Bayestar

Using Bayesian Networks to identify galaxies from image features

Julius Stopforth
University of Cape Town
Private Bag X3
Cape Town, Western Cape 7701
stpjul004@myuct.ac.za

ABSTRACT

Galaxy classification from images is implemented using a Bayesian network that incorporates shape descriptors in order to identify the galaxy type in terms of the Hubble sequence. The system was trained on various sized datasets using Variational Bayesian approximation in order to test over 10 thousand examples from the Galaxy Zoo dataset. When compared to a neural network, the Bayesian network performed poorly in predicting the values of the test data regardless of the size of the dataset used to learn the parameters of the network.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; *Bayesian network models*; • **Applied computing** → *Astronomy*;

1 INTRODUCTION

Astronomy is a field that actively accumulates and processes vast amounts of data with structure and information that may only be extracted using computational means [16]. For instance, large sky surveys such as the Sloan Digital Sky Survey (SDSS) [32] and the LINEAR survey [26] produce datasets many terabytes in size with tens of millions of data points [17] that require intense computer pre-processing and analysis [16]. South Africa will soon be one of the leading contributors to the field of radio astronomy with the construction of the Square Kilometre Array (SKA) [11]. The telescope will produce data on the order of exabytes in a single year of operation, at a rate of 2 terabytes per second during operation [11]. Therefore, analysis of this data will require the use of automated techniques such as machine learning in order to reduce data into consumable and useful knowledge as well as to perform analyses such as astronomical phenomenon classification [11, 29] in order to assist human processing of the data.

A cognitive vision system is one possible approach to processing the data. Cognitive vision systems are an attempt at merging current image processing, bottom-up machine learning methods, and top-down Knowledge representation and reasoning strategies [20]. These systems have strong roots in computer vision as they perform feautre extraction that is then used by the other two modules of the system. A good analogy is to view the system as a pipeline, with the feature extraction at the front and the two machine learning components processing the extracted features to perform tasks on the content of the images such as classification.

Knowledge representation and reasoning is a field of artificial intelligence that is able to utilise expert knowledge on a subject or area and enabling computer systems to solve complex issues. This differs from bottom-up AI because the system is initially given expert knowledge and then reasons about issues, or attempts to answer queries about the related field. For example, medical diagnosis can be achieved by representing the knowledge of diseases and their symptoms and then providing the system with a set of symptoms and have it reason as to which disease it is most likely to be[5]. Knowledge representation and reasoning can be achieved with many different strategies including Description Logics[1], and Bayesian networks[13].

A Bayesian network (otherwise known as a belief network[9, 13]) is usually implemented as a probabilistic directed acyclic graph. The graph functions in such a way that a traversal through it can represent a probable (plausible) deduction of a series of dependent variable (represented as nodes in the graph). For example, this would be akin to the steps of finding the most probable deduction as to what disease a person is suffering from depending on what symptoms they are presenting with. Bayesian network is able to act as both a knowledge representation as well as an inference engine [5, 13]. Furthermore, extending or modifying a Bayesian network is relatively easy [6, 18].

2 BACKGROUND

2.1 Bayesian networks

Bayesian network classifiers has successfullly been used for a variety of problems including facial expression recognition [8], medical diagnosis[23], and semantic image understanding [18].

2.1.1 Bayes Theorem. The statistical basis upon which Bayesian networks is built is that of Bayes Theorem. Bayes Theorem is a method of estimating the posterior probability of an event (A) occuring given that another event (B) is observed[30]. This is done by using the prior probability $P(A)$ and the amount of support that observing event B lends to A represented by $\frac{P(B|A)}{P(B)}$. Put more simply, Bayes Theorem enables one to represent how likely it is for an event to occur given some evidence and accounting for historical data. Thus, the probability of A occuring if B occurs is calculated using the following formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

With Bayesian networks, each node in the network represents an event that can be observed.

2.1.2 Inference and Inference Methods. Exact methods for calculating the priors and conditional probabilities for the nodes in the network exist as well as approximate methods [?]. The Junction tree algorithm is one example where the graph structure is decomposed by exploiting the Markov blanket of nodes that allows the probability of a node to be calculated based only on its direct parents and children.

Exact methods can be very computationally costly and so approximate methods were developed[12].

2.2 Hubble Classification Scheme

The Hubble classification scheme is central to the classification of galaxy morphologies in Astronomy today[15]. The scheme itself distinguishes between four major galaxy types: Elliptical, Lenticular, Spiral and Spiral Bar. Within each of these there are further sub types based on visual features and properties of the galaxy.

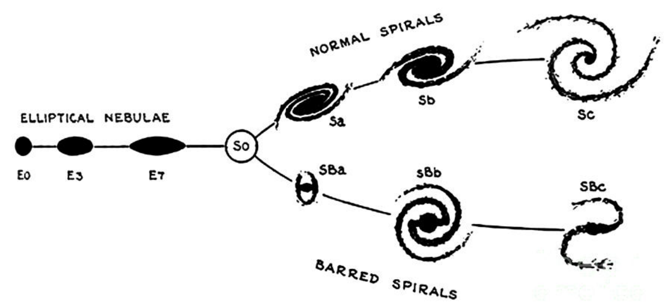

Figure 1: The tuning fork diagram for the Hubble sequence

2.3 Data

The data consists of the Kaggle Galaxy Zoo dataset[17] and a set of shape descriptors provided by a feature extractor. The shape descriptors that were extracted are those outlined by Goderya and Lolling [14].

2.4 Cognitive Vision Systems

There is some contention over the use of the term "cognitive system" and its meaning. Therefore, the term "cognitive vision system" will refer a system structured as a pipeline with the feature extraction and selection at the beginning followed by the automated machine learning and, knowledge representation and reasoning modules.

3 RELATED WORK

3.1 Existing Galaxy Classifiers

While extensive literature exists in the classification of galaxies in astronomy, most implement some form of neural network[2, 4, 27] or other bottom up technique. These approaches are very good at classifying galaxies, but fail to address the issue of reasoning or deduction in order to add knowlegde about the observation.

3.2 Kaggle Galaxy Zoo

The Galaxy Zoo project utilised crowd-sourcing in order to classify images from the SDSS [32]. The participants were not required to be educated astronomers or have any prior knowledge of astronomy before participating so that Galaxy Zoo (as the project was named) could attract as many possible participants. They admit that the screening for allowing the participants (referred to as "classifiers") was intentionally low, requiring that they only correctly classify 11 out of 15 galaxies from a control set determined by the members of Galaxy Zoo. Classification criteria was also simplified for the layman to more easily classify galaxies by their shape alone and not other factors (such as color). Furthermore, Linott et. al needed to negate a lot of human error when inputting answers that culminated in a roughly 4% loss of classifications. When bias testing was attempted, [17] confirms that the bias survey was affected by the Hawthorne effect that caused classifiers to be more cautious in classification when they suspected they were being tested for bias. They evaluated the sample generated from the SDSS by comparing it to three other datasets that had been prepared by expert astronomers (mention or cite those datasets) which suggests that in this project may be able to do the same to establish the confidence in classification. Finally, they conclude that harnessing the general public to participate in "citizen science" can provide reliable classification of galaxies. It is interesting to note that Linott et al mention that merely classifying galaxies based on morphology with automated systems (such as artificial neural networks) is limited in that they are not able to correctly investigate the influence that start formation could have on galaxy formation. This is something the reasoner may be able to address and fill the knowledge gap.

3.3 Semantic Image Understanding

Bayesian networks have been used to perform semantic scene understanding using extracted features from the image. These networks have been able to label images based on their contents [7]. Furthermore, frameworks have been developed with the intent to drive Bayesian image understanding[18].

4 SYSTEM DEVELOPMENT AND IMPLEMENTATION

The core of the implementation is based on the Bayespy library[19] written in the Python language. Python was chosen due to its many data manipulation libraries as well as its file utilities such as JSON and CSV file parsing. Furthermore, it was convenient to write analysis scripts in order to collect the necessary data for comaprison with relative ease.

4.1 Variational Bayesian Approximantion

The Bayespy library uses Variational Bayesian (VB) approximation [3] as its method of inference. This is an appropriate method to apply in the case of this project since the dataset is very large with over 60 000 individual entries. While a method such as expectation maximization (EM) [22] is similar to VB it performs poorly on large datasets or complex models. Similarly, Markov chain Monte Carlo (MCMC) methods also suffer when applied to large amounts of data. Variational Bayesian approximation also has the benefit of being a deterministic optimisation algorithm that is guaranteed to converge after a number of iterations [12, 28].

4.2 Model design

The topology of the network was constructed by drawing heavily from the decision tree provided in the Galaxy Zoo 2 paper (figure 3)[17]. Care was taken to ensure that the nodes were all correctly *d-separated* as suggested by [6]. Finally, this model ensures that there were no uneccesary nodes added to the network that may increase the complexity of the network without benefit. Therefore, it was decided to collapse certain nodes together (such as the "edge on" node) due to some classses specifying "true" for and others specifying "false" for the exact same node.

5 EXPERIMENT DESIGN AND EXECUTION

Since the Bayesian network is using Variational Bayesian approximation as its infernce method, the investigation aims to compare the performance of the Bayesian network when trained using different sized data batches.

In order to test the ability of the Bayesian network to identify galaxy types from the given shape descriptors in Section 2, the network will be tested by providing the shape descriptors, performing inference and then measuring the mean squared error (MSE) between the probabilities inferred and the actual probabilities for the classes of the Galaxy Zoo dataset.

To achieve this, the network was constructed with the same structure each time see figure 2, then it was given a batch of data containing both the shape descriptor values and the classes values. For each item in the training batch, the relevant nodes in the network were then observed, and inference was made to update the parameters of the network.

After the network had finished learning the parameters from the data batch, it was given a set of 10 thousand shape descriptors. The probabilities of the class nodes in the network were recorded after the relevant shape descriptors in the test case were observed and infernce was performed.

Once the testing had completed, the class probabilities of the network were compared to that of the Galaxy Zoo dataset and the mean squared error (MSE) was computed on a case by case basis. Finally, the average MSE value was calculated and compared to

5.1 Data

In total, there were more than 60 thousand sets of classes as well as 60 thousand sets of shape descriptors. While the class dataset originally contained 37 classes, only 16 were used to test the network as the other classes were determined to have a negligible influence on the type of the galaxy. Of the shape descriptors described in [14], 5 descriptors were chosen as they had the most influence in the classification of a galaxy type and were derived from the other shape descriptors such as "galaxy area" or "galaxy perimeter".

6 RESULTS AND FINDINGS

When the Bayesian network was given 50 thousand examples to learn from and then tested on approximately 11 thousand cases, it achieved an average MSE value of 0.284. In terms of the quality of the estimation, a value of 0.284 suggests that the Bayesian network is not particularly effective in predicting the correct value.

Figure 4 shows the plot of the MSE for each of the test cases. It was assumed that the MSE value would decrease over time, as the network would continue to update the network parameters as the testing proceeded. However, the plot gives no indication that the MSE of the predicitons improved as testing proceeded. Plotting the density distibution of the MSE for the test cases (figure 5) shows that the MSE values rarely dipped below 0.2, but at the same time would fluctuate as high as 4.5 or above.

Batch size	Avg. MSE
3125	0.236
6250	0.241
12500	0.275
25000	0.231
50000	0.284

Table 1: The average mean squared error for large batch sizes

Batch size	Avg. MSE
10	0.234
20	0.278
30	0.290

Table 2: The average mean squared error for small batch sizes

There was no real statistical difference in the average MSE for the various batch sizes as can be see in Table 1. Therefore, smaller batch sizes were tested with only 100 test examples to see if there was a significant difference between using very small batch sizes. As can be seen from Table 2 there is no significant difference between the smaller batch sizes of 10, 20, and 30. Finally, between the large and small batch sizes the average MSE did not show any significant discrepancy. This could suggest that the network learned very quickly, or that the network did not learn the probabilities very well. This was then compared to a neural network to see if network was performing well or not.

6.1 Comparison to Neural Network

Neural networks have often been used in galaxy classification [4, 27] (see also Section 3) and are a popular approach for solving multiple regression problems [25]. Since both Bayesian networks and neural networks are DAGs, it was a worthwhile investigation into how the Bayesian network would perform against a neural network on the Galaxy Zoo dataset.

The same mean squared error metric mentioned in section 5 was used to compare the Bayesian networks to the neural network as it is a good measure of the quality of an estimator [10].

The neural network achieved an average MSE value of 0.027 when tested against the same test cases as those for the large batch size Bayesian networks. In terms of quality of estimation the neural network was highly accurate in predicting the correct proability of the chosen classes. However, the neural network MSE data did contain one significant outlier value of 493.564 that, when taken into consideration, shifted the average MSE from 0.027 to 0.0695. Even when factoring in the outlier, the neural network handily outperforms the Bayesian network regardless of training batch size by at least one order of magnitude.

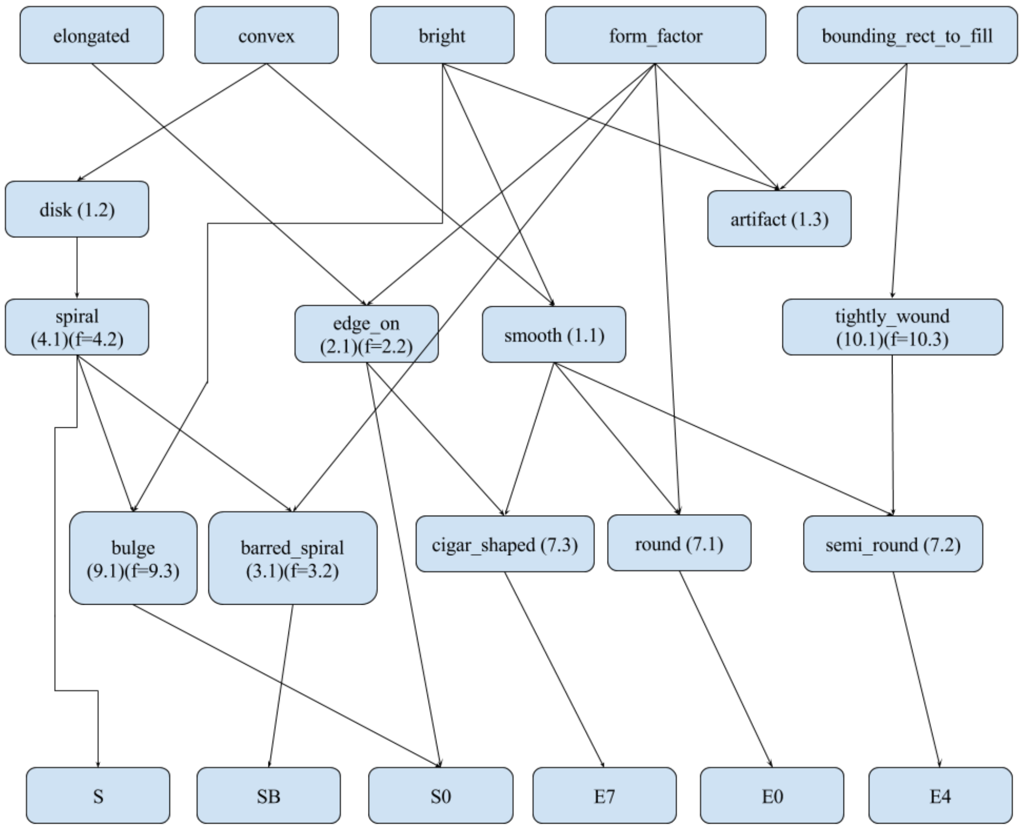

Figure 2: The topology of the network

Comparing a plot of the MSE values for the neural network (figure 7) to that of the Bayesian network (figure 4), it seems like the Bayesian network behaved similarly to the neural network in that it clustered around lower MSE values, but also maintained a consistent fluctutaion throughout testing. Additionally, the density plot for the neural network in figure 7 shows a sharp smooth peak, suggesting that the neural network was also more consistent in its predicitons.

6.2 Issues and Pitfalls

Ultimately, the tests and comparisons conducted were representative of a regression problem, and not a identification problem. Instead, the Bayesian network could be compared against a naive Bayes classifier instead as they compete well with other classifiers and are also in the Bayesian family[24].

Furthermore, use of an exact and not approximate inference method may have been able to produce more accurate results, and could be trained on smaller datasets than those used for the Variational Bayesian approximation method.

Additonally, the use of the mean square error as a metric for measuring the quality of predictions made by the Bayesian network may have been a poor choice as [31]. However, the minimum mean squared error is a metric for measuring the classification error in discrete classification[10].

One result that was rather worrying is that the MSE values showed no improvement during the testing. Even with the small batch sizes such as in figure 8 there seemed to be no obvious improvement during testing. It should be noted that the inference method used was built in to the library and was assumed to be correct.

Another issue regards the observation of the variables. The shape descriptor and class values for the learning of network parameters were mapped from continuous values onto the discrete values required by the nodes. This may have introduced some assumptions about the data that could have harmfully biased the results.

7 CONCLUSIONS AND FUTURE WORK

A Bayesian network that indetifies galaxy types from the shape descriptors in [14] perfoms poorly against a neural network when compared using the mean squared error. However, this discrepancy can be attributed to the nature of the tests and comparisons conducted were geared towards regression problems, while the design of the Bayesian network was for the identification of the galaxy type.

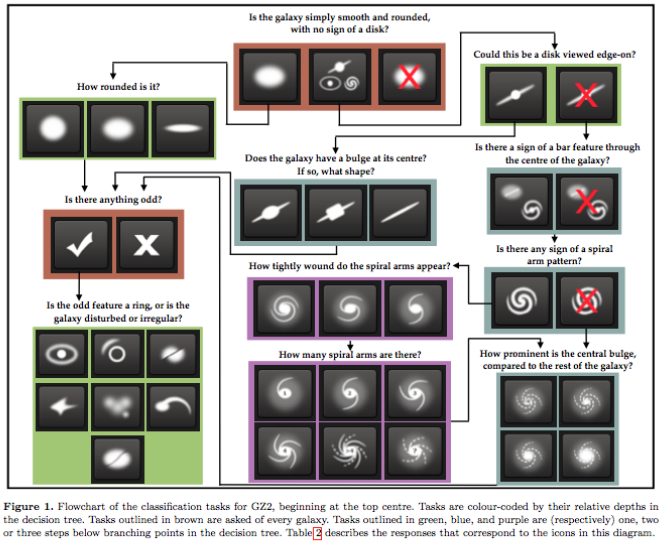

Figure 3: Visual representation of the Galaxy Zoo decision tree

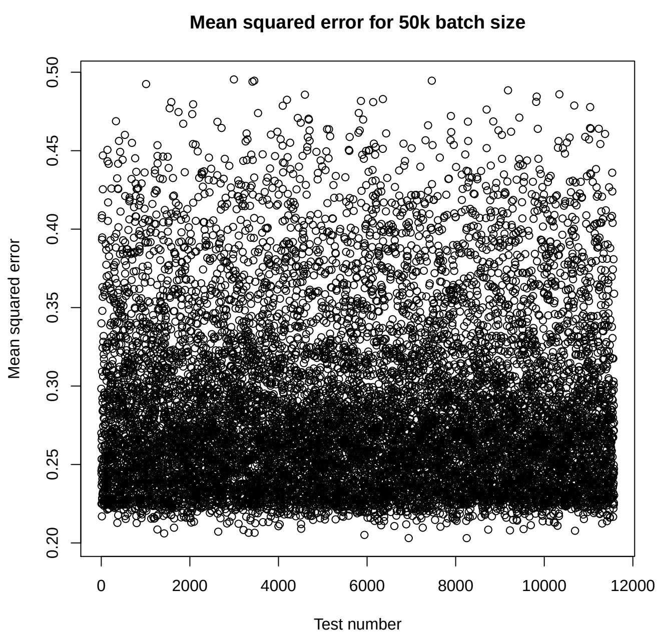

Figure 4: Plot of the mean squared error for batch size of 50k

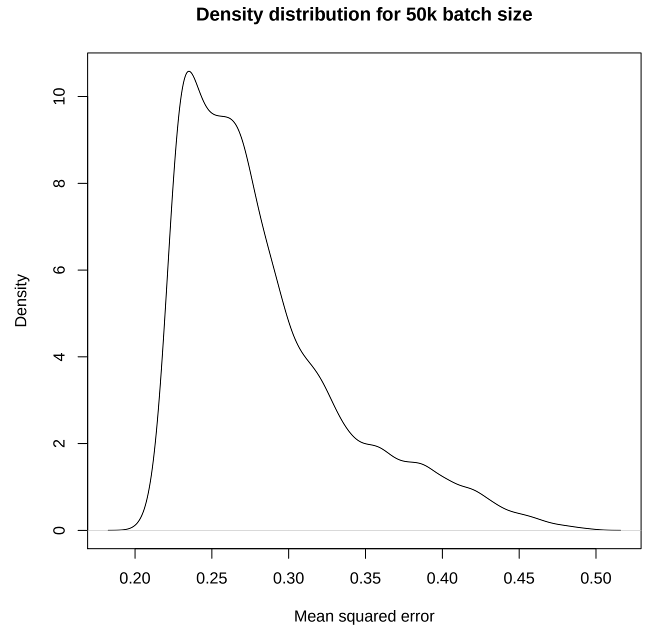

Figure 5: Plot of the density distribution for batch size of 50k

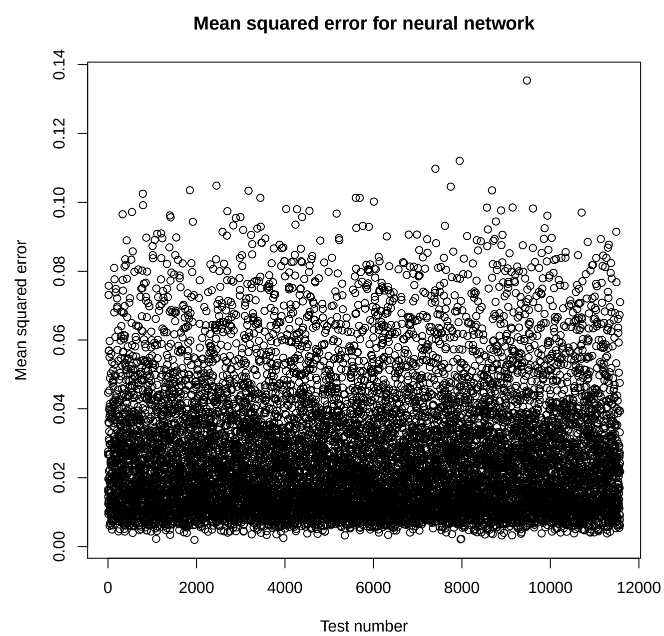

Figure 6: Plot of the mean squared error for the neural network

Further investigation should be made by using data that will allow an analysis of the Bayesian network in terms of a classification problem. Furthermore, the network itself can be analysed using the metrics described in [21] to provide a better context for the performance of the network. Finally, the trade-offs made by using Variation Bayesian approximation can be investigated more appropriately by comparing the results against a Bayesian network using a method of exact inference for the very large dataset.

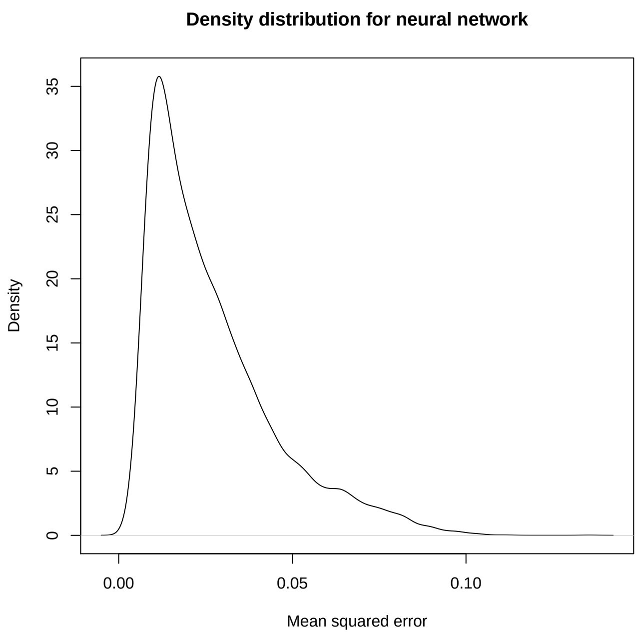

Figure 7: Plot of the density distribution for the neural network

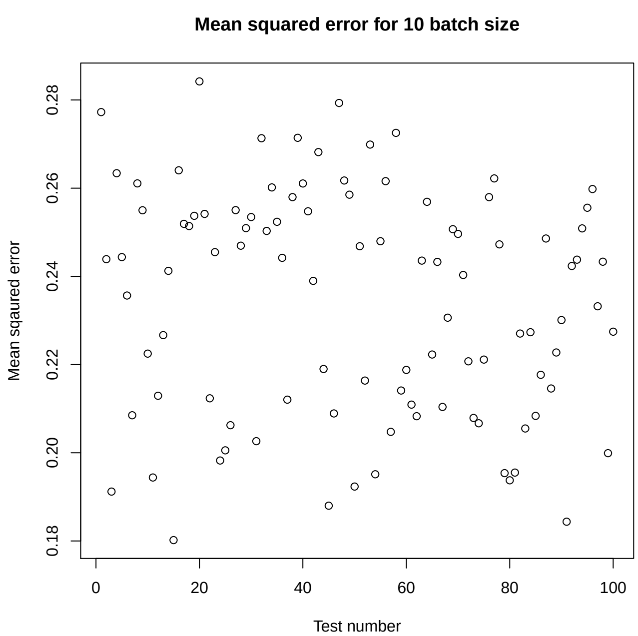

Figure 8: Plot of the mean squared error for batch size 10

8 ACKNOWLEDGEMENTS

I would like to thank Roy Henha Eyono for extracting the shape descriptors from the Galaxy Zoo images.

I would like to thank Victor Gueorgiev for providing me with an optimised neural network to test against.

Finally, I would like to thank my supervisor Deshen Moodley who graciously provided me with guidance and advice.

REFERENCES

[1] Franz Baader. 2003. *The description logic handbook: Theory, implementation and applications.* Cambridge university press.

[2] Nicholas M Ball, Jon Loveday, Masataka Fukugita, Osamu Nakamura, Sadanori Okamura, Jon Brinkmann, and Robert J Brunner. 2004. Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society* 348, 3 (2004), 1038–1046.

[3] Matthew James Beal. 2003. *Variational algorithms for approximate Bayesian inference.* University of London London.

[4] Emmanuel Bertin and Stephane Arnouts. 1996. SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series* 117, 2 (1996), 393–404.

[5] Eugene Charniak. 1991. Bayesian networks without tears. *AI magazine* 12, 4 (1991), 50.

[6] Serena H Chen and Carmel A Pollino. 2012. Good practice in Bayesian network modelling. *Environmental Modelling & Software* 37 (2012), 134–145.

[7] Paul B Chou and Christopher M Brown. 1990. The theory and practice of Bayesian image labeling. *International journal of computer vision* 4, 3 (1990), 185–210.

[8] Ira Cohen, Nicu Sebe, FG Gozman, Marcelo Cesar Cirelo, and Thomas S Huang. 2003. Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, Vol. 1. IEEE, I–I.

[9] Gregory F Cooper. 1990. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence* 42, 2-3 (1990), 393–405.

[10] Lori A Dalton and Edward R Dougherty. 2011. Bayesian minimum mean-square error estimation for classification errorâĂŤPart I: Definition and the Bayesian MMSE error estimator for discrete classification. *IEEE Transactions on Signal Processing* 59, 1 (2011), 115–129.

[11] Peter E Dewdney, Peter J Hall, Richard T Schilizzi, and T Joseph LW Lazio. 2009. The square kilometre array. *Proc. IEEE* 97, 8 (2009), 1482–1496.

[12] Charles W Fox and Stephen J Roberts. 2012. A tutorial on variational Bayesian inference. *Artificial intelligence review* (2012), 1–11.

[13] Nir Friedman, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian network classifiers. *Machine learning* 29, 2-3 (1997), 131–163.

[14] Shaukat N Goderya and Shawn M Lolling. 2002. Morphological classification of galaxies using computer vision and artificial neural networks: A computational scheme. *Astrophysics and space science* 279, 4 (2002), 377–387.

[15] John Kormendy and Ralf Bender. 1996. A proposed revision of the Hubble sequence for elliptical galaxies. *The Astrophysical Journal Letters* 464, 2 (1996), L119.

[16] Jan Kremer, Kristoffer Stensbo-Smidt, Fabian Gieseke, Kim Steenstrup Pedersen, and Christian Igel. 2017. Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy. *IEEE Intelligent Systems* 32, 2 (2017), 16–22.

[17] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 3 (2008), 1179–1189.

[18] Jiebo Luo, Andreas E Savakis, and Amit Singhal. 2005. A Bayesian network-based framework for semantic image understanding. *Pattern recognition* 38, 6 (2005), 919–934.

[19] Jaakko Luttinen. 2016. BayesPy: variational Bayesian inference in Python. *The Journal of Machine Learning Research* 17, 1 (2016), 1419–1424.

[20] Nicolas Maillot, Monique Thonnat, and Alain Boucher. 2004. Towards ontology-based cognitive vision. *Machine Vision and Applications* 16, 1 (2004), 33–40.

[21] Bruce G Marcot. 2012. Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological modelling* 230 (2012), 50–62.

[22] Thomas P. Minka. 2001. Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI'01).* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 362–369. http://dl.acm.org/citation.cfm?id=2074022.2074067

[23] Daniel Nikovski. 2000. Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. *IEEE Transactions on Knowledge and Data Engineering* 12, 4 (2000), 509–516.

[24] Irina Rish. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. IBM, 41–46.

[25] Donald F Specht. 1991. A general regression neural network. *IEEE transactions on neural networks* 2, 6 (1991), 568–576.

[26] Grant H Stokes, Frank Shelly, Herbert EM Viggh, Matthew S Blythe, and Joseph S Stuart. 1998. The lincoln near-earth asteroid research (linear) program. *Lincoln Laboratory Journal* 11, 1 (1998).

[27] MC Storrie-Lombardi, O Lahav, L Sodre Jr, and LJ Storrie-Lombardi. 1992. Morphological classification of galaxies by artificial neural networks. *Monthly Notices of the Royal Astronomical Society* 259, 1 (1992), 8P–12P.

[28] Dimitris G Tzikas, Aristidis C Likas, and Nikolaos P Galatsanos. 2008. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine* 25, 6 (2008), 131–146.

[29] Jake Vander Plas, AJ Connolly, and Z Ivezic. 2014. AstroML: Python-powered Machine Learning for Astronomy. In *American Astronomical Society Meeting Abstracts# 223*, Vol. 223.

[30] Vladimir Naumovich Vapnik and Vlamimir Vapnik. 1998. *Statistical learning theory*. Vol. 1. Wiley New York.

[31] Cort J Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research* 30, 1 (2005), 79–82.

[32] Donald G York, J Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. 2000. The sloan digital sky survey: Technical summary. *The Astronomical Journal* 120, 3 (2000), 1579.