EFFECTIVE OUTLIER DETECTION IN SCIENCE DATA STREAMS

KIRK BORNE[1], ARUN VEDACHALAM[1]

ABSTRACT. The growth in data volumes from all aspects of space and earth science (satellites, sensors, observatory monitoring systems, and simulations) requires more effective knowledge discovery and extraction algorithms. Among these are algorithms for outlier (novelty / surprise / anomaly) detection and discovery. Effective outlier detection in data streams is essential for rapid discovery of potentially interesting and/or hazardous events. Emerging unexpected conditions in hardware, software, or network resources need to be detected, characterized, and analyzed as soon as possible for obvious system health and safety reasons, just as emerging behaviors and variations in scientific targets should be similarly detected and characterized promptly in order to enable rapid decision support in response to such events. We describe a new algorithm for outlier detection (KNN-DD: K-Nearest Neighbor Data Distributions) and we presents results from preliminary experiments that compare KNN-DD with a previously published algorithm, to determine the effectiveness of the algorithms. We evaluate each of the algorithms in terms of their precision and recall, and in terms of their ability to distinguish between characteristically different data distributions among different classes of objects.

1. INTRODUCTION

Novelty and surprise are two of the more exciting aspects of science – finding something totally new and unexpected. This can lead to a quick research paper, or it can make your career, or it can earn the discoverer a Nobel Prize. As scientists, we all yearn to make a significant discovery. Petascale databases potentially offer a multitude of such opportunities. But how do we find that surprising novel thing? These come under various names: interestingness, outliers, novelty, surprise, anomalies, or defects (depending on the application). We are investigating various measures of interestingness in large databases and in high-rate data streams (e.g., the Sloan Digital Sky Survey [SDSS][2], 2-Micron All-Sky Survey [2MASS][3], and GALEX[4] sky survey), in anticipation of the petascale databases of the future (e.g., the Large Synoptic Survey Telescope [LSST][5]), in order to validate algorithms for rapid detection and characterization of events (i.e., changes, outliers, anomalies, novelties).

In order to frame our scientific investigation of these algorithms, we have been focusing on a specific extragalactic research problem. We are exploring the environmental dependences of hierarchical mass assembly and of fundamental galaxy parameters using a combination of large multi-survey (multi-wavelength) object catalogs, including SDSS (optical) and 2MASS (NIR: near-infrared). We have generated and are now studying a sample of over 100,000 galaxies that have been identified and catalogued in both SDSS and 2MASS. The combination of multi-wavelength data in this cross-matched set of 100,000 galaxies from these optical and NIR surveys will enable more sophisticated characterization and more in-depth exploration of relationships among galaxy morphological and dynamical parameters. The early results are quite tantalizing. We have sliced and diced the data set into various physically partitioned large subsamples (typically 30 bins of more than 3000 galaxies each). We initially studied the fundamental plane of elliptical galaxies, which is a tight correlation among three observational parameters: radius, surface brightness, and velocity dispersion [13, 14]. This well known relation now reveals systematic and statistically

[1] Department of Computational and Data Sciences, George Mason University, Fairfax, VA, USA
{kborne, avedacha}@gmu.edu

[2] www.sdss.org

[3] www.ipac.caltech.edu/2mass/

[4] galex.stsci.edu

[5] www.lsst.org

significant variations as a function of local galaxy density [9]. We are now extending this work into the realm of outlier/surprise/novelty detection and discovery.

2. MOTIVATION

The growth in massive scientific databases has offered the potential for major new discoveries. Of course, simply having the potential for scientific discovery is insufficient, unsatisfactory, and frustrating. Scientists actually do want to make real discoveries. Consequently, effective and efficient algorithms that explore these massive datasets are essential. These algorithms will then enable scientists to mine and analyze ever-growing data streams from satellites, sensors, and simulations – to discover the most "interesting" scientific knowledge hidden within large and high-dimensional datasets, including new and interesting correlations, patterns, linkages, relationships, associations, principal components, redundant and surrogate attributes, condensed representations, object classes/subclasses and their classification rules, transient events, outliers, anomalies, novelties, and surprises. Searching for the "unknown unknowns" thus requires unsupervised and semisupervised learning algorithms. This is consistent with the observation that *"unsupervised exploratory analysis plays an important role in the study of large, high-dimensional datasets"* [30].

Among the sciences, astronomy provides a prototypical example of the growth of datasets. Astronomers now systematically study the sky with large sky surveys. These surveys make use of uniform calibrations and well engineered pipelines for the production of a comprehensive set of quality-assured data products. Surveys are used to collect and measure data from all objects that are visible within large regions of the sky, in a systematic, controlled, and repeatable fashion. These statistically robust procedures thereby generate very large unbiased samples of many classes of astronomical objects. A common feature of modern astronomical sky surveys is that they are producing massive catalogs. Surveys produce hundreds of terabytes (TB) up to 100 (or more) petabytes (PB) both in the image data archive and in the object catalogs. These include the existing SDSS and 2MASS, plus the future LSST in the next decade (with a 20-40 Petabyte database). Large sky surveys have enormous potential to enable countless astronomical discoveries. Such discoveries will span the full spectrum of statistics: from rare one-in-a-billion (or one-in-a-trillion) type objects, to the complete statistical and astrophysical specification of a class of objects (based upon millions of instances of the class).

With the advent of large rich sky survey data sets, astronomers have been slicing and dicing the galaxy parameter catalogs to find additional, sometimes subtle, inter-relationships among a large variety of external and internal galaxy parameters. Occasionally, objects are found that do not fit anybody's model or relationship. The discovery of Hanny's Voorwerp by the Galaxy Zoo citizen science volunteers is one example [22, 23]. Some rare objects that are expected to exist are found only after deep exploration of multi-wavelength data sets (e.g., Type II QSOs [27, 35]; and Brown Dwarfs [5, 29]). These two methods of discovery (i.e., large-sample correlations and detection of rare outliers) demonstrate the two modes of scientific discovery potential from large data sets: (1) the best-ever statistical evaluation and parametric characterization of major patterns in the data, thereby explicating scaling relations in terms of fundamental astrophysical processes; and (2) the detection of rare one-in-a-million novel, unexpected, anomalous outliers, which are outside the expectations and predictions of our models, thereby revealing new astrophysical phenomena and processes (the "unknown unknowns"). Soon, with much larger sky surveys, we may discover even rarer one-in-a-billion objects and object classes.

LSST (www.lsst.org) is the most impressive astronomical sky survey being planned for the next decade. Compared to other sky surveys, the LSST survey will deliver time domain coverage for orders of magnitude more objects. The project is expected to produce ~15-30 TB of data per night of observation for 10 years. The final image archive will be ~70 PB, and the final LSST astronomical object catalog (object-attribute database) is expected to be ~20-40 PB, comprising over 200 attributes for 50 billion objects and ~10 trillion source observations.

Many astronomy ***data mining use cases*** are anticipated with the LSST database [6], including:

- Provide rapid probabilistic classifications for all 10,000-100,000 LSST events each night;
- Find new correlations and associations of all kinds from the 200+ science attributes;
- Discover voids in multi-dimensional parameter spaces (e.g., period gaps);
- Discover new and exotic classes of objects, or new properties of known classes;
- Discover new and improved rules for classifying known classes of objects;
- Identify novel, unexpected behavior in the time domain from time series data;
- Hypothesis testing – verify existing (or generate new) astronomical hypotheses with strong statistical confidence, using millions of training samples;
- Serendipity – discover the rare one-in-a-billion type of objects through outlier detection.

We are testing and validating exploratory data analysis algorithms that specifically support many of these science user scenarios for large database exploration.

3. RELATED WORK

Various information theoretic measures of interestingness and surprise in databases have been studied in the past. Among these are Shannon entropy, information gain [19], Weaver's Surprise Index [34], and the J-Measure [32]. In general, such metrics estimate the relative information content between two sets of discrete-valued attributes. These measures can be used to identify interesting events in massive databases and data streams (through efficient interestingness metrics).

We have used PCA to identify outliers [16, 17]. In particular, we have been studying cases where the first two PC vectors capture and explain most (>90%) of the sample variance in the fundamental plane of elliptical galaxies. Consequently, in such a case, the component of a data record's attribute feature vector that projects onto the third PC eigenvector will provide a measure of the distance z_3 of that data record from the fundamental plane that defines the structure of the overwhelming majority of the data points. Simply sorting the records by z_3, and then identifying those with the largest values, will result in an ordered set of outliers [15] from most interesting to least interesting. We have tested this technique on a small cross-matched test sample of ellipticals from SDSS and 2MASS [16]. We will research the scalability of this algorithm to larger dataset sizes, to higher dimensions (i.e., number of science parameters), and to a greater number of principal components.

In many cases, the first test for outliers can be a simple multivariate statistical test of the "normalcy" of the data: is the location and scatter of the data consistent with a normal distribution in multiple dimensions? There are many tests for univariate data, but for multivariate data, we will investigate the Shapiro-Wilk test for normalcy and the Stahel-Donoho multivariate estimator for outlyingness [24, 31]. The Stahel-Donoho outlyingness parameter is straightforward to calculate and assign for each object: it is simply the absolute deviation of a data point from the centroid of the data set, normalized to the scale of the data. These tests should not be construed as proofs of non-normalcy or outlyingness, but as evidence. For petascale data, even simple tests are non-trivial in terms of computational cost, but it is essential to apply a range of algorithms in order to make progress in mining the data. Several other algorithms and methods have been developed, and we will investigate these for their applicability and scalability to the large-data environment anticipated for LSST: *"Measures of Surprise in Bayesian Analysis"* [1], *"Quantifying Surprise in Data and Model Verification"* [2], and *"Anomaly Detection Model Based on Bio-Inspired Algorithm and Independent Component Analysis"* [33]. Such estimators can also be used in visual data mining – to highlight the most interesting regions of the dataset – this provides yet another tool for visual exploration and navigation of large databases for outliers and other interesting features [18, 25; cf. 11, 21].

4. NEW ALGORITHM FOR OUTLIER DETECTION: KNN-DD

We have implemented a new algorithm for outlier detection that has proven to be effective at detecting a variety of novel, interesting, and anomalous data behaviors [7]. The *"K-Nearest Neighbor Data Distributions"* (KNN-DD) outlier detection algorithm evaluates the local data distribution around a test data point and compares that distribution with the data distribution within the sample defined by its K nearest neighbors. An outlier is defined as a data point whose distribution of distances between itself and its K-nearest neighbors is measurably different from the distribution of distances among the K-nearest neighbors alone (i.e., the two sets of distances are not drawn from the same population). In other words, an outlier is defined to be a point whose behavior (i.e., the point's location in parameter space) deviates in an unexpected way from the rest of the data distribution. Our algorithm has these advantages: it makes no assumption about the shape of the data distribution or about "normal" behavior, it is univariate (as a function only of the distance between data points), it is computed only on a small-N local subsample of the full dataset, and as such it is easily parallelized when testing multiple data points for outlyingness. The algorithm is specified by the following rules, slightly updated from our previous results [7], as a consequence of our new experimental results (§7.1):

Algorithm – Outlier Detection using K-Nearest Neighbor Data Distributions (KNN-DD)

```
Find the set S(K) of K nearest neighbors to the test data point O.
Calculate the K distances between O and the members of S(K). These distances define $f_K(d,O)$.
Calculate the K(K-1)/2 distances among the points within S(K). These distances define $f_K(d,K)$.
Compute the cumulative distribution functions $C_K(d,O)$ and $C_K(d,K)$, respectively, for $f_K(d,O)$ and
    $f_K(d,K)$.
Perform the K-S Test on $C_K(d,O)$ and $C_K(d,K)$. Estimate the p-value of the test.
Calculate the Outlier Index = 1-p.
If Outlier Index $\geq$ 0.98, then mark O as an "Outlier". The Null Hypothesis is rejected.
If 0.90 < Outlier Index < 0.98, then mark O as a "Potential Outlier".
If p $\geq$ 0.10, then the Null Hypothesis is accepted: the two distance distributions are drawn from
    the same population. Data point O is not marked as an outlier.
```

Here, $f(d,x)$ is the distribution of distances d between point x and a sample of data points, $f_K(d,O)$ is the distribution of distances between a potential outlier O and its K-nearest neighbors, and $f_K(d,K)$ is the distribution of distances among the K-nearest neighbors. The algorithm compares the two distance distribution functions $f_K(d,O)$ and $f_K(d,K)$ by testing if the two sets of distances are drawn from the same population.

The KNN-DD algorithm is different from the Distribution of Distances algorithm for outlier detection [28], in which the comparison is between the local data distribution around a test data point and a uniform data distribution. Our algorithm is also different from the k-Nearest Neighbor Graph algorithm for outlier detection [20], in which data points define a directed graph and outliers are those connected graph components that have just one vertex. Furthermore, our algorithm appears similar but is actually different in important ways from the incremental LOF (Local Outlier Factor) algorithms [8, 26], in which the outlier estimate is density-based (determined as a function of the data point's local reachability density), whereas our outlier estimate is based on the full local data distribution. Finally, we report that the KORM (K-median OutlieR Miner) approach to outlier detection in dynamic data streams [12] is most similar to our algorithm, except that their approach is cluster-based (based on K-medians) whereas our approach is statistics-based.

To test the KNN-DD algorithm and to evaluate its effectiveness, we compared experiment results from outlier detection tests using two algorithms: KNN-DD and PC-Out [18]. We briefly summarize below the essential characteristics of the PC-Out algorithm. For more details, the reader is urged to consult the original paper [18].

As part of our algorithm validation process, we examined the separation of the true outliers from the training data and the separation of the false outliers from the training data using a standard unsupervised cluster evaluation metric: the Davies-Bouldin Index [10]. These results are described in §7.3.

5. THE PC-OUT ALGORITHM

Statistical methods for outlier detection often tend to identify as outliers those observations that are relatively far from the center of the data distribution. Several distance measures are used for such tasks. For the multivariate case, the Mahalanobis distance provides a well known criterion for outlier detection. Astronomy databases (object catalogs) are generally high-dimensional (i.e., hundreds of attributes per object). Often it is desirable in such cases to reduce the number of dimensions for easier analysis. Principal Component Analysis (PCA) is one such common method used for dimension reduction. PCA identifies a smaller number of new variables that are linear combinations of the original variables and that preserve the covariance structure of the data. The PC-Out algorithm is one of these PCA-based methods, specifically used for outlier detection. The algorithm detects both location and scatter outliers. As these authors explain, *"Scatter outliers possess a different scatter matrix than the rest of the data, while location outliers are described by a different location parameter."* The PC-Out algorithm starts by performing PCA on the data scaled by the median absolute deviations (MAD) in each of the dimensions. It then retains those components that preserve 99% of the total variance in the data. For the first phase of the algorithm, each one of the principal components is weighted with a robust kurtosis measure that captures the significance of each component in identifying location outliers. The Euclidean distance from the center on this principal component space is equivalent to the Mahalanobis distance since the data have been scaled by the MAD. A translated biweight function is used to down-weight points with large distances. This function also allows the portion of points closer to the center to receive full weights and those points that are farther away from the center get zero weight. These weights are then used as a measure to detect location outliers. The same steps are then repeated in the second phase of the algorithm to detect scatter outliers, except that the kurtosis weighting scheme has been ignored. Weights for each observation are obtained as before, which are then used to identify scatter outliers. Finally, both sets of weights are then combined to get the final weights. By definition, outliers are those points that have final weights less than a default threshold weight value, which is set to be 0.25 initially, though we have experimented with this value and we find that a weight of 0.80 gives the best results for our galaxy-star outlier dataset (§7.2).

6. EXPERIMENTAL DATA SET

For the preliminary experiments reported here, we used a very small set of elliptical galaxies and stars from the combined SDSS+2MASS science data catalogs. We used 1000 galaxies as the training set (i.e., as the set that represents "normal" behavior). We then used 114 other galaxies and 90 stars as test points (i.e., to measure and test for outlyingness). The galaxies represent "normal" behavior, and the stars are intended to represent outlier behavior. We chose 7 color attributes as our feature vector for each object. The 7 colors are essentially the 7 unique (distance-independent, hence intrinsic) flux ratios (i.e., "colors") that can be generated from the 8 (distance-dependent, hence extrinsic) flux measures from SDSS and 2MASS: the ugriz+JHK flux bands (which astronomers call "magnitudes"). Hence, we are exploring outlier detection in a 7-dimensional parameter space. In reality, there is some overlap in the colors of galaxies and stars, since galaxies are made up of stars, which thereby causes the stars to have much less outlyingness measure than we would like. On the other hand, this type of star-galaxy lassification/segregation is a standard and very important astronomy use case for any sky survey, and therefore it is a useful outlier test case for astronomy. The data distribution overlap among the stars and galaxies in our 7-dimensional parameter is somewhat ameliorated by the following fact. The flux of a galaxy GAL(flux) in one waveband is approximately the linear combination of its 10 billion constituent stars' fluxes SUM*(flux) in that same waveband (modulo other effects, such as dust absorption and reddening, which are minimal in elliptical galaxies). Hence the colors of a galaxy are formed from the ratios of these linearly combined SUM*(flux) values. Consequently, the 7-dimensional feature vector of a galaxy need not align with any particular combination of stars' feature vectors. To illustrate this point, we consider a "toy" galaxy comprised of just 2 stars, with red band fluxes R*1 and R*2 and ultraviolet band fluxes U*1 and U*2. The U-R color (i.e., flux ratio) of the galaxy (modulo a logarithm and scale factor that astronomers like to use) is essentially (U*1+U*2)/(R*1+R*2), which cannot be represented by any simple linear combination of the stars' U-R colors: U*1/R*1 and U*2/R*2. Therefore, the actual distributions of stars and galaxies in our parameter space are sufficiently non-overlapping to allow us to perform reasonable outlier tests using stars as the outlier test points with regard to the "normal" galaxy points.
For our distance metric, we used a simple Euclidean (L2-norm) distance calculated from the 7 feature vector attributes. Since they are all flux ratios, the 7 attributes are already physically similar in terms of their mean, variance, and scale factor. No further data normalization or transformation is required.

Though the total numbers of galaxies and stars in our experiments are quite small, especially compared to the massive databases from which they were extracted, they actually do represent a somewhat typical data stream use case, in which a small number of incoming observations are tested against a small comparison set of "local" measurements, to search for and to detect outlier behaviors among the incoming measurements. In the future, we will expand our experiments to much greater numbers of test and training points.

7. RESULTS

7.1 **KNN-DD algorithm results**. We found the following results for the KNN-DD algorithm [7]. We measured the Recall-Precision metrics and produced a ROC curve (Figure 1) for continuously varying p-values (1-p is the Outlier Index, as defined in the Algorithm definition in §4). In these experiments, Recall is calculated from the ratio of (number of stars correctly classified as outliers)/(total number of stars), and Precision is calculated from the ratio of (number of stars correctly classified as outliers)/[(number of galaxies misclassified as outliers)+(number of stars correctly classified as outliers)]. The variation in Precision as a function of p-value is illustrated in Figure 2. The maximum precision (99%) for our test dataset is reached when the p-value reaches the limiting value 0.02. We establish this p-value (0.02) as the critical value used in the KNN-DD algorithm definition (§4). Note that the "knee" (i.e., the discrimination point) in the ROC curve (Fig. 1) occurs at p-value ≈ 0.05, which was the value originally used in our first experiments with the KNN-DD algorithm [7].

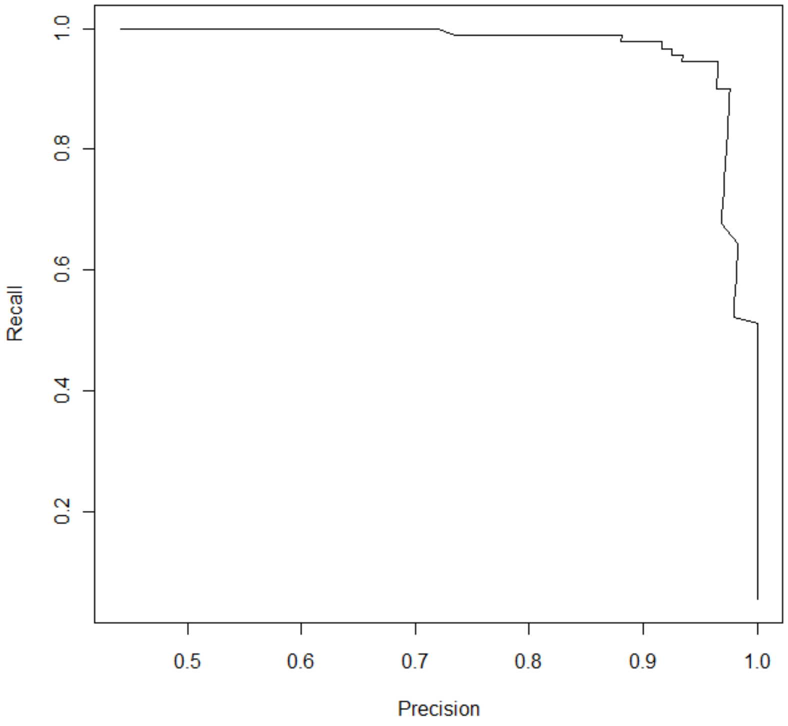

Figure 1. ROC curve for Precision and Recall measured from the KNN-DD algorithm for outlier detection (§4).

We see in Figure 1 that the Recall is nearly 100% over most of the range of the ROC curve. This is illustrated more emphatically in Figure 3, which presents the variation in Recall as a function of p-value. This clearly corroborates the point that we made in the first part of §5, that the data distribution of stars in our 7-dimensional feature space is mostly distinct from the data distribution of galaxies in that same parameter space. We will say more about this below, when we discuss the application of the DBI (Davies-Bouldin Index, [10]) evaluation metric for measuring the distinctness (i.e., separation) of the star and galaxy data distributions.

For p-value=0.02, we find the following results: (1) for the 114 test galaxies, 89 are correctly classified (78%), and 25 are incorrectly classified as outliers (22%); and (2) of the 90 stars, 89 are correctly classified as outliers (99%), and one is misclassified as "normal". Hence, in this case, Recall=99% and Precision=78% (=89/(89+25)).

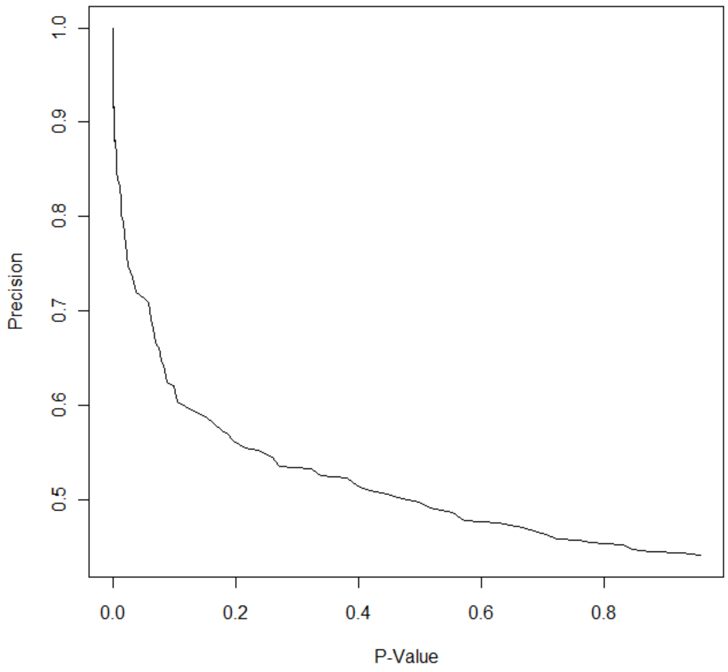

Figure 2. Variation in the Precision of the outlier experiments using the KNN-DD algorithm, as a function of the p-value (where Outlier Index = 1-p; see §4).

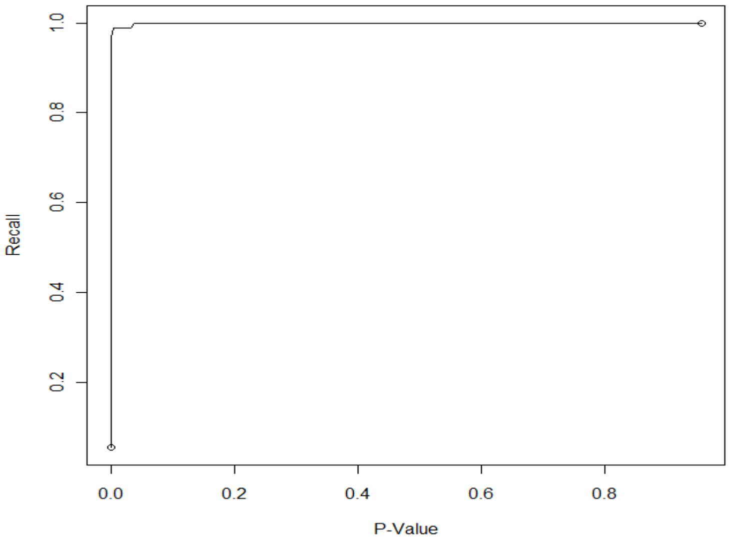

Figure 3. Variation in the Recall of the outlier experiments using the KNN-DD algorithm, as a function of the p-value (where Outlier Index = 1-p; see §4).

7.2 **PC-Out algorithm results**. We found the following results for the PC-Out algorithm [18]. In this case, there is no concept of a training set. (We note that this is actually also true for the KNN-DD algorithm, which can test any data point in a data stream relative to the other data points in the data set. For this paper, we used a training set to evaluate the ROC curve and the Recall/Precision metrics shown in Figure 1 in order to evaluate the effectiveness of the KNN-DD algorithm.) For PC-Out testing, therefore, all 1114 galaxies constituted our "normal" behavior objects and the 90 stars represented our outlier test cases. However, for our calculation of Precision and Recall, we used the same 114 galaxies and 90 stars that we used above for the Precision and Recall calculations for the KNN-DD algorithm. The PC-Out algorithm allows the user to adjust a threshold parameter. We experimented with a few values of this threshold in order to produce the ROC curve shown in Figure 4. In particular, though the original paper [18] recommended a threshold weight of 0.25, we found that a threshold weight of 0.80 provides the optimum results. The ROC curve has a non-monotonic behavior because the Precision curve is non-monotonic (Figure 5), though the Recall curve behaves monotonically (Figure 6)

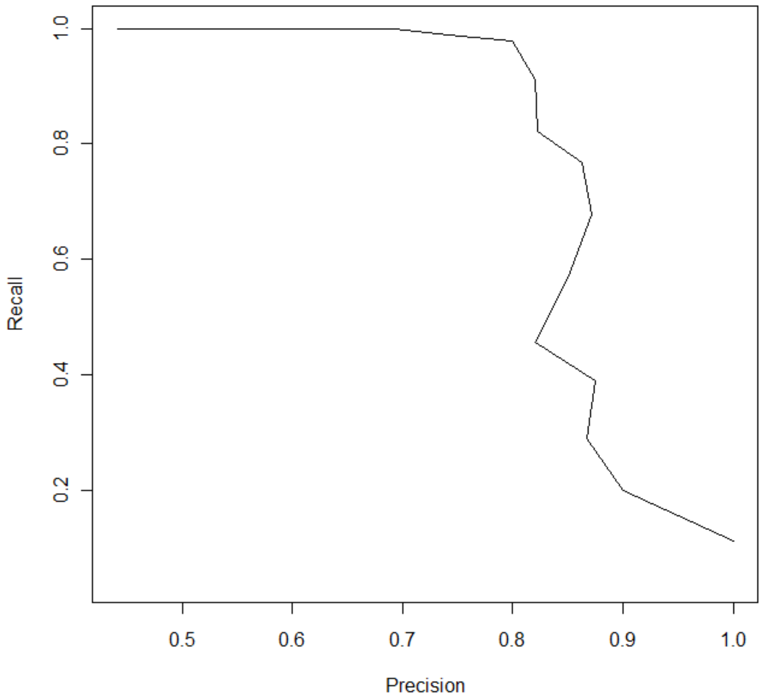

Figure 4. ROC curve for Precision and Recall measured from the PC-Out algorithm for outlier detection (§5).

For a threshold weight of 0.25, we found Recall=18% and Precision=89%. This unsatisfactory value for Recall persuaded us to search for a better choice of the threshold weight used by the PC-OUT algorithm. We settled on a threshold weight value=0.80 that led to the following results: (1) for the 114 galaxies, 96 are correctly classified (84%), and 18 are incorrectly classified as outliers (16%); and (2) of the 90 stars, 78 are correctly classified as outliers (87%), and 12 are incorrectly labeled as non-outliers (13%). Hence, in this case, Recall=87% and Precision=81% (=78/(78+18)). The Recall performance is still lower than the KNN-DD algorithm, while the Precision is a little higher than KNN-DD. In addition, we note that the PC-Out algorithm requires a full PCA eigen-analysis of the complete (big-N) data set, which involves a massive matrix inversion, whereas the KNN-DD algorithm only involves distance calculations of local (small-N) subsets of the data set. For cases where efficiency is critical (e.g., in space-borne sensors, low-power sensor nets, and remote detector platforms), KNN-DD would be both an efficient and an effective algorithm for finding (with high Recall and good Precision) true outliers, anomalies, and surprises in the data.

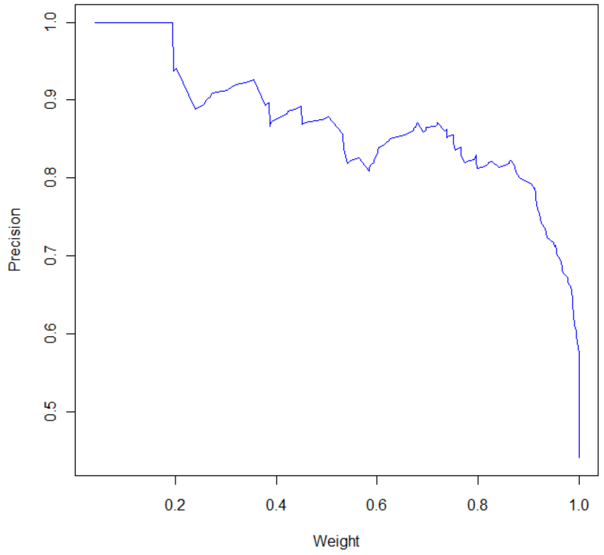

Figure 5. Variation in the Precision of the outlier experiments using the PC-Out algorithm, as a function of the threshold weight [18].

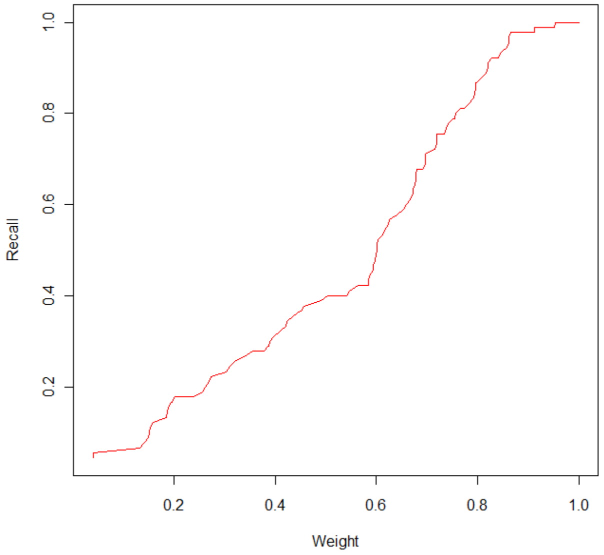

Figure 6. Variation in the Recall of the outlier experiments using the PC-Out algorithm, as a function of the threshold weight [18].

7.3 **Evaluation of results**. We evaluated our outlier test results using the DBI clustering evaluation metric [10]. DBI basically measures the ratio (D1+D2)/D12, where D1 and D2 represent the "diameters" of two data distributions (clusters) and D12 represents the distance between the two clusters. Note that these distances and diameters could be calculated from a variety of different algorithms (e.g., for cluster distance, one could use the single link, complete link, average link, or centroid distance; for cluster diameter, one could use the RMS separation among all members of the cluster, or use the mean, median, or maximum distance of the points from the centroid of their distribution; and for the distance metric, one could use Euclidean distance, Cosine similarity, or any other such metric for these calculations).

Clustering is considered effective if the DBI metric is a small number (i.e., the sizes of assigned clusters are small compared to the distances between the clusters, so that the clusters are compact). We find that this is a useful concept for our outlier experiments also, in the following sense. We measured DBI for 5 pairs of data groupings: [Case 0] set A1 consisting of the original 90 stars versus set A2 consisting of the original 114 galaxies (as classified in the original published data catalogs); [Case 1] set B consisting of all objects classified as stars (outliers, according to the selected algorithm) versus set C consisting of all objects classified as galaxies (non-outliers, according to the selected algorithm); [Case 2] set A1 versus set C consisting of all galaxies that were misclassified as outliers; [Case 3] set A1 versus set D consisting of all galaxies that were correctly classified as non-outliers; and [Case 4] set C versus set D. Note that set A2 = set C + set D. In these comparisons we were hoping to confirm several expectations about the galaxy and star data distributions. In Case 0, we expect that stars are reasonably well differentiated from galaxies in our 7-dimensional feature space (DBI < 1). In Case 1, we expect some overlap between sets B and C (DBI > 1), since set B includes some real galaxies mixed in with the stars and set C includes some real stars mixed in with the galaxies. In Case 2, we expect that the distribution of real stars and misclassified galaxies would occupy similar (overlapping) regions of feature space (DBI > 1). In Case 3, we expect that stars are well separated from galaxies that are correctly classified as non-outliers (DBI < 1). Finally, in Case 4, we expect that the two sets of galaxies (those classified incorrectly as outliers versus those classified correctly as non-outliers) will have essentially the same centroid position (i.e., small D12) in feature space, since they are all elliptical galaxies (i.e., intentionally a very homogeneous sample with uniform average galaxy properties), while the outlier distribution will have a greater extent than the non-outliers (D2 > D1), as measured by their distance from the centroid (hence, DBI >> 1).

For the KNN-DD algorithm, we find the following values for the DBI metric:
Case 0: DBI = 0.86
Case 1: DBI = 1.27
Case 2: DBI = 0.81
Case 3: DBI = 0.92
Case 4: DBI = 8.74

For the PC-Out algorithm, we find the following values for the DBI metric:
Case 0: DBI = 0.86 (same as above)
Case 1: DBI = 1.42
Case 2: DBI = 0.87
Case 3: DBI = 0.84
Case 4: DBI = 3.31

We observe from these results some interesting and some peculiar patterns. The good news is that the DBI metrics for Cases 0, 1, 3, and 4 all behave as we would expect. The (possibly) bad news is that Case 2 yields problematic values for the DBI metric. The Case 2 DBI scores (0.81 and 0.87) are among the lowest of all of the DBI values that we measured, indicating that these two data distributions are among the most cleanly separated in feature space: the stars (true outliers) and the galaxies that were misclassified as outliers. We think that one explanation for this is that the "outlier" galaxies really are correctly labeled as outliers, but they are outlying in all directions (roughly isotropically) in feature space, in contrast with the stars, which are outlying in some preferred direction in feature space: for example, see the schematic diagram in Figure 7. If this is the correct explanation, which we will investigate in our future work, then KNN-DD actually discovered some new and interesting galaxies (true outliers relative to the normal galaxy population), and thus the KNN-DD algorithm is vindicated – it actually fulfilled its objective to discover surprises in scientific datasets.

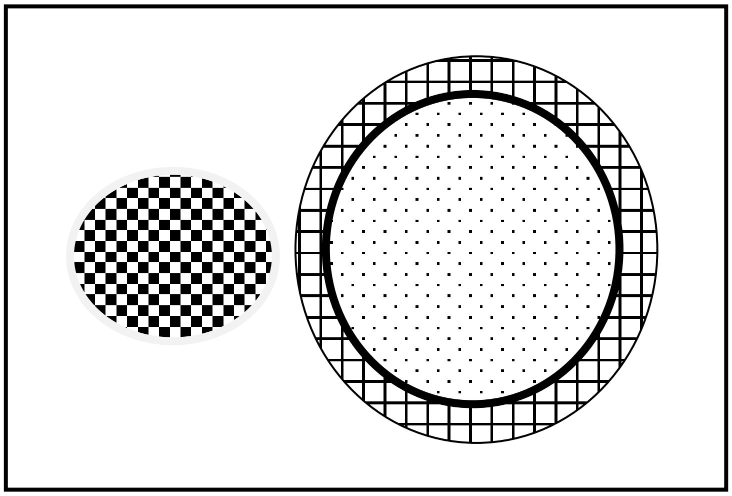

Figure 7. Possible distribution of stars, galaxies, and outlier galaxies that can explain why the Case 2 DB index is so low and answer this question: why are the galaxies that are misclassified as outliers and the stars (which are true outliers) so distinctly separated in feature space? In this schematic diagram, the feature space distribution of normal galaxies is represented by the large circle (filled with dots), the distribution of stars is represented by the large oval (filled with solid squares), and the distribution of outlier galaxies is represented by the large annulus (with grid lines). In this example, the outlier galaxies are outlying in all directions in feature space compared with the stars, which are outlying in some preferred direction. This interpretation is also consistent with the high DBI values in Case 4, and it is further consistent with the very similar DBI values between Case 0, Case 2, and Case 3.

8. CONCLUDING REMARKS AND FUTURE WORK

We find that our new KNN-DD algorithm is an effective and efficient algorithm for outlier detection. It has similar Precision and Recall accuracies relative to the PCA-based PC-Out algorithm [18], while KNN-DD operates efficiently on small-N local data points and PC-Out operates intensively on the full (large-N) set of global data. We therefore see the value of further experimentation with the KNN-DD algorithm on larger, more complex data streams. We also found some interesting behavior in high-dimension feature spaces regarding the region occupied by the outlier stars, compared with the region occupied by the outlier galaxies, compared with the region occupied by normal (non-outlier) galaxies. Further investigation of these surprising results is also warranted, which may already be yielding some scientific discoveries from these simple experimental test cases. We will also extend our KNN-DD comparison tests to include additional published outlier detection algorithms (in addition to the PC-Out algorithm).

As part of our research program in outlier (novelty / surprise / anomaly) detection and discovery, we are planning to evaluate a new approach to discovering surprising correlations and features in large data streams. In particular, we anticipate a new and different vision for exploration of large catalogs: ***Machine Vision***. This is not a new field, but it has been traditionally applied primarily to image processing and image analysis, particularly in the field of robotics. MV algorithms include edge-detection, gradient-detection, motion-detection, change-detection, segmentation, template-matching, and pattern recognition. Many of these can be applied in higher-dimensional data streams, not simply 2-dimensional images. In particular, they can be used to detect interesting features in high-dimensional sky survey catalogs. Specifically, MV techniques have already been used to discover tidal stellar streams around the Milky Way – the "field of streams" [3, 4] – the remnants of tidally shredded dwarf galaxies that were the hierarchical building blocks of mass assembly that has produced our home galaxy. These tidal streams are distinguishable against the rich background of Milky Way stars because they are "cold" – dynamically cold (low velocity dispersion across the stream), spatially cold (low spatial cross-section perpendicular the extremely long narrow stream), and photometrically "cold" (low dispersion in colors [similar age and metallicity] relative to the diverse populations in a random Milky Way star field. Searching for these tidal streams is therefore an example of MV – finding edges, narrow features, and sharp (cold) patterns against the relatively smooth stellar background, which has high dispersion in velocity, color, and spatial extent. It is precisely this distinction between the cold (low variance) and hot (high variance) components in the data distribution that enables MV to discover interesting features in the data. MV techniques can be used to track down cases of cold (astrophysically confined) features in large sky survey catalogs. These features may include restricted tracks (or confined hyperplanes) of astrophysical parameters in some parameter spaces (akin to the field of streams), or discovery of unknown classes of new objects, or unusual subclasses of known classes of objects, or unusual behaviors of known objects. This will be especially important in time-domain studies (e.g., LSST), as we search for interesting (new, unexpected) temporal events or for changes in the temporal behavior (stationarity) of known variable objects.

9. ACKNOWLEDGEMENTS

This research is supported in part by NASA AISR grant number NNX07AV70G.

REFERENCES

[1] M. J. Bayarri and J. O. Berger. Measures of Surprise in Bayesian Analysis. Downloaded from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.6365, 1997.

[2] M. J. Bayarri and J. O. Berger. Quantifying Surprise in the Data and Model Verification. Downloaded from http://citeseer.ist.psu.edu/old/401333.html, 1998.

[3] V. Belokurov, et al. The Field of Streams: Sagittarius and Its Siblings. Astrophysical Journal, vol. 642, pp. L137-L140, 2006.

[4] V. Belokurov, et al. An Orphan in the "Field of Streams. Astrophysical Journal, vol. 658, pp. 337-344, 2007.

[5] B. Berriman, D. Kirkpatrick, R. Hanisch, A. Szalay, and R. Williams. Discovery of Brown Dwarfs with Virtual Observatories. IAU Joint Discussion 8: Large Telescopes and Virtual Observatory: Visions for the Future. http://adsabs.harvard.edu/abs/2003IAUJD...8E..60B

[6] K. Borne. Scientific Data Mining in Astronomy. Next Generation Data Mining. CRC Press: Taylor & Francis, Boca Raton, FL, pp. 91-114, 2009.

[7] K. Borne. Effective Outlier Detection using K-Nearest Neighbor Data Distributions: Unsupervised Exploratory Mining of Non-Stationarity in Data Streams. Submitted to the Machine Learning Journal, March 2010.

[8] M. Breunig, H.-P. Kriegel, R. Ng, and S. Sander. LOF: Identifying Density-Based Local Outliers. ACM SIGMOD Record, vol. 29, pp. 93-104, 2000.

[9] K. Das, K. Bhaduri, S. Arora, W. Griffin, K. Borne, C. Giannella, and H. Kargupta. Scalable Distributed Change Detection from Astronomy Data Streams using Eigen-Monitoring Algorithms. 2009 SIAM International Conference on Data Mining (SDM09), 2009.

[10] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2): 224-227, 1979.

[11] M. Debruyne. An Outlier Map for Support Vector Machine Classification. Annals of Applied Statistics, 3(4): 1566-1580, 2009.

[12] P. Dhaliwal, M. Bhatia, and P. Bansal, P. A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: K-median OutlieR Miner). Journal of Computing, vol. 2, pp. 74-80, 2010.

[13] S. G. Djorgovski and M. Davis. Fundamental Properties of Elliptical Galaxies. Astrophysical Journal, vol. 313, pp. 59-68, 1987.

[14] A. Dressler, D. Lynden-Bell, D. Burstein, R. L. Davies, S. M. Faber, R. Terlevich, and G. Wegner. Spectroscopy and Photometry of Elliptical Galaxies. I - A New Distance Estimator. Astrophysical Journal, vol. 313, pp. 42-58, 1987.

[15] H. Dutta. Empowering Scientific Discovery by Distributed Data Mining on the Grid Infrastructure. Ph.D. dissertation, UMBC, 2007.

[16] H. Dutta, C. Giannella, K. Borne, and H. Kargupta. Distributed Top-K Outlier Detection from Astronomy Catalogs using the DEMAC System. 2007 SIAM International Conference on Data Mining, 2007.

[17] H. Dutta, C. Giannella, K. Borne, H. Kargupta, and R. Wolff. Distributed Data Mining for Astronomy Catalogs. IEEE Transactions in Knowledge and Data Engineering, 2009.

[18] P. Filzmoser, R. Maronna, and M. Werner. Outlier Identification in High Dimensions. Computational Statistics and Data Analysis, 52, pp. 1694-1711, 2008.

[19] A. Freitas On Objective Measures of Rule Surprisingness. LNCC, 1510, pp. 1-9, 1998.

[20] V. Hautamaki, I. Karkkainen, and P. Franti. Outlier Detection Using k-Nearest Neighbour Graph. Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04), 2004.

[21] C. R. Johnson, M. Glatter, W. Kendall, J. Huang, and F. Hoffman. Querying for Feature Extraction and Visualization in Climate Modeling. ICCS 2009, Part II, LNCS 5545, pp. 416-425, 2009.

[22] G. I. G. Jozsa, M. A. Garrett, T. A. Oosterloo, H. Rampadarath, Z. Paragi, H. van Arkel, C. Lintott, W. C.Keel, K. Schawinski, and E. Edmondson. Revealing Hanny's Voorwerp: Radio Observations of IC 2497. Astronomy and Astrophysics, vol. 500, pp. L33-L36, 2009.

[23] C. J. Lintott, et al. Galaxy Zoo: Morphologies Derived from Visual Inspection of Galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, vol. 389, pp. 1179-1189, 2008

[24] R. A. Maronna and V. J. Yohai. The Behavior of the Stahel-Donoho Robust Multivariate Estimator. Journal of the American Statistical Association, vol. 90, pp. 330-341, 1995.

[25] D. Pena and F. J. Prieto. Multivariate Outlier Detection and Robust Covariance Matrix Estimation. Technometrics, vol. 43, pp. 286-301, 2001.

[26] D. Pokrajac, A. Lazarevic, and L. Latecki, L. Incremental Local Outlier Detection for Data Streams. IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 2007.

[27] G. T. Richards, et al. Eight-Dimensional Mid-Infrared/Optical Bayesian Quasar Selection. Astronomical Journal, vol. 137, pp. 3884-3899, 2009.

[28] V. Saltenis. Outlier Detection Based on the Distribution of Distances between Data Points. Informatica, 15(3): 399-410, 2004.

[29] R.-D. Scholz, M. J. McCaughrean, N. Lodieu, and B. Kuhlbrodt. Epsilon Indi B: A New Benchmark T Dwarf. Astronomy and Astrophysics, vol. 398, pp. L29-L33, 2003.

[30] A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel. Finding Large Average Submatrices in High Dimensional Data. Annals of Applied Statistics, 3(3): 985-1012, 2009.

[31] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). Biometrika, vol. 52, pp. 591–611, 1965.

[32] P. Smyth and R. M. Goodman. Rule Induction Using Information Theory. Knowledge Discovery in Databases, pp 159-176, AAAI/MIT Press, 1991.

[33] S. Srinoy and W. Kurutach. Anomaly Detection Model Based on Bio-Inspired Algorithm and Independent Component Analysis. TENCON 2006, IEEE Region 10 Conference proceedings, pp. 1-4, 2006.

[34] Weaver's Surprise Index. Encyclopedia of Statistical Sciences (Wiley), vol. 9, pp. 104-109, 1988.

[35] N. Zakamska, et al. Candidate Type II Quasars from the Sloan Digital Sky Survey. I. Selection and Optical Properties. Astronomical Journal, vol. 126, pp. 2125-2143, 2003.