Hypothesis Formation and Tracking in ARGUS

B. Cenk Gazen, Jaime G. Carbonell, Philip J. Hayes, Chun Jin, Eugene Fink
Carnegie Mellon University and DYNAMiX Technologies
NIMD PI Meeting
November, 2004

Introduction

New Hypothesis formation may be an analyst-initiated activity, an automated process whereby a novel trend is discovered and tracked, or a hybrid one. In the hybrid case, the system offers its discovery of novel, potentially interesting patterns for analyst review, leading to new hypothesis being formed and tracked, or to discarding the novelty as coincidental or uninteresting. The ARGUS project assumes the third paradigm, where a combination of analysis of massive data – both historical and real-time streams – leads to automated creation of potential hypothesis for analysts to consider, discard, embellish, combine, and/or instruct ARGUS to track as a new persistent interest profile.

Novelty Detection

ARGUS assumes that the genesis of a new hypothesis is essentially data driven, often as a need to explain the onset of a new pattern or a previously unforeseen change in an established one. For instance, if shipments of certain precursor chemicals consistent with the production of nerve agents directed to a new location in a potentially hostile country start at a certain date and were not observed before, a hypothesis that something new, possibly of a nefarious or perilous nature, is being produced at that location needs to be considered. In ARGUS, this would be detected as a new emerging cluster distinct from background clusters in a data stream.

As a different example consider the outbreak of a disease like SARS, which the medical community was slow to recognize because SARS symptoms clusters were masked by similar cold or influenza symptoms. In this case we need to detect a change in existing clusters – a much greater percentage of patients do not recover within the expected time frame, for instance. This requires detection of change in the density function of a cluster, rather than the onset of a clear new one – i.e. we need to perform a de-convolution process to detect a new component in a mixture of observations. We believe that the second case may be more common, either though accidental masking (as in SARS) or intentional obfuscation, such as combining legitimate medical facilities with potential bio-weapon R&D.

Cluster Density Functions

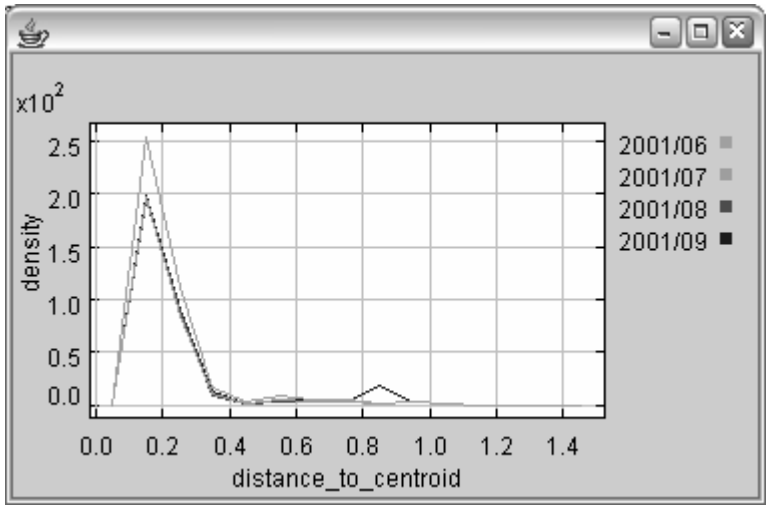
To detect changes in the shape and density of clusters, we analyze the density of points within a cluster as a function of the distance to the cluster's centroid. More formally, we define the density function as $f(r) = dM(r)/dV(r)$ where $M(r)$ is the number of points within a sphere of radius r and $V(r)$ is the volume of that sphere. Another way to view the density $f(r)$ is as a spatial differential, i.e. the number of points per unit volume on the

thin shell of a hollow, n-dimensional sphere of radius r centered at the cluster centroid. This gives us a way to approximate efficiently the density function: we quantize the sphere into a number of shells, count the number of points that fall into each shell, and finally divide this count by the volume of the shell. The density function characterizes the shape and density of the cluster. The peaks and valleys of the density function correspond to dense and sparse regions within the cluster. For example, the density function of a cluster whose points are uniformly distributed with a density of c within a sphere of radius r is simply f(x)=c for 0<=x<=r and f(x)=0 otherwise.

By tracking changes in the density function over time, we can detect changes in both the shape and density of clusters. For example, if a new cluster forms but is masked by a larger cluster, the density of the points around the new cluster increases. The increase in the local density shows up as a new peak in the density function of the larger cluster.

Epidemiology Example

To demonstrate novelty detection by density function analysis, let us look at a hypothetical SARS outbreak. For this example, we picked a large cluster of influenza patients and added a set of new patients who live around the same area and were diagnosed with similar symptoms in September 2001. The clustering algorithm places these patients in the same cluster as influenza patients making it difficult to detect the outbreak by simply looking for outliers or new clusters – there are none. However, analysis of the density function shows that a new peak is formed (at radius 0.9 units) that potentially leads to a novel-event alert. Essentially, the process de-convolves events, where the larger more common event might otherwise mask the novel and probably rarer event.



Novelty + Analyst interest → Profile/Hypothesis Tracking

Once a novel event is detected, the next step is to determine whether the analyst wishes to track it going forward. For instance, in the above example, the system generated the hypothesis that there is a new disease outbreak whose symptoms might be masked by those of influenza. If the analyst is not interested – e.g. it is off-topic, or already known via other means – then no further action is taken. However, if the new event generates a hypothesis of direct or potential interest, then a new persistent hypothesis tracker is generated, and the input streams are filtered for information pertinent to this hypothesis using the Rete algorithm to correlate data efficiently, as reported in our paper in the last NIMD PI meeting. Hence, novel event detection adds a new dimension by providing hypothesis genesis in a semi-automated manner – where the analyst remains in the driver's seat to guide which hypotheses are tracked, which are promoted, and which are eliminated due to lack of supporting data or lack of topical pertinence.