Surprise Discovery in Scientific Databases: A Framework for Data Intensive Science Utilizing the Power of Citizen Science

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Arun Vedachalam
Master of Science
George Mason University, 2012
Bachelor of Science
Anna University, 2008

Director: Dr. Kirk Borne, Professor
Department of Computational and Data Sciences

Spring Semester 2016
George Mason University
Fairfax, VA

Dedication

My brother, Vinodh Vedachalam. The pursuit for this advanced degree would not have been possible without you. I have also sustained unwavering support, motivation and love from my mother Latha, my father Vedachalam and my wife Vinodhini. It is therefore with great pride and affection that I dedicate this dissertation to my family.

Acknowledgments

The pursuit for this advanced degree would not have been possible without the support and prayers of my family, professors, the administrative staffs of Mason and friends. this very important section is one of the very few occasions where I have the opportunity to formally acknowledge and thank those responsible for the success of this work.

First of all, I would like to thank Dr. Kirk Borne for agreeing to advice me in this important capacity. When I first met Dr. Borne, I had very limited idea of what I wanted to do for my PhD. I was uncertain whether research in machine learning was a fit for me. However, after few months working with him on exploratory search for fundamental plane like structure like in other parameter space of galaxies, it became clear to me the research direction I needed to pursue for my PhD. I enjoyed all the intense discussion that we had during our meeting and I always walk out of his office with plenty of ideas and smile. I owe a great part of my exposure to Astronomy and expertise in machine learning to him.

I would like to thank Dr. Edward Wegman for accepting to serve as my dissertation chair and his demeanor. I learnt most of the statistical techniques from him. Also, I would like to thank Dr. James Gentle for all his insightful teaching and feedback on the many courses I took with him. It was with Dr. Gentle and Dr. Wegman that I learnt most of the statistics. Additionally, I would like to thank Dr. Igor Griva and Dr. Fernando Camelli for agreeing to serve on my dissertation committee. I truly appreciate all of the time, assistance, comments and suggestions as I worked my way through this process. My special thanks to Dr. Dimitrios Papaconstantopoulos for his guidance and help during my initial days in CDS and Dr. Estela Blaisten-Barojas for being the super awesome graduate coordinator.

Many friends have helped me survive and stay sane in graduate school. Many thanks to my friends/Roommates: Guru Kannan, Aravind Srinivasan, Ajay Nagarajan, Sivasundar Gurubaran, Swaninathan Venkatramanan, Venkatraman Kalpathy Balan, Kannan Venkatachalam for your support outside of grad school. I will never forget the happy hours, cricket and flag football game, and all the fun activities that we have done together. Also I would like to thank my CDS friends Che Ngufor, Yang Xu, Steven Baehr, Fuxin Huang for all the technical help and numerous ideas shared on the different topics that we were working together. Also special thanks to Mazhalai Chellathurai, Andrew Corrigan and Byeonghwa Park for the initial help and guidance they provided me during my initial days in grad school/CDS. Also, special thanks to Kathleen Enos, Executive Assistant for all the helped that she did throughout my stay in Mason. Last, but certainly not least, to my family, I say thank you for your unbelievable patience, trust, and encouragement. I am especially grateful for the emotional and financial support you provided me through out my studies even when you had little idea of what exactly I was studying.

Table of Contents

List of Tables

List of Figures

Abstract

SURPRISE DISCOVERY IN SCIENTIFIC DATABASES: A FRAMEWORK FOR DATA INTENSIVE SCIENCE UTILIZING THE POWER OF CITIZEN SCIENCE

Arun Vedachalam, PhD

George Mason University, 2016

Dissertation Director: Dr. Kirk Borne

The ability to collect and analyze massive amounts of data is rapidly transforming science, industry and everyday life. Too often in the real world, information from multiple sources such as humans, experts, agents need to be integrated to provide support for a making any scientific discovery. This holds true for modern sky surveys in Astronomy where the common theme is that they produce hundreds of terabytes (TB) up to 100 (or more) petabytes (PB) both in the image data archive and in the object catalogs. For example, the LSST will produce a 2040 PB catalog database. Such large sky surveys have enormous potential to enable countless astronomical discoveries. The discoveries will span the full spectrum of statistics: from rare object types, to complete statistical and astrophysical specifications of many classes of objects. The challenges faced by this data driven approach often revolves around two major issues: 1) The lack of the expert labels present in the database and 2) The lack of sufficient knowledge in the database for identifying the known expert labels.

In this dissertation, first we will discuss novel approach to finding interesting (novelty/surprise/anomaly detection) objects that enable scientists to discover the most interesting scientific knowledge hidden within large and high-dimensional datasets.

Then will move on towards utilizing the power of citizen science in identifying features where the goal is to determine indicators, based solely on discovering those automated pipeline-generated attributes in the astronomical database that correlate most strongly with the patterns identified through visual inspection of galaxies by the Galaxy Zoo volunteers. Further expanding this the capability to Latent variable models where the hidden/latent variables extracted from the citizen science data help bridge the gap between the human generated classifications and the features not captured by the astronomy data pipeline. Proper utilization of these latent variables helped unearth new classes or in some cases most representative/interesting sample that are previously unknown to the astronomers. These interesting objects act as a training set for the machine learning algorithms and can be used to build automated models to classify the galaxies from the future sky surveys such as LSST.

Chapter 1: Introduction

In many real world problems, due to the advancement of technology and Internet, the way science being conducted, businesses operate, governments function and people live has changed visibly. But a new, less visible technological trend is just as transformative: Big Data or Data Science. Big data starts with the fact that there is a lot more information floating around these days than ever before, and it is being put to extraordinary new uses. Big data is distinct from the Internet, although the Web makes it much easier to collect and share data. Big data is about more than just communication: the idea is that we can learn from a large body of information things that we could not comprehend when we used only smaller amounts.

In Scientific discover, big data has lead to a new paradigm of science called the Fourth Paradigm, which is especially powerful in some of the observational science such as astronomy. In this chapter we will see what the fourth paradigm of science is and how it influence some scientific disciplines and briefly list some the discoveries made with the data. Later in this chapter we will see some of the challenges faced by this big data avalanche in astronomy and moving forward how Citizen Scientist play a major role in making new and interesting scientific discovery.

1.1 BigData Rising Challenges

Big Data, after all, is the biggest buzzword of the new millennium. Its everywhere, from genomics, biomics and a bunch of others. Social networks, financial networks, ecological networks all contain vast amounts of data that no longer overwhelm computer hard drive storage capabilities. Scientists are now swimming in a superocean of endless information, fulfilling their wildest dreams. Scientists usually celebrate the availability of a lot of data

and they have been extolling all the research opportunities that massive databases offer. Scientific advances are becoming more and more data-driven and the massive amounts of data bring both opportunities and new challenges to data analysis. For one thing, huge datasets are seductive. They invite aggressive analyses with the hope of extracting prizewinning scientific findings. Wringing intelligent insights from Big Data poses formidable challenges for computer science, statistical inference methods and even the scientific method itself.

Computer scientists, of course, have made the accumulation of all this big data possible by developing exceptional computing power and information storage technologies. But collecting data and storing information is not the same as understanding it. Figuring out what Big Data means isnt the same as interpreting little data, just as understanding flocking behavior in birds doesnt explain the squawks of a lone seagull. Standard statistical tests and computing procedures for drawing scientific inferences were designed to analyze small samples taken from large populations. But Big Data provides extremely large samples that sometimes include all or most of a population. The magnitude of the task can pose problems for implementing computing processes to do the tests. Many statistical procedures either have unknown runtimes or runtimes that render the procedure unusable on large-scale data. Also High dimensionality in the data may lead to wrong statistical inference.

Many computer scientists and statisticians are aware of these issues, and a lot of work is under way to address them. But there is more to it than just mashing up some more sophisticated statistical methodologies. Scientists also need to confront some biases, rooted in the days of sparse data, about what science is and how it should work. Old style science coped with natures complexities by seeking the underlying simplicities in the sparse data acquired by experiments. Big Data forces scientists to confront the entire repertoire of natures nuances and all their complexities. Consequently, science cannot rely on the strictly empirical approach to answer questions about complex systems. There are too many possible factors influencing the system and too many possible responses that the system might make in any given set of circumstances. To use Big Data effectively, science might just have to learn to subordinate experiment to theory.

1.2 Data Intensive Science and Fourth Paradigm

In scientific discovery, the first three paradigms were experimental, theoretical and more recently computational science. A book published by Microsoft [1] argues that a fourth paradigm of scientific discovery is at hand: the analysis of massive data sets. The basic idea being discussed in this is that the capacity for collecting scientific data has far outstripped our capacity to analyze it, and the focus should be on developing technologies that will make sense of this deluge of data. The late Microsoft Research Dr. Jim Gray called "to have a world in which all of the science literature is online, all of the science data is online, and they inter-operate with each other". This dream of him is close to reality in some scientific domains like astronomy, where advanced telescopes generate petabytes of data available for public analysis. With further advancements in the distributed and high-performance computing framework like Hadoop, and with advanced open-source analysis tools like R rapidly adapting to the scales of these data sets, the fourth paradigm is certain to become a mainstream reality in other scientific domains as well.

1.3 Scientific Data Mining in Astronomy

Data mining has always been fundamental to astronomical research, since data mining is the basis of evidence based discovery. Astronomers have been doing it for centuries in terms of characterizing the objects in the sky, assign the new objects to a set of previously known classes and discover the unknowns. These skills are becoming more critical than ever due to the advancement in sensor technologies that can collect vast amount of data and the availability of computing resources to analyze them. A brief survery of some the popular machine learning algorithms to research problems in astronomy is provided in [2].

1.4 Rise of Crowdsourcing

Crowdsourcing is a term coined in 2005 and made popular a decade ago [3], is the process of obtaining needed services, ideas and content by contributions from a large group of people,

especially from an online community. This distributed problem solving and production model is put to use by for-profit organizations such as InnoCentive [4] and iStockphoto [5]. iStockphoto is the web's original resource started in 2000 for crowd sourced royalty-free stock images, media and design elements. The site's tag-line says "by creatives, for creatives ", is a pioneer micro-payment photography site, freeing creative people around the world to create beautiful communications on a budget. InnoCentive crowdsource innovative solutions from the world's smartest people, who compete to provide ideas and solutions to important business, social, policy, Scientific, and technical challenges. [3, 5], address the implications for future research into crowdsourcing, regarding notions of professionalism and investment in online communities.

1.4.1 Citizen Science

Citizen science (also called Crowd science, crowd-sourced science, civic science) is a scientific research conducted in whole or part by amateur or non professional scientists. Formally, citizen science has been defined as "the systematic collection and analysis of data; development of technology; testing of natural phenomena; and the dissemination of these actives by researchers on primary avocational basis" [6]. Oxford English Dictionary [7] defined citizen scientists as "a member of general public who engages in scientific work, often in collaboration with or under the direction of professional scientists and scientific institutions". Certain scientific disciples such as archaeology, astronomy and natural history require observation skills that are more important than the expensive equipments. Some early citizen science projects such as Christmas Bird Count [8], The Evolution MegaLab [9], Open Air Laboratories(OPAL) [10] has been started. Numerous other successful examples in other areas of science include eBird [11], an online database of bird observations, EyeWire [12], a game to map the 3D structure of neurons in the brain. The latest of them is TurtleSAT, that is mapping freshwater turtle deaths throughout Australia utilize smartphone and tablet technology to create better engagement of the citizen science community.

1.4.2 Galaxy Zoo

The Sloan Digital Sky Survey (SDSS) has compiled a list of more than 1 million galaxies. To glean information about galaxy evolution, astronomers need to know what type of galaxy each one is: spiral, barred spiral, elliptical, or something else. The only reliable way to classify galaxies is to look at each one, but all the world's astronomers working together couldn't muster enough eyeballs for the task. A volunteer online effort called Galaxy Zoo, launched in 2007, has classified the entire catalog years ahead of schedule, bringing real statistical rigor to a field used to samples too small to support firm conclusions. The Galaxy Zoo team went on to ask more-complicated classification questions that led to studies they hadn't thought possible. And in a discussion forum on the Galaxy Zoo Web site, volunteers have pointed to anomalies that on closer inspection have turned out to be genuinely new astronomical objects [13].

In order to understand the formation and subsequent evolution of galaxies on must distinguish between the two main morphological classes: Spirals and early-type systems. The galaxy zoo project[14] that went live in 2007, provides visual morphological classifications of nearly one million galaxies, extracted from the Sloan Digital Sky Survey (SDSS). This was made possible by general public who voluntarily inspect and classify these galaxies. The project has obtained more than 800 million classifications made by a million participants. It is noted that these classifications obtained from the volunteers are consistent with the classifications obtained by the professional Astronomers. In addition, this provides a catalog that can be used to directly compare SDSS morphologies with older data sets. In the forthcoming sections of this chapter we will see some examples of some interesting scientific discoveries made utilizing these classifications. Also this leads to interesting set of citizen science projects under the Zooniverse framework [15], that extend the concept of crowd sourced science to other disciplines including biology, lunar science, solar science and the humanities. The goal of Zooniverse is to create a sustainable future of large scale, internet based citizen science, utilizing the human cognition of a community of citizen scientist in an innovative manner that impacts the knowledge discovery process.

1.5 Scientific Discoveries with Galaxy Zoo

Several researchers tried to analyze the data from galaxyzoo to reproduce the classification obtained by the human eye. In [16], Artificial Neural Network was trained on a subset of objects classified by the human eye and tested whether the machine learning algorithm can reproduce the human classifications. In this paper they also stressed on the parameter selection problem the classification task as the accuracy of the model heavily depends on the parameters used. They also conclude that it is promising to use machine learning algorithms to perform morphological classification for the next generation of wide-field imaging surveys and that the Galaxy Zoo catalogue provides an invaluable training set for such purposes.

The Galaxy zoo project also led to the identification of merging galaxies [17]. This presents the largest, most homogeneous catalog of merging galaxies in the nearby universe obtained through the Galaxy zoo project. The suggested method in this paper convert a set of visually inspected classifications for each galaxy into a single parameter which describes the confidence that the object is an ongoing merger.

It has been recently demonstrated that using the broad bandpass photometry of the Sloan Digital Sky Survey in combination with with precise knowledge of galaxy morphology should help in estimating more accurate photometric redshifts for galaxies [18]. Using the Galaxy Zoo separation for spirals and ellipticals in combination with Sloan Digital Sky Survey photometry the attempt is to calculate photometric redshifts. In the best case the root mean square error for Luminous Red Galaxies classified as ellipticals is as low as 0.0118. Given these promising results a better photometric redshift estimates for all galaxies in the Sloan Digital Sky Survey (350 million) can be calculated utilizing the power of Machine Learning. This provides promising results for scientist interested in estimating Weak-Lensing, Baryonic Acoustic Oscillation, and other fields dependent upon accurate photometric redshifts.

Galaxy zoo has also lead to some scientific discover of unusual objects in the sky. One such discovery is Hanny's Voorwerp [19] which is an unusual object near the spiral galaxy IC 2497, discovered by the visual inspection of the SDSS as part of Galaxy zoo project. The

object, is bright in SDSS g band due to unusually strong OIII 4959-5007 emission lines. The result of the first targeted observations of the object in the optical, UV and X-ray, which show that the object contains highly ionized gas. Subsequent study leads to a discovery that this object may represent the first direct probe of quasar history.

1.6 Thesis Overview

The rest of the thesis is arranged as follows: a standalone surprise discovery algorithm to find interesting objects in a data set has been provided in chapter 2. A brief description about the algorithm and its application in several domains has been provided in this chapter. Chapter 3 describes a methodology to identify interesting features are sub sample of features that are important for the classification of merging or colliding galaxies that utilizes the power of citizen science data has been provided. Chapter 4 moves on to the more complex task of classifying galaxies and address the complexity of the data and demands the need for more sophisticated analysis. Later, chapter 5 discusses the idea of latent variable analysis and discusses the application of non-parametric Bayesian framework that identifies the most representative sample of galaxies that simplifies the classification task. Chapter 6 offers conclusion and discussion of results.

Chapter 2: Surprise Detection in Science Data Streams Using K-Nearest Neighbor Data Distributions

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.

2.1 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][1], 2-Micron All-Sky Survey [2MASS][2], and GALEX[3] sky survey), in anticipation of the petascale databases of the future (e.g., the Large Synoptic Survey Telescope [LSST][4]), 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 [20, 21]. This well known relation now reveals systematic and statistically significant variations as a function of local galaxy density [22]. We are now extending this work into the realm of outlier/surprise/novelty detection and discovery.

[1] www.sdss.org
[2] www.ipac.caltech.edu/2mass/
[3] galex.stsci.edu
[4] www.lsst.org

2.2 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 [23].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 anybodys model or relationship. The discovery of Hannys Voorwerp by the Galaxy Zoo citizen science volunteers is one example [24,25]. Some rare objects that are expected to exist are found only after deep exploration of multi-wavelength data sets (e.g., Type II QSOs [26, 27]; and Brown Dwarfs [28, 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 [30], including:

1. Provide rapid probabilistic classifications for all 10,000-100,000 LSST events each night;

2. Find new correlations and associations of all kinds from the 200+ science attributes;

3. Discover voids in multi-dimensional parameter spaces (e.g., period gaps);

4. Discover new and exotic classes of objects, or new properties of known classes;

5. Discover new and improved rules for classifying known classes of objects;

6. Identify novel, unexpected behavior in the time domain from time series data;

7. Hypothesis testing verify existing (or generate new) astronomical hypotheses with strong statistical confidence, using millions of training samples;

8. 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.

2.3 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 [31], Weaver's Surprise Index [32], and the J-Measure [33]. 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 [22,34]. 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 records attribute feature vector that projects onto the third PC eigenvector will provide a measure of the distance z3 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 z3, 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 [34]. 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 [25, 35]. 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 [36], Quantifying Surprise in Data and Model Verification [37], 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.

2.4 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.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 points 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.

Algorithm 1 Outlier Detection using K-Nearest Neighbor Data Distributions (KNN-DD)

1: Find the set S(K) of K nearest neighbors to the test data point O.
2: Calculate the K distances between O and the members of S(K). These distances define fK(d,O).
3: Calculate the K(K-1)/2 distances among the points within S(K). These distances define fK(d,K).
4: Compute the cumulative distribution functions CK(d,O) and CK(d,K), respectively, for fK(d,O) and fK(d,K).
5: Perform the K-S Test on CK(d,O) and CK(d,K). Estimate the p-value of the test.
6: Calculate the Outlier Index = 1-p.
7: If Outlier Index > 0.98, then mark O as an Outlier. The Null Hypothesis is rejected.
8: If $0.90 <$ Outlier Index < 0.98, then mark O as a Potential Outlier.
9: If $p > 0.10$, 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, fK(d,O) is the distribution of distances between a potential outlier O and its K-nearest neighbors, and fK(d,K) is the distribution of distances among the K-nearest neighbors. The algorithm compares the two distance distribution functions fK(d,O) and fK(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 [38], 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 [39], 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 [40,41], in which the outlier estimate is density-based (determined as a function of the data points 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 [42] 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 [43]. We briefly summarize below the essential characteristics of the PC-Out algorithm. For more details, the reader is urged to consult the original paper [43].

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 [44]. These results are described in section 2.7.1.

2.5 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 (2.7.2).

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.

2.7 Results

2.7.1 KNN-DD algorithm results

We found the following results for the KNN-DD algorithm. We measured the Recall-Precision metrics and produced a ROC curve figure 2.1 for continuously varying p-values (1-p is the Outlier Index, as defined in the Algorithm 1). 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 2.2. 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. Note that the knee (i.e., the discrimination point) in the ROC curve (2.1) occurs at p-value 0.05, which was the value originally used in our first experiments with the KNN-DD algorithm.

We see in 2.1 that the Recall is nearly 100% over most of the range of the ROC curve. This is illustrated more emphatically in figure 2.3, 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

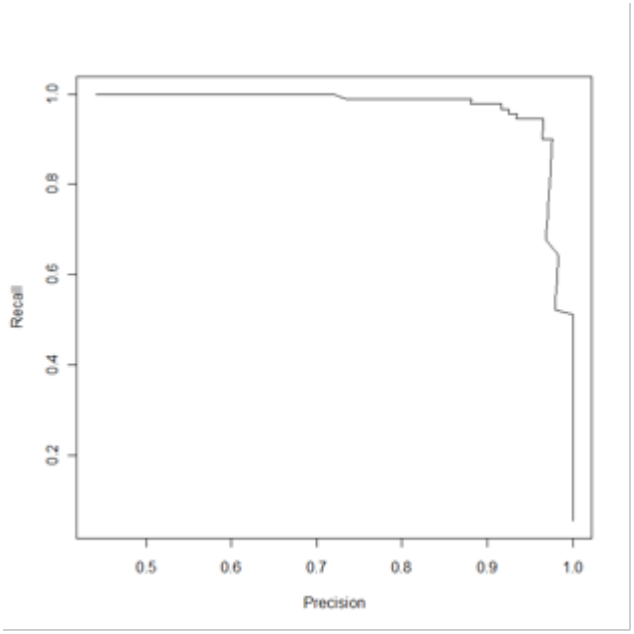

Figure 2.1: ROC curve for Precision and Recall measured from the KNN-DD algorithm for outlier detection.

about this below, when we discuss the application of the DBI (Davies-Bouldin Index, [44]) 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)).

2.7.2 PC-Out algorithm results

We found the following results for the PC-Out algorithm [43]. 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 2.1 in order to evaluate the effectiveness of

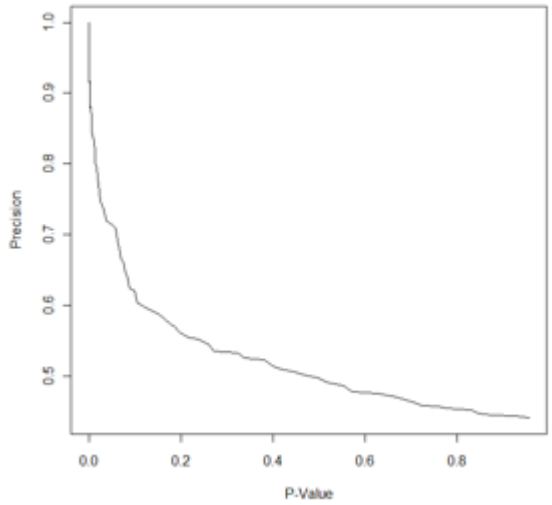

Figure 2.2: Variation in the Precision of the outlier experiments using the KNN-DD algorithm, as a function of the p-value.

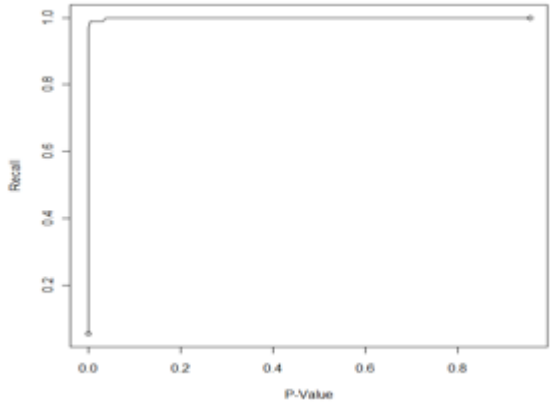

Figure 2.3: Variation in the Recall of the outlier experiments using the KNN-DD algorithm, as a function of the p-value.

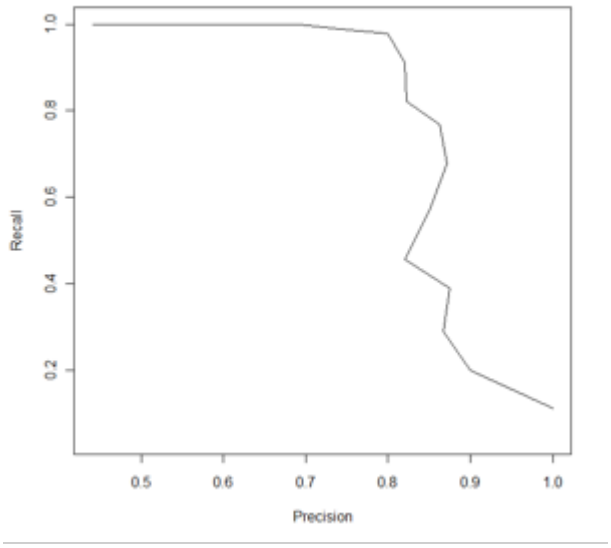

Figure 2.4: ROC curve for Precision and Recall measured from the PC-Out algorithm for outlier detection.

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 2.4. In particular, though the original paper [43] 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 2.5), though the Recall curve behaves monotonically (figure 2.6)

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

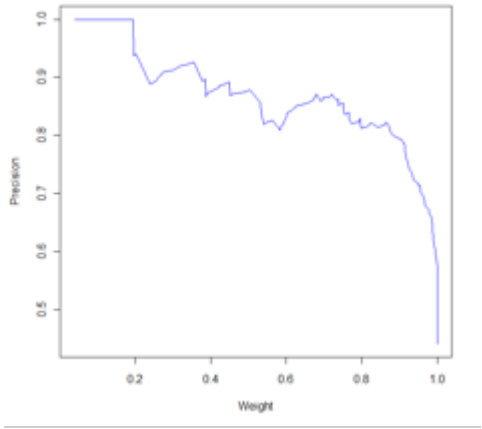

Figure 2.5: Variation in the Precision of the outlier experiments using the PC-Out algorithm, as a function of the threshold weight .

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.

2.7.3 Evaluation of results

We evaluated our outlier test results using the DBI clustering evaluation metric [44]. 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

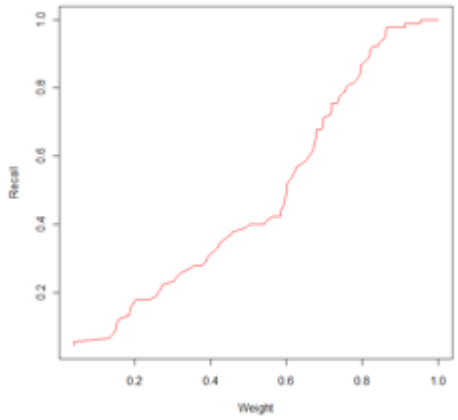

Figure 2.6: Variation in the Recall of the outlier experiments using the PC-Out algorithm, as a function of the threshold weight.

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 Case1, 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

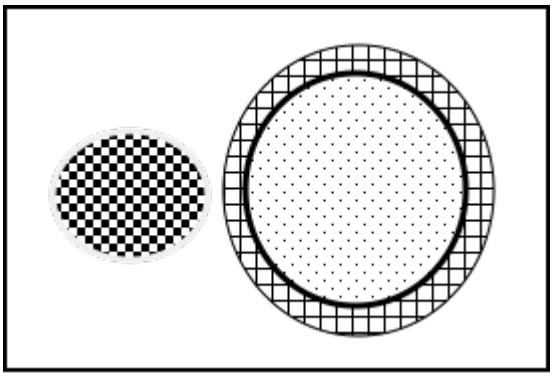

Figure 2.7: Possible distribution of stars, galaxies, and outlier galaxies that can explain why the Case 2 DB index is so low and answer this question: why are the galaxies that are misclassified as outliers and the stars (which are true outliers) so distinctly separated in feature space? In this schematic 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.

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 2.7. 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.

2.8 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).

The main advantages of our KNN-DD algorithm are:

1. It is based on the non-parametric K-S test
2. It makes no assumption about the shape of the data distribution or about normal behavior (of non-outliers).
3. It compares the cumulative distributions of the test data (i.e., the set of inter-point distances), without regard to the nature of those distributions.
4. It operates on multivariate data, thus solving the curse of dimensionality.
5. It is algorithmically univariate, by estimating a function that is based entirely on the scalar distance between data points (which themselves occupy highdimensional parameter space).
6. It is simply extensible to higher dimensions.

7. The KNN-DD distance distributions are computed only on small-K local subsamples of the full dataset of N data points (K¡¡N).

8. The algorithm is easily (embarrassingly) parallelizable when testing multiple data points for outlyingness.

The major deficiencies of the KNN-DD algorithm that need attention, as the algorithm is currently defined, and areas for future work include:

1. The choice of K (see Sect. 26.4) is not determined or justified. We need to validate our choice of K, or else find a justifiable selection criterion for particular values.

2. The choice of p (Sect. 26.4) is only weakly determined.

3. We need to measure the learning times of the KNN-DD algorithm.

4. We need to determine (and validate) the complexity of the KNN-DD algorithm.

5. We need to compare the KNN-DD algorithm against a larger set of other outlier detection algorithms.

6. We need to evaluate KNN-DD algorithms effectiveness and efficiency on much larger datasets.

7. We aim to demonstrate the usability of the KNN-DD algorithm on streaming data, not just with static data (as used in this papers experiments).

As part of further research 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 the use of croudsouced labelled data set from galaxyzoo for exploration of large catalogs. In particular, they can be used to detect interesting features in high-dimensional sky survey catalogs. 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 of known variable objects.

Chapter 3: Data Mining the Galaxy Zoo Mergers

3.1 Summary

Collisions between pairs of galaxies usually end in the coalescence (merger) of the two galaxies. Collisions and mergers are rare phenomena, yet they may signal the ultimate fate of most galaxies, including our own Milky Way. With the onset of massive collection of astronomical data, a computerized and automated method will be necessary for identifying those colliding galaxies worthy of more detailed study. This chapter researches methods to accomplish that goal. Astronomical data from the Sloan Digital Sky Survey (SDSS) and human-provided classifications on merger status from the Galaxy Zoo project are combined and processed with machine learning algorithms. The goal is to determine indicators of merger status based solely on discovering those automated pipeline-generated attributes in the astronomical database that correlate most strongly with the patterns identified through visual inspection by the Galaxy Zoo volunteers. In the end, we aim to provide a new and improved automated procedure for classification of collisions and mergers in future petascale astronomical sky surveys. Both information gain analysis (via the C4.5 decision tree algorithm) and cluster analysis (via the Davies-Bouldin Index) are explored as techniques for finding the strongest correlations between human-identified patterns and existing database attributes. Galaxy attributes measured in the SDSS green waveband images are found to represent the most influential of the attributes for correct classification of collisions and mergers. Only a nominal information gain is noted in this research, however, there is a clear indication of which attributes contribute so that a direction for further study is apparent.

3.2 Introduction

3.2.1 Scientific Rationale

Current computational detection of a galaxy merger in astronomical data is less than ideal. However, human pattern recognition easily identifies mergers with varied, but strong, levels of accuracy. If this superior human input can be incorporated into the automated data pipeline detection scheme, informed by machine learning models, then a more accurate assessment of merger presence can be gained automatically in future large sky surveys. These improvements could potentially lead to more powerful detection of various astronomical objects and interactions.

Our goal was to generate merger classification models using two prominent machine learning approaches, as a preliminary exercise toward the incorporation of human input into future automated pipeline classification models.

3.2.2 Citizen Science

Citizen Science refers to the involvement of layperson volunteers in the science process, with the volunteers specifically asked to perform routine but authentic science research tasks that are beyond the capability of machines. Complex pattern recognition (and classification) and anomaly detection in complex data are among the types of tasks that would qualify as Citizen Science activities. The Galaxy Zoo project (galaxyzoo.org) presents imagery from the Sloan Digital Sky Survey (SDSS) to laypersons for classification (e.g., whether a galaxy is of the elliptical or spiral type) via a web interface. The project went live in 2007, and already over 200 million classifications have been provided by more than 260,000 individuals. During the classification process, volunteers can flag a particular image as depicting a merger of two or more galaxies. Approximately 3000 prominent mergers in the SDSS (Sloan Digital Sky Survey) have been identified[45].

3.2.3 Related Work

Image recognition has long been a major deficiency in computation. Classification tasks such as facial recognition, trivially exercised with great accuracy and precision by living organisms, have been predominantly inaccurate and slow when attempted using computers. While current algorithms are fairly capable of recognizing substructures and details in imaging data, recognition of gestalt in the data has proved more elusive. This shortcoming, combined with the contemporary unyielding influx of data in the natural sciences and the vastness of a data domain such as astronomy, has led to the necessity of attempting to tap into the effortless capability of human cognition.

The Galaxy Zoo web application has as its goal the collection and application of human classifications applied to images of galaxies from the SDSS. Efforts have been made to use human input to reinforce existing machine learning models such as artificial neural networks and genetic algorithms[46]. Additionally, work has been done using supervised learning algorithms to classify galaxy type (non-merging), with considerable success using spectroscopic data for training[47] and data derived from human cognition[48]. It has been found that the introduction of parameters chosen using human input shows great promise for improving current detection and classification of astronomical objects.

3.3 Defining the Data

To help us identify the SDSS photometric attributes that show promise in merger classification, data from the SDSS survey were collected in two distinct groups, one group chosen as a representative sample of galaxy objects in SDSS, and the other to represent known mergers.

3.3.1 Data Sources

We utilized data strictly from the Galaxy Zoo project and SDSS. Galaxy Zoo was used to obtain SDSS ID's for merger objects, along with an attribute representing the users'

confidence in the classification as a merger. All photometric data, merger or non-merger, was obtained from the SDSS.

Mergers

The data chosen to represent known merging galaxies were represented by 2,810 of the 3,003 SDSS mergers presented in [45] (i.e., those that had the full set of attributes that we examined).

These objects are known to be involved in mergers and to represent objects with relatively high surface brightness (making human classification possible).

Non-Mergers

To build classification models, galaxies assumed to be predominantly non-mergers were also needed as training examples.

As the vast majority of the 100 million SDSS galaxies are not mergers, a representative random sample of SDSS galaxies was chosen for this role.

The sample (initially comprised of 3500 instances) was chosen at random from objects of galaxy type within the SpecPhotoAll view in the SDSS database. This view represents objects that have spectral data associated with them. The spectral data was necessary to obtain object redshift, which was needed to remove distance dependence from the gathered attributes.

Utilizing objects with spectral data also had the ancillary impact of restricting the non-mergers to those with similar surface brightness to the mergers.

3.3.2 Data Cleaning and Pre-Processing

Upon completion of these steps, the sample consisted of 6,310 objects with 76 attributes, including the nominal attribute "merger/non-merger." Considerable pre-processing was necessary to ready the data for use as the training set for classifiers. Some pre-processing steps were necessary for both of the two algorithms utilized. All attributes that did not

represent morphological characteristics were removed. For example, the SDSS object ID's, measurement error magnitudes, and attributes representing location or identity, rather than morphology, were among those removed. In Astronomical Catalog missing values occurs for variety of reason from. It is not possible to estimate these values, as these values may be physically meaningful. Therefore instances with placeholder values (in SDSS, "-9999") in any attribute were removed. Since data were gathered from bright objects, most objects did not require this removal. Distance-dependent attributes were transformed, using redshift, to be distance-independent. A concentration index was also generated, using the ratio of the radii containing 50% and 90% of the Petrosian flux within each galaxy.

3.4 Machine Learning

3.4.1 Decision Trees

Decision trees are a straightforward machine learning algorithm that produces a classifier with numerical or categorical input, and a single categorical output (the 'class'). Decision trees have several advantages:

- The resulting tree is equivalent to a series of logical 'if-then' statements, and is therefore easy to understand and analyze.
- Missing attribute values can be incorporated into a decision tree, if necessary.
- Easy to implement as a classifier.
- Computationally cheap to 'train' and use in classification.

The most popular decision tree algorithm, C4.5, was published by Ross Quinlan in 1993 [8]. To generate a decision tree, the Weka data mining software suite was utilized. Weka is a robust and mature open source Java implementation of many prominent machine learning algorithms. It also automates many pre-processing tasks, including transformations of parameters and outlier detection/removal. Weka refers to its C4.5 implementation as J48. This is the routine we used to build a decision tree for classification.

Table 3.1: Important attributes from SDSS Catalog and their description

Attribute	Description
$petroMag\_ug$	Petrosian magnitude colors. A color was calculated for four independent pairs of bands in SDSS (u-g, g-r, r-i, and i-z).
$petroRad\_u * z$	Petrosian radius, transformed with redshift to be distance-independent.
$invConIndx\_u$	Inverse concentration index. The ratio of the 50% flux Petrosian radius to the 90% flux Petrosian radius.
$isoRowcGrad\_u * z$	Gradient of the isophotal row centroid, transformed with redshift to be distance-independent.
$isoColcGrad\_u * z$	Gradient of the isophotal column centroid, transformed with redshift to be distance-independent.
$isoA\_u * z$	Isophotal major axis, transformed with redshift to be distance-independent.
$isoB\_u * z$	Isophotal minor axis, transformed with redshift to be distance-independent.
$isoAGrad\_u * z$	Gradient of the isophotal major axis, transformed with redshift to be distance-independent.
$isoBGrad\_u * z$	Gradient of the isophotal minor axis, transformed with redshift to be distance-independent.
$isoPhiGrad\_u * z$	Gradient of the isophotal orientation, transformed with redshift to be distance-independent.
$texture\_u$	Measurement of surface texture.
$lnLExp\_u$	Log-likelihood of exponential profile fit (typical for a spiral galaxy).
$lnLDeV\_u$	Log-likelihood of De Vaucouleurs profile fit (typical for an elliptical galaxy).
$fracDev\_u$	Fraction of the brightness profile explained by the De Vaucouleurs profile.

Decision Trees in Weka

The Weka J48 algorithm has several arguments. The relevant arguments for our exploration are:

- **binarySplits**: If set to true, the generated tree will be binary. A binary tree is simpler to interpret.
- **confidenceFactor**: The lower this is set, the more pruning that will take place on the tree. More pruning can result in a simpler tree, at the expense of predictive power. However, too little pruning can contribute to overfitting.
- **minNumObj**: The minimum number of instances required in each tree leaf. The higher this is set, the simpler the resulting tree.

As the goal of this work is primarily to explore the strength of SDSS attributes in merger classification, emphasis in tree generation was on generating simple trees, and examining the strongest predicting attributes. In particular, we are searching for those database attributes that contain the most predictive power: those that show the highest correlation with Galaxy Zoo volunteer-provided classification as a merger. These would be the attributes that match most strongly with the outputs of human pattern recognition.

Information Gain

In the C4.5 and J48 algorithms, the tree design is predicated upon maximizing information gain (a measurement of entropy in the data). Using Weka, the information gain was calculated for each of the attributes, using the 6310 instances referenced in section 3.3.2 with tenfold cross-validation. The top five attributes are listed below. Notably, 4 of these top 5 attributes are related to the SDSS observations in the green waveband. These are the attributes that have the highest predictive power in merger classification accuracy.

Decision Tree Results

We decided to generate three different trees, with the following characteristics:

Table 3.2: Important attributes from SDSS Catalog with high Information Gain

Attribute	Information Gain
$lnLExp_g$	0.099
$texture_g$	0.074
$lnLDeV_g$	0.068
$petroMag_{gr}$	0.065
$isoAGrad_u * z$	0.057

1. A tree that is trained on all instances. This tree should use all mergers, regardless of the vote of merger confidence given by Galaxy Zoo users.

2. A tree that is trained on merger instances with stronger Galaxy Zoo user confidence. This tree was to be generated with only mergers that a majority of Galaxy Users flagged as such. These instances are assumed to be the mergers that are, in some sense, 'obvious.'

3. A tree that is trained on merger instances with less than a majority of Galaxy Zoo users indicating then as such. These instances are assumed to be less than obvious to the layperson.

If one simply classifies all galaxies as non-mergers, a predictive accuracy of 55% is obtained. In the simplest tree with one split (seen in figure 3.1), a 66% correct classification occurs, so there is a modest but definite information gain. The attribute $lnLExp_g$ is at the root node with values at or below -426.586609 indicating a merger and all others classified as non-mergers.

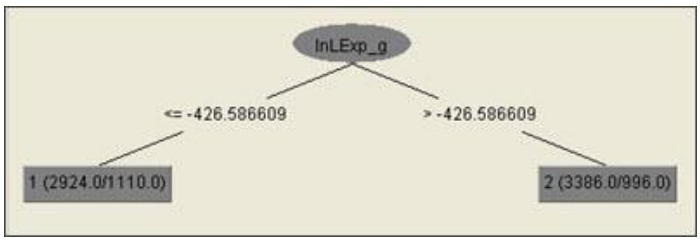

Figure 3.1: Visualization of decision tree with a single node.

Table 3.3: Output from the build Random Forest model

	Precision	Recall	F-Measure
Merger	0.659	0.682	0.670
Non-Merger	0.734	0.714	0.724
Weighted Avg.	0.700	0.699	0.700

When the minimum number of leaf instances is set to 500, and the confidence factor to 0.001, a relatively simple tree is obtained that still has a reasonable predictive power of 70%. A 66%/34% training/test set split was used. A portion of the model output is shown below.

The root node of this tree (as seen in figure 3.2) is $lnLExp_g$, which is not a wholly unexpected result, as will be discussed later in this paper.

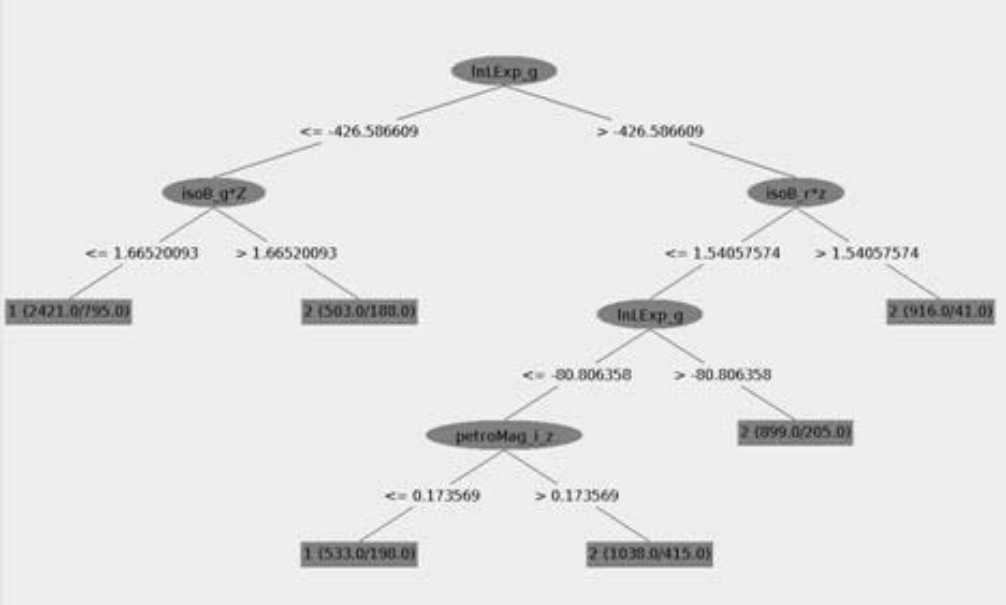

Figure 3.2: Visualization of decision tree built using all mergers.

After removing merger instances with a user confidence of less than 0.50 (with the number of leaf instances set to 200 to produce a simple tree and a 66%/34% split),we measured the precision, recall and F-measure for each of the two classes to determine the

Table 3.4: Output from a average sized Random Forest model

	Precision	Recall	F-Measure
Merger	0.657	0.456	0.538
Non-Merger	0.766	0.882	0.820
Weighted Avg.	0.730	0.741	0.726

Table 3.5: Output from a Smaller Random Forest model

	Precision	Recall	F-Measure
Merger	0.416	0.167	0.238
Non-Merger	0.796	0.933	0.859
Weighted Avg.	0.712	0.762	0.721

accuracy of the model. For mergers, recall is calculated as the proportion of the number of mergers correctly classified as such out of the total number of mergers. Precision is calculated as the proportion of the number of mergers correctly classified as such out of all instances classified as mergers (correctly or not). The F-measure is a commonly reported measure intended to incorporate both precision and recall into a single measure. It is defined as $\frac{2 \cdot precision \cdot recall}{precision + recall}$.

Contrary to intuition, while the overall classification accuracy increases, the recall of the model for mergers decreased significantly. With this approach, $petroMag_{gr}$ is now the strongest predictor at the root of the tree. This can be seen in figure 3.3. $lnLExp\_g$ is still a key attribute, but it is no longer at the root. This model has very strong predictive power for non-mergers, but quite weak recall for mergers.

After removing merger instances with a user confidence of more than 0.50 (with the number of leaf instances set to 200 to produce a simple tree and a 66%/34% split), we achieve the output shown below.

The users' confusion seems to be expressed in the resulting model, which has high overall accuracy, but a very weak recall. This poor performance is due to its excessive tendency to classify as Non-Merger, as the data set now is only comprised of objects that are not

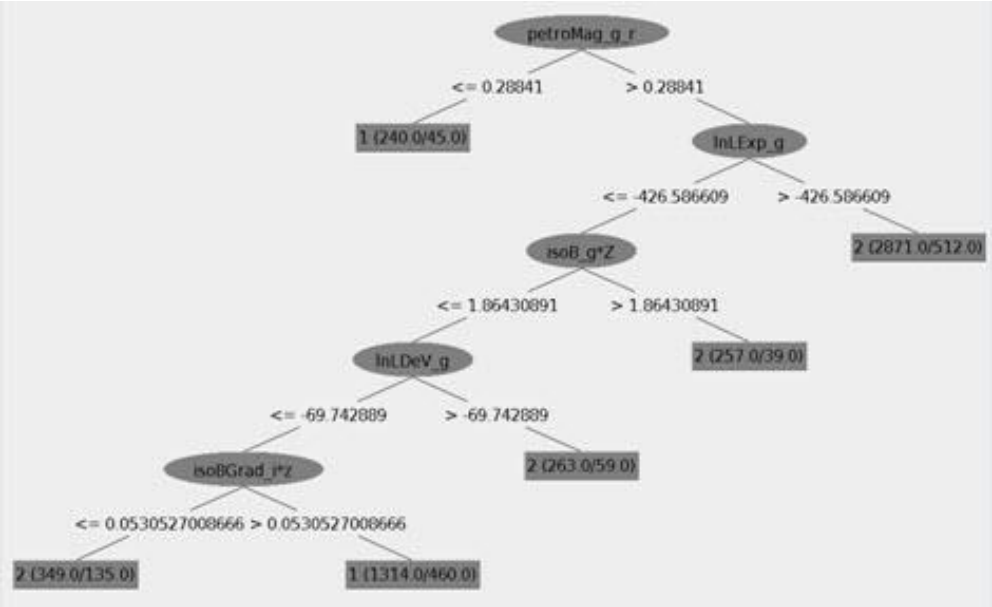

Figure 3.3: Visualization of decision tree built using the strongest mergers.

obviously mergers. Using these weaker voted mergers, the model is rooted on $petroMag_{ui}$, as seen in figure 3.4.

Tree Strengths and Weaknesses

The trees generated are of varying usefulness.

The tree generated using all of the mergers exhibited an overall accuracy of about 70%, with precision of 66% and recall of 68%. This is above average predictive power, but not incredibly useful.

The trees generated using the stronger and weaker mergers separately seem to indicate two things:

1. The user confusion over some mergers appears to be manifested in the resulting model, as the parameters that are influential in the model are not strongly morphological, indicating that the objects may be missing strong visual cues of merging.

2. The confidence of users in some merger classifications results in a tree that incorporates

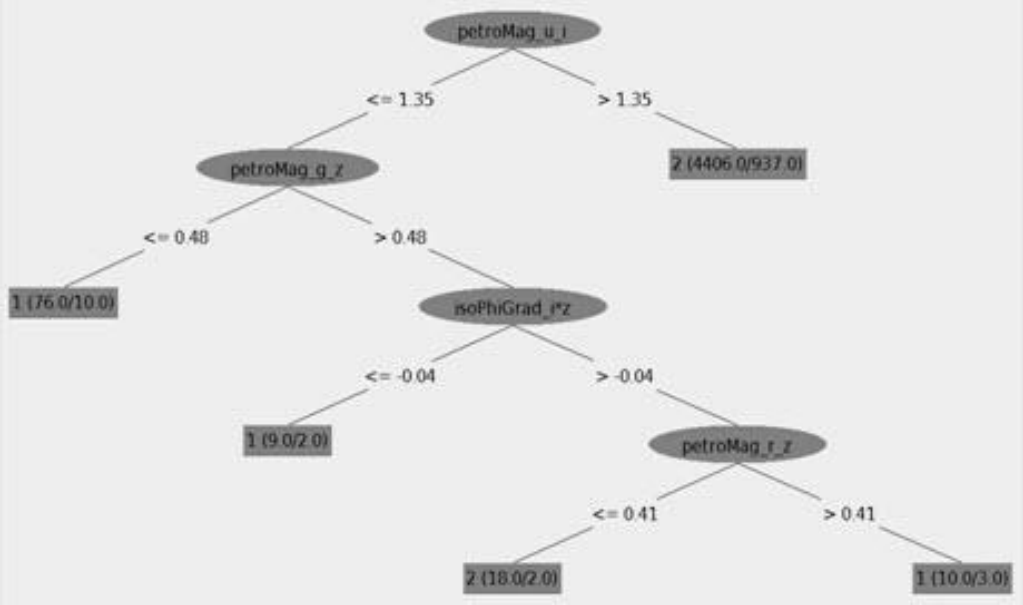

Figure 3.4: Visualization of decision tree built using the weakest mergers.

more strongly morphological attributes, but has diminished recall power. We feel that this merits further investigation.

There are two especially interesting things about the decision trees generated from this data:

- The strongest predicting attributes seem to be associated with the SDSS green filter waveband.
- Poor exponential fit and small isophotal minor axis are among the strongest indicators of merger presence.

Significance of the Green Band

The strongest predicting attributes seem to be associated with the green band. In the tree generated using all merger instances, The two strongest attributes for merger prediction are associated with the green band, and fully half of the top ten information gaining attributes are associated with this band. The green band seems to carry a disproportionate amount

of information relative to the other four bands measured in SDSS photometry.

Upon investigation, we discovered that strong green spectral lines are associated with stellar formation via doubly ionized oxygen, and stellar formation is itself unusually abundant in galactic mergers[**?**]. So it is not surprising that the green band seems to be important in the classification models we have generated.

Significance of $lnLExp$ and $isoB$ Attributes

The attributes $lnLExp$ and $isoB$ both featured prominently in the decision tree approach as influential values for merger detection.

The $isoB$ attribute represents the length of the minor axis of the isophote of the galaxy's surface brightness in a given band. It is a reasonable expectation that tidal distortion from merger involvement may influence an axis of such an isophote.

The $lnLExp$ attribute represents the extent to which the galaxy object has a brightness profile that is fit well by an exponential fit, the details of which can be found in [9]. It is not surprising that this measure of morphology would be an influential factor in merger classification, as tidal distortion would almost certainly affect the brightness profile of a galaxy involved in a merger and thereby reduce the likelihood that the galaxy brightness profile would be fit by a standard non-distorted spiral galaxy exponential function. It should also be noted that another measure of brightness profile fit was featured among attributes with the highest information gain: $lnLDeV$. $lnLDeV$ is a measure of goodness of fit with the De Vaucouleur profile (which is the functional form of the brightness profile in elliptical galaxies), and this would also be expected to exhibit irregularities in the presence of tidal distortion in true colliding/merging galaxies.

Future Direction for Decision Trees

Given the modestly strong evidence that we have generated for the quality of green-band morphological attributes as merger predictors, a promising avenue for further development

of classifiers may be other attributes in this band. These may be novel image characterization parameters or simply transformations of existing database parameters.

The inclusion of isophotal axis length among the influential parameters seems to indicate that more examination of isophotal properties may be fruitful in this area.

3.5 Cluster Analysis

Identifying groups of similar observations in a dataset is a fundamental step in any data analysis task. Classification and clustering are the two main approaches used to identify similar groups of data instances. Whereas classification attempts to assign instances to one of several known classes, clustering attempts to derive the classes themselves. In the case of one or two dimensions, visual inspections of the data such as scatter plots can help to quickly and accurately identify the classes. Datasets in astronomy are generally comprised of many more dimensions. With advancements in astronomical data collection technology, astronomers are able to collect several hundred variables for millions of observations. Not all these collected variables are useful for a given classification task. There typically are many insignificant attributes that might prevent us from identifying the structure of the data.

With the knowledge of class labels from the Galaxy Zoo catalog of merging and interacting galaxies, we would like to be able to identify which morphological and photometric attributes in the SDSS data correlate most strongly with the user-selected morphological class. These variables can be identified by measuring the separation of the instances in the attribute feature space in which the data reside: which attributes provide the best discriminator between different human-provided patterns and classes? Measures like Dunn's Validity Index[49] and Davies-Bouldin Validity Index[50] are two metrics by which to achieve this.

3.5.1 The Davies-Bouldin Index

Davies-Bouldin Validity Index (DBI) is a function of the ratio of *intra*-cluster instance separation to *inter*-cluster instance separation. This is given by:

$$DB = \frac{1}{n}\sum_{i=0}^{n} max_{i \neq j} \frac{S_n(Q_i) + S_n(Q_j)}{S(Q_i, Q_j)}$$

...where n is the number of clusters, $S_n(Q_i)$ is the average distance of all objects from the cluster to their cluster center, and $S(Q_i, Q_j)$ is the distance between clusters centers. Good clusters (i.e., compact clusters with respect to their separation) are found with low values of DBI, and poor clusters (i.e., strongly overlapping groupings) have high values of DBI. For the inter-cluster distance function S one could use single linkage, complete linkage, average linkage, centroid linkage, average of centroids linkage, or Hausdorff metrics and for the intra-cluster distance function S one could use complete diameter, average diameter, or centroid diameter[49]. For purposes of experimentation, we picked used the centroid linkage and the centroid diameter as our measures to calculate the DB index.

3.5.2 Approach

To determine the database attributes that influence the separation of the human-provided galaxy classes (merger versus non-merger) most strongly, we first calculated the DB index for the two clusters (i.e., the cluster of mergers versus the cluster of non-mergers) using each one of variables individually. We then ranked the variables based on these calculated DBI values. The variable that tops this list is the most important variable for instance separation, at least according to this metric. This single variable of course cannot necessarily provide us with the best separation. So we looked for any higher dimensional subset of the feature space that has improved separation for these two classes of objects. To this end, we selected the top ten individual variables and calculated the DB index of all possible combinations of these ten variables and ranked the combinations to identify the subset of the original

Table 3.6: List of DBI values in different parameter space

10 Best Separating Individual Attributes	10 Best Separating of all 1014 Subsets of Best 10 Attributes
$isoAGrad_u * z$	$isoAGrad_u * z$
$petroRad_u * z$	$petroRad_u * z$
$texture_u$	$texture_u$
$isoA_z * z$	$isoA_z * z$
$lnLExp_u$	$lnLExp_u$
$lnLExp_g$	$lnLExp_g$
$isoA_u * z$	$petroRad_u * z$, $isoB_z * z$, $isoBGrad_u * z$, $lnLExp_g$
$isoB_z * z$	$isoAGrad_u * z$, $lnLExp_g$
$isoBGrad_u * z$	$petroRad_u * z$, $isoA_u * z$, $isoB_z * z$, $lnLExp_g$
$isoAGrad_z * z$	$isoAGrad_u * z$, $isoBGrad_u * z$, $lnLExp_g$

attribute set that provides the best separation.

3.5.3 Results

The following is the list of the top 10 features and subsets with the lowest DB index:

Features such as $isoPhiGrad_i * z$, $isoColcGrad_g * z$, $isoColcGrad_u * z$, $petroMag_{ug}$, $isoColcGrad_i * z$, and $fracDev_z$ have a significantly large DBI and are therefore do not appear to be useful for clustering. These features seem to be of little significance for decision tree classification as well, since they were not present in any of the trees we generated. Also, visual inspection of the attributes using histograms revealed that with the four individual attributes with lowest DB Index values (seen in figure 3.5), little to no separation can be seen.

In the scatter plot (seen in figure 3.6) of mergers and non-mergers in $isoAGrad_u * z$, $lnLExp_g$ feature space shows slight separation between these two classes.

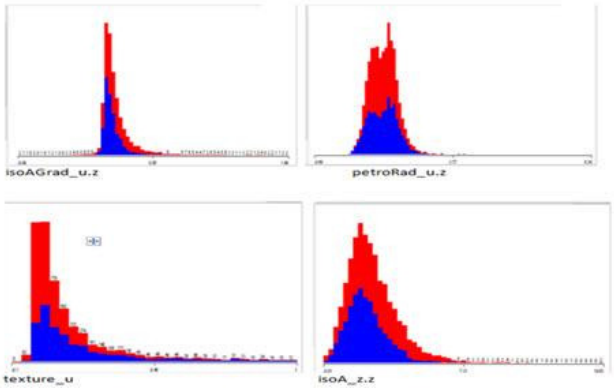

Figure 3.5: Histograms of the four lowest attributes according to DBI.

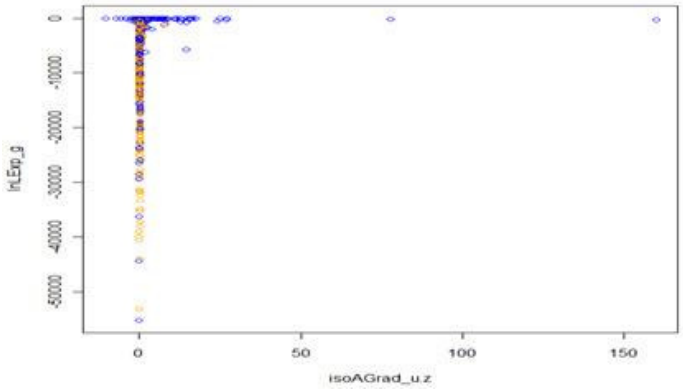

Figure 3.6: Merger and non-merger classes in $isoAGrad_u * z$, $lnLExp_g$ space.

3.5.4 Future Direction for Cluster Analysis

From the plots it is evident that there is not a clear separation between mergers and non-mergers in the subsets of the feature space that we have explored. This is also evident from the fact that the minimum value of all DBI's that we calculated is 2.19, which is substantially greater than the ideal value of 1. This is an indication of relatively weak clustering. The value 2.19 is the local minimum of the parameter-space. With further analysis of all the possible (75-factorial!) combinations of the 75 numerical attributes, we might be able to find the global minimum value where the clusters have the strongest separation. However, finding

the global minimum in this way would be extremely (in fact, prohibitively) computationally intensive. It is, however, important to note that two of the top ten features according to individual DBI are $isoAGrad_u * z$ and $lnLExp_g$, which are also among the top five features in information gain. Therefore, our approach to feature extraction is to some degree consistent with the information gain-based decision tree approach. With limited computation time and resources, only certain combinations of the best ten attributes could be examined. Use of optimal search algorithms (such as genetic algorithms) and use of a massively parallel computational environment (such as Cloud computing) could empower us to discover the best separating subset of the attributes and provide some interesting results.

3.6 Summary of Outcomes

We were able to generate a decision tree with accuracy of approximately 70%, including recall for merger detection of approximately 66%. Two classes of morphological attributes were identified as potentially having promise in future work on decision tree analysis:

- Attributes related to the SDSS green waveband, specifically brightness profile fits in this band. This result is validated by the known characteristics of star formation emissions in merging galaxies.

- Attributes related to the galaxy isophotes. This has validity due to the tidal distortions of isophotes that are typically present in galactic mergers.

Results from the cluster analysis also indicate the significance of these two feature-types, providing more evidence of their importance in merger classification. Further analysis might lead to combinations of features that greatly improve the classification accuracy of mergers and non-mergers. Mathematically derived or entirely novel features (especially of a more morphological nature) could also be a promising avenue for improving merger classification, as success with the chosen features was modest. Utilizing a combination of cluster-based feature extraction and decision tree analysis will likely aid in further improvements to classification accuracy, and more generally, to the identification of the salient features that

will enable automated pipelines to emulate human cognitive powers and pattern recognition abilities, and thereby automatically indicate the presence of such events in massive petascale sky surveys of the future.

Chapter 4: Initial Experiments with classification of Galaxies

In this chapter we will move on to more harder task of classifying galaxies. Each galaxy is being labeled into one of the three different classes namely elliptical, spiral and merger. There is an option for the fourth class under the galaxyzoo framework called "don't know" to account any stars, artifacts or outliers. The main goal here is to use most post popular machine learning algorithms like random forest and support vector machines to see how these algorithms perform with the same classification task using the photometric attributes from the Sloan Digital Sky Survey. The class labels used for these algorithm are based on the classifications provided by the galaxy zoo volunteers.

For these experiments, We limited ourselves to only big, bright galaxies where the Galaxy Zoo volunteers still had problems providing the identifications. We sampled galaxies that are bright by placing a limit of 17 mag, with spectroscopic redshifts, and a Petrosian radius greater than 70 in R band. Also we sampled galaxies on the number of classifications provided by the galaxyzoo volunteers under the zooniverse framework. We picked the galaxies that receive more than 30 classifications to account for any erroneous classification from the citizen science volunteers. The results from these algorithms are summarized using classification error and confusion matrix has been used to analyze the performance of these algorithms.

4.1 Initial Experiments with the data

Data set, used for our experiments consists of samples of galaxies in two categories: Elliptical(E) and Spiral(S) from the Galaxy Zoo with three different levels of confidence, namely 90% ($\geq$ 85%), 70% (65 to 75%), and 50% (45 to 55%) in which the Galaxy Zoo volunteers have agreed. There are generally a few thousand of galaxies in each category. We utilized

data from the Galaxy zoo [51] and extracted the photometric features for each galaxy from Sloan Digital Sky Survey catalog (SDSS).

4.2 Data Pre-processing

We have the attributes that define the Galaxy Zoo Volunteers confidence in classification and the photometric attributes from SDSS that represent morphological characteristics of galaxies. Few distance dependent attributes were transformed, using redshift, to be distance-independent. Table 1 describes the list of important attributes and transformation applied to it. After preprocessing, the data set has 47 attributes. There are 16 attributes that are from Galaxy Zoo and the remaining 30 are from the SDSS catalog. Table 1 contains the list of the few important attributes and their description. Note that each of the SDSS attributes typically exists for the five SSDS filter wavebands.

4.3 Classifying Galaxies Using Some Popular Algorithms

4.3.1 One-class Support Vector Machines

One-class classifiers [52][53] aim at distinguishing a single class from the rest of the classes. These classifiers are used when there are unknown numbers of classes in the data but the user is interested in the rules for one particular class. This is different from the usual classifier, which tries to distinguish between two or more classes. In general, they are used for outlier/anomaly detection. These classifiers are perfect for the task that we are dealing since we are not aware of the type of the galaxies that are present in the samples with 70% and 50% confidence levels. One-class Support Vector Machine [54] is one of the most post popular one-class classifier and it will be used in our experiments. The algorithm, like the traditional SVM for classification maps the data using an appropriate kernel function into a feature space H, and then trying to separate the mapped data from the origin with maximum margin. In our experiments below we used the Gaussian Radial Basis kernel function. Misclassification error rate (MSE) is used as the measure to evaluate the models.

Table 4.1: Important attributes from SDSS Catalog and their description

Attribute	Description
petroMag_ug	Petrosian magnitude colors. A color was calculated for four independent pairs of bands in SDSS (u-g, g-r, r-i, and i-z).
petroRad_u*z	Petrosian radius, transformed with redshift to be distance-independent.
invConIndx_u	Inverse concentration index. The ratio of the 50% flux Petrosian radius to the 90% flux Petrosian radius.
isoRowcGrad_u*z	Gradient of the isophotal row centroid, transformed with redshift to be distance-independent.
isoColcGrad_u*z	Gradient of the isophotal column centroid, transformed with redshift to be distance-independent.
isoA_u*z	Isophotal major axis, transformed with redshift to be distance-independent.
isoB_u*z	Isophotal minor axis, transformed with redshift to be distance-independent.
isoAGrad_u*z	Gradient of the isophotal major axis, transformed with redshift to be distance-independent.
isoBGrad_u*z	Gradient of the isophotal minor axis, transformed with redshift to be distance-independent.
isoPhiGrad_u*z	Gradient of the isophotal orientation, transformed with redshift to be distance-independent.
texture_u	Measurement of surface texture.
lnLExp_u	Log-likelihood of exponential profile fit (typical for a spiral galaxy).
lnLDeV_u	Log-likelihood of De Vaucouleurs profile fit (typical for an elliptical galaxy).
fracDev_u	Fraction of the brightness profile explained by the De Vaucouleurs profile.

One-class SVM Results

We sampled our data set into two different categories, E90s and S90s based on the type of galaxy and the confidence level associated with the classification from the Galaxy Zoo volunteers. Our goal is to develop a classifier to identify the pure elliptical and spiral galaxies, E90s and the S90s respectively.

One-class SVM for Elliptical Galaxies

We developed a one-class SVM (SVME90) model to classify the E90s in the data set. Out of the 9115 galaxies in the E90 sample 8579 galaxies has been identified as pure elliptical galaxies by the galaxy zoo team. We used this as our class labels in building the one-class SVM classifier. We used 70% of the sample for training the SVME90 model and the remaining 30% of our sample for testing. The generated model is evaluated using misclassification rate. Also, Confusion matrix of the predicted class has been documented in Table 2(a). The SVME90 model generated has a predictive accuracy of 20% and 24% on the training and test set respectively. Using the SVME90 model, we predicted the classes of galaxies in the E70 and E50 sample and the details of the predictions are shown in Table 2(b), 2(c). In order to evaluate the difference between the two types of galaxies we applied the SVME90 model to our S90 sample and the result are shown in Table 2(d).

Table 4.2: Confusion matrix with different data sets using SVME90

Applying SVME90 to test data		Predicted class	
		E90	NotE90
Actual class	E90	2040	122
	Not E90	543	30

Applying SVME90 to E70 sample		Predicted class	
		E90	NotE90
Actual class	Not E90	506	589

Applying SVME90 to E50 sample		Predicted class	
		E90	Not E90
Actual class	Not E90	286	768

Applying SVME90 to S90 sample		Predicted class	
		E90	Not E90
Actual class	Not E90	279	14598

One-class SVM for Spiral Galaxies

We repeated the same experiment with the spiral galaxies. We developed a SVMS90 model and tested with the S90, S70, S50 and the E90 sample. Results from these experiments are summarized in table 3. We can see the results follow similar trend.

Table 4.3: Confusion matrix with different data sets using SVMS90

Applying SVMS90 to test data		Predicted class	
		S90	NotS90
Actual class	S90	4061	1098
	Not S90	165	54

Applying SVMS90 to S70 sample		Predicted class	
		S90	NotS90
Actual class	Not S90	1333	354

Applying S90 to S50 sample		Predicted class	
		S90	Not S90
Actual class	NotS90	801	342

Applying SVMS90 to E90 sample		Predicted class	
		S90	NotS90
Actual class	Not S90	1779	7336

Discussion of one-class SVM results

The confusion matrices in tables 2 and 3 clearly addresses the main problem with our galaxy zoo data set. The one class SVM classifiers trained using the 90% sample for the elliptical and spiral galaxies can differentiate the E90s and the S90s with high accuracy. But these classifiers when applied to to samples with less confidence like the E70s and E50s performs badly. This is clearly evident from the confusion matrices. This suggests that these galaxies (with low confidence) are quite different from the 90% sample and the galaxy zoo volunteers help us in identifying these galaxies. Also, the current SDSS parameters space is not sufficient enough for the machines to train better classifiers involving these galaxies.

4.3.2 Random Forest

Random forest [55] is an ensemble classifier that consists of several decision trees that try to classify the same task and combines the output like other ensemble classifiers to come up with rules that help in the classification task. It is one of the most accurate and popular algorithms that work well with most data sets where the class population is unbalanced. We

assumed that galaxies with different levels of voter confidence belong to different classes, and so we built a random forest model that classifies these galaxies.

Random Forest Results

Our data set now consists 6 different classes and we built a simple random forest model to classify them. The generated random forest model has a training and test error of 18%. This clearly shows that it has good overall performance on our data set. Table 4 shows the confusion matrix respect to the random forest model. The last column in table 4 has the error associated with each class. This clearly shows that even though the random forest model has a good overall accuracy of 88%, it suffers in identifying the E70, E50, S70 and S50 type galaxies. This clearly confirms the trend that we saw with the one-class SVM classifiers. The machine learned classifiers trained using SDSS parameters has trouble in classifying these galaxies similar to the Galaxy Zoo volunteers. This indicates that these galaxies are different from the type of galaxies (elliptical/spiral galaxies) that are previously known to the astronomers.

These results clearly suggest that the sample with low confidence among the volunteers (E70, E50, S70, S50) are different from the galaxy samples with high confidence in the feature space. The results from one-class SVM and random forest confirms that the galaxies with low confidence occurs in our dataset not because of erroneous classification by the volunteers. They are truly different from the galaxies with high confidence in the feature space as well. One main goal of the future work with this dataset will be to develop new methods to analyze these galaxies with low confidence for the presence of true elliptical or spiral galaxies and try to separate them. Also they need to be analyzed for the presence of any new classes which can help us understand the confusion among the galaxy zoo volunteers in trying to classify them.

Table 4.4: Confusion matrix on the test set using random forest

		Predicted class						
		E90	E70	E50	S90	S70	S50	Class error(%)
Actual Class	E90	2702	14	4	37	5	1	2
	E70	215	29	14	46	9	8	90
	E50	78	20	12	62	17	126	96
	S90	40	5	3	4389	29	10	2
	S70	30	11	7	478	44	27	92
	S50	33	7	112	130	34	27	95

4.4 Summary

These results clearly suggest that the popular classification algorithms such as one-class SVM and Random Forest do not work well with our complex data set. This could be the trend with other such rule based or space partitioning algorithms. The classification task that we are dealing with is quite different from other traditional data mining tasks. The SDSS catalog has photometric features extracted from the high resolution images from the telescopes. These include scientific parameters such as Petrosian radius containing 90% and 50% of Petrosian flux, Petrosian magnitude, DeVaucouleurs fit, DeVaucouleurs fit a/b, Exponential fit a/b, Exponential fit scale radius, measurement of surface texture etc. Machine learning algorithms used for galaxy classification utilizes these photometric attributes from the SDSS catalog and generates rules for classification. Whereas the Galaxy Zoo volunteers utilize their visual perception to extract information from the images and use their cognitive skills to come with with a decision. The result of applying human cognition to these galaxy images results in the class labels for each of the galaxies. Apart from the class labels, we have photometric features extracted from the SDSS catalog. When we build a classifier we utilize only these photometric parameters recorded by the SDSS.This is analogous to the classification task by the galaxy zoo volunteers but it is not the same. Instead they use human vision. Here human cognitive skills processes the image as whole and comes up with a decision. This demands for the need of latent variable models. Latent variable models have been widely used in machine learning, but their applications have been

widely limited. Latent variables models used wisely can help us extract the parameters used by the galaxy zoo volunteers and using these latent variables when used along with the photometric attributes from the SDSS catalog will help us model the human decision making skills. Modeling these hidden features and extracting the hidden rules calls out the need for new-sophisticated methods. Therefore, in the following chapters we will see the use the hidden variable models to build a classifier that mimics human decision making capabilities.

Chapter 5: Bayesian Nonparametric Analysis of Crowdsourced Citizen Science Data, with application to Interestingness Discovery

In this chapter we will address most of the problems that we encountered in the previous chapter in developing automated classification models from the SDSS parameters that match with the galaxyzoo volunteer provided classifications. The problem with the models generated in the previous chapter is mostly with respect to the grouping of galaxies into specific category by manual threshold. In this chapter we developed a non-parametric Bayesian framework and let the framework group them into different categories from the data and see how the groupings affect our automated classification task using photometric features from SDSS catalog. Also, we will closely look into the misclassified instances from such classifier and discuss how we arrived at a meaningful conclusion about these instances. Towards the end will discuss about how this could be applied to similar problem faced in other domains with some illustrations.

5.1 Introduction

Cluster analysis is the identification of groups of observations that are cohesive and separated from other groups. Interest in clustering has increased recently due to the emergence of several new areas of application. These include data mining, which started from the exploratory search for groupings of customers and products in massive retail datasets; document clustering and the analysis of web use data; gene expression data from micro-arrays; and image analysis, where clustering is used for image segmentation and quantization. Most clustering done in practice is largely based on heuristics. One widely used class of methods involves hierarchical agglomerative clustering, in which two groups chosen to optimize some

criterion are merged at each stage of the algorithm. Another common class of methods is based on iterative partitioning, in which data points are moved from one group to another until there is no further improvement in some criterion. K-means clustering in one such popular algorithm that falls under this class [56].

Cluster analysis can also be based on probability models. [57, 58] provide a brief survey about such techniques. This has provided insight to when a particular clustering method can be expected to work well and has led to the development of new clustering methods. It has also been shown that some of the most popular heuristic clustering methods are approximate estimation methods for certain probability models. For example, standard k-means clustering is equivalent to known procedures for approximately maximizing the multivariate normal classification likelihood when the covariance matrix is the same for each component and proportional to the identity matrix.

Many popular clustering algorithms require the number of clusters to be known a priori to choose an approximate number. By contrast, Dirichlet process mixture models (DPMMs) provide a non-parametric Bayesian framework to describe distributions over mixture models with an infinite number of mixture components. In this chapter we will define a DPMM based clustering approach on the galaxy classifications acquired from the galaxyzoo dataset to identify the actual number of classes in the dataset. Later we will build a classification model based on these clusters and explore the potential of such classifier in automated discovery of unlabeled galaxies. To aid thinking about our algorithm, the number of clusters present in the dataset can be viewed as latent class that are implicitly defined in the classifications provided by the galaxyzoo volunteers.

5.2 Background and Related Work

5.2.1 Mixture Model Based Clustering

In statistics, mixture model is a probability model for representing the presence of subpopulation within an overall population, without requiring that an observed data set should

identify the subpopulation to which an individual observation belongs. Formally this represents the mixture distribution that represents the probability distribution of observations in the overall population.

Given the data y with independent multivariate observations $y_1, ...y_n$, the likelihood for a mixture model with G components is

$$\mathcal{L}_{MIX}(\theta_1, ...\theta_G; \tau_1, ...\tau_G|\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(y_i|\theta_k). \tag{5.1}$$

where f_k and θ_k are the density and parameters of the k^{th} component in the mixture and τ_k is the probability that an observation belongs to the k^{th} component ($\tau_k \geq 0; \sum_{k=1}^{G} \tau_k = 1$).

Most commonly f_k is the multivariate normal(Gaussian) density and ϕ_k, parameterized by its mean μ_k and covariance matrix Σ_k,

$$\phi_k(\mathbf{y}_i|\mu_k, \Sigma_k) \equiv \frac{exp(-\frac{1}{2}(\mathbf{y}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{y}_i - \mu_k)}{\sqrt{det(2\pi\Sigma_k)}} \tag{5.2}$$

Data generated by the mixture of multivariate normal densities are characterized by groups or clusters centered at the means μ_k, with increased density for points nearer the mean. The corresponding surfaces of constant density are ellipsoidal. Geometric features (shape, volume, orientation) of the clusters are determined by the covariances Σ_k, which may also be parameterized to impose cross-cluster constraints. Common instance include $\Sigma_k = \lambda I$ constant across clusters, where all clusters are spherical and of same size; $\Sigma_k = \Sigma$ contains cross clusters, where all clusters have same geometry but need not be spherical. A general framework [59] for geometric cross-cluster constraints in multivariate normal mixtures by parameterizing covariance matrices is of the form

$$\Sigma_k = \lambda_k D_K A_k {D_K}^T. \tag{5.3}$$

where D_K is the orthogonal matrix of eigenvectors, A_K is a diagonal matrix whole elements are proportional to eigenvalues and λ_K is an associated constraint of proportionality. These represent a set of independent parameters that define the geometric properties of the clusters. More extensive enumeration of possible models resulting from this formation can be found in [60]

5.2.2 Nonparametric Mixture Model based clustering using Dirichlet Process

Latent variables in statistics are variables that are not being observed but are inferred from the variables that are observed. Latent variables are widely used in psychology, economics, life sciences and machine learning. In machine learning, many problems involve collection of high-dimensional multivariate observations and then hypothesizing a model that explains them. An appealing representation for such a model is a latent variable model. The role of latent variable is to represent the properties of objects that have not been directly observed. Bayesian statistics is often used for inferring latent variables. Examples of popular latent variable models include graphical models and dynamical system models. Discovering latent variables in graphical models relies heavily on local search heuristics such as expectation maximization (EM). Bayesian non-parametric latent feature models [12] is one of the recent approach to latent variable modeling in which the number of latent variables in unbounded i.e. there is no upper limit on the number of latent variables. Each data point can be associated with a set of possible latent variables.

Assume we have N objects, represented by an $N \times D$ matrix X, where the i^{th} row of this matrix, x_i, consists of measurements of D observable properties of the i^{th} object. In a latent feature model, each object is represented by a vector of latent feature values f_i, and the properties x_i are generated from a distribution determined by those latent feature values. Latent feature values can be continuous, as in principal component analysis, or discrete. For our discussion the latent features are assumed to be continuous. Let F be the matrix that indicates the latent feature values for all N objects, the model is specified by

a prior over features, p(F), and a distribution over observed property matrices conditioned on those features, p(X|F). Here p(F) specifies the number of features, their probability, and the distribution over values associated with each feature, while p(X|F) determines how these features relate to the properties of objects. In non-parametric Bayesian latent variable models, p(F), is defined without placing an upper limit on the number of features. Feature matrix F is considered to have two components: a binary matrix Z indicating which features are possessed by each object, with $z_{ik} = 1$ if object i has feature k and 0 otherwise, and a second matrix V indicating the value of each feature for each object. In sparse latent feature models only a subset of features take on non-zero values for each object, and Z picks out these subsets.

A prior on F can be defined by specifying priors for Z and V separately, with p(F) = P(Z)p(V). Nonparametric Bayesian latent feature models focus on defining a prior on Z. This is done by defining a prior over infinite binary matrices. The literature on nonparametric Bayesian models suggest starting with a model that assumes a finite number of features, and consider the limit as the number of features approaches infinity. This is done by simple generative process called the Indian buffet process for the distribution which is analogous to the Chinese restaurant process. The posterior can derived using Markov chain Monte Carlo algorithms.

5.3 Dirichlet Process Mixture Model (DPMM)

A Dirichlet process (DP) [61] [62] [63], parameterized by a base distribution G_0 and a concentration parameter *aplha*, is used as a prior over the distribution G of mixture components. For data points X, mixture component parameters *theta*, and a parameterized distribution F, the DPMM can be written as [64]

$$
\begin{aligned}
G|\alpha, G_0 &\sim DP(\alpha, G_0) \\
\theta_i|G &\sim G \\
x_i|\theta_i &\sim F(\theta_i).
\end{aligned}
\tag{5.4}
$$

One type of DPMM can be implemented as an infinite Gaussian mixture model in which all parameters are inferred from the data [63]. The generic function fits a Dirichlet process mixture of normal model for density estimation [65]

$$
\begin{aligned}
y_i|\mu_i, \Sigma_i &\sim \mathcal{N}(\mu_i, \Sigma_i), i = 1, ..., n \\
(\mu_i, \Sigma_i)|G &\sim G \\
G|\alpha, G_0 &\sim DP(\alpha G_0)
\end{aligned}
\tag{5.5}
$$

where, the baseline distribution is the conjugate normal-inverted-Wishart,

$$
G_0 = \mathcal{N}(\mu|m_1, (1/k_0)\Sigma) IW(\Sigma|\nu_1, \psi_1) \tag{5.6}
$$

To complete the model specification, independent hyper-priors are assumed (optional),

$$
\begin{aligned}
\alpha|a_0, b_0 &\sim Gamma(a_0, b_0) \\
m_1|m_2, s_2 &\sim \mathcal{N}(m_2, s_2) \\
k_0|\tau_1, \tau_2 &\sim Gamma(\tau_1/2, \tau_2/2) \\
\psi_1|\nu_2, \psi_2 &\sim IW(\nu_2, \psi_2)
\end{aligned}
\tag{5.7}
$$

where a_0, b_0 giving hyper-parameters for prior distribution of the precision parameter of the Dirichlet process prior, α giving the value of the precision parameter, ν_2 and ψ_2^{-1} giving the hyper-parameters of the inverted Wishart prior distribution for the scale matrix, ψ_1, of the inverted Wishart part of the baseline distribution, τ_1 and τ_2 giving the hyper-parameters for the gamma prior distribution, m_2 and s_2 giving the mean and the covariance of the normal prior for the mean, m_1, of the normal component of the baseline distribution, respectively, ν_1 and $\psi_1 - 1$ giving the hyper-parameter of the inverted Wishart part of the baseline distribution and, m_1 giving the mean of the normal part of the baseline distribution (it must be specified if m_2 is missing) and, k_0 giving the scale parameter of the normal part of the baseline distribution (it must be specified if τ_1 is missing). Note that the inverted-Wishart prior is parameterized such that if $A \sim IW_q(\nu, \psi)$ then $E(A) = \psi^{-1}/(\nu - q - 1)$.

5.3.1 Prior Specification

We hold little prior information about the distribution for parameters that are nested two or more layers in our modeling hierarchy, so we select lower magnitude hyper-parameter values under proper prior distributions that may be readily updated by the data. For the specific model that we implemented for our experiments in this chapter we set the values for the following hyper-parameters described in (refer equation here)

1. $m_2 = (180, 3)$ and $s_2 = \left(\begin{smallmatrix} 10000 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ giving the mean and covariance of the normal prior for the mean m_1.

2. $\psi_2^{-1} = \left(\begin{smallmatrix} 10000 & 0 \\ 0 & 1 \end{smallmatrix}\right)^{-1}$ giving the hyper-parameters of the inverted Wishart prior distribution for the scale matrix, ψ_1, of the inverted Wishart part of the baseline distribution.

3. $\alpha = 0.5$ giving the value of the precision parameter.

4. $\tau_1 = 1$ and $\tau_2 = 100$ giving the hyper-parameters for the gamma prior distribution.

5. $\nu_1 = 4$ giving the hyper-parameter of the inverted Wishart part of the baseline distribution.

6. $\nu_2 = 4$ giving the hyper-parameter of the inverted Wishart prior distribution for the scale matrix.

5.3.2 Inference under model parameterization

We have described above a class of Bayesian Non-parametric mixture models. This model posit a generative probabilistic process of a collection of observed data that includes hidden structure. We analyze data with these models by examining the posterior distribution of the hidden structure given the observations. This gives us a distribution over which latent structure likely generated our data. Thus the basic computation problem in Bayesian nonparametric modeling is computing the posterior which is not available in closed form. The most widely used posterior inference methods are Markov Chain Monte Carlo (MCMC) methods. The idea in MCMC methods is to define a Markov Chain on the hidden variables that has the posterior as its equilibrium distribution. By drawing samples from this Markov chain, one eventually obtains samples from the posterior. A simple form of MCMC sampling is Gibbs sampling, where the Markov chain is constricted by considering the conditional distribution of each hidden variable given the others and the observations.

5.3.3 Latent Class Discovery

To aid thinking about the algorithm, the types of galaxies in our galaxyzoo dataset can be viewed as samples from the latent class that are implicitly defined by the distribution of votes (number of volunteers agreeing upon particular task). Specifically, we define the latent class as sample of galaxies that received diverse set of classifications from the volunteers. We cannot reason directly about the class of galaxies since we do not know what they are a priori. This leads to the idea of studying them with the scientific parameters from the SDSS catalog. We assume that clusters identified by the DPMM form a contiguous region on some manifold in the 2-D vote space. Under this assumption, clustering of the galaxies in the vote space can be used to approximate the latent number of classes present in the dataset. We later analyzed the clusters parameterized by the Gaussians into galaxy classes

using traditional classification algorithm. Next section briefly describes this process with some sample dataset and then the results from our galaxyzoo dataset.

5.4 Experiments

This section explains the performance of Dirichlet Process Mixture Model clustering developed in previous section. Before we apply the model to the Galaxyzoo dataset, several experiments were performed to understand the performance of the model and understand the intuition behind the misclassified instance. First, we will test this on the island of games dataset and then move on to a simulated data where we will introduce the notion of most representative sample or the support vectors and then carry that notion to the most important classification task with Galaxyzoo dataset.

5.4.1 Island of Games Dataset

In this section, we will study the performance of Dirichlet Process Mixture Model clustering to Island of games datasets. Prior to applying DPMM, we restricted ourself from any visual inspection of this dataset in-order to analyze the true performance of the model. Later we will see how some systematic visualizations can reveal the actual structure in the dataset. The dataset was downloaded from [66]. As mentioned in [67]this dataset is special in its own way as it eludes several off-the-shelf data mining tools. These tools fail to capture the underlining structure of this data. The dataset consists of thousand of inhabitants that enjoy competing at chess, checkers and Rubik's Cube puzzle. The islanders are rated based on their skills at each of the three games and the ratings fall between 0 and 1. For better understanding of the data, we have provided some visualizations even though this wasn't the case with our actual exercise. Figure 5.1 shows the sample histogram for chess ratings. It is evident that each category seems to follow uniform distribution. Initial analysis of this dataset reveal that there isn't any correlation among the attributes. Figure 5.2, 5.3, 5.4 confirms this lack of correlation and requires further analysis. Table 5.1 shows the summary of the correlations. The structure is not apparent if only 2 dimensions are contained at a

Table 5.1: Correlations between Skills

	Chess	Checkers	Rubik's Cube
Chess	1.0000		
Checkers	0.0530	1.0000	
Rubik's Cube	0.0452	-0.0049	1.0000

time. The real structure in this dataset can be revealed by coloring Figure 5.2 using the Rubik's cube ratings. In Figure 5.5 green and red colors indicate Rubik's cube ratings above 0.5 and blue and yellow colors for Rubik's cube ratings below 0.5. Figure 5.6 shows the 3-D view of the ratings with the same color-coding. The x-y-z coordinates are the three game ratings for each islander. The points are framed by large cube, divided into eight small cubes. We can see that these data points are well contained within four small cubes. This data set is a simple and perfect example where standard statistical analysis nor visualizations are guaranteed to reveal the true structure in the data. We will now see the performance of DPMM with this data.

Dirichlet Process Mixture Model (Equation 5.7) when applied to this data set without any assumption about the number of clusters, reveals the presence of six clusters in the data even-though it could well be represented by 4 clusters (Figure 5.6). Figure 5.7 shows the results from DPMM. Though the number of clusters is not the exact reflection of true clusters, the results from DPMM are quite satisfactory. The three clusters represented by blue, red and pink reflects the actual cluster in the data. But DPMM sub divides the fourth cluster into 3 different clusters. This is the result of DPMM accounting for some additional mixture component to explain the actual data distribution within that particular cluster.

5.4.2 Another Toy Example - Simulated Dataset

We now test the DPMM model with another simulated dataset. Similar to the Island of games dataset we restricted ourself from any visual inspection of the data before applying DPMM. The dataset was generated from multivariate normal with different means and variance. It consists of two different classes representing the two normal distributions from

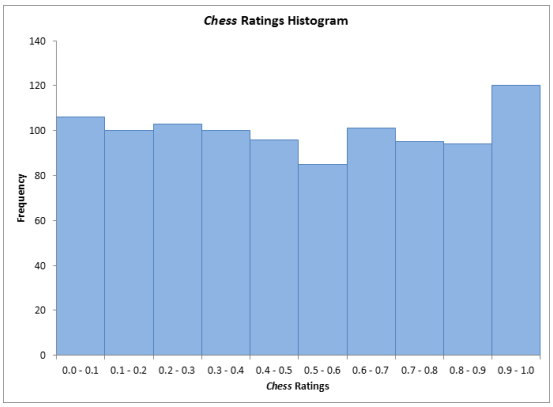

Figure 5.1: Chess Ratings Histogram.

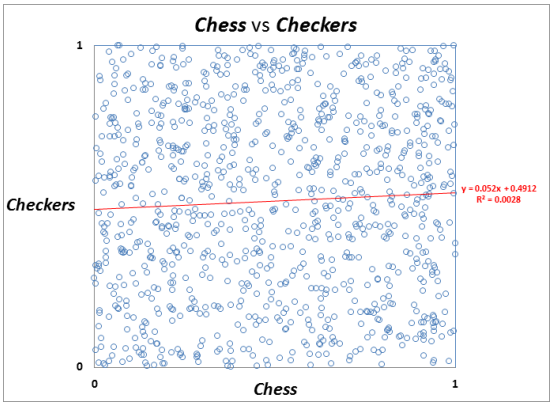

Figure 5.2: Scatter plot of Chess and Checkers ratings.

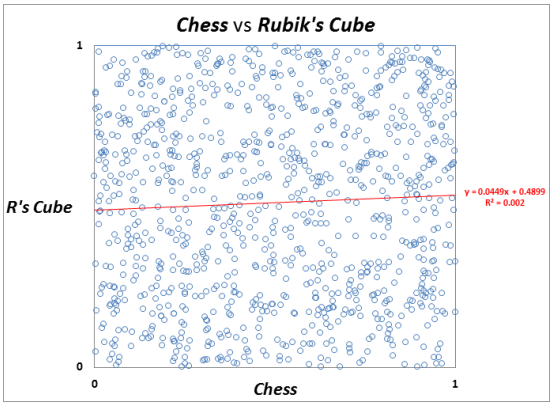

Figure 5.3: Scatter plot of Chess and Rubik's Cube ratings.

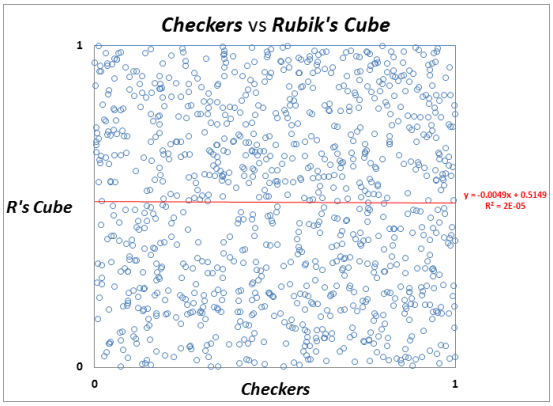

Figure 5.4: Scatter plot of Checkers and Rubik's Cube ratings.

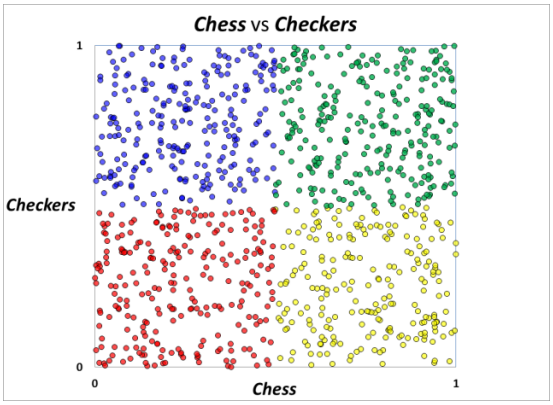

Figure 5.5: Scatter plot of Chess and Checkers ratings color coded.

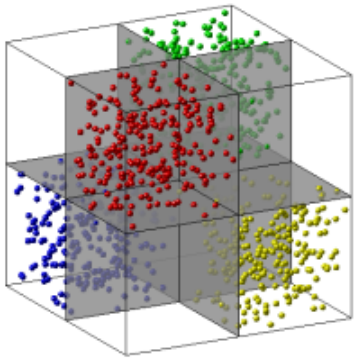

Figure 5.6: Three dimensional look at Island of Games dataset

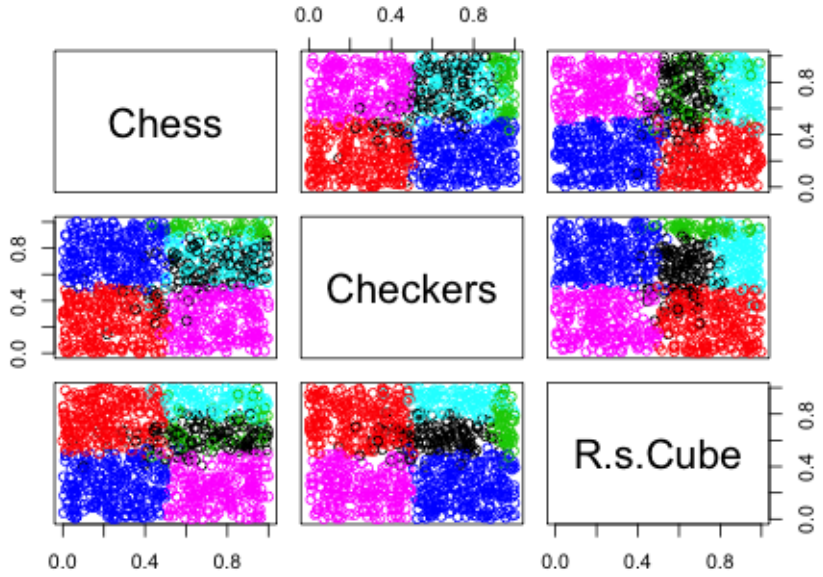

Figure 5.7: Scatter plot with clustering results.

Table 5.2: Confusion Matrix showing predicted class from DPMM

		Predicted Class	
		Class 1	Class 2
True Class	Class 1	6060	7
	Class 2	2	5851

which the data was generated. We ran the DPMM to this data without any prior knowledge on the actual number of classes. As expected the DPMM model identified the two clusters in the data. Figure 5.8 shows the cluster outputs from the two different classes. As expected, you can notice the presence of two well separated clusters. We evaluated the performance of DPMM by comparing the results from DPMM to the ground truth which is the actual distribution from the data being generated. Figure 5.9 compares the results from DPMM to the ground truth. Similar to any machine learning algorithms, DPMM is subject to errors. Here the misclassified instances are the ones to which DPMM assigns different class labels. Table 5.2 presents the confusion matrix of the results from DPMM. Among the 2000 instances DPMM misclassified only 8 of them which is negligible. In Figure 5.9 you can see the misclassified instances from DPMM being colored in blue and green. We were surprised by the location of these misclassified instances in the 2-D space feature space as we can clearly see that theses misclassified instances lie at the boundary of their respective clusters. This leads to a different angle of viewing the misclassified instances from DPMM as this can be the most representative points or the support vectors that define these cluster boundaries. It is this notion about the misclassified instances that we carry on to the next section when we apply the DPMM to the real world dataset from GalaxyZoo. This helps us understand about the galaxies that causes confusion among the GalaxyZoo volunteers. Also, this can address the poor accuracy with the popular machine learning models when applied to the GalaxyZoo dataset that we identified in our previous chapter.

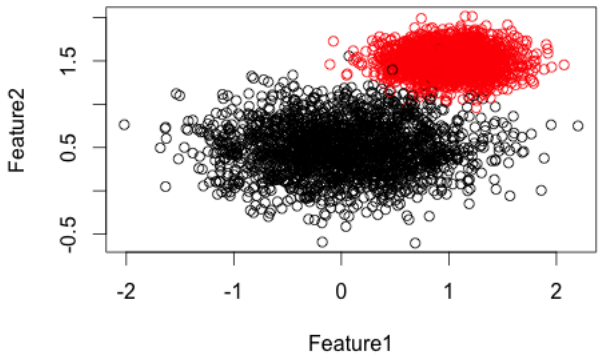

Figure 5.8: Scatter of Simulated Dataset with Two Classes.

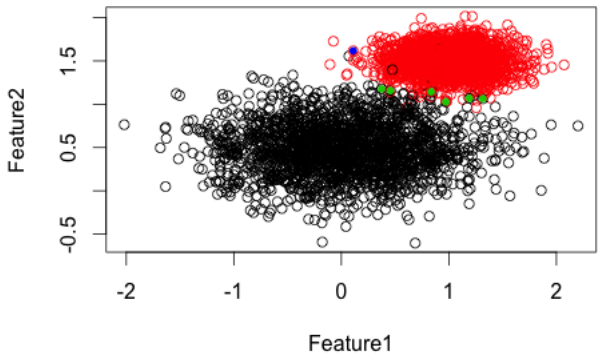

Figure 5.9: Scatter of Simulated Dataset with DPMM Clustering.

5.4.3 Galaxy Zoo Dataset

In this section we will apply the DPMM to the galaxyzoo dataset and see how it performs. The dataset being used for this has the sample of galaxies used in the previous chapter. The only difference here is that instead of photometric features of galaxies we used the information gathered from galaxyzoo framework. The notion behind the selection of galaxyzoo parameters is to bridge the gap between the photometric features and the classifications provided by the citizen science volunteers as the photometric features alone fail at explaining the complex nature of the galaxy classifications. Table 5.4 contains the list of parameters used for clustering with DPMM and their description. As each galaxy in galaxyzoo framework is labeled by multiple volunteers, these attributes represent the percentage of people who agreed upon a particular class. In the previous chapter we grouped these galaxies into Elliptical and Spiral galaxies with multiple levels of confidence namely E90, E70, E50, S90, S70 and S50 respectively by simple thresholding on these two parameters. Here in this experiment we let the DPMM work with the diverse opinions from the volunteers and choose the number of classes by clustering on these opinion. DPMM identifies the number of clusters in the 2-D vote space and returns the cluster results.

The results from DPMM are plotted in Figure 5.10. From the figure it is clear that DPMM identified four clusters. These four clusters are colored in black, green, blue and red. Figure 5.17 plotted these clusters separately in the 2-D vote space. By looking at the location of the clusters in the 2-D space in these plots it is clear that black and the green clusters are the one in which majority of the people agree on a class. So the black and the green cluster are considered to be true spiral and true elliptical galaxies. Also red and the blue clusters are the ones in which most people seem to have problem in agreeing to a single class. These are the one that get very diverse opinions from the citizen science volunteers and towards the rest of this chapter we focus our attention towards these set of galaxies.

In order to better understand the sample of galaxies in these red and blue cluster, we reverted back to similar analysis that we performed in our previous chapter to see if there is anything peculiar about these red and blue clusters. We are moving on from unsupervised

learning on 2-D vote space to more traditional supervised learning on the photometric features. We make this transition smooth by assuming the four clusters identified by DPMM to be our ground truth for the classifier. We trained a random forest model on the photometric features from the SDSS catalog to classify these four classes.

The trained random forest model has the overall accuracy of about 72% on the test dataset. Confusion matrix of the results from the test data are shown in Table 5.3. From the table it is clear that the random forest model is able to classify the black and the green clusters which correspond to the true elliptical and true spiral galaxies with minimal error (15% and 7% respectively) even though the overall accuracy is only 72%. The overall accuracy is affected by the poor performance in classifying the red and the blue clusters.

Further we moved our attention to the galaxies that are misclassified by the random forest model. Figure 5.12 5.13 plots the random forest provided classifications of the galaxies in the black and green cluster respectively on the 2-D vote space to compare it with the galaxyzoo volunteer provided classifications. The location of these misclassified instances are being consistent with the citizen scientist's classifications. These mis-classifications are purely due to the error in the random forest model. We moved our attention to the two other clusters colored in red and blue. It is these galaxies that receive diverse set of classifications and our goal is understand the nature of these galaxies. Also from the Table 5.3 it is clear that the the random forest suffers greatly in classifying these galaxies. Further analysis of these classifications in Figure 5.14, 5.15 leads us to a new set of conclusion with respect to these mis-classifications. The galaxies that are identified to be in the red and blue clusters are truly different from the true elliptical and true spiral galaxies. This could be the results of artifact or the object in the images could be a star or something else. We are able to identify this from our analysis.

Further we are interested in the misclassified instances from the random forest model. The location of these instances in the vote space tell us that these galaxies are quite quite closer to the black and the blue clusters in the 2-D vote space. This leads us to a new conclusion that these galaxies and similar to the true elliptical and true spiral galaxies that

Table 5.3: Confusion Matrix showing predicted Galaxy class from DPMM

		Predicted Class				
		Class 1 (Black)	Class 2 (Red)	Class 3 (Green)	Class 4 (Blue)	Class Error(%)
True Class	Class 1 (Black)	3460	39	12	583	0.15
	Class 2 (Red)	260	478	356	381	0.67
	Class 3 (Green)	8	49	2736	141	0.067
	Class 4 (Blue)	1034	114	286	1400	0.50

Table 5.4: Description of Galaxyzoo Attributes used with DPMM

Attribute	Description
p_el	Percent Elliptical
p_cs	Percent Spiral

are most representative to their class (elliptical and spiral) and we were able to identify these set of galaxies. We correspond this to notion of support vectors in support vector machines that truly help is identifying the classes. Also, these set of galaxies could used for active learning purposes where the goal is to identify the most informative data for training. This leads us to major breakthrough where we were able to identify the most representative samples using when the number of labeled samples available for machine learning is very limited. It is this results that is quite different and a major improvement from what is shown in the previous chapter where the machine learning algorithms suffered to be consistent with that of the volunteers.

Further we extended this analysis to other clusters we see that the results are consistent with what we encountered before with the black clusters. Figures, shows this trend.

5.4.4 Dependency to Baseline Distribution

In the previous section we saw the results of the DPMM pertaining to a set of hyperparameters. The visual evidence in the cluster results indicate the chances for further more

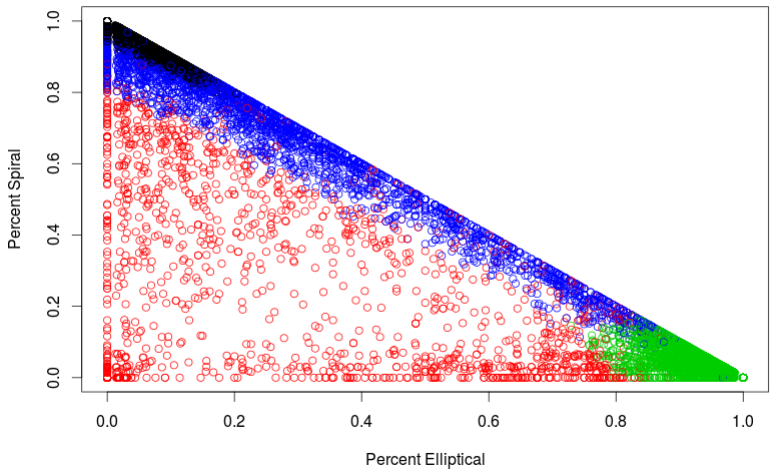

Figure 5.10: Clusters in 2-D Vote Space discovered by DPMM.

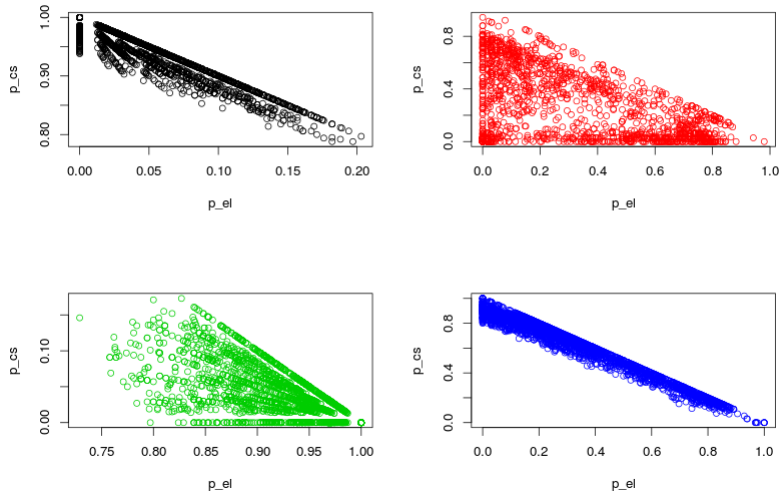

Figure 5.11: Clusters in 2-D Vote Space (from DPMM) Plotted Separately .

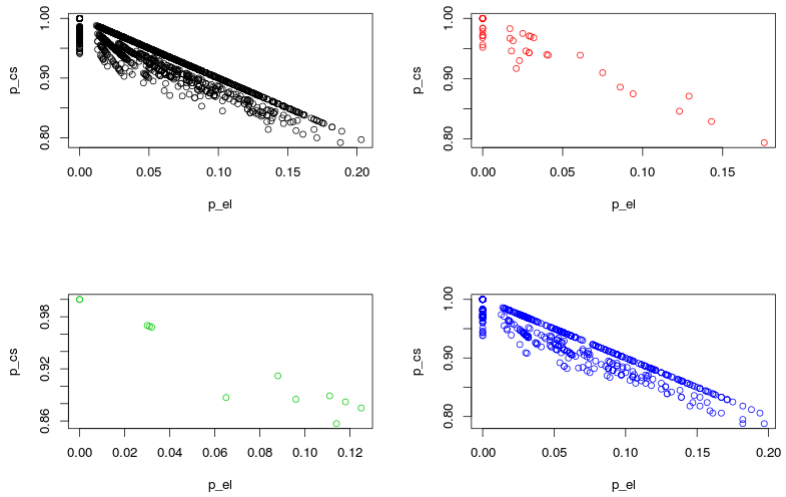

Figure 5.12: Black Cluster: assigned labels from Random Forest model.

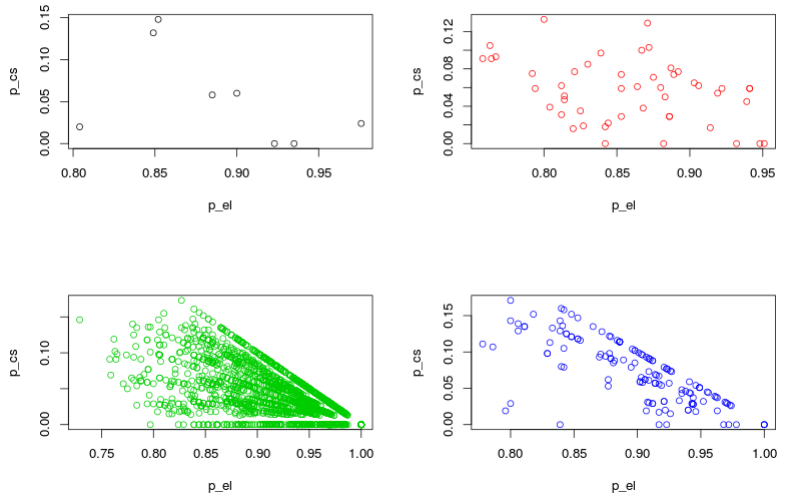

Figure 5.13: Green Cluster: assigned labels from Random Forest model.

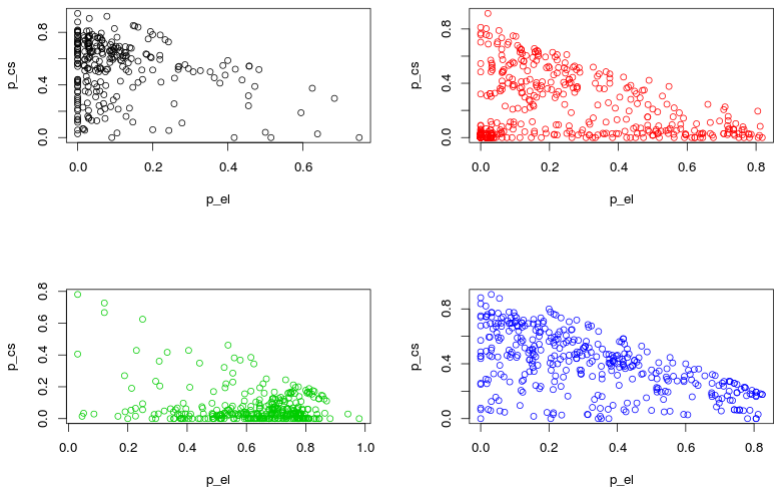

Figure 5.14: Red Cluster: assigned labels from Random Forest model.

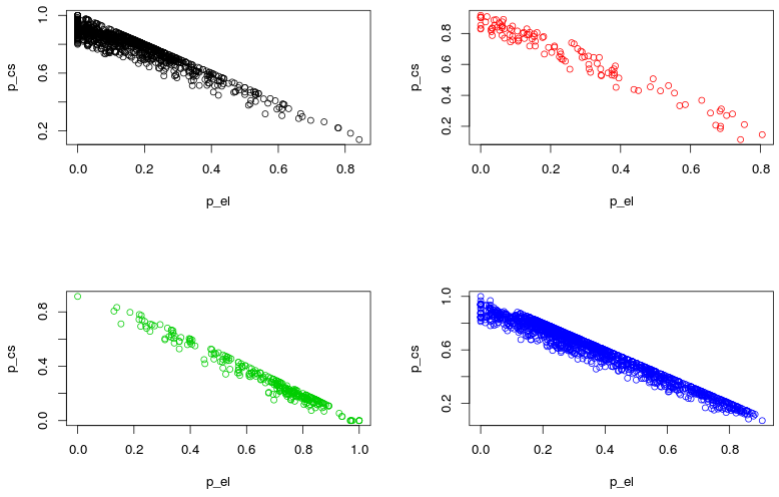

Figure 5.15: Blue Cluster: assigned labels from Random Forest model.

clusters. So several models of DPMM are tried by varying the prior for the baseline distribution and the choice of baseline distribution impacts the results of DPMM. This is being widely discussed in [68] . Specifically, for our model the Normal-inverse-Wishart prior have some unappealing properties with prior dependencies between the mean and covariance parameters and is discussed in [69] . The baseline distribution of the Dirichlet process pertains to the uncertainty about the between group population distribution in these models. It is this uncertainty that leads to different clusters. In Nonparametric Bayesian literature, Deviance Information Criterion (DIC) is being widely used to determine the accuracy of the model.

The same procedure explained in the previous section has been repeated with the following set of hyperparameters.

1. $m_2 = (180, 3)$ and $s_2 = \left(\begin{smallmatrix} 10000 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ giving the mean and covariance of the normal prior for the mean m_1.

2. $\psi_1^{-1} = \left(\begin{smallmatrix} 0.5 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ giving the hyper-parameters of the inverted Wishart prior distribution for the scale matrix, ψ_1, of the inverted Wishart part of the baseline distribution.

3. $\alpha = 0.5$ giving the value of the precision parameter.

4. $\tau_1 = 1$ and $\tau_2 = 100$ giving the hyper-parameters for the gamma prior distribution.

5. $\nu_1 = 4$ giving the hyper-parameter of the inverted Wishart part of the baseline distribution.

6. $\nu_2 = 4$ giving the hyper-parameter of the inverted Wishart prior distribution for the scale matrix.

This is same as the prior hyperparameters in the previous section expect for the parameters of the inverted Wishart distribution which acts as the scale parameters of the Dirichlet distribution. The impact of this can been seen in Figure 5.16. This is slightly different than the one in Figure 5.10 where there seems to be two clusters in the region identified by color red. This gets sorted out if increase the amount of uncertainty between the group

Table 5.5: Confusion Matrix showing predicted Galaxy class from 5 cluster DPMM Model

		Predicted Class					
		Class 1 (Black)	Class 2 (Red)	Class 3 (Green)	Class 4 (Blue)	Class 5 (Cyan)	Class Error(%)
True Class	Class 1 (Black)	1438	106	308	933	19	0.48
	Class 2 (Red)	323	257	14	399	33	0.74
	Class 3 (Green)	170	0	2470	5	43	0.08
	Class 4 (Blue)	517	57	11	3350	5	0.14
	Class 5 (Cyan)	94	8	458	23	296	0.66

population distribution. Based on visual evidence and also the DIC we conclude that this being the correct model for this clustering talk.

The same random forest model is applied to the galaxy dataset with the labels changed to this new cluster allocation (5 clusters). The results from the random forest can be in Table 5.5. Similar plots of the predictions from random forest for each cluster/class in shown in Figures 5.18, 5.20, 5.19, 5.21, 5.22 respectively.

5.4.5 Discussion of Results

The combination of supervised algorithm such as random forest and unsupervised clustering technique like DPMM helped unearth the two major drawbacks that we faced right through the thesis. The success of DPMM revolves around the two factors listed below :

1. The presence of the expert labels present in the database which the human vision helped to identify using the crowd-sourcing techniques. But feeding this information for future machine learning algorithm requires identifying these unknown expert labels from the list of available feature in the database.

2. The lack of sufficient knowledge in the database is another reason affecting the accuracy of the classifiers. DPMM helped us bridge this lack of information utilizing the cognitive power of citizen scientists.

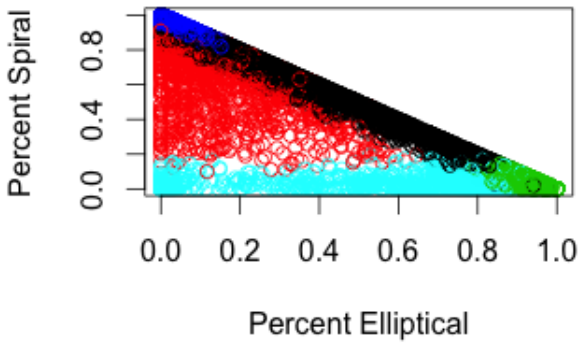

Figure 5.16: Five Clusters in 2-D Vote Space discovered by DPMM.

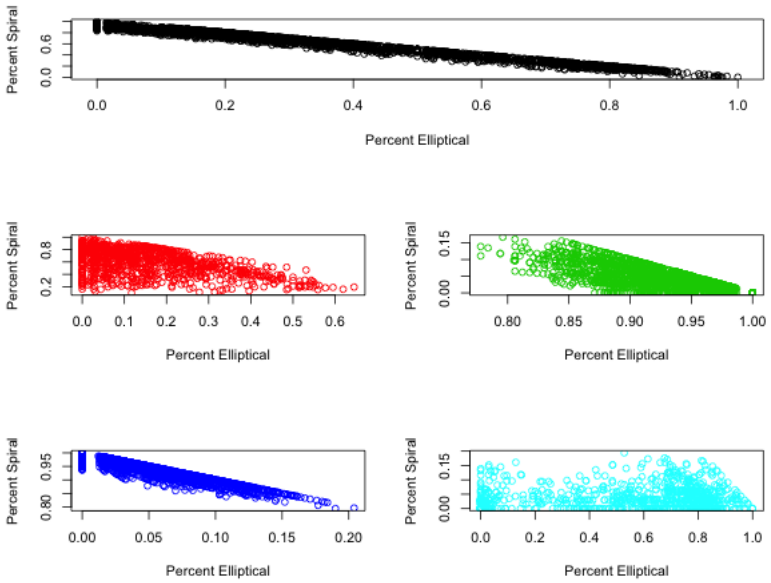

Figure 5.17: Five Clusters in 2-D Vote Space (from DPMM) Plotted Separately .

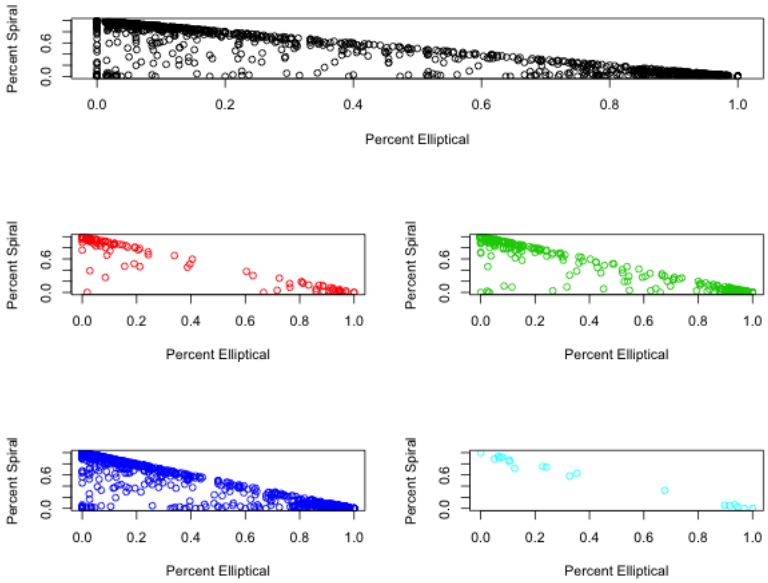

Figure 5.18: Black Cluster: assigned labels from Random Forest model.

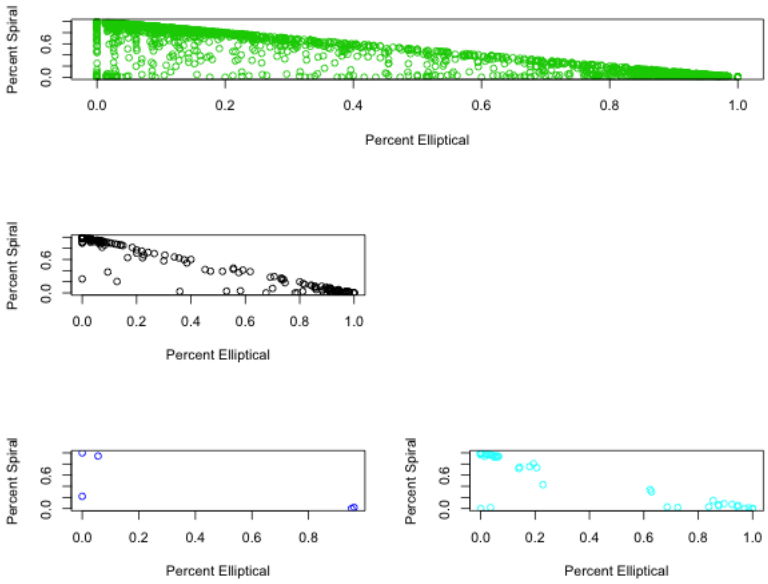

Figure 5.19: Green Cluster: assigned labels from Random Forest model.

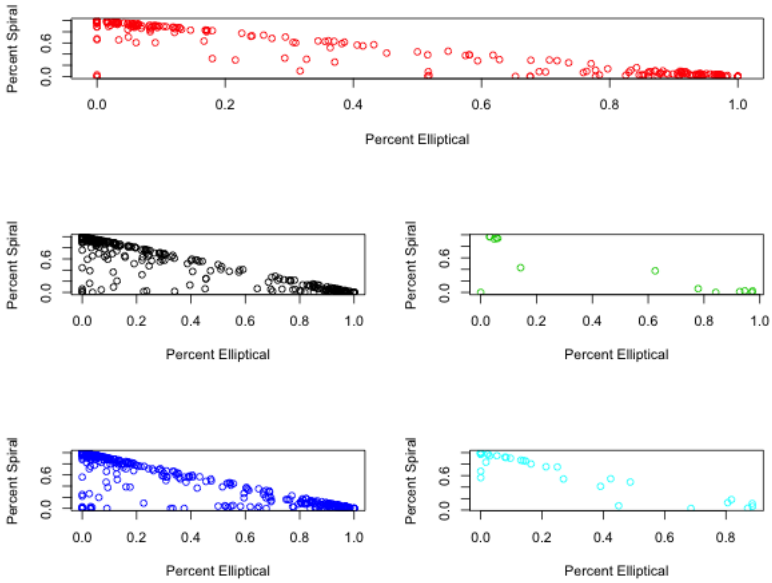

Figure 5.20: Red Cluster: assigned labels from Random Forest model.

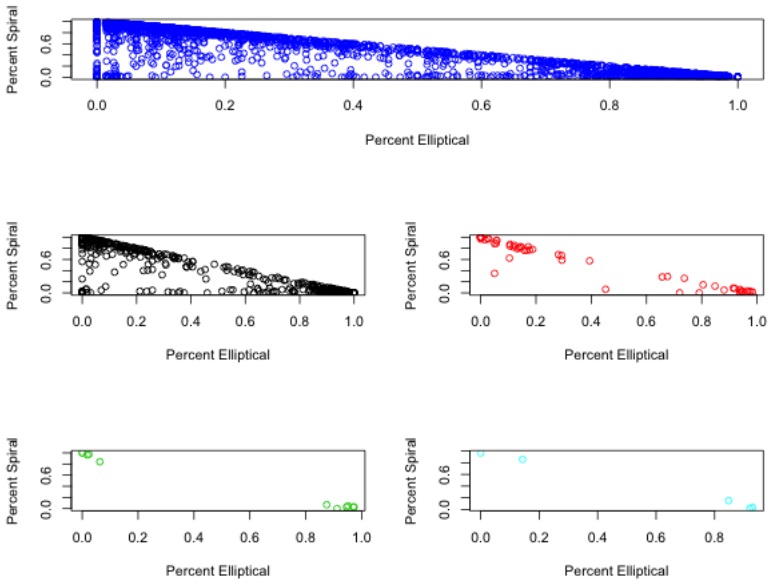

Figure 5.21: Blue Cluster: assigned labels from Random Forest model.

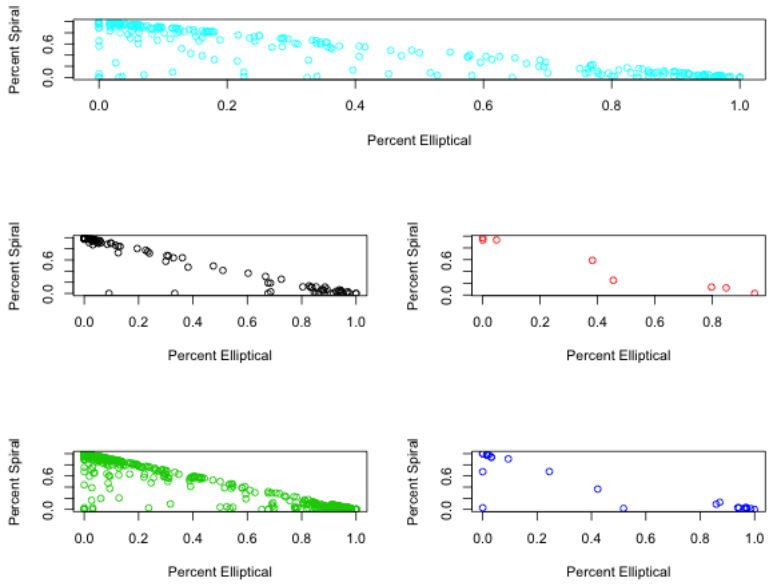

Figure 5.22: Cyan Cluster: assigned labels from Random Forest model.

5.5 Veterans Hospital Application: Suicide Prevention

A similar work [70] deployed on line for the elicitation of the opinions among the experts who derive from a diversity of knowledge areas with the goal of prioritizing a set of objects that are relevant for policy making in government, science and industry sectors. The main goal is to develop consensus between a group of experts by employing multiple scoring rounds. Multiple stakeholders (patients, doctors, clinicians, counselors, policymakers, family members, DoD, VA administrators) are asked to vote on a list of 12 research goals in the area of Suicide Prevention, producing a scored ranking of the goals. Each list and the final scored list here represent a projection of the real research goals that need to be implemented: these are the observed vectors. The real ranks of the goals represent the latent variable vector or the truth is derived from the from the various observed vectors. A Bayesian framework employing Dirichlet Process Mixture Models (DPMM) similar to the one in equation 5.7 was developed to find the real ranking among the 12 set of goals.

5.6 Summary

We have adapted a version of Dirichlet process mixture model clustering to identify the number of clusters in the galaxy classifications provided by the citizen science volunteers that helps in identifying the number of classes in the dataset outside of the two types of galaxies. We trained a random forest model from the photometric features extracted from the SDSS catalog that helps in identifying these clusters from the photometric features. This model could help to remove the "human in the loop" for any future sky surveys such as LSST. Also, comparing the results from the random forest model we were able to identify the most representative instances or what is called support vectors in the support vectors machines classifiers. These misclassified galaxies from the random forest model are the galaxies that lie closest to the decision surface and are most difficult to classify. This leads us to finding the training sample of galaxies that can be used in finding any decision function for the big data from LSST.

Chapter 6: Conclusion and Future Work

This thesis has described a overall framework for finding interestingness in the data that comes in different form. First in terms of interesting observations that can lead to scientific discovery. A general algorithm based on K-Nearest Neighbors was presented for the identifying such interesting objects.

This chapter reviews these results in the light of the original research questions and goals of the thesis. It re-examines the open issues and challenges of such data oriented discovery and shows how the presented work meets the stated goals of the thesis. The chapter concludes with other possible applications and future directions of the research.

6.1 Challenges and Solutions

A general discussion of the open issues and challenged in making data oriented scientific discovery was presented in chapter 1. The three main challenged were of interest to the thesis: First, development of techniques that can help in finding interesting objects in any given data sets. Second, identifying key features from the datasets that can explain the underlying reason behind certain scientific phenomenon. Third, how to find unknown features that were not captured in the datasets that can be utilized in extracting the most interesting or the representative sample belonging to any particular class. All these are aimed at preparing the scientists, especially astronomers deal with the data challenges faced by the future sky survey telescopes such as LSST.

6.1.1 Thesis Generated Publications

Designing machine learning algorithms that can be used to make scientific discovery is a central motivation of the work presented in this thesis. In the course of writing the thesis,

a number of algorithms were developed and several interesting applications to already existing machine learning algorithms were proposed that can harness the power of big data.

List of Thesis generated articles and Talks:

1. Baehr, Steven, et al. "Data Mining the Galaxy Zoo Mergers." CIDU. 2010.

2. Borne, Kirk D., and Arun Vedachalam. "Surprise detection in multivariate astronomical data." Statistical Challenges in Modern Astronomy V. Springer New York, 2012. 275-289.

3. Vedachalam, Arun. "Machine Learning Explorations of Citizen Science Data." Talk given at Chapman University Symposium on Big Data and Analytics: 44th Symposium on the Interface of Computing Science and Statistics

6.2 Suggestions for Future Work

A few research issues were stated in some of the chapters of the thesis but were reserved for future work. This section briefly expands on some of these issues and other research topics stemming out from this thesis and warranting future investigation.

Chapter 2 presented a novel algorithm developed in the course of writing this thesis. For better results from this algorithm, a theoretical study of how the value of K and choice of p-value affect the effectiveness of the algorithm need to studied. Thus rigorous theoretical investigations are required to establish stability, generalization bounds of the methods. Also, applicability of this method to the results produced in chapter 4 needs to evaluated to confirm the other set of results i.e the true outliers in the dataset.

Chapter 5 presented a whole new dimension towards the idea of surprise discovery or representative learning by extracting latent features. This approach is feasible in this thesis because of the presence of crowd sourced citizen science information. Extending this approach to other traditional data sets require some tweaking to the model. A theoretical study of the needed adjustments the prior and the assumed baseline distribution need

to be further understood. Also, the relationship of the baseline distribution to the model accuracy need to be studied. Computationally, applicability of more advanced Variational Bayesian as an alternative to MCMC methods, for approximating intractable integrals arising in Bayesian inference need to be evaluted. Also these Non-parametric Bayesian latent variable models rely heavily on the prior distribution to incorporate the domain knowledge for discovering latent variables. Recently, [71] proposed methods that can overcome this idea of using the prior imposing constraints on the posterior distribution.The literature on nonparametric Bayesian has made some recent advancements to extracting multiple latent variables from the data. This can be done by replacing the Dirichlet Process used in our model by some simple generative process called the Indian buffet process or the Chinese restaurant process.

Second part of the problem is to understanding the latent variables. This can be done exploring the relationship with the observed variables that are in the SDSS catalog. Several correlation measures and dimensionality reduction methods such as PCA can be used in getting the relationship. Since we arent aware of the relationship that exists between the latent variables and the observed variables, more generalized measures like the Maximal information coefficient (MIC)[72] or distance correlation [73] can be used.The fundamental plane of elliptical galaxies is one the well know problem in astronomy that explains the relationship between the effective radius, surface brightness and the velocity dispersion of elliptical galaxies. The elliptical galaxies lie on a plane in the three dimensional space. It has been well demonstrated in the astronomy literature that this relationship when properly utilized can help in classifying elliptical galaxies. There are other examples in machine learning literature as well that utilize correlations among features for classifications. Having latent variables extracted from the galaxy zoo data set provides us with few parameters that are previously unknown or unmeasured. Searching these new parameter space for any such relationship like the fundamental plane opens up the door for scientists for more detailed analysis. Since we are not aware of what the latent variable actually represent, it is not possible to predict the type of relationship that exists the observed variables. So MIC like

measure makes sense and can help astronomers understand the type of relationship between the latent variables and existing variables. This might help in extracting some new scientific features from the images that can lead to novel science discovery.

Also for the problems explained in Chapter 3 and Chapter 5 the applicability of other machine learning methods such as Deep Learning(DL) [74] need to be evaluated as it attempts to model high-level abstractions in data.

Bibliography

Bibliography

[1] A. J. Hey, S. Tansley, K. M. Tolle *et al.*, *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research Redmond, WA, 2009, vol. 1.

[2] K. Borne, "Scientific data mining in astronomy," *arXiv preprint arXiv:0911.0505*, 2009.

[3] J. Howe, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[4] K. R. Lakhani, "Innocentive. com (a)," *Harvard Business School Case*, no. 608-170, 2008.

[5] D. C. Brabham, "Moving the crowd at istockphoto: The composition of the crowd and motivations for participation in a crowdsourcing application," *First monday*, vol. 13, no. 6, 2008.

[6] "OpenScientist: Finalizing a Definition of "Citizen Science" and "Citizen Scientists"." [Online]. Available: http://www.openscientist.org/2011/09/finalizing-definition-of-citizen.html

[7] "citizen, n. and adj." [Online]. Available: http://www.oed.com/view/Entry/33513

[8] J. Sauer, S. Schwartz, and B. Hoover, "The christmas bird count home page," 1996.

[9] J. P. Worthington, J. Silvertown, L. Cook, R. Cameron, M. Dodd, R. M. Greenwood, K. McConway, and P. Skelton, "Evolution megalab: a case study in citizen science methods," *Methods in Ecology and Evolution*, vol. 3, no. 2, pp. 303–309, 2012.

[10] L. Davies, J. Bell, J. Bone, M. Head, L. Hill, C. Howard, S. Hobbs, D. Jones, S. Power, N. Rose *et al.*, "Open air laboratories (opal): A community-driven research programme," *Environmental Pollution*, vol. 159, no. 8, pp. 2203–2210, 2011.

[11] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling, "ebird: A citizen-based bird observation network in the biological sciences," *Biological Conservation*, vol. 142, no. 10, pp. 2282–2292, 2009.

[12] H. Seung and L. Burnes, "Eyewire," *Available a t eyewire. org*, 2012.

[13] D. Clery, "Galaxy zoo volunteers share pain and glory of research," *Science*, vol. 333, no. 6039, pp. 173–175, 2011.

[14] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu *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, no. 3, pp. 1179–1189, 2008.

[15] K. D. Borne and Zooniverse Team, "The Zooniverse: A Framework for Knowledge Discovery from Citizen Science Data," *AGU Fall Meeting Abstracts*, p. C650, Dec. 2011.

[16] M. Banerji, O. Lahav, C. J. Lintott, F. B. Abdalla, K. Schawinski, S. P. Bamford, D. Andreescu, P. Murray, M. J. Raddick, A. Slosar, A. Szalay, D. Thomas, and J. Vandenberg, "Galaxy Zoo: reproducing galaxy morphologies via machine learning," *Monthly Notices of the Royal Astronomical Society*, vol. 406, no. 1, pp. 342–353, Jul. 2010. [Online]. Available: http://mnras.oxfordjournals.org/content/406/1/342

[17] D. Darg, S. Kaviraj, C. Lintott, K. Schawinski, M. Sarzi, S. Bamford, J. Silk, R. Proctor, D. Andreescu, P. Murray *et al.*, "Galaxy zoo: the fraction of merging galaxies in the sdss and their morphologies," *Monthly Notices of the Royal Astronomical Society*, vol. 401, no. 2, pp. 1043–1056, 2010.

[18] M. Way, "Galaxy zoo morphology and photometric redshifts in the sloan digital sky survey," *The Astrophysical Journal Letters*, vol. 734, no. 1, p. L9, 2011.

[19] C. J. Lintott, K. Schawinski, W. Keel, H. Van Arkel, N. Bennert, E. Edmondson, D. Thomas, D. J. Smith, P. D. Herbert, M. J. Jarvis *et al.*, "Galaxy zoo:hanny's voorwerp, a quasar light echo?" *Monthly Notices of the Royal Astronomical Society*, vol. 399, no. 1, pp. 129–140, 2009.

[20] S. Djorgovski and M. Davis, "Fundamental properties of elliptical galaxies," *The Astrophysical Journal*, vol. 313, p. 59, Feb. 1987. [Online]. Available: http://adsabs.harvard.edu/doi/10.1086/164948

[21] R. L. Davies, D. Burstein, A. Dressler, S. M. Faber, D. Lynden-Bell, R. J. Terlevich, and G. Wegner, "Spectroscopy and photometry of elliptical galaxies. II - The spectroscopic parameters," *The Astrophysical Journal Supplement Series*, vol. 64, p. 581, Aug. 1987. [Online]. Available: http://adsabs.harvard.edu/doi/10.1086/191210

[22] K. Das, K. B. Sug, C. Giannella, and H. Kargupta, "Scalable Distributed Change Detection from Astronomy Data Streams using Local, Asynchronous Eigen Monitoring Algorithms," in *In Proceedings of SDM09 (accepted*, 2009.

[23] A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel, "Finding large average submatrices in high dimensional data," *The Annals of Applied Statistics*, vol. 3, no. 3, pp. 985–1012, Sep. 2009, arXiv: 0905.1682. [Online]. Available: http://arxiv.org/abs/0905.1682

[24] C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. v. d. Berg, "Galaxy Zoo : Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179–1189, Sep. 2008, arXiv: 0804.4483. [Online]. Available: http://arxiv.org/abs/0804.4483

[25] R. A. Maronna and V. J. Yohai, "The Behavior of the Stahel-Donoho Robust Multivariate Estimator," *Journal of the American Statistical Association*, vol. 90, no. 429, pp. 330–341, Mar. 1995. [Online]. Available: http://www.jstor.org/stable/2291158

[26] G. T. Richards, R. P. Deo, M. Lacy, A. D. Myers, R. C. Nichol, N. L. Zakamska, R. J. Brunner, W. N. Brandt, A. G. Gray, J. K. Parejko, A. Ptak, D. P. Schneider, L. J. Storrie-Lombardi, and A. S. Szalay, "Eight-Dimensional Mid-Infrared/Optical Bayesian Quasar Selection," *The Astronomical Journal*, vol. 137, no. 4, p. 3884, Apr. 2009. [Online]. Available: http://iopscience.iop.org/1538-3881/137/4/3884

[27] N. L. Zakamska, M. A. Strauss, J. H. Krolik, M. J. Collinge, P. B. Hall, L. Hao, T. M. Heckman, Z. Ivezic, G. T. Richards, D. J. Schlegel, D. P. Schneider, I. Strateva, D. E. V. Berk, S. F. Anderson, and J. Brinkmann, "Candidate Type II Quasars from the Sloan Digital Sky Survey: I. Selection and Optical Properties of a Sample at 0.3<Z<0.83," *The Astronomical Journal*, vol. 126, no. 5, pp. 2125–2144, Nov. 2003, arXiv: astro-ph/0309551. [Online]. Available: http://arxiv.org/abs/astro-ph/0309551

[28] B. Berriman, D. Kirkpatrick, R. Hanisch, A. Szalay, and R. Williams, "Discover of Brown Dwarfs with Virtual Observatories," vol. 8, 2003, p. 60. [Online]. Available: http://adsabs.harvard.edu/abs/2003IAUJD...8E..60B

[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, no. 3, pp. L29–L33, Feb. 2003, arXiv: astro-ph/0212487. [Online]. Available: http://arxiv.org/abs/astro-ph/0212487

[30] K. Borne, "Scientific Data Mining in Astronomy," *arXiv:0911.0505 [astro-ph, physics:physics]*, Nov. 2009, arXiv: 0911.0505. [Online]. Available: http://arxiv.org/abs/0911.0505

[31] A. A. Freitas, "On Objective Measures of Rule Surprisingness." in *Proceedings of the Second European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'98*. Springer-Verlag, 1998, pp. 1–9.

[32] W. Weaver, "Probability, rarity, interest, and surprise," *The Scientific Monthly*, vol. 67, no. 6, pp. 390–392, Dec. 1948.

[33] P. Smyth and R. M. Goodman, "An Information Theoretic Approach to Rule Induction from Databases," *IEEE Trans. on Knowl. and Data Eng.*, vol. 4, no. 4, pp. 301–316, Aug. 1992. [Online]. Available: http://dx.doi.org/10.1109/69.149926

[34] H. Dutta, C. Giannella, K. Borne, and H. Kargupta, "Distributed Top-K Outlier Detection from Astronomy Catalogs using the DEMAC System," in *Proceedings of the 2007 SIAM International Conference on Data Mining*, ser. Proceedings. Society for Industrial and Applied Mathematics, Apr. 2007, pp. 473–478. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/1.9781611972771.47

[35] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, Dec. 1965. [Online]. Available: http://www.jstor.org/stable/2333709

[36] M. J. Bayarri and J. O. Berger, "Measures of Surprise in Bayesian Analysis," in *Duke University*, 1997.

[37] M. Bayarri and J. Berger, "Quantifying surprise in the data and model verification," in *Bayesian Statistics 6*, J. Bernardo, J. Berger, A. Dawid, and A. Smith, Eds. Oxford University Press, 1999, pp. 53–82.

[38] V. altenis, "Outlier Detection Based on the Distribution of Distances between Data Points," *Informatica*, vol. 15, no. 3, pp. 399–410, Jan. 2004. [Online]. Available: http://iospress.metapress.com/content/HXM4R66MEWYYJ6KL

[39] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*, vol. 3, Aug. 2004, pp. 430–433 Vol.3.

[40] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-based Local Outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '00. New York, NY, USA: ACM, 2000, pp. 93–104. [Online]. Available: http://doi.acm.org/10.1145/342009.335388

[41] D. Pokrajac, A. Lazarevic, and L. Latecki, "Incremental Local Outlier Detection for Data Streams," in *IEEE Symposium on Computational Intelligence and Data Mining, 2007. CIDM 2007*, Mar. 2007, pp. 504–515.

[42] P. Dhaliwal, M. P. S. Bhatia, and P. Bansal, "A Cluster-based Approach for Outlier Detection in Dynamic Data Streams (KORM: k-median OutlieR Miner)," *arXiv:1002.4003 [cs]*, Feb. 2010, arXiv: 1002.4003. [Online]. Available: http://arxiv.org/abs/1002.4003

[43] P. Filzmoser, R. Maronna, and M. Werner, "Outlier Identification in High Dimensions," *Comput. Stat. Data Anal.*, vol. 52, no. 3, pp. 1694–1711, Jan. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.csda.2007.05.018

[44] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[45] D. W. Darg, S. Kaviraj, C. J. Lintott, K. Schawinski, M. Sarzi, S. Bamford, J. Silk, R. Proctor, D. Andreescu, P. Murray, R. C. Nichol, M. J. Raddick, A. Slosar, A. S. Szalay, D. Thomas, and J. Vandenberg, "Galaxy Zoo: the fraction of merging galaxies in the SDSS and their morphologies," *arXiv*, vol. 401, pp. 1043–1056, Jan. 2010.

[46] M. Banerji, O. Lahav, C. J. Lintott, F. B. Abdalla, K. Schawinski, S. P. Bamford, D. Andreescu, P. Murray, M. J. Raddick, A. Slosar, A. Szalay, D. Thomas, and J. Vandenberg, "Galaxy Zoo: reproducing galaxy morphologies via machine learning," *arXiv*, pp. 663–+, Apr. 2010.

[47] N. M. Ball, R. J. Brunner, A. D. Myers, and D. Tcheng, "Robust Machine Learning Applied to Astronomical Data Sets. I. Star-Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees," *apj*, vol. 650, pp. 497–509, Oct. 2006.

[48] A. Gauci, K. Zarb Adami, and J. Abela, "Machine Learning for Galaxy Morphology Classification," *ArXiv e-prints*, May 2010.

[49] D. Davies and D. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 224–227, 1979.

[50] J. Dunn, "Well separated clusters and optimal fuzzy-partitions," *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.

[51] "http://www.galaxyzoo.org/."

[52] D. M. J. Tax, *One-class Classification: Concept-learning in the Absence of Counter-examples*, 2001.

[53] D. M. J. Tax and R. P. W. Duin, "Uniform Object Generation for Optimizing One-class Classifiers," *J. Mach. Learn. Res.*, vol. 2, pp. 155–173, Mar. 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=944790.944809

[54] B. Schlkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001. [Online]. Available: http://dx.doi.org/10.1162/089976601750264965

[55] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://link.springer.com/article/10.1023/A%3A1010933404324

[56] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[57] H. H. Bock, "Probabilistic models in cluster analysis," *Computational Statistics & Data Analysis*, vol. 23, no. 1, pp. 5–28, 1996.

[58] ——, "Probabilistic aspects in classification," in *Data Science, Classification, and Related Methods*. Springer, 1998, pp. 3–21.

[59] J. D. Banfield and A. E. Raftery, "Model-based gaussian and non-gaussian clustering," *Biometrics*, pp. 803–821, 1993.

[60] G. Celeux and G. Govaert, "Gaussian parsimonious clustering models," *Pattern recognition*, vol. 28, no. 5, pp. 781–793, 1995.

[61] T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *The annals of statistics*, pp. 209–230, 1973.

[62] S. Jain and R. M. Neal, "A split-merge markov chain monte carlo procedure for the dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, vol. 13, no. 1, 2004.

[63] C. E. Rasmussen, "The infinite gaussian mixture model." in *NIPS*, vol. 12, 1999, pp. 554–560.

[64] R. M. Neal, "Markov chain sampling methods for dirichlet process mixture models," *Journal of computational and graphical statistics*, vol. 9, no. 2, pp. 249–265, 2000.

[65] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the american statistical association*, vol. 90, no. 430, pp. 577–588, 1995.

[66] W. Smith. Puzzle #9: Island of games. [Online]. Available: https://welltemperedspreadsheet.wordpress.com/2013/03/07/puzzle-9-island-of-games/

[67] K. Borne. Top 10 capabilities for exploring complex relationships in data for scientific discovery. [Online]. Available: http://www.datasciencecentral.com/profiles/blogs/top-10-capabilities-for-exploring-complex-relationships-in-data

[68] D. Görür and C. E. Rasmussen, "Dirichlet process gaussian mixture models: Choice of the base distribution," *Journal of Computer Science and Technology*, vol. 25, no. 4, pp. 653–664, 2010.

[69] M. West and M. D. Escobar, *Hierarchical priors and mixture models, with application in regression and density estimation*. Institute of Statistics and Decision Sciences, Duke University, 1993.

[70] T. D. Savitsky and S. R. Dalal, "Bayesian non-parametric analysis of multirater ordinal data, with application to prioritizing research goals for prevention of suicide," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 63, no. 4, pp. 539–557, 2014.

[71] J. Zhu, N. Chen, and E. P. Xing, "Bayesian inference with posterior regularization and applications to infinite latent svms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1799–1847, 2014.

[72] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[73] G. J. Székely, M. L. Rizzo, N. K. Bakirov *et al.*, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[74] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

Curriculum Vitae

Arun Vedachalam received his Bachelor of Engineering (B.E.) in Computer Science and Engineering from the Anna University, Chennai, India in 2008. He then moved to the US where he obtained his M.S in Computational Sciences from George Mason University, Fairfax, VA in 2012. In the Spring of 2009, Arun joined the PhD program in Computational Sciences and Informatics offered by the Department of Computational and Data Sciences (CDS) at George Mason University, Fairfax, VA. His concentration in the program is Computational Learning. Currently his main areas of research include: Computational Statistics, Astroinformatics, Distributed and Parallel Machine Learning, and Big Data Analytics. Arun plans to continue doing research and in particular on innovative ideas on Big Data. He is passionate about designing, improving, and using state of the art machine learning techniques to solve outstanding Big Data problems in areas such as Science, Computational Linguistics, operations research, and many other areas where he believes machine learning can make a difference. In this direction, after graduation, his immediate plan is to find a position as a data scientist in an environment that promotes scientific research.