Bayesian Decision Tree Averaging for the Probabilistic Interpretation of Solar Flare Occurrences

Vitaly Schetinin [1], Valentina Zharkova [2], and Sergei Zharkov [3]

[1] Computing and Information Systems, University of Bedfordshire, Luton, Park Square, LU1 3JU, UK
Vitaly.Schetinin@luton.ac.uk
[2] Cybernetics Department, Bradford University, Bradford, BD7 1DP, UK
V.V.Zharkova@bradford.ac.uk
[2] Department of Applied Mathematics, University of Sheffield, Sheffield, S3 7RH, UK
S.Zharkov@sheffield.ac.uk

Abstract. Bayesian averaging over Decision Trees (DTs) allows the class posterior probabilities to be estimated, while the DT models are understandable for domain experts. The use of Markov Chain Monte Carlo (MCMC) technique of stochastic approximation makes the Bayesian DT averaging feasible. In this paper we describe a new Bayesian MCMC technique exploiting a sweeping strategy allowing the posterior distribution to be estimated accurately under a lack of prior information. In our experiments with the solar flares data, this technique has revealed a better performance than that obtained with the standard Bayesian DT technique.

Keywords: Machine learning, uncertainty, Bayesian averaging, decision tree, Markov Chain Monte Carlo, solar flare.

1 Introduction

Solar activity is characterized by patterns which can be represented by features of active regions, sunspots, solar flares, coronal holes, and/or filaments which are observable on full disk solar images taken from the ground and space-based instruments [1 - 3]. In the theory, these patterns have a complex dynamics associated with an 11 year solar activity cycle represented by the features dependent on the time and location on the solar disk [3 - 5]. The observed full-disk images are stored and then used in order to discover new knowledge about unknown phenomenon of solar activity. During the last decade many research has been done in order to discover phenomenon models in the observed data by using the machine learning paradigms which potentially are capable of providing a high performance in terms of predictive accuracy [2 - 8].

In this paper we describe a new machine learning method developed for an automated classification of solar activity which is associated with solar flares. This method is based on the methodology of Bayesian model averaging which under some conditions is able to provide the best performance [9 - 12].

The Bayesian model averaging methodology has revealed promising results when the uncertainty in classification outcomes has to be estimated [13, 14]. The use of Bayesian averaging over decision trees (DTs) allows domain experts to understand the nature of a phenomenon by treating the features as explanation variables involved in a model of probabilistic inference [14 - 18].

In practice, a prior information, which is required within the Bayesian methodology, can be distorted or unavailable in a full volume. For example, the nature of flare occurrences is not yet fully understood in terms of such features as brightness, magnetism, and topology of the observable regions of solar activity. Obviously, in such cases the domain experts cannot provide a high volume of a prior information, and therefore the Bayesian methodology cannot yield the optimal estimates [11, 13]. However, when the prior information is fully available, the Bayesian methodology provides the best performance, although this technique is still computationally expensive. Fortunately, this obstacle can be overcome by using Markov Chain Monte Carlo (MCMC) technique of stochastic approximation [13, 14].

In this paper we aim to explore the potential of the Bayesian DT MCMC technique on the benchmark solar flares data taken from the Machine Learning Repository [19]. The comparisons are made in terms of the predictive accuracy and uncertainty in classification outcomes estimated within the Uncertainty Envelope Technique described in [20].

Further in Section 2 we describe the bases of the MCMC sampling for the Bayesian averaging over DT models which can be easily interpreted by domain experts. In Section 3 we explore the conditions under which the Bayesian DT averaging allows the posterior distribution to be estimated accurately. The application of the Bayesian DT technique to the classification of solar flares is described in Sections 4 and 5, and finally Section 6 concludes the paper.

2 The Bayesian Decision Tree Technique

Bayesian model averaging methodology allows the uncertainty in classification outcomes to be evaluated. In this section first we consider how the Bayesian approach can be practically implemented on the base of the MCMC technique of stochastic approximation. Then we consider the use of DT models for probabilistic interpretation of the classification outcomes which gives domain experts useful information for understanding.

2.1 Bayesian Averaging over Decision Trees

The main idea of DT classification models is to recursively partition data points in an axis-parallel manner. Such models provide natural feature selection and uncover the features which make the important contribution to the classification. The resultant DT classification models can be easily interpretable by domain experts [11, 17].

By definition, DTs consist of splitting and terminal nodes, which are also known as tree leaves. DTs are binary if the splitting nodes ask a specific question and then

divide the data points into two disjoint subsets, assigned to the left and the right branches [11, 17].

The number of the data points in each split should not be less than that predefined by a user, which has to properly specify this number; otherwise DT model can lose the ability to generalise well. All data points fallen in a terminal node are assigned to a class of majority of the training data points residing in this terminal node. Within a Bayesian framework, the class posterior distribution is calculated for each terminal node [11 - 14].

The Bayesian MCMC methodology has revealed promising results in the applications to some real-world problems [13 - 16]. To deal with large DTs, Chipman *et al.* [13] and recently Denison *et al.* [14] have developed the MCMC techniques using the RJ extension suggested by Green [21]. These techniques make the moves such as *birth* and *death* in order to induce large DTs under the priors given on the shape or size of the DTs. In the theory, RJ MCMC technique exploring the posterior distribution has to keep the balance between the birth and death moves which is required to obtain the desired estimates of the posterior unbiased [13, 14, 21].

Within the existing RJ MCMC techniques the proposed moves are assigned unavailable when the number of data points, falling in one of splitting nodes, becomes less than the given number. In practice, a user can improperly set up an acceptable number of data points in splits as well as the priors on favourite shape of the DTs. In such cases, the resultant estimates of class posterior distributions become biased [12, 13].

Moreover, within the standard RJ MCMC technique suggested for Bayesian DT averaging, the desired balance between the birth and death moves practically cannot be achieved as shown in [16]. This observation is based on the facts that the RJ MCMC technique exploring DTs makes some moves unavailable because of an unacceptable number of data points in splits. As a result, such moves cause a disproportion in the given probabilities of moves. Next we describe the standard Bayesian RJ MCMC technique.

2.2 The Methodology of Bayesian Averaging

In general, the class posterior distribution we are interested in is written as an integral over parameters $\boldsymbol{\theta}$ of the classification model

$$p(y \mid \mathbf{x}, \mathbf{D}) = \int_{\theta} p(y \mid \mathbf{x}, \theta, \mathbf{D})\, p(\theta \mid \mathbf{D}) d\theta, \tag{1}$$

where y is the predicted class (1, …, C), $\mathbf{x} = (x_1, \ldots, x_m)$ is the m-dimensional input vector, and $\mathbf{D}$ denotes the given training data.

In practice, except some simple cases, the posterior density $p(\boldsymbol{\theta} \mid \mathbf{D})$ cannot be evaluated analytically. However, when values $\boldsymbol{\theta}^{(1)}$, …, $\boldsymbol{\theta}^{(N)}$ are drawn from the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{D})$, we can write:

$$p(y \mid \mathbf{x}, \mathbf{D}) \approx \sum_{i=1}^{N} p(y \mid \mathbf{x}, \theta^{(i)}, \mathbf{D}) p(\theta^{(i)} \mid \mathbf{D}) = \frac{1}{N} \sum_{i=1}^{N} p(y \mid \mathbf{x}, \theta^{(i)}, \mathbf{D}). \tag{2}$$

This is the basis of the MCMC technique for approximating integrals (1) [13, 14]. To perform such an approximation, we need to run a Markov Chain until it converges to a stationary distribution. After this we can draw *N* samples from the Markov Chain and calculate the class posterior density (2).

2.3 Reversible Jump MCMC

To sample models of a variable dimensionality, it has been suggested to extend the MCMC method by the Reversible Jumps (RJ) [21]. The RJ MCMC technique allows large DT models to be sampled from real data [13 - 15]. Within this technique the posterior probability is explored by using the following types of moves.

1. *Birth*. Randomly split the data points falling in one of the terminal nodes by a new splitting node with the variable and rule drawn from the corresponding priors.
2. *Death*. Randomly pick a splitting node with two terminal nodes and assign it to be one terminal with the united data points.
3. *Change-split*. Randomly pick a splitting node and assign it a new splitting variable and rule drawn from the corresponding priors.
4. *Change-rule*. Randomly pick a splitting node and assign it a new rule drawn from a given prior.

The first two moves, birth and death, are reversible and change the dimensionality of $\boldsymbol{\theta}$ as described in [13, 14, 21]. The remaining moves make jumps within the current dimensionality of $\boldsymbol{\theta}$. Note that the change-split move is included to make "large" jumps which can increase the chance of sampling from a maximal posterior, whilst the change-rule moves do "local" jumps.

Because of a hierarchical structure of DTs, the changes at the nodes located at the upper levels can significantly change the location of data points at the lower levels. For this reason there is a very small probability of changing and then accepting a DT split located near a root node. As a result, the RJ MCMC algorithm cannot explore a full posterior distribution properly.

One way to extend the search space is to restrict DT sizes during a given number of the first burn-in samples as described in [14]. This strategy, however, requires setting up in an *ad hoc* manner the additional parameters such as the size of DTs and the number of the first burn-in samples.

Alternatively, the search space can be extended by using a restarting strategy described in [13]. Clearly, both strategies cannot guarantee that most of DTs will be sampled from a model space region with a maximal posterior. In the next section we describe our approach based on a sweeping strategy.

3 The Bayesian Averaging with a Sweeping Strategy

The main idea of using a sweeping strategy is to assign the prior probability of further splitting DT nodes to be dependent on the range of values within which the number of

data points will be not less than a given number of points. Such a prior is explicit because at the current partition the range of such values is unknown.

Formally, the probability $P_s(i, j)$ of further splitting at the ith partition and variable j can be written as:

$$P_s(i, j) = \frac{x_{\max}^{(i,j)} - x_{\min}^{(i,j)}}{x_{\max}^{(1,j)} - x_{\min}^{(1,j)}}, \quad (3)$$

where $x_{\min}^{(i,j)}$ and $x_{\max}^{(i,j)}$ are the minimal and maximal values of variable j at the ith partition level.

For all the partition $i > 1$ we can see that $x_{\max}^{(i,j)} \le x_{\max}^{(1,j)}$ and $x_{\min}^{(i,j)} \ge x_{\max}^{(1,j)}$, and therefore there is partition k at which the number of data points becomes less than a given number p_{min}. Therefore, probability P_s ranges between 0 and 1.0 for any variable j, and its value is dependent on the level i of partitioning a data set.

Form this point of view, prior (3) favors splitting the terminal nodes which contain a large number of data points. This allows accelerating the convergence of Markov chain and, therefore, the RJ MCMC technique can explore an area of a maximal posterior in more detail.

To make the birth and change moves within such a prior, the new splitting values $s_i^{\text{rule,new}}$ for the ith node and variable j are drawn from a uniform distribution $s_i^{rule,new} \sim U(x_{\min}^{1,j}, x_{\max}^{1,j})$. Likewise, for the change-split moves, a new variable is assigned as follows $s_i^{var,new} \sim U\{S_k\}$, where S_k is the set of features except variable s_i^{var} currently used at the ith node.

For the change-rule moves, the value $s_i^{\text{rule,new}}$ is drawn from a Gaussian with a given variance σ_j:

$$s_i^{rule,new} \sim N(s_i^{rule}, \sigma_j), \quad (4)$$

where $j = s_i^{\text{var}}$ is the variable used at the ith splitting node.

For some moves, the number of data points becomes less than a predefined number p_{min}. Within the existing Bayesian DT techniques such moves are assigned unavailable [13, 14].

Within our approach after making the birth or change move, three possible cases arise. In the first case, the number of data points in all the partitions is larger than p_{min}. In the second case, the number of data points in one partition becomes larger than p_{min}. In the third case, the number of data points in two or more partitions becomes larger than p_{min}. These cases are processed as follows.

For the first case, the RJ MCMC algorithm runs as usual. For the second case, a node with an unacceptable number of data points in the split is removed from the current DT. If the move was of the birth type, the RJ MCMC just resamples the current DT; otherwise, this move is considered as the death move. For the last third case, the RJ MCMC algorithm simply resamples the DT.

Because the unacceptable nodes are removed from the DT, we named such a strategy *sweeping*. Next we describe the application of the Bayesian DT using such a strategy to the solar flares data.

4 Application to Solar Flares Data

The solar flares data were taken from the Machine Learning Repository [19]. These data contain three classes of the observations represented by the number of times of a certain type of solar flares occurred in a 24 hour period in various active regions allocated to different groups of sunspot complexities. Each observation represents captured features for one active region on the Sun. The total number of the observations is 1066, and each observation is presented by 10 features listed in Table 1.

Table 1. The solar flares data features.

#	Features	Values
1	Code for class (modified Zurich class)	{A,B,C,D,E,F,H}
2	Code for largest spot size	{X,R,S,A,H,K}
3	Code for spot distribution	{X,O,I,C}
4	Activity	{1 = reduced, 2 = unchanged}
5	Evolution	{1 = decay, 2 = no growth, 3 = growth}
6	Previous 24 hour flare activity code	{1 = nothing as big as an M1, 2 = one M1, 3 = more activity than one M1}
7	Historically-complex	{ 1 = Yes, 2 = No}
8	Did region become historically complex on this pass across the sun's disk	{1 = yes, 2 = no}
9	Area	{1 = small, 2 = large}
10	Area of the largest spot	{1 = <=5, 2 = >5}

Three classes of flares are predicted as C, M, and X classes. However, in our first experiments this domain problem was considered as a 2-class problem of predicting either flare or non-flare outcome.

In these experiments, no prior information on the preferable DT shape and size was available. The minimal number of data point allowed being in the splits was set to 1. The proposal probabilities for the death, birth, change-split and change-rule moves were set to 0.1, 0.1, 0.1, and 0.7, respectively. The numbers of burn-in and post burn-in samples were set to 50000 and 5000, respectively.

All these parameters of the MCMC sampling were the same for the standard and proposed Bayesian DT techniques. The performance of these techniques was evaluated within 5 fold cross-validation and 2σ intervals. The uncertainty in classification outcomes was evaluated within the Uncertainty Envelope technique providing the rate of sure correct classifications as described in [20]. Next we present the experimental results.

5 Experimental Results

Both Bayesian DT techniques with the standard (DBT1) and the suggested (BDT2) strategies have correctly recognized 82.1% and 82.5% of the test examples, respectively. The average number of DT nodes was 17.5 and 10.1, respectively. Table

2 shows the obtain results. This table shows also the rate of sure correct classifications which in accordance with the Uncertainty Envelope Technique is proportional to the classification confidence.

Table 2. The performance and size of the BDT1 and BDT2 on the Solar Flares Data.

Strategy	Number of DT nodes	Perform, %	Sure correct, %
BDT1	17.5±1.5	82.1±4.5	67.4±3.3
BDT2	**10.1**±1.6	82.5±3.8	**70.2**±4.2

From Table 2, we can see that both strategies reveal the same performance on the test data. However, the number of DT nodes induced by the suggested BDT2 strategy is much less than that induced by the standard BDT1 strategy. Besides, the BDT2 strategy provides more reliable classifications than the BDT1 strategy: the rate of sure correct classification provided by the BDT2 is higher than that of the BDT1.

Fig. 1 depicts the samples of log likelihood and numbers of DT nodes as well as the densities of DT nodes collected during the burn-in and post burn-in phases for the suggested BDT2 strategy. From the top left plot of these figures we can see that the Markov chain very quickly converges to the stationary value of log likelihood near to –320. During the post burn-in phase the values of log likelihood slightly oscillate around this value that allows us to conclude that the samples of DTs are drawn from a stable Markov Chain.

Fig. 2 depicts the contributions of the 10 features to the classification outcome. The feature importance is estimated in terms of the posterior weights with which the features were used in the DT models collected during the post burn-in phase.

From Fig. 2 we see that the most important contribution are made by features x_1 (Code for class) and x_2 (Code for largest spot size). Much less contribution are made by features x_8 (the region complexity), x_9 (Area) and x_{10} (Area of the largest spot).

Thus, based on these experimental results, we can conclude that the suggested BDT2 strategy using a sweeping strategy allows the classification uncertainty to be decreased without affecting the classification accuracy. At the same time, the suggested Bayesian strategy provides the shortest DTs which are easy-to-understand by domain experts.

As the Bayesian DT techniques require extensive calculations, the computational time required to run these techniques becomes important for real-world applications. Having compared the computational time in our experiments, we found that for the suggested BDT2 technique the computational time is less than that for the standard BDT1 technique on average on 20%. We can explain this by reducing the size of DTs induced by the suggested BDT2 strategy.

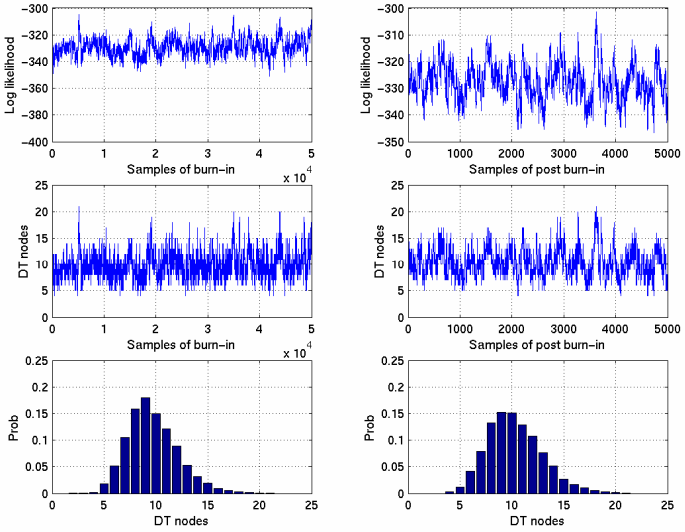

Fig. 1. The BDT2 strategy. Samples of the burn-in and post burn-in.

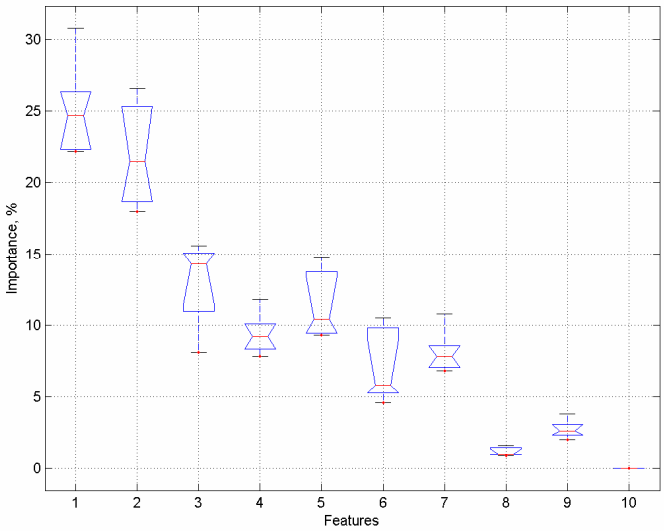

Fig. 2. The importance of the 10 features.

6 Conclusions

The classification technique we developed on the basis of Bayesian DT methodology has revealed promising results in our experiments on predicting the occurrences of solar flares. Both the standard and suggested Bayesian DT techniques provide a high performance in terms of predictive accuracy. However, the suggested technique using a sweeping strategy outperforms the standard Bayesian DT technique in terms of classification uncertainty. The suggested technique also provides shortest DTs which can be easily interpreted by domain experts.

The implementation of the Bayesian DT model averaging methodology is still computationally expensive and, therefore, further research should be done in this direction in order to reduce the computational expenses and make this methodology applicable to large-scale problems. Further research should be also done in order to verify the advantages of the proposed technique on new domain problems.

Overall, based on the obtained results, we believe that the Bayesian DT technique presented in this paper can be successfully applied to solar data. This technique is able to provide high performance, accurate estimates of uncertainty in classification outcomes, as well as the interpretability of models.

Acknowledgments. This work was partially (V. Zharkova) supported by the project European Grid of Solar Observations (EGSO), funded by the European Commission, Grant IST-2001-32409, and (V. Schetinin) by EPSRC, Grant GR/R24357/01. The authors also are thankful to the two anonymous reviewers for their constructive comments.

References

1. Bentley R.D. *et al*: The European Grid of Solar Observations. In: Proc. Solar Cycle and Space Weather Euro Conference, Vico Equense, Italy (2001) 603
2. Turmon M., Pap J., Mukhtar S.: Automatically Finding Solar Active Regions Using SOHO/MDI Photograms and Magnetograms. In: Proc. Structure and Dynamics of the Interior of the Sun and Sun-like Stars, Boston (1998)
3. Zharkova V.V., Ipson S.S., Zharkov S.I., Benkhalil A., Aboudarham J., Bentley R.D.: A Full Disk Image Standardization of the Synoptic Solar Observations at the Meudon Observatory. Solar Physics 214/1 (2003) 89
4. Zharkova, V.V., Ipson. S.S., Zharkov, Aboudarham, J., Benkhalil, A.K., Fuller, N.: Solar Feature Catalogues in EGSO. Solar Physics, 228/1 (2005) 139-150
5. Zharkova V.V., Ipson S. S., Qahwaji R., Zharkov S., Benkhalil A.: An Automated Detection of Magnetic Line Inversion and Its Correlation with Filaments Elongation in Solar Images. In: Proc. SMMSP-2003, Barcelona, Spain (2003) 115-121
6. Bader D.A., Jaja J., Harwood D., Davis L.S: Parallel Algorithms for Image Enhancement and Segmentation by Region Growing with Experimental Study. In: Proc. IEEE IPPS'96 (1996) 414
7. Gao J., Zhou M., Wang H.: A Threshold and Region Growing Method for Filament Disappearance Area Detection in Solar Images. In: Proc. Information Science and Systems, Johns Hopkins University (2001)

8. Turmon M., Mukhtar S., Pap J.: Bayesian Inference for Identifying Solar Active Regions. In: Proc. Knowledge Discovery and Data Mining (1997)
9. Kopparapu S., Desai U.: Bayesian Approach to Image Interpretation. Kluwer (2002)
10. Duda, R.O., Hart, P.E.: Pattern Classification. Wiley Interscience (2001)
11. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Belmont, Wadsworth (1984)
12. Buntine, W.: Learning Classification Trees. Statistics and Computing 2 (1992) 63-73
13. Chipman, H., George, E., McCullock, R.: Bayesian CART Model Search. J. American Statistics 93 (1998) 935-960
14. Denison, D., Holmes, C., Malick, B., Smith, A.: Bayesian Methods for Nonlinear Classification and Regression. Wiley (2002)
15. Schetinin, V., Partridge, D., Krzanowski, W.J., Everson, R.M., Fieldsend, J.E., Bailey, T.C., Hernandez, A.: Experimental Comparison of Classification Uncertainty for Randomized and Bayesian Decision Tree Ensembles. J. Math. Modeling and Algorithms 4 (2006) forthcoming
16. Schetinin V., Fieldsend J.E., Partridge D., Krzanowski W.J., Everson R.M., Bailey T.C., Hernandez A.: The Bayesian Decision Tree Technique with a Sweeping Strategy. In: Advances in Intelligent Systems - Theory and Applications, In cooperation with the IEEE Computer Society, Luxembourg (2004)
17. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
18. Kuncheva, A.: Combining Pattern Classifiers: Methods and Algorithms. Wiley (2004)
19. Blake, C.L., Merz, C.J.: UCI Repository of Machine Learning Datasets. www.ics.uci.edu/~ mlearn/MLRepository. Irvine, University of California (1998)
20. Fieldsend J.E., Bailey T.C., Everson R.M., Krzanowski W.J., Partridge D., Schetinin V.: Bayesian Inductively Learned Modules for Safety Critical Systems. In: Proceedings of the 35th Symposium on the Interface, Computing Science and Statistics, US, Salt Lake City (2003)
21. Green, P.: Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika 82 (1995) 711-732