those anomalies that are really of interest, and those anomalies that are valid variations from normal input behaviour.

This is an interesting proposition as it permits the general problem of novelty detection to approached in at least two broad ways. The first is the already mentioned two-set problem of normal and abnormal, where all abnormal input is of interest, and not all normal input is of interest. The second assumes some overlap between the sets and breaks the problem down into a two-step process. The first classifies unknown input as normal or abnormal, the second process operates only on the abnormal input and determines if the abnormal input is of interest or not of interest.

The first approach assumes that there is sufficient information to identify all interesting abnormal input in a single decision. This may be because the problem domain defines all anomalies as interesting. The second approach is a specialisation of the first and assumes that anomalies can be detected but not directly separable as interesting and not interesting. A second process is used to identify those abnormal input cases that of interest. The ultimate result of the system is the same as that of the first; only it is achieved in a more precise manner. The effect is that the process requires additional modelling. The selection of the specific implementation is problem dependent, though it is interesting to note that the second process can be setup to emulate the first, by having a simple passthrough in the second process.

This problem of normal-abnormal, interesting-uninteresting can be described as two sets (normal-abnormal) each with the subsets (interesting-uninteresting) as follows:

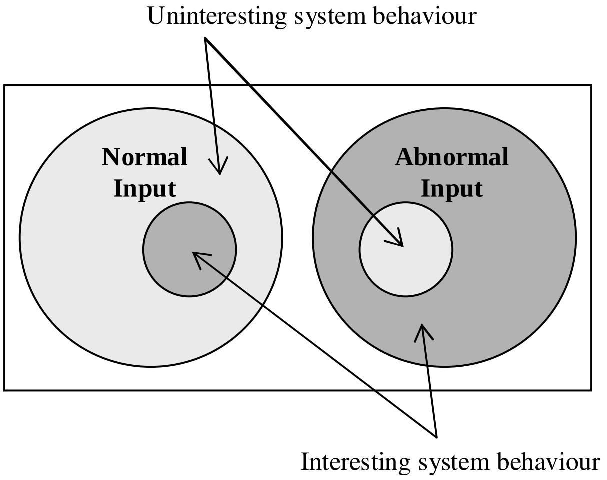

Figure 5 - Venn diagram of the general problem of novelty detection

Given that the detection problem can be broken down into that of classification, it is important to discuss what classification is. Discrimination between two classes can be taken as a probabilistic decision that is ultimately can result in a crisp decision. The probability of an unknown sample belonging to either the normal or the abnormal class is based on the probability distribution of similar input samples captured in an internal model. Thus, the process of classification can be reduced to that of modelling probability density functions (PDF's) of each class in the problem domain. In the case where the PDF is known or is simple, then explicit rules can be used. In cases where the PDF is complex or unknown then it can be approximated. This is exactly what an internal model of input data captures, where the accuracy of approximation affects accuracy of the discrimination between the classes.

It showed that novelty detection can both specialised and reduced to common problem description. As mentioned, this is useful for relating the problem to existing work both in this field and in other fields. Further, it provides a useful tool for both understanding the nature of the selected domain and in defining problem constraints and ultimately solution requirements. If novelty detection is simply a special case of function approximation then there must be sufficient justification for using an immunological approach for a specific problem over other classification techniques.

The results of reducing and specialising the selected general problem of novelty detection can be summarised using the following figure.

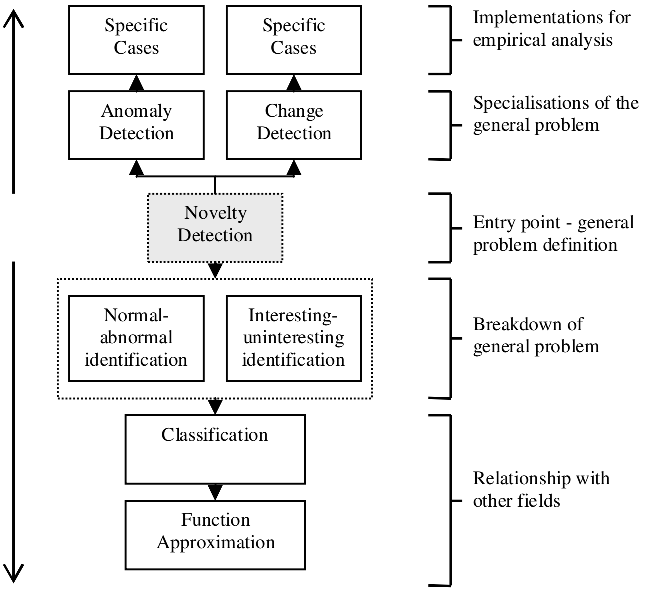
Specific Cases
Specific Cases
Anomaly Detection
Change Detection
Novelty Detection
Normal-abnormal identification
Interesting-uninteresting identification
Classification
Function Approximation
Implementations for empirical analysis
Specialisations of the general problem
Entry point - general problem definition
Breakdown of general problem
Relationship with other fields

Figure 6 - Overview of the specialisation and reduction of the novelty detection problem

3.2 Constraints

As has been described, the problem of novelty detection is ultimately that of classification and function approximation. The critical distinction of novelty detection from classification is that the specifics of what is novel are not known specifically other than newness and unusualness. Thus, the problem can be summarised as follows:

Identify interesting-abnormal input data given a model of uninteresting-normal/abnormal input data

The premise of the problem is that the underlying probability density function for the classification task is unknown. The problem of calculating a PDF can be difficult, so much so that for some problem domains it can only be approximated. Thus, the question becomes what are some reasons why a PDF cannot be calculated for a given problem instance.

There are two primary reasons for using an approximation technique:

1. Incomplete Information
2. Intractable Computation

Incomplete information implies that for some reason the information required to calculate the function is unavailable. There are many reasons for this to occur in practice such as; the input data is noisy, the sample available is limited or that the underlying function is subject to change (dynamic). Intractable computation implies that the task of explicitly calculating a PDF is too difficult to manage in terms of computational complexity (space or time). There are many reasons for this to occur in practice including an extensive volume of data, high dimensionality in multivariate input data, and non-linear joint dependencies between attributes. It is possible for a given difficult novelty detection problem to have some or all of the mentioned incomplete information and intractable computation concerns.

The following figure summarise some of the mentioned concerns.

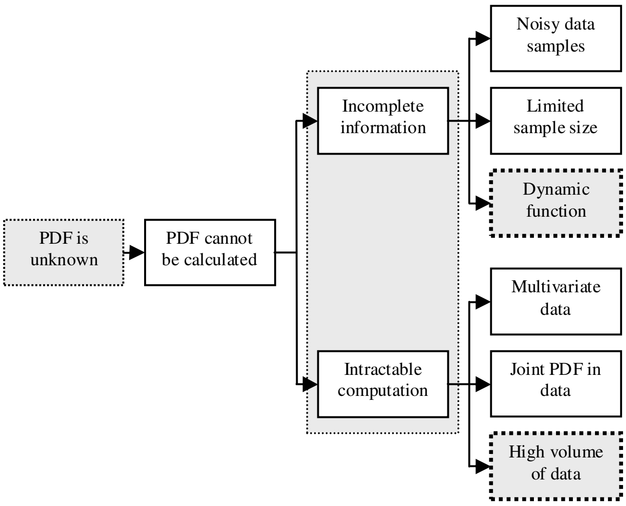

Figure 7 - Shows some of the reasons why a PDF cannot be calculated and this must be approximated

Classification is a mature field in statistics and machine learning, and many very good classification algorithms exist for specific problem constraints. Classical problem constraints are limited to things such as data dimensionality, complex attribute interdependencies, noisy data, and limited samples. The resulting effect is that many algorithms are capable of performing well on common problem constraints, though performance degrades on some measure under certain conditions. Two specific cases of such constraints include non-static underlying PDF's and high volumes of input data. This does not imply that good techniques do not exist for these constraints, rather that they are less commonly addressed, given the complexity involved.

For the selected problem domain, the broader problem constraints include the following:

1. Novelty detection using an approximate model of non interesting-abnormal input data
2. Attention to classical classification constraints such as complex interdependencies, multivariate data, noise and limited samples
3. Operation in large-scale (**high-volume**) changing (**dynamic**) environments

A solution to the selected problem is expected to be approximate because of the two selected constraints cause it to be both intractable and incomplete.

3.3 Requirements

No single technique can do well for all problem domains, the technique must be specialised in some way. Nonetheless, it is desirable to have a technique that facilitates specialisation and yet is able to meet the fundamental requirements of approximation-based classification. Each of the specified constraints is sufficiently difficult to be research problems in their own right. Although the problem requires a solution that pays attention to these conventional concerns, the focus of these problems requirements is the two constraints as follows: high-volume of input data and dynamic underlying PDF.

Requirements can be broken down by the two selected high-level reasons for using an approximation based classification approach. This list of requirements is by no means complete, rather it provides general areas of interest that are likely beneficial to address in a proposed solution. The benefit of defining even broad requirements is that it provides an initial framework to be matched by a solution, and provides something a solution can be tested against as a measure of effectively addressing a problem.

3.3.1 Incomplete Information

Constructing a model from incomplete information implies some kind of interpolation and or extrapolation from available input data. This implicitly implies error, and thus the ultimate goal of such an approach system efficacy - that is the minimisation of measurable error in the interpolation to a tolerable level within the context of the specific problem domain.

Robustness Samples taken from the domain can be inherently noisy in practice. The system must provide some robustness in terms of sensitivity to input data samples. Further, the system must be adaptive to unexpected change in the underlying function being approximated.

Learning Complete information is not available and may never be available, thus the system must be able to integrate new information in an incremental manner that is learn from experience. Generally, this is the ongoing acquisition and useful application of knowledge. This implies both extracting information from the domain and the application of specific feedback provided to the system.

Plasticity Assumptions must be made by the system and those assumptions may be incorrect. This implies that the system must be flexible enough to redefine its assumptions. Moreover, given the underlying probability distribution is changing (dynamic environment), the system must be able to track the changes in the PDF.

Reliability Given that the system has inherent error; there must be means of instilling trust in the results produced. This is something that has to be added to the system in the context of the problem. Once a level of trust has been established, the system must be consistent in behaviour so that trust can be maintained.

3.3.2 Intractable Computation

Intractable computation implies a sufficient amount of work that is beyond the capacity of the system to perform. Thus, the ultimate goal of this constraint is system efficiency whilst maintaining a defined level of system efficacy.

Complex Modelling Constructing a model of multivariate input data stream that has complex and unknown interdependencies is non-trivial. The technique must be suitable for such complexity and be able to both operate on a non-parametric premise, as well as be able to take advantage of any available information. Classification labels are not available in the data, thus the technique must extract features from the data and associate them with class labels in an automatic manner. In addition, variates can be of differing types (examples include nominal, real, ordinal, strings, etc), and given the multivariate nature, outliers can be very difficult to identify as dimensionality increases.

Scalable A high-volume of input data implies a single large stream, which needs to be processed efficiently and or multiple data streams to be processed in parallel. The system must be scalable in terms of capacity of information that it can process within a reasonable time (capacity must be defined and measurable). Further, the system should be modular enough to have scale adjusted in response to changes in the nature of the problem domain (increases and decrease in load).

Online and Batch The system should capable of operating in both an online and batch manner. This flexibility in application permits the system to be used in a wide range of practical problem domains.

3.4 Examples

The problem of novelty detection is broad, and has been shown can be specialised at least to two problems that are more specific: anomaly detection and change detection. This section seeks to further specialise the general problem to practical case studies that both reinforce the defined constraints and requirements, and show the selected problem to be real world. Only brief overviews of each problem are provided, though each is sufficiently detailed to relate the nature of the case to the broad problem of novelty detection.

3.4.1 Network Intrusion Detection

A network intrusion detection system (NIDS) addresses the problem of identifying attempts to gain access to resources on or cause malicious activity to systems on a network. Attacks can be either internal or external and typically include such suspicious activity as scanning the network (port scanning), disrupting services in the network (denial of service attacks), and gaining entry to systems. Typically, such attacks operate at the transport (network packet) level evaluating such attributes as host addresses, ports, and application data contained within the packets.

Solutions commonly involve static rule based systems that operate on input data attributes. Rules are specifically tailored to an organisation and require explicit management. Information regarding common attacks can be captured and distributed by a

NIDS vendor and distributed to customers via an automated update procedure. Alternatively rules describing common attacks can be captured internally and distributed within an enterprise. The primary concern in both cases is that an attack must occur, be identified and be describable within the context of the NIDS. An area of interest in NIDS is the ability to detect and respond to what are termed “zero-day” attacks. These are previously unknown attack/intrusion attempts that has no existing rule or behavioural pattern recorded in the protective system.

Question	**Answer**
What is the ultimate goal?	Detect network intrusion attempts.
What is the input data?	Network traffic – likely transport (packet) level or application protocol level (see OSI model)
Where does feedback come from?	Validation of previously made decisions checked manually – querying user or relevant investigation (etc…)
How is it Dynamic?	Complete knowledge of what is normal network usage is unknown and is changing with computer usage habits (new applications and new people)
How is it High volume?	Operates at various levels of abstraction from the individual host level to the enterprise network level where volume of packets is beyond practical evaluation.
What is novelty?	New methods of network intrusion are continually being devised

Table 1 - Summary of the network intrusion detection problem domain

3.4.2 Identity Theft

Identity theft, also called host intrusion detection, is the detection of unauthorised access or misuse on a computer system. Such misuse may include access to restricted information or malicious activity such as disrupting services or destroying sensitive files. The problem commonly involves the use of a skilled social engineer, the theft of user login details or the use of automated tools such as scripts, viruses, or Trojan horse programs.

Like network intrusion detection systems, common solutions include both explicit rule based systems that monitor normal system behaviour, as well as template or signature based approaches that use knowledge from previously identified attacks. Like NIDS detection zero-day (new automated techniques) infiltrations are of interest, as are more subtle variations in usage behaviour that go unnoticed by conventional static (parametric) approaches and signature-based solutions.

Question	**Answer**
What is the ultimate goal?	Detect when identity (system access) is misused.
What is the input data?	System usage behaviour – application usage, temporal usage, geographical usage
Where does feedback come from?	Validation of previously made decisions checked manually - querying user or checking surveillance (etc…)
How is it Dynamic?	Complete profile of user behaviour is unavailable, profile changes with time, valid sharing of identity may occur
How is it High volume?	Amount of information captured for a single user, amount of information captured for a common group of users
What is novelty?	It is unknown what a user will do with a stolen identity

Table 2 - Summary of the identity theft problem domain

3.4.3 Antivirus

A computer virus is a computer program that replicates itself and infects computer systems. The result of receiving such a virus can range from the benign to the malicious. Some virus programs modify or disrupt critical system files, thus one of the tasks of antiviral software is to detect malicious activity in important system files.

Commonly this is achieved through signature based means where the antivirus software has a library of known virus behaviours. The protection software detects such known behaviour and takes action. Further more adaptive approaches are used to augment the signature method including taking snapshots of system files and performing peridotic checks for change, as well as restricting access to sensitive system directories.

Question	Answer
What is the ultimate goal?	Detect malicious changes to system data (files).
What is the input data?	Structure of system files and behavioural changes to files
Where does feedback come from?	Validation of previously made decisions checked manually – responsible program investigated
How is it Dynamic?	Nature of valid changes to system files can vary over time, files themselves can change with time in valid ways
How is it High volume?	Amount of checking to perform for a single file, vast number of individual system files on a host, or on a collection of hosts
What is novelty?	Changes to a file that are malicious or can result in malicious activity (key logger, etc…). The methods used for such activities are vast and under continuous development

Table 3 - Summary of the antivirus problem domain

3.4.4 Bill Validation

It is common in many industries to have services delivered by a service provided, sold by a service retailer and in some cases created or prepared by a service manufacturer. Services are provided at one level and then on-sold by the next level until ultimately the customer (end-user) receives the service. Example industries include utilities such as natural gas, electricity, water, and broadband internet. Retailers send bills out to customers and providers send bills out to retailers. Given that it is quite common for each link in the chain to be a different specialised company, the systems of the retailers, providers, and manufactures are not shared. The result is that usage between levels needs to be ratified at each level back up the chain.

The effect is a high-volume bill validation problem. In the case of a retailer, a bill is received from the provider for a large number of individual customers across multiple customer groupings or tariffs. The solution is to perform a matching process between the electronic bill data and the retailers own system to determine which bills to pay and not pay the retailer. Given the volume and customer-based nature of the data many errors in the data, for example human errors (typos) and machine errors (customers no longer exist). The problem is that given minor differences in bill charge lines exact matches cannot be made for large percentages of the data. Existing solutions involve simple statistic based approaches and human-expert validation. Ultimately, large numbers of unverified bills are simply payed and the company looses money.

This problem of bill validation can be extended to wilful fraudulent acts in the case of identity theft for bills/statements or individual charge lines in financial industries such as credit card systems.

Question	Answer
What is the ultimate goal?	Detect various types of errors in bills
What is the input data?	Transactions from individual bills or total bill amounts and various bill and user details
Where does feedback come from?	Validation of previously made decisions checked manually using validation process
How is it Dynamic?	Services usage changes with time, customer behaviour changes with time (cyclic with seasons, major events, etc…)
How is it High volume?	Hundreds to millions to hundreds of millions of individual users, new transitions or bills numerous times per year
What is novelty?	Large variations may be highly anomalous for a user group yet still valid

Table 4 - Summary of the bill validation problem domain

3.4.5 Discussion

The four basic computer science examples provide some concrete understanding of the nature of the selected problem domain. There are numerous other specialised examples of anomaly and change detection including by not limited to the following:

1. Spam email detection
2. Hardware fault detection
3. Machine vision domains such as traffic and security video monitoring
4. Click fraud (fraudulent clicking of pay-per-click advertising)

It is very clear that all the examples provided meet some or all of the constraints regarding incomplete information and intractable computation. Further, it is interesting to note that the example problems all have existing solutions that assume a static problem or are static by nature using simple statistical and signature based methods. Given the ubiquitous nature of the internet, problems such as intrusion detection and antivirus have extended static approaches by including automatic update procedures that distribute updated rules and behaviour signatures. Obviously, the solutions employed by industry are more advanced than those described here, though it is clear that so-called intelligent and adaptive techniques are not in wide spread use.

Further, for specific implementation (or embedding) of a solution, it is clear that a domain expert is required. Some tasks that require an expert include the identification of relevant variates in data, data preparation, system configuration (essentially requirements for system architecture) and so on. For practical application a solution to novelty detection problems is not fire and forget which is consistent with most other problem domains, although this does not mean that the core processes of a solution cannot be generalised.

In the above cases, normal input behaviour is ever-present and anomalies or novelty is infrequent. Some of the existing approaches are good at catching many instances, though the occasional anomaly still slips through the gaps, other solutions fall far short of the expectations. In all of the above solutions, this idea of slipping through the gaps is known and factored in whether the losses are acceptable or not. Perhaps one natural approach to solving this problem is to leverage existing trusted solutions and augmenting them with an adaptive solution that can catch novelty that is currently missed.

4. Biological Metaphor

Evolution is a ruthless problem solver that is able to refine designs using directed a generate-and-test strategy over millennia. Biologically inspired computation implies a biological process or mechanism to base a solution upon, preferably a solution found in nature to a problem that closely resembles the engineering problem of interest. The immune system found in vertebrates addresses the problem of novelty or anomaly detection among other things, in what is called self-nonself discrimination.

Most complex multicellular organisms have some kind of system that protects the host from potentially harmful foreign material called an immune system. Both vertebrates (organisms with a spinal column) and invertebrates have a collection of protection mechanisms that provide the first line of defence called the innate immune system. These mechanisms include things like protective barriers like skin or armour and environments harmful or preventative to pathogens (potentially harmful material) from entering the host. The innate immune system also contains simple pattern matching or signature mechanisms for detecting pathogens. This static approach is hard-coded so to speak in the hosts genetic material (DNA) and is selected upon by natural selection at a generational level. Thus, change occurs at a species level and is somewhat slow to change relative to the lifetime of an individual. The innate immune system is fast acting, typically detecting, and responding to pathogenic material within minutes to hours.

At some point in evolutionary history, an additional complementary system was developed in vertebrates that seem to address some of the shortcomings of the innate immune system. Unlike the innate immune system that has its defence processes manipulated by natural selection at a generational level, the acquired immune system is adaptive over the course of the host's lifetime permitting somatic selection at the cellular level. The host is provided with a general system consisting of specialised organs and processes that produce cells that organise a defence for the host given a specific environment. The result is that each host as a unique environment-based defence system which ultimately provides a selective advantage for a species at a group level (heterogeneous defence across the population) and a potentially selective advantage for a host at an individual level.

The adaptive or acquired immune system is a complex biological system with many open questions that can be comparable to the complexity of the central nervous system. The problem of identifying potentially harmful material and ultimately protecting a host organism was identified as a useful analogue to the problem of novelty detection in computer science thus only a broad overview of the relevant processes and theories will be discussed.

Self-nonself is the discrimination process that the system must perform in identifying material that belongs to the host and should not be responded to (self-antigens), and foreign material that does not belong to the host and is potentially harmful (nonself antigens). An antigen is something that has the potential to elicit an immune response. Self-nonself discrimination is addressed by the acquired immune system using two primary mechanisms. These included the cellular immune system for detecting and responding to cells that have been infected by pathogenic material, and the humoral immune system that is responsible for detecting smaller pathogenic extrusions.

The following figure shows the decision processed undergone for a given unknown antigen (taken from “The Immune System”, Langman, 1989).

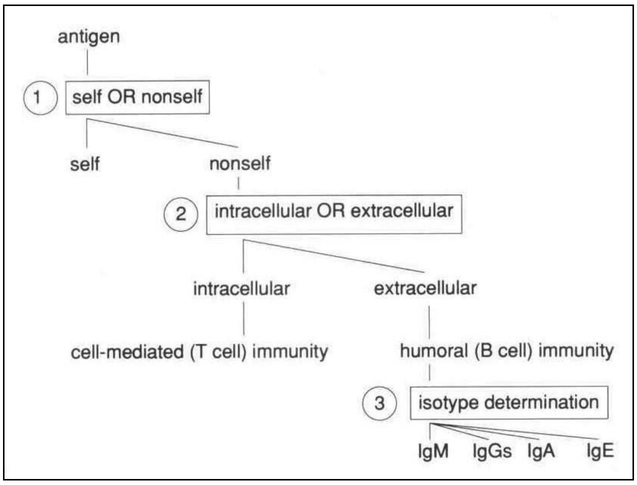

Figure 8 - Provides a simple overview of the immune system function

4.1 Negative Selection

As mentioned the cellular immune system is responsible for identifying those cells in the host that do not belong or have been overrun by pathogens such as virus or bacteria. The process involves the use of T lymphocyte cells which are a specialised type of detector cell that move throughout the circulatory system and tissues of the host seeking infected cells. The T-cells are prepared in a specialised organ in the host called the thymus. Each T-cell is unique in that is configured to recognise a certain feature of an infected cell. This preparation and configuration process of T-cells in the thymus is referred to as negative selection.

Initially the detector cells are blank with no specific configuration. The cells migrate to the thymus for maturation and specialisation before being released into the host. The cells specialise and first undergo a positive selection step – any cells that do not meet specific chemical requirements are removed (asked to commit suicide). Next, a negative selection step is performed, where the population of T-cells is exposed to self-material. Any cells that respond that is that identify self as nonself are removed. The result is a small collection of remaining detector cells that only detect and respond to infected cells.

4.2 Clonal Selection Theory

Before a cell can be infected or damaged, there is a chance of encountering and detecting the pathogen – this is the task of the humoral immune system. Again, specialised detector cells are used referred to as B-lymphocytes, which are responsible for releasing specialised detectors called antibodies that detect and bind to pathogenic molecules. Like the T-cells, each B-cell has receptors on its surface that detects a specific feature of an antigen. The cells go though a maturation process similar to the T-cell negative selection to ensure the receptors on the cells surface do not respond to self antigens. The process that permits the immune system to specialise and remember previous pathogenic encounters is descried predominantly with B lymphocyte cells and is called the clonal selection theory.

For a detector lymphocyte cell to detect a pathogen it must come into physical contact with said pathogen. A detector cell has only one type of receptor and a receptor can bind to only one determinant of a pathogen. The interesting point is that a pathogen can have many different determinants, thus a given detector can detect groups of pathogens that have similar characteristics (share the same determinant). This resulting ability to generalise called “cross reactivity” is the feature used as the basis in disease vaccination.

When an antigen is matched (selects) a detector cell, it causes the cell to chemically bind to the antigen, replicate and produce more cells with the same receptor. During this cell proliferation stage, genetic mutations occur in the clone of cells produced, some of which improve the match or affinity with the antigen. This allows the binding ability of the cell population to improve with time and exposure to the antigen. The selection of detector cells by antigens can be viewed as a type of Darwinian microcosm where the fittest cells (best match with antigens) are selected for survival through proliferation, and genetic mutation provides cell variation. Those cells that are not selected are removed from the system (eventually die).

The figure below (taken from “Introduction to Immunology”, Kimball, 1983) provides a good basic overview of the clonal selection process. The image shows B lymphocyte cells at the top that bind to specific antigens. Once bound, the cell proliferates (divides or mitosis) and produces many B lymphoblasts which differentiate into either plasma cells that produce antibodies (effector of the immune response), or long lived memory cells (used if the antigen reappears).

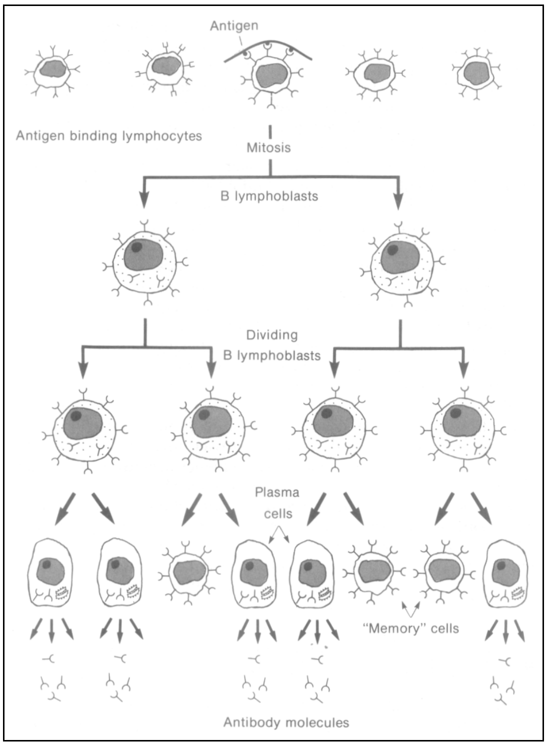

Figure 9 - Provides a simple overview of the clonal selection process

4.3 Discussion

The overview provided of the immune system is a gross oversimplification of the actual biochemical processes that occur and was intended as such to provide a crash course for the selected metaphor. Regardless, there are a number of interesting take-home points regarding the described biological system.

To summarise the immune system consists of an innate or built in component and acquired or learned component. The built in component is the primary defence and catches or protects against the majority of potentially harmful antigens. Those foreign antigens that make it past the innate system are handled by the acquired immune system. The acquired immune system is the interesting adaptive component and has processes for preparing cell populations and processes for maintaining cell populations.

The acquired immune system is a distributed system in two senses. Knowledge in the system, that is successful detector cells are distributed and move throughout the host organism. New detectors are being generated and old detectors are being removed all the time providing partial coverage of the pathogen space everywhere and complete coverage in no single location in the host (a pathogen can enter anywhere in the host). The system takes a bottom up approach to control meaning that the system is controlled through local interaction or local rules rather than in a central location. This is a description of a robust distributed system of redundant individual components that facilitates innate parallelism for the task of self-nonself discrimination and subsequent response.

The system is somewhat effective in biology in that it was deemed required and discovered by natural selection, and exists in vertebrates such as humans that live a reasonable length of time. Innate immunity came first permitting the acquired immune system to evolve, suggesting that perhaps a static solution over a host's lifetime may not have been sufficient for the environment in which vertebrates lived. It is further interesting to note that the immune system is a solution for a set of related problems though it is not necessary an optimal solution. The reason for this lack of optimality for this is that hosts die from many different causes and thus the scope for natural selection is too broad to optimise a single biological subsystem.

It is also interesting to note that there is an evolutionary arms race between invading pathogenic material such as viruses and parasites and the immune system. This arms race or "red queen" race in which each side is required to constantly and quickly adapt to overcome or out perform its competitor. The result is a system that may be optimised to exploit specific features of the hosts biological nature of the environment in which the host lives. An example of such an environmental bias or specialisation was the susceptibility to disease of native peoples in South America and Australia from white settlers in which large percentages of native peoples had no innate defence.

Another interesting feature of the system is that of feedback. Given an organism with only an innate immune system, the species or groups of individuals receive feedback from natural selection as to the effect of genetic variations. This idea of feedback through survival can be extended to the success or failure of properties of the acquired immune system. This is a critical point that needs careful consideration when using such a system as a metaphor for solutions to engineering problems. The point being that feedback through death or survival of an organism (system) is too drastic for practical consideration.

The system is adaptive in that it can learn or specialise a defence to an environment. This means that the more times a host is exposed to a specific pathogen the better improved the response of the host to the pathogen. It also provides a useful level of efficiency in that the system maintains a specialised defence or memory against those attacks that are likely or have already occurred, rather than requiring specific machinery such as mechanisms in the innate immune systems against every possible attack. The obvious trade-off is that between being able to cover enough of the pathogen space in a general manner as to not leave too many or too large a holes in the defence open to attack.

It is clear that the immune system is a collection of powerful mechanisms and the acquired immune system in particular has many desirable properties. There are number of open questions as to how to best generalise and apply the metaphor as a solution to the selected problem domain of novelty detection. The concerns highlighted here show that perhaps understanding the metaphors evolutionary origin and asking questions as to why specific features were selected in the biological system rather than blindly adopting or modelling said features may provide useful tools when applying the metaphor .

5. Thesis

Research Problem: Given the stated problem of dynamic discrimination of novelty from non-novelty on a high-volume input scenario, propose a solution inspired by features of the acquired immune system. Demonstrate that the characteristics of the proposed solution meet the requirements and constraints of the selected problem domain.

Existing solutions for specific cases of the selected general problem of novelty detection can be considered equivalent to an innate system. They are explicitly designed and tested, they have evolved through trail and error by domain experts which can be considered a type of natural selection.

Given that an innate solutions exist and can be adopted by new problem cases (same problem in different specific implementation), and given that innate solutions still permit novelty to go undetected, I propose an adaptive complementary system. This analogy is attractive to a point. The biological innate immune system has a high-cost associated with it in that it is hard-coded in DNA and changes in the system are generated and tested at a generational level. The same high-cost occurs in practice in that rules and signatures needs to be investigated, recognised from past events, captured, described, and propagated to areas of interest for implementation. This work is performed by domain experts and is costly in monetary and time concerns. A process that is autonomous and is able to adapt and acquire the same or similar knowledge required with less cost is highly desirable.

5.1 Problem and Metaphor

Self-nonself discrimination in the biological acquired immune system is a specific type of novelty detection and like other specialisations mentioned above such as spam email detection and hardware fault detection, it has constraints specific to the implementation. This is an important observation as it clearly indicates that replication of immunological mechanisms in a computer science domain directly is questionable. Thus using the argument "*because the biological immune system does it*" to defend a system characteristic is not a valid argument.

The benefit of building a system that is inspired by a biological system rather than modelling a biological system is that it is open to interpretation and modification. Those processes or mechanisms that are not appropriate can be dropped, and those inadequacies identified in the biological theory for the system can be circumvented. The result should be a system with purely engineering level limitations, again negating the argument "*the biological immune system is flawed in this way thus so is the inspired system*".

The above points somewhat clarify the line between modelling and inspired solution and provide useful tools in selecting how much or how little of the biological metaphor to bring to a practical implementation. The selected metaphor is a good fit for the selected problem domain given the closeness in underlying problem already shown. Further, many of the characteristics of the acquired immune system are desirable characteristics for a solution to the proposed novelty detection problem given the constraints specified and the defined requirements for a given solution.

As mentioned the two primary concerns with this work are the two constraints; large volume of data and a dynamic underlying probability density function. It is useful to highlight the ways in which the selected metaphor addresses these constraints as follows.

5.1.1 Large Volume of Data

A large volume of input data implies that the proposed solution must be scalable, that it is able to increase in capacity in proportion to increases in data volume and maintain an acceptable level of accuracy. One way in which the biological immune system demonstrates this is though the various forms that have such a system. From elephants to mice, vertebrates of all sizes have the same fundamental system architecture. A larger host organism means that there are more foreign antigens entering the system and a wider area to physically defend. This shows that by allocating more resources for larger scale problems or fewer resources for smaller systems that the same architecture can adequately perform the same task. The difficulty is that it is not reasonable to correlate effectiveness of the system with size of host organism as organisms are not directly comparable given the many differences in environment and behaviours.

Another area related to scale is that of the size of the problem domain, that is the size of the normal and abnormal multivariate spaces. This does not have a bearing with the volume of input data though is an important feature nonetheless. As the size of the problem space increases the probability of useful coverage by the system is expected to decrease thus, the number of detectors must be increased. This too could lead to a situation where the input data load from a single data stream may be too much and thus must be split between a collection of novelty detection systems. It is interesting that the proposed distributed architecture somewhat resembles the biological immune system in that the collection of all novelty detection systems together can be considered the host, where individual detector units are capable of migrating freely throughout the distributed system.

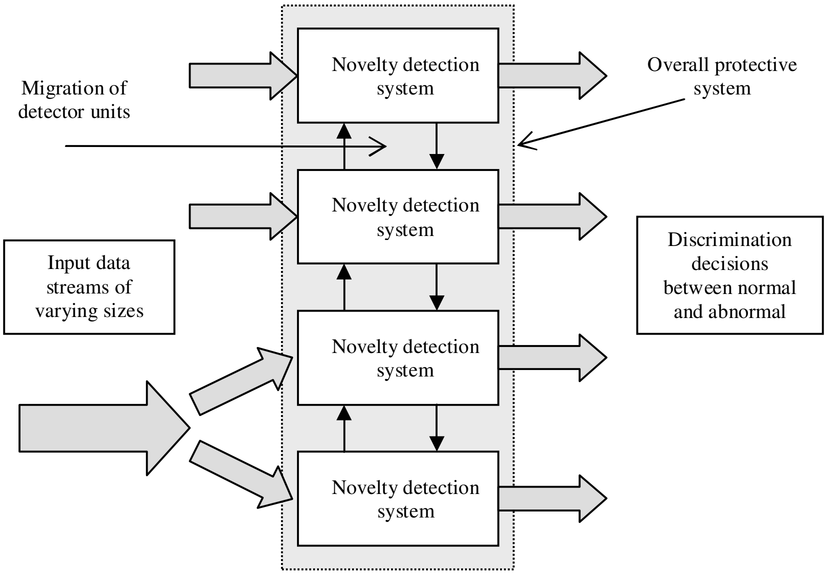
Migration of detector units
Novelty detection system
Novelty detection system
Novelty detection system
Novelty detection system
Overall protective system
Input data streams of varying sizes
Discrimination decisions between normal and abnormal

Figure 10 - Shows an example of a scalable system architecture with multiple input data streams and detector systems and

5.1.2 Dynamic Problem

A dynamic problem means that the underlying probability density function is moving or changing in some way over time. Fundamentally, this does not occur for the biological immune system. The model the biological system has of normal is complete and static, thus the task of the system is to anticipate attacks and remember past attacks by pathogens. The obvious question has to be asked – that is if the immune system has complete information regarding self, why not use complete information to detect any pathogen. The answer is less than simple, though one big factor is likely to be that of cost. The question as to why the biological system models the problem in the complement space is an open question though is likely based on the evolution of the strategy and the restrictions of interactions between cells and antigens.

To understand the cost argument, one must understand the nature of complete information in the biological immune system. A naïve understanding would equate complete information to the definition of all self-antigens and all potential non-self antigens. This is an unreasonably large amount of information as the spectrum of nonself antigens not only covers all known pathogenic material, but all pathogens not discovered and those not in existence yet (the system has demonstrated the capability to detect man made pathogens that have not previously existed on earth). A more succinct definition is to take complete information as the collection of all self-antigens and only those pathogens a host encounters during its lifetime.

This is a much smaller and thus simpler problem to model in terms of cost, and is precisely what the immune system attempts to do. It is a question of coverage. It is less costly in terms of resources and less risky in terms of provided protection to model expectation than it is to model the entire possibility space. This is the crux of the solution and shows that the problem of self-nonself discrimination faced by the immune system is dynamic in that the pathogens faced by the host change with the environment and behaviours of the host. Thus, the approach used by the immune system could be considered the modelling or anticipation of the most likely pathogenic events, yet still providing some level of general coverage for outlier events.

It has been shown that the system has the capacity to learn though experience. Further, the system has shown a useful level of plasticity in its ability to specialise a defence for a host for a given environment (and speculated as given a predisposed bias for that specialised defence). The biological system is capable of modelling a dynamic problem space, though the critical point in which the metaphor and defined problem domain separate is in the model of self. As mentioned the model of normal behaviour in the biological immune system does not change thus the chance of attacking self (autoimmune disease) is not very likely. This is not the case in the domain of novelty detection described. Not only does the model of likely attacks change with time, the model of normal behaviour of the system is subject to change. The degree in which the changes to normal input behaviour are tracked by the system is problem specific, though must be recognised and explicitly handled.

5.2 The Proposed System

To summarise, the premise is that existing novelty detection problems can be equated to an organism with only an innate immune system desiring further defence. It was proposed that natural selection came to a similar juncture and developed the acquired immune system. In using this juncture as an analogy of goal of this work there are a number of ways to proceed:

1. Reverse engineer the solution found by natural selection and use it as a general template
2. Setup conditions for such a system to develop and let such a system design itself
3. Some hybrid of the first and second approaches

The first approach has been the focus of work in the area of artificial immune systems to date, more specifically; looking at sub-processes of such a system as will be described shortly. Letting a system develop itself is an interesting idea, though is quite open-ended. The third is very attractive, specifically the constraining of some processes or structures of the system whilst allowing other processes to develop via adaptation. It must be clarified that what is being proposed is a complete system. The system itself is adaptive over the lifetime of that system. The requirements of the proposed general system are clear, thus taking the role of intelligent designer is appropriate. What is being proposed is a system-based solution and not an algorithm. In this case, an algorithm is considered a single sub-process of the proposed system, whereas the system is complete and is embeddable for a specific problem domain. In the role of intelligent designer the task is to somewhat reverse engineering the acquired immune system – that is to design a solution based on theoretic acquired immunology, and not model such theories.

The system has input data in the form of samples from some problem space. Output is required from the system for each provided input sample in the form of a classification decision as either normal or abnormal. An acquired immunological based system proposes the use of a discrete detector based framework. Three principle processes exist for the framework that have already been lightly discussed in a biological context:

1. *Detector Preparation* – A process for the preparation of detectors inspired by the negative selection theory for preparing lymphocyte cells
2. *Detector Maintenance* – A process for the maintenance of detectors (detector migrations, creating new detectors, and killing old detectors) inspired by equivalent processes used to maintain lymphocyte cells
3. *Learning Process* – A process to coordinate the activation and specialisation of detectors inspired by the clonal selection theory for lymphocyte cells.

Other required processes can be inferred from the proposed system based on arguments already presented, as follows:

1. *Discrimination Process* – Each unit of input data must pass through a matching process and be evaluated as either normal or abnormal (some degree of novelty).
2. *Model of Normal Maintenance* – If the model of self can change then the changes that occur must be tracked.

The two most important processes from a problem perspective are matching and learning. The system exists to discriminate between normal and abnormal – to detect novelty. The manner in which this task is performed is secondary within the scope of the specific problem domain. The second important process is learning. This process is not as important as the first, though exists to permit the decisions made by the system to improve with experience. All other processes exist to support the first process and to a lesser degree the second process.

5.2.1 Discrimination Process

Broadly speaking, the system receives input data and makes a decision as to whether the input is normal or abnormal. This discrimination process involves evaluating input data against an internal model and assigning some kind of confidence as to the degree to which the data belongs to each class. Such a system must make use of as much domain specific information as possible to be practically useful. This includes making use of existing innate processes. A likely breakdown of this discrimination process is as follows:

1. Input data enters the system in discrete units and each input unit is processed independently
2. Data enters the innate systems for evaluation. If the data is novel a concern is raised to an external process and feedback is received
3. If the data is perceived by the innate process as normal then it enters the proposed acquired system
4. Previous experience is used to determine if the input data matches previous identified novel pattern, if so a concern is raised and feedback is received

5. If the data is perceived by the memory units to be normal then it enters the current detector section
6. Input data is compared to the current detector units, if the data matches any units then a concern is raised and feedback is received
7. If the current detector units determined the input data to be normal then a system output of normal is provided

The following figure shows an overview of the described process.

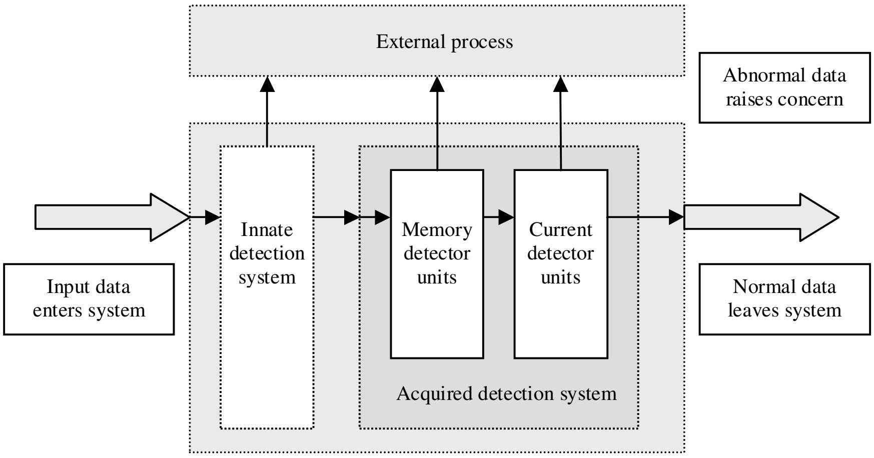

Figure 11 - Overview of the discrimination process for a single novelty detection system

5.2.2 Learning Process

The learning process employed is based on the process used by the biological acquired immune system to improve its specificity to encountered pathogenic material. As mentioned the theory used to describe the biological process is clonal selection theory. Using this theory as a base an abstract overview of the process can be defined as follows, assuming a point of input data is identified by current detector cells to be novel:

1. A concern is raised with an external process regarding the input data point
2. The process provides feedback as to the validity of the claim of novelty
3. Depending on the validity of the raised concern, two actions can occur
    a. If incorrect, then the current detector systems are wrong and the offending detector is discarded. This may also involve starting a process that seeks or filters detectors in other systems and repairs the inaccuracy
    b. If correct, the detector goes through a proliferation and maturation process.
4. A clone of detectors are prepared where some progeny are mutated to improve the specificity to the causal input
    a. The detector unit that made the identification or one of its more specific progeny are duplicated and become memory cells capable of more effectively recognising offending input data in the future
    b. The better detector progeny are added back into the detector pool for active service

The following figure provides an overview of the detector specialisation (learning system) process:

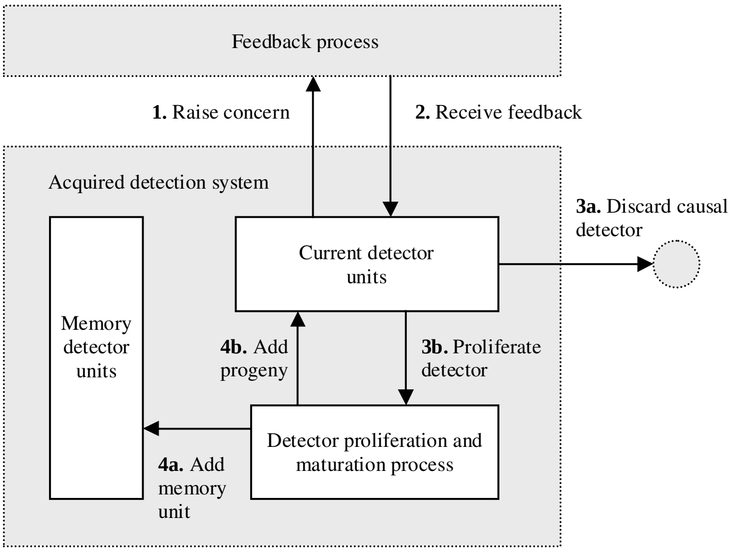

Figure 12 - Overview of the learning process for a single acquired detection system

The explicit definition of the process highlights a number of interesting features; the integration of feedback as knowledge and the feedback agent.

5.2.2.1 Feedback Integration

The process shows that knowledge is captured in a “memory detector pool”, though this does not have to be the case. It will be shown that it is possible to integrate knowledge into the detector preparation process. Further, it is possible that the knowledge could be directly captured into the existing innate processes. This is a powerful idea because it implies that the proposed adaptive system could provide a means for automatically identifying new rules or behaviour signatures for innate systems, which in turn could be distributed to other interested parties. A final interesting way in which the knowledge from feedback could be captured is in a decision making process within the current detector pool. This new sub-process could capture the knowledge in a data structure of some kind, which is consulted by the current detector pool before raising a concern to the external feedback agent.

To summarise, three means of feedback integration have been specified, though the system is clearly not limited to just these three. The default process of promoting successful units to memory cells will be taken as a default process.

5.2.2.2 Feedback Agent

It has been discussed that the biological immune system receives feedback on decisions at a generational level. All responses are taken as correct. If the system is in error then it is up to internal regulatory processes to fix the problem or the host ultimately dies. This permits the strategies used by hosts to be general, yet still possess some kind of locality bias. Such a feedback mechanism is not useful in an inspired system. It is too late if the system requires itself to fail before changes or useful knowledge can be captured. Instead, some kind of real-time feedback agent is required to assess the accuracy of decisions made.

The nature of the feedback agent constraints the nature of the solution, thus the identification of an agent or agents is crucial. For example if the feedback agent is a human, then the number of false positives must be minimised to instil trust in the system. A confidence measure or some problem specific measure such as monetary cost will have to be used to order possible concerns to ensure the agent spends time on those concerns that add value in the context of the problem. An alternative approach would be to have some automated process such as a database lookup or costly computation. There may or may not be some kind of answer oracle for a given problem, though its automated checking procedure permits higher volumes of concerns in which the "humanity" constraints of agent does not need to be taken into consideration.

It is interesting to note that in the case of a non-oracle feedback agent, there is chance that the feedback received has error and this must be taken into consideration when integrating knowledge into the system. The system has to be able to reinforce knowledge, though must also be able to forget or loose past knowledge. This will be further discussed in the model of normal behaviour process, though it is clear that again, the specifics of the problem domain define the constraints and thus configuration of the system.

5.2.3 Detector Preparation

Detectors are immutable after being produced, this means those detectors that are created must be accurate and useful. As mentioned the detector generation process is modelled on a theory used to describe the biological equivalent in the acquired immune system called negative selection. This process can be generalised to the following steps:

1. Using some stochastic process generate a new detector unit
2. Check that the detector unit is useful (positive selection)
3. Check that the detector unit is accurate (negative selection)
4. Add the remaining detectors to the pool of current units

An implementation of this process may be as follows. Generate a batch of random detector units. Using some prepared rules ensure that the detectors are in or close to regions of expected novelty (bias using domain knowledge). Check that the detector units discriminate correctly by using a prepared model of normal input behaviour. Finally add the prepared units into the pool with previously prepared detector units. It is important to note that some equivalence of this process will be used on progeny detector units created during the learning process described previously.

Two interesting points are highlighted from the description of this process: bias in detector generation and model of normal behaviour.

5.2.3.1 Bias in detector generation

Each detector has some specific coverage in the multivariate problem space. This coverage may be a single point or some area, though each detector is pseudo-unique in this regard. In step two, the process imposes a positive selection on the pool of prepared detectors, selecting for desirable properties. As mentioned, these properties can be as simple as a domain-knowledge imposed bias towards high-probability areas for novelty in the problem space.

The power of such a bias is subtle. An intuitive example is to implicitly encode learned knowledge into the population, and thus allow the explicit memory pool to be removed from the system. Knowledge is captured in the process used to generate new pseudo-random detectors. The process then generates batches of detectors, many of which exhibit learned variates or combinations or learned variates. The same degree of centralised control is held over the knowledge provided though the feedback agent, although specialised processes may be required to manipulate detector populations if certain knowledge requires an immediate purge. Perhaps some combination of both an explicit memory and implicitly memory would provide a desired level of flexibility.

5.2.3.2 Model of normal behaviour

The detector preparation requires some model of normal behaviour. This fact is captured in the name of the immunological theory that describes the biological process – negative selection. That is, detectors are selected for death if they respond to normal input patterns. The extent and accuracy and maintenance processes for such a model are described as a standalone governing process further on.

5.2.4 Detector Maintenance

Thus far, two specific detector populations have been described for a single novelty detection system, and the architecture proposed facilitates multiple parallel detector systems operating on independent or split, high-volume data streams. Processes have been described for detector construction and specialisation, though what is required are processes for maintaining detectors over the course of their lifecycle.

As described, discrete detector units are designed to be immutable, though the knowledge captured by the system regarding the underlying problem domain is mutable. Thus, the system needs to be able to not only devise new strategies over time, it also needs to remove or forget recently redundant knowledge. This section describes a strategy called general coverage that is inspired by the biological immune systems strategy to provide robust coverage using randomly generated discrete units using high numbers of detectors, detector lifecycles, and detector migration.

5.2.4.1 Arms Race – Red Queen Race

It has been shown that the natural acquired immune system is in an arms race with pathogenic material. The problem is not fixed, it is always moving, a situation that also describes the dynamic novelty detection problem being addressed. One strategy used to address the continuous change is to have the system in a state of flux.

As mentioned, integrated feedback is stored in the system as knowledge or memory and is fully exploited. Domain specific knowledge is exploited to provide a bias to generating detectors in regions or with coverage with higher probabilities of usefulness. Given limited resources, it is not possible to try all possible detector combinations, especially as the dimensionality of the problem increases. It is proposed that a strategy of approximate coverage is used where the system is continuously generating new detector units for application. Thus, the pool of current detectors in each novelty detection system is in a state of constant change, having both its specificity increased through the learning process and new detectors continuously added.

5.2.4.2 Limited Resources and Detector Death

The biological immune system does not have unlimited resources to address the problem of self-nonself discrimination, and nor does an implementation of the proposed system to address specific problem cases. Thus, the described pools of detector units must have a regulation process to control their size. A process is required for both the memory detector pool as well as the current detector pool. Further, the processes must be able to move the system to a state of equilibrium in times of rapid change.

A example of such a state of disequilibrium is when there is a sudden increase in novelty in the input data stream. This will result in a large number of memory cells being created as a well as an increase in specific detector units into the current memory pool. Each pool will have a preconfigured ideal size, which the process will seek to return to, though given the dynamic nature of the underlying problem, hard limits for maximum resource consumption will be high or removed.

A suitable process would evaluate the nature of newly added detector units in the context of detector units already present in the pool. Only those detectors that the process deems to add value (are sufficiently different for example) from those detectors in the pool will be added. An alternative approach may be to replace like-detectors thus maintaining a desired level of detector heterogeneity within the pool. Another example is that of normal system usage where the arms race process provides a periodic increase in detectors. To handle this case, each detector in the current pool could have a last used timestamp or creation timestamp and the process could simply retire unused or elderly detector units.

5.2.4.3 Robust Coverage though Detector Migration

The strategy for the proposed system is approximate coverage though continuous change. This is a robust strategy that still relies on large numbers of detectors. The beauty of this strategy as personified in the proposed system is that the collection of distributed novelty detection systems represents a single host organism. Like the biological immune system from which the system is inspired, the detectors require physical contact with input data to detect it. The second point of this strategy is that detectors move around the host organism to provide general coverage everywhere and specialist coverage nowhere.

The migration of detectors provides another means of gaining new units to both the memory and current detector pools. Moving around memory units permits learned knowledge to be exploited across the entire system. Migration of newly generated detector units may or may not be useful, though the sharing of specialised detectors from the learning process may prove beneficial. A migration process is expected to duplicate randomly selected detector units and distribute them to neighbour novelty detection systems.

Such a process raises some interesting concerns. In the case of memory detector units, it is desirable to ideally have all learned knowledge available everywhere. Using a lazy migration process in this case may provide results less optimal than a process that explicitly replicates all captured knowledge throughout the novelty system network. Concerns are also raised regarding the connectivity between such systems and the locality of such systems. Many connectivity patterns exist to choose from in network theory, and a pattern should be selected that best fits a given problem domain and available resources. In terms of system locality, it may be desirable to partition data streams so that only systems working on similar domains (such as on feeds in the same geographical area) share knowledge. Further, it may be desirable to partition data using a stream level filter so that an upfront process dispatches input data to the relevant system or system cluster.

It is clear that the selected strategy of generalised coverage, in conjunction with the selected modular system architecture is remarkably flexible.

5.2.5 Model of Normal Behaviour

The detector generation process requires a model of self to perform the negative selection step – that is to prepare detectors that do not respond to normal input patterns. Thus, the accuracy in terms of false positives can be attributed to the accuracy of the model of internal model of normal activity. In the biological immune system, the model is static, and is maintained in a single location – the thymus in which the negative selection process for T lymphocyte cells occurs. The process described for detector preparation does not specify whether that process is centralised or decentralised.

It is attractive to have the model centralised, and it is possible to permit an implementation of the detector preparation process that is local to each novelty detector system that uses a centralised version of the model. Again, this shows that the proposed architecture if flexible to the whims or requirements of a specific implementation. The benefits of a centralised model of normal behaviour become apparent when a problem domain is considered that does not have a static model. In this case, a process is required to sample proven normal input data and update the model accordingly. Further, in the case that negative feedback is received from the feedback agent (a concern regarding novelty is shown to be false), the model of normal behaviour must be updated accordingly. This is to ensure that detectors are not produced that make the same flawed discrimination.

There is also the concern of model size in terms of allocated resources. This too will require maintenance processes that like the similar processes on the detector pools, cull unused or decayed elements. Distributing, the model of self opens the system to further holes in its protective coverage. It implies that both models of normal and abnormal are approximations at a single attack entry point, it implies a multiplicative effect for errors from approximations, and further implies that both systems suffer from delays in knowledge propagation throughout the detector system network.

5.3 Discussion

A framework has been proposed for a novelty detection system that is expected to meet the constraints and requirements proposed for the selected problem domain. The system architecture has been defined in terms of a set of required processes, each of which can be addressed by a specialised algorithm and data structures. It is worth again highlighting some of the finer desirable properties of such a system in terms of addressing the specific requirements of interest: large data volume and dynamic domain.

The proposed distributed system has borrowed some of the desirable characteristics of the biological metaphor from which it was inspired. In constructing a system of discrete read-only autonomous building blocks, the processes performed on the system are inherently parallelisable. By distributing control and distributing knowledge in discrete redundant units, it is possible for portions of the system to fail or be destroyed, leaving the remaining modules to naturally pick up the slack. Further, the modular nature of the system permits resource capacity of individual modules to be manipulated in real time, and entirely new detector systems to be integrated into the detector system network while the system is live.

Algorithms that implement described processes are expected to take such advantageous characteristics into account, facilitating the most desirable features from the underlying model. From a engineering vantage, it is an exciting architecture because it supports the easy integration of additional complementing distributed computing patterns. For example, it is possible to implement various processes using distributed queuing systems and use a dispatcher (producer-consumer) structures that permit dynamic load balancing of data streams across the novelty detector network. Further, multicast technologies can be used in the detector network that support features such as automatic discovery of additional system nodes as well as recovery and failover of dynamically removed nodes. These are just a few implementation-specific benefits of the proposed system framework.

6. Analysis as Evidence

A lot of conjecture has been provided as to the ability and function of the proposed system and arguments have been made related back to defined problem constraints and requirements. The arguments provided must be supported by evidence, and as defined in the methodology evidence can be either qualitative (analytical techniques) or quantitative (empirical techniques), and that both types of evidence are required for a given argument. This section describes tools and potentially useful analysis techniques for the prosed system that are expected to provide relevant supporting evidence.

6.1 Analytical Analysis (Qualitative)

In the field of biologically inspired computation, it is common for the analysis to be biased towards empirical observations, more than quantitative approaches. The reason for this bias is that can be very difficult to analyse such techniques because of both their stochastic elements and their inherent complexity. Regardless, there are still qualitative analyses that can be used to both model specific system behaviours and elucidate concerns such as efficiency. The majority of these analytical approaches are founded in discrete mathematics and applied probability.

This section will not cover all possible useful analytical techniques, rather just those few that have been identified as being relevant to the selected problem domain. The intent of using analytical approaches is to formally propose descriptive models of system behaviour. Models are typically exploratory and inductive, and are used to provide a general sense of what is happening.

6.1.1 Computational Complexity

In this case computational complexity theory can be used to address two specific concerns: the efficiency of a given algorithm, and the difficulty of a given problem.

Complexity theory provides tools to evaluate how much resources are required by an algorithm to address a specific problem. The asymptotic behaviour of an algorithm can be defined in a formal notation in terms of algorithm bounds for space and time complexity. Space complexity refers to the amount of memory used to during the execution of the algorithm, and time complexity is the number of steps the algorithm takes over the course of its execution. Common asymptotic measures include the big-**O** (upper bound), big-**Θ** (average case) and big-**Ω** (lower bound). Measures are reduced to one of a common set of forms, thus permitting algorithms to be directly comparable for a given specific problem.

A given problem can always be reduced to a decision problem (a problem with a yes or no answer), and complexity theory deals primarily with decision problems. Decision problems can be reduced to one of a number of complexity classes that provide an indication of the difficulty of the problem. Example is P for problems that can be solved in polynomial time complexity and NP for those problems that only have no solutions that are nondeterministic polynomial time complexity. That is, NP type problems are hard and can only be solved using a nondeterministic machine in polynomial time. Examples include those n factorial problems or problems with complexity with n as the exponent.

These tools can be used to evaluate algorithm implementations of defined processes within the system, or the system itself. Further, these tools provide a means of demonstrating either the difficulty or intractability of the problem domain or a specific case of the selected problem.

6.1.2 Competitive Analysis

Online problems are those problems, which cannot be solved offline that is there is insufficient data or information to address the problem in an offline manner. Further, online problems have data provided in discrete pieces, one at a time. Algorithms that address such problems are called online algorithms. The selected problem domain can be considered an online problem, given that complete information is not available for offline computation, the domain is dynamic, and information regarding the underlying probability distribution is provided in discrete units.

A technique used to assess the performance of online algorithms is called competitive analysis. Competitive analysis is a data dependent means of measuring the worst-case performance of an online or stochastic algorithm. The theory assumes that an optimal offline algorithm (algorithm with all available data) exists for the problem that can be used as a base measure of comparison. The intent is for the online algorithm to perform as close to the optimal offline algorithm as possible. The technique assumes that an adversary exists which provides difficult data to the algorithm. Different types of adversaries are used from those with knowledge of the algorithm being used, and those that have knowledge of the specific state and function of the algorithm. Thus, online algorithms commonly have a randomised component to overcome the imaginary adversary, and typically outperform a deterministic algorithm for the same problem.

This analysis approach provides a c-complexity measure that can be used to rank online algorithms for a specific set of data. The theory is attractive given its direct applicability to the chosen problem domain and proposed solution. Further the adversary concept provides a useful metaphor to be analysed for embedded systems for such problems as antivirus and intrusion detection as it provides a way of thinking about possible weaknesses of such an embedded system.

6.1.3 Applied Probability

Probability provides a powerful tool in modelling and analysing a complex system. In the case of the proposed system, probabilistic models can be used describe behaviours of specific processes in the system over time, such as protective coverage, holes in coverage, system equilibrium, and models for system performance given various configurations. Markov chains and hidden Markov models can be used to model the stochastic properties of the system, again at a process or system wide level.

6.2 Empirical Analysis (Quantitative)

Empirical evidence by definition is confirmatory. This form of quantitative analysis is implicitly deductive in nature and us used to prove or validate analytical models and theories. In the case of the proposed solution, empirical testing involves isolating portions of the system and testing them against specific constraints and requirements. For example, a bottom line concern for such a system is accuracy – how well does the system perform? This is a broad question, and can only be satisfied by providing arguments and evidence as to the conditions under which the system will perform well and to what measurable level of accuracy the system can achieve.

Performance is the bottom line concern, though must be demonstrated in relation to the constraints of the problem domain. Examples include the performance of the system under data loads of varying size, and system performance with a PDF with certain characteristics. It is clear that specialised problem domains must be constructed for these specific tests. It is useful to test for specific functionalities both independently and combined. Testing of the system on data of which little is known about its underlying features, adds little value to any argument.

It is proposed that each analytical model be backed by empirical observation (if possible), and each requirement of the solution be demonstrated with empirical observation. Further, empirical analysis can provide a basis for implementation based heuristics for configuring the solution in the general case.

7. Conclusions

The problem of high-volume dynamic novelty detection was clearly defined. A metaphor was identified that addressed a closely related problem and a solution was proposed that is expected to address the constraints and requirements of the selected problem. Some arguments and a lot of conjecture was provided as to the performance and satisfiability of the proposed solution, all of which and more must be addressed with evidence and further refinement.

The contention is that an immunological inspired solution can address practical cases of high-volume dynamic novelty detection. The proposed solution is a general system or framework that has numerous processes that control its function. The framework is a skeleton that needs filling. There are many open questions as to how the defined processes should be implemented and configured, as well as the performance of such an implemented system. This work will occupy the proposed research investigation.

The methodology defined proposed a continuous surface which defined the area of study. This work has lightly visited both axes in both directions and has suggested some tools and areas for further research that may aid in the tasks ahead. By no means has the surface being covered, and by no means will all the regions of the surface that are covered be included in the final dissertation. The methodology simply provides a framework for organising knowledge for the selected project and identifies strengths and weakness in arguments for a proposed contention. Figure 13 rounds up the discussion highlighting some of the visited landmarks for the selected area of study.

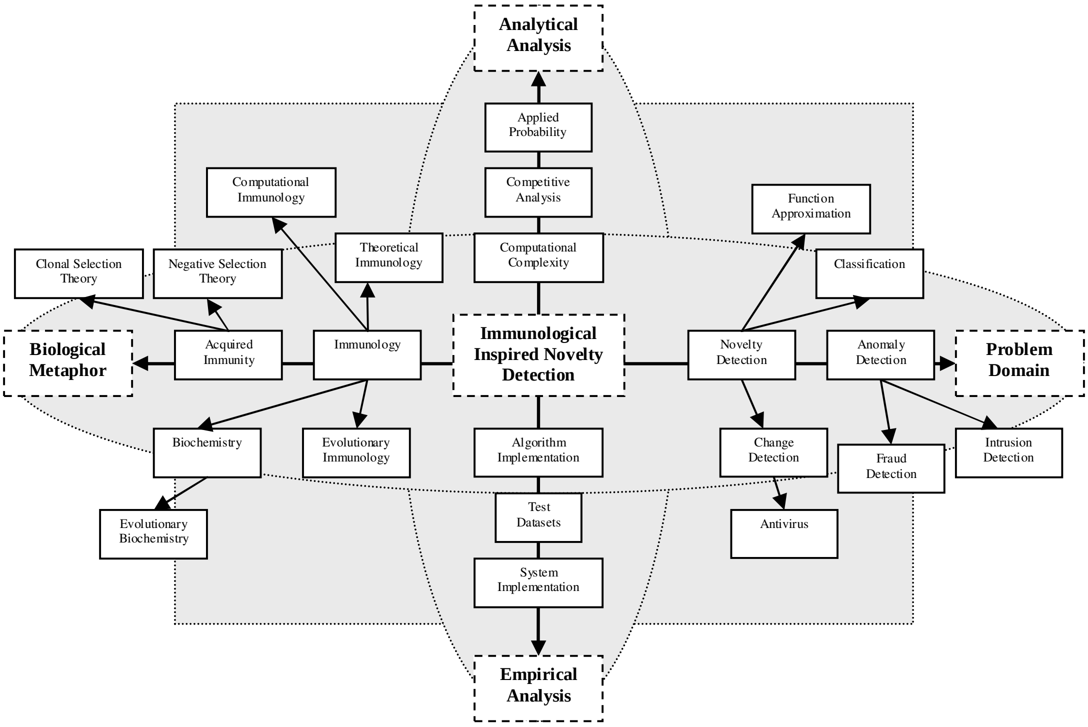

Figure 13 - Overview of the proposed research project in terms of general areas that may be useful