Predicting the Days of the Week Using Machine Learning Techniques

Rui Hu
The University of Melbourne

1. Introduction

Back when chaos theory was in fashion, aficionados used to talk about butterflies flapping their wings over Peking and causing cyclones in Fiji. The study of apparently innocuous events and their effect on the weather has once again come to the fore with researchers from Arizona State University finding some interesting correlation between pollution and rainfall. The study, published in *Nature*, suggests that rain is most likely to occur along the US Atlantic coast on the weekend and the weather is most likely to be better on a Monday, Tuesday or Wednesday.

This interesting discovery implies that the days of the week are possible to be predicted by extracting the knowledge from the related meteorological data. Therefore, in this report, I present machine learning techniques to analyse the predictive capacity of meteorological data on a particular social construct: the days of the week. To scope the topic, my analysis is typically focused on solving the problem of how to build an accurate classification model to predict the days of the week based on the meteorological data of Brisbane.

In this report, Weka workbench is employed to perform all the machine learning processes including data pre-processing, classifier building, days predictions and result evaluations.

The remainder of this report is organised as follows. An overview of related work is provided in section 2. Section 3 introduces the data sets and machine learning software that are utilised to proceed the analysis. Section 4 describes the methodology that is employed to experiment with different meteorological data sets and machine learning algorithms. A critical discussion on the experiment results are presented in Section 5. Finally, the report concludes in section 6.

2. Related Work

A literature survey showed that machine leaning research is becoming a common complement to many scientific areas such as medicine, biotechnology and meteorology. For example, there have been many studies on the prediction of breast cancer, storm. solar generation and even stock price trend. Even though there is almost no studies related to the prediction on the days of the week, those existing studies can provide many effective machine learning approaches and methodologies of deriving patterns and rules from large sets of data.

[1] introduces a common workflow of machine learning processes, which mainly consists of six basic steps: defining the problem; collecting the data; pre-processing data; training the model; testing and evaluating the model. Additionally, it also raises some challenges that need to be addressed to improve the accuracy of machine learning prediction, such as how to identify a good discretisation algorithm and what is the best way to dealing with missing attribute values in imperfect real-life data. A number of possible solutions to this discretisation issue are proposed in [2], which particularly reviews all the existing work on continuous feature discretisation, identifies characteristics of the methods and conducts an empirical evaluation of several methods. On the other hand, nine different approaches to missing attribute values are presented and compared in [3]. And it concludes that the C4.5 approach and the method of ignoring instances with missing values are the best among all nine approaches.

3. Methodology

There are many possible ways to improve the prediction accuracy of machine learning, such as employing an effective classification algorithm that are well suited to the input data, or performing appropriate data pre-processing before building the model. In this report, I investigate two classification algorithms: the Naive Bayes and the J48 decision tree. In the meantime, four data pre-processing techniques are attempted and experimented, which are dataset splitting, attributes selection, missing values handling and continuous attributes discretisation. My aim is to identify the most suitable classification algorithm as well as desirable data pre-processing techniques for predicting the days of week on meteorological data.

The workflow of my experiments is as followings. Firstly, Brisbane dataset is decomposed into several parts, based upon internal data characteristics and my observations of Brisbane weather patterns. Secondly, each possible attributes

subset is experimented together with the forementioned three classification algorithms, the objective here is to find out the best attributes subset which achieves the highest prediction accuracy. Thirdly, different techniques dealing with missing values are attempted. Fourthly, attribute discretisation is applied on continuous values. Finally, the model achieving highest accuracy is tested on another development dataset to evaluate its capability of generalisation.

3.1. Splitting Dataset

From a statistical point of view, it is difficult to derive an accurate prediction pattern from a dataset containing extremely fluctuated data values. To illustrate, a rule based on the average precipitation could be extracted from Figure 1: Most likely, if the precipitation value of a record is higher than 400, it would be inferred as a Saturday. It looks reasonable; however, in fact most of the Saturdays in training data have an average precipitation lower than 100, only because several torrential rains swept a small number of those Saturdays, the average precipitation value is pulled up to 400. Additionally, we have to note that if each day's feature values are almost identical, it is also hard for machine learner to differentiate 7 days. Hence, in order to build an accurate model, I should first find out a partition containing relatively balanced weather data. A simple analysis of the local weather shows that from November to March, thunderstorms are common over Brisbane, with the more severe events accompanied by large damaging hail stones, torrential rains and destructive winds, whereas in early spring and winter days, the weather is extremely stable.

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Average Precipitation	23	22	39	50	147	400	100

Based on above analysis, I hypothesise that October would be a good month to predict the day of week due to the relative balance that accompanies Brisbane October climate. Thus, the rest of the report is focused on October's data.

Another simple analysis shows that the Brisbane dataset has missing information in the fields of minimum observed sky ceiling (F9), maximum sky ceiling (F10), average sky ceiling (F11), minimum barometric pressure (F18), maximum barometric pressure (F19) and the average barometric pressure (F20), for almost 31%. It can be seen from Table 1 that these missing value attributes are regularly distributed based on the gathering time of records. According to this distributed regularity, I split the Brisbane dataset into three single sets, which contain the records of 1944 - 1946, 1949 - 1969 and 1973 - 2013, respectively.

	F9	**F10**	**F11**	**F18**	**F19**	**F20**
1944-1946	0%	0%	0%	0%	0%	0%
1949-1969	100%	100%	100%	100%	100%	100%
1973-1993	100%	100%	100%	0%	0%	0%
1993-2013	100%	100%	100%	0%	0%	0%

Table 1 Missing Value Percentage

3.2. Attribute Selection

After splitting dataset, a common analysis would be determining the effect of the attributes on the prediction, or attribute selection. The Weka Classifier Subset Evaluator is applied to search through the space of possible features and evaluate each subset by running a particular model on the it. In this step, Naive Bayes and J48 are tested, respectively. By applying recommendations provided by Weka, I can sort out the most suitable attributes subset which achieves the highest accuracy.

3.3. Handling Missing Values

In this step, the following techniques are analysed and experimented, respectively, to handle missing values in Brisbane dataset.

3.3.1. Ignoring Data Row or Attribute

This technique is usually applied when the class label is missing, or many attributes are missing from the row. However, it is obvious that this kind of technique is not suitable to a dataset where the percentage of such rows is high. For example, in 1949 - 1969 partition, almost 100% of the values of F9, F10 and F11 are missing, we definitely cannot simply remove all of those rows. Hence, instead of removing all the data tows, I decide to remove attributes and prediction will depend on other attributes.

3.3.2. Using Attribute Mean

The most common way is to replace missing values of a certain attribute by the mean or median by looking at all the rows in a database. However, this method does not make sense in some case, such as replacing Decembers' temperature missing values with the mean of all the twelve months. Therefore, in my experiment, a more accurate method is applied: to use attribute mean for all samples belonging to the same class. For instance, Decembers' temperature missing values are replaced by the mean of all the December's data. In weka, we can simply use the unsupervised *ReplaceMissingValues* filter to achieve so.

3.4. Attribute Discretisation

Many algorithms developed in the machine learning community focus on learning in nominal feature spaces. According to an empirical evaluation done by James, Ron and Mehran (1995), the performance of Naive Bayes algorithm was significantly improved when features were discretised using an entropy-based method, such as Fayyad and Irani's algorithm. This is mainly because this kind of method excels at choosing partition point(s) in a sorted set of continuous values to minimise the joint entropy of the continuous variable and the classification variable. In my experiment, Weka supervised discretisation is utilised to achieve effective discretisation, which applies Fayyad and Irani's MDL method as a default algorithm.

3.5. Performance Evaluation

I use the performance accuracy of all the techniques to find out the best way to predict the days of the week. In order to have a fair measure of the performance of the classifier; I use a cross-validation with 10 folds. In its most elementary form, cross-validation consists of dividing the data into k subgroups. Each subgroup is predicted via the classification rule constructed from the remaining subgroups, and the estimated error rate is the average error rate from these k subgroups. In this way, the error rate is estimated in an unbiased way. The Weka toolkit can calculate all these performance metrics after running a specified k-fold cross-validation. After cross-validation, the final classifier is also tested on another new dataset, which is the development dataset, to ensure that the final classifier truly has a good capability of generalisation.

4. Experiment Results

The experiment results are shown in the Appendix section. Table 1 presents the prediction results on Octobers' data gathering from 1944 to 1946. Since complete information is provided in this partition, it is not needed to deal with missing values. Table 2, 3 and 4 shows the perdition results of each attempt of improving classifier performance.

5. Discussion

As it is highlighted as red in Table 1, all the experiments produce a good prediction accuracy when applying cross-validation, whereas in test dataset, the performance is very poor. The clear contrast indicates that these classifiers are not reliable. This is mainly due to the insufficient data provided by both training dataset (44 records) and testing dataset (11 records). Hence, I would not choose any of these models as the final model even though they achieve a good accuracy and it can be concluded that providing sufficient data is crucial for a machine learner to produce reliable and accurate pattern.

The data partitions used in Table 2, 3 and 4 have sufficient records for building a reliable model; however, it can be seen from the results that by handling missing values and discretising continuous attributes, the performance is just slightly improved or even worse than predicting without any pre-processes. This does not imply that these techniques are useless. The reasons why they are not working in my case are as followings. Firstly, the Brisbane dataset is partitioned based on the distribution regularity of missing values (section 3.1), which means that it is highly possible that in one dataset, there is almost no missing values whereas in another dataset, certain attributes miss all the values. For instance, during 1993 and 2013, 100% values of attribute F9, F10 and F11 are missing. When building models on these data, classification algorithms such as Naive Bayes would just ignore those attributes. Moreover, when applying Weka attribute subset selection technique, all of these useless attributes would be filtered out as well. Thus, handling missing values does not really impact the performance. Secondly, in my experiment, the Weka supervised discretisation is implemented but it categorises all the values into one group and cannot find any proper cut points that minimise the joint entropy of the continuous variable and the classification variable. Therefore, it can be said that discretisation is not always desirable when dealing with continuous attribute values.

From the results, we can also see that attribute selection is an important and useful step in data pre-processing, especially when dealing with a dataset containing many attributes. But it is worth noting that the selection of attributes greatly depend on the classification algorithm as well, which means that an attribute subset working well with Naive Bayes does not mean it must work well with J48 decision tree.

Finally, as it is coloured as green in table 4, the best classification model is generated by applying Naive Bayes on the data partition covering the October's records gathering from 1993 to 2013, without discretising continuous values and handling missing values.

6. Conclusions and Future Works

This report is focused on solving the problem of how to build an accurate classification model to predict the days of the week based on

meteorological data. It specifically describes the experiment workflow, and details the entire attempt process. The contribution of this report is that it attempts various data pre-processing techniques for machine learning and finally finds out a proper way of implementing these techniques to improve the performance of the model.

My work assumes that October is a good month for predicting the days of work, but it does not fully test the other 11 months. Future work will compare the model performance of each month and the accuracy might be increased by applying proper pre-processing techniques on the data of another month.

References

[1] Bellaachia, A., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. *Age*, *58*(13), 10-110.

[2] Dougherty, J., Kohavi, R., & Sahami, M. (1995, July). Supervised and unsupervised discretization of continuous features. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE*- (pp. 194-202). Morgan Kaufmann Publishers, Inc..

[3] Grzymala-Busse, J. W., & Hu, M. (2001, January). A comparison of several approaches to missing attribute values in data mining. In *Rough sets and current trends in computing* (pp. 378-385). Springer Berlin Heidelberg.

Appendix

Discretised	Handling Missing Values	Classification Algorithm	Attribute Subset	Cross Validation Accuracy	Testing Dataset Accuracy
No	No	Naive Bayes	F6, F11, F16, F20, F21	18.18%	2.0%
No	No	J48	F3, F8, F12, F14, F21	9.10%	1.3%
Yes	No	Naive Bayes	F3, F5, F6, F7, F11, F15, F18, F19, F21	20.45%	3.1%
Yes	No	J48	F3, F5, F6, F11, F12, F16, F21	20.45%	2.3%

Table 1: 1944 - 1946 October

Discretised	Handling Missing Values	Classification Algorithm	Attribute Subset Recommendation	Cross-Validation Accuracy	Testing Dataset Accuracy
No	No	Naive Bayes	F3, F5, F8, F12, F15, F21	15.39%	16.49%
No	No	J48	F1, F4, F7, F8, F12, F14, F15, F17	13.31%	13.41%
No	Yes	Naive Bayes	F1, F5, F8, F12, F14, F17, F21	13.51%	14.50%
No	Yes	J48	F1, F5, F8, F12, F14, F17, F21	13.51%	15.21%
Yes	No	Naive Bayes	F4, F6, F7, F8, F12, F13, F15, F16, F17, F21	15.80%	13.22%
Yes	No	J48	F1, F3, F8, F15, FF21	12.27%	13.20%
Yes	Yes	Naive Bayes	F1, F3, F6, F7, F8, F12, F13, F14, F15, F16, F21	13.72%	13.68%
Yes	Yes	J48	F1, F3, F8, F15, F21	12.47%	13.87%

Table 2: 1949 - 1969 October

Discretised	Handling Missing Values	Classification Algorithm	Attribute Subset	Cross-Validation Accuracy	Testing Dataset Accuracy
No	No	Naive Bayes	F1, F8, F10, F13, F19, F21	14.76%	15.76%
No	No	J48	F1, F7, F8, F11, F14, F17, F20, F21	13.20%	14.20%
No	Yes	Naive Bayes	F1, F10, F13, F15, F19	14.37%	15.30%
No	Yes	J48	F1, F3, F7, F11, F13, F14, F15, F17, F18, F20	13.20%	13.90%
Yes	No	Naive Bayes	F4, F5, F6, F9, F10, F11, F12, F13, F14, F16, F18, F20	13.98%	14.51%
Yes	No	J48	F1, F3, F4, F7, F9, F12, F13, F14, F16, F17, F18	16.51%	16.72%

Yes	Yes	Naive Bayes	F4, F11, F12, F14, F16, F18	16.50%	16.89%
Yes	Yes	J48	F1, F3, F4, F7, F9, F12, F13, F14, F16, F17, F18	16.50%	16.77%

Table 3: 1973 - 1993 October

Discretised	Handling Missing Values	Classification Algorithm	Attribute Subset	Cross-Validation Accuracy	Testing Dataset Accuracy
No	No	Naive Bayes	F1, F6, F8, F12, F17, F20, F21	17.11%	18.18%
No	No	J48	F3, F7, F11, F14, F16, F18	15.27%	16.78%
No	Yes	Naive Bayes	F1, F6, F8, F12, F17, F20, F21	17.11%	18.20%
No	Yes	J48	F3, F7, F11, F14, F16, F18	15.27%	16.37%
Yes	No	Naive Bayes	F4, F5, F6, F8, F9, F10, F11, F12, F13, F15, F16, F17, F20	15.68%	16.72%
Yes	No	J48	F1, F3, F5, F6, F7, F10, F12, F15, F17, F18, F19 , F20 , F21	16.70%	16.89%
Yes	Yes	Naive Bayes	F4, F5, F6, F8, F9, F10, F11, F12, F13, F15, F16, F17, F20	15.68%	16.72%
Yes	Yes	J48	F1, F3, F5, F6, F7, F10, F12, F15, F17, F18, F19 , F20 , F21	16.70%	16.89%

Table 4: 1993 - 2013 October