Work In-progress: Mining the Student Data for Fitness

Yunshu Du and Matthew E. Taylor

School of Electrical Engineering and Computer Science,
Washington State University,
Pullman, WA 99163, US
{ydu1,taylorm}@eecs.wsu.edu

Abstract. Data mining-driven agents are often used in applications such as waiting times estimation or traffic flow prediction. Such approaches often require large amounts of data from multiple sources, which may be difficult to obtain and lead to incomplete or noisy datasets. University ID card data, in contrast, is easy to access with very low noise. However, little attention has been paid to the availability of these datasets and few applications have been developed to improve student services on campus. This work uses data from CougCard, the Washington State University official ID card, used daily by most students. Our goal is to build an intelligent agent to improve student service quality by predicting the crowdedness at different campus facilities. This work in-progress focuses on the University Recreation Center, one of the most popular facilities on campus, to optimize students' workout experiences.

Keywords: Agent Mining, Data Mining, Exploratory Data Analysis, Machine Learning, Decision Tree, Recommender System, Fitness

1 Introduction

This in-progress work uses existing data mining-driven approaches to build an intelligent agent with the goal of optimizing students' workout experiences at Washington State University's (WSU) Recreation Center (the Rec). The Rec is among the most frequently visited campus facilities. However, students may prefer to avoid the Rec when it is most crowded. This work aims to solve this problem by predicting how crowded the Rec will be at different times using CougCard, the WSU official ID used by all students when entering the Rec. Similar data analyses and applications have been deployed in areas such as waiting times estimation [8, 13, 18, 23] and traffic flow prediction [2, 7, 12, 17]. However, little attention has been paid to the availability of university ID cards and these sets of data have often been overlooked. We believe that by analyzing the activities of CougCard, we will be able to understand how student exercise activities are distributed over time. This work was approved by our IRB in July 2015.

First, we performed *Exploratory Data Analysis (EDA)* [21] to discover interesting patterns in collected CougCard data. We used the three basic EDA tools,

plots, graphs and summary statistics, and were able to find out general student exercise trends, such as the yearly/monthly/daily frequency of students visiting the Rec and the peak hours during a day. With limited space and fitness equipment in the Rec, some students may not be able to do the exercise they want to if all of the equipment or spaces have been taken. If students can know whether the Rec will be crowded in advance, they will be able to make better workout plans based on the information, leading to better user experiences. Therefore, we apply a *Decision Tree* [20] algorithm to build a predictive model that can make high-accuracy predictions on when the Rec would be more or less crowded.

This paper is structured as follows: In the next section we briefly introduce the concept of agent mining and discuss some of the existing related works. Section 3 describes the characteristics of our dataset. Section 4 shows our methodology on exploratory data analysis. Preliminary results of the decision tree model will be present in Section 5. An outline of future developments of this work concludes the paper.

2 Related Work

Agent Mining refers to the interdisciplinary approaches that integrate multiagent systems, data mining and knowledge discovery, machine learning and other related areas such as statistics and math. It has provided more efficient ways in solving problems that arise [6]. The agent mining area can be categorized into two main cases: agent-driven data mining and data mining-driven agents [5]. The former make use of agents in data mining, especially in distributed data mining (DDM) to cope with challenges of autonomy, interaction, dynamic selection and gathering, scalability, multistrategy, and collaboration and lead to better knowledge extraction [15]. In contrast, data mining-driven agents deploy data mining approaches in order to build more robust agent systems. An agent system can make use of the knowledge from data mining to construct its intelligent components, while mining the agent behavioral data can again benefit the knowledge extracting process. It is possible to achieve a balance of agent autonomy and supervised evaluation throughout this approach [5]. Our work focuses on building data mining-driven agents. The remainder of this section introduce a number of crowd-predicting applications that used this approach.

There are a number of extant applications that can estimate waiting times at certain places and have brought many conveniences into our daily life. A good example is *NoWait* [18], a mobile app that estimates waiting times at restaurants. Instead of waiting in a crowded lobby, users can search for nearby restaurants and join a waiting list via their smart phones. A text message will be sent to the user once their table is ready. NoWait is also engaged with the concept of hospitality business and encourages restaurants to improve their services based on customer ratings. Similar crowd-predicting applications such as the *Orlando Undercover Tourist APP* [13] allows tourists to see the current waiting times for Disney World, Universal Orlando and SeaWorld so that they could plan tours more efficiently.

Our work was also inspired by *traffic flow prediction* techniques. Novel algorithms have been developed to predict traffic flow and suggest more efficient routes [2, 7, 12, 17]. They are relevant because our work aims to predict student flow based on student card data. In addition, a traffic-prediction and route-recommendation mobile application, *Waze* [23], gives us insights on how to deploy this paper into real-world applications. After logging into the app, drivers can see millions of activities from other drivers and communicate traffic situations with each other in real-time. This community-based approach brings highly accurate estimates of road conditions.

Perhaps the most relevant work to this paper is the new *Popular Times* function in Google Maps [8]. When searching for a place, Google shows the most popular times and users will see when this place will be most crowded. This location-based service can assist users in determining when would be the best time to go visit a place.

These works can be viewed as data mining-driven agents since they all follows the process of extracting knowledge from data (e.g., the number of people at a restaurant or park, the current traffic situation, etc), transferring knowledge to an agent (e.g., estimating waiting times or traffic), an agent acting in the environment (e.g., meal or tour planning, route suggesting), and an agent changing behaviors based on feedback (e.g., customer rating, driver activities). However, most of them require massive amounts of data from multiple sources which could lead to incomplete or noisy datasets. Our work builds a data mining-driven agent with CougCard data. This resource is easy to obtain and since the card users are all students, the dataset is complete and with very low noise.

3 Data

We collected CougCard data from across campus. In this work, we focused on records from the Rec. Over 2 million fitness-related records were extracted from the collected dataset. The structure of data consists of time stamps for each card swipe when entering the Rec. This gives us a snapshot of the cumulative fitness data at the Rec from August 2012 to August 2015. However, one downside is that these data only allow us to know when students enter the Rec, but not when they leave — we are only able to estimate how long they stay at the Rec based on expert knowledge from the Rec staff. For example, it is common that people stay for about an hour to do aerobic exercise, but if they do weight-lifting or swimming, the stay time could be more than two hours. Despite this drawback, we still believe we can see how student exercise trends change over time and make predictions because we will know the relative crowdedness.

4 Exploratory Data Analysis

The first step of our work is to obtain a high-level understanding of the collected data. We use Exploratory Data Analysis (EDA) approach to summarize and visualize the main characteristics of our data.We are interested in finding out

the long-term trend of student activities and detecting the factors that cause this trend.

4.1 Methodology

Exploratory Data Analysis (EDA) is a data mining approach to explore the characteristics of a given data set. Tools such as plots, graphs and summary statistics are used during EDA process to see what the data itself can tell us before it fits to a formal model or tests a hypothesis. The objectives of EDA can be defined as [21]: 1) Suggest hypotheses about the causes of observed phenomena, 2) Assess assumptions on which statistical inference will be based, 3) Support the selection of appropriate statistical tools and techniques, 4) Provide a basis for further data collection through surveys or experiments.

This work adopts EDA techniques, in particular plots and graphs, to better understand student usage patterns at the Rec. We started with plotting overall yearly trend by summing up each months' total number of visits, obtaining a whole picture of our dataset. Then, we examined how the trends change over each day of a week and different time of a day and was able to figure out peak hours at the Rec.

After discovering the general usage patterns at the Rec, the question then become"what factors can affect those trends?". Therefore, we considered possible external factors, such as school events, holidays/vacations, and weather, making plots to see if the trends would show significant changes when those factors occur. Results show there is a relationship between those factors and usage trends at the Rec. Detailed findings and discussions are presented in the next section.

4.2 Descriptive Analysis

First we looked at overall student exercise trends by summing up each month's total number of visits at the Rec, and plotted them as yearly trends. Figure 1 shows 1) each year shares similar patterns, there were not much variations and 2) there were significantly fewer people during Summer (May to August) than during Fall and Spring semesters (September to April). We learned that a large proportion of students leave campus for summer break and so the Rec is likely to be uncrowded most of the time. It could be unnecessary for our model to predict the Rec crowdedness during summer time.

We are also interested in finding out which day of the week and what time of day would be relatively busier than another. Figure 2 illustrates the former. To our surprise, there were fewer people at the Rec during weekend than on weekdays. At the beginning we assumed there will be more people working out on weekend since they have more free times. Figure 3 shows which hour of a day is more crowded. The shape on different days of the week represent the distribution of visit frequencies over each hour of a day; the thicker means the more crowded. We can see that during weekdays the peak-hour appears after around 3pm and lasts until the closing time, while during weekends the busiest time tends to occur in morning and early afternoon.

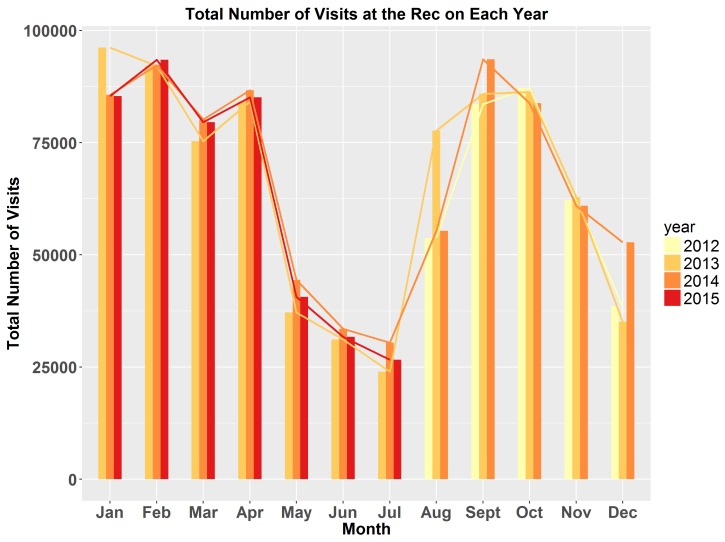

Fig. 1. Yearly and monthly trend

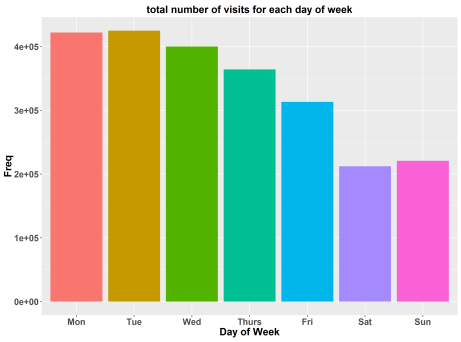

Fig. 2. Weekly trend

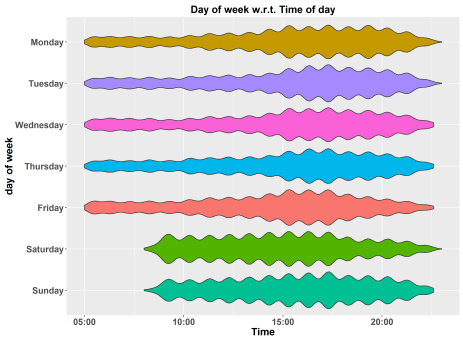

Fig. 3. Weekly trend w.r.t. hourly trend

Finding out the reasons behind those trends is essential for this work, thus calling for the exploration of possible external factors. Recall Figure 1 showed there were a lot less demands on Rec facilities during the Summer semester than in the Fall and Spring. In this case, as an initial step of this work, we will focus on analyzing the factors that influence the trend on Fall and Spring semesters. We have a total of six semesters: Fall 2012-2014 and Spring 2013-2015.

We explore the overall trend for both semesters (see Figure 4). It is easy to observe that there is a rapid frequency of decreases at a certain time interval for both Fall (around the 90^{th} days of school) and Spring (around the 60^{th} days of school) semesters. By checking the WSU academic calendar we learned that it was because of Thanksgiving break and Spring break, respectively. Another

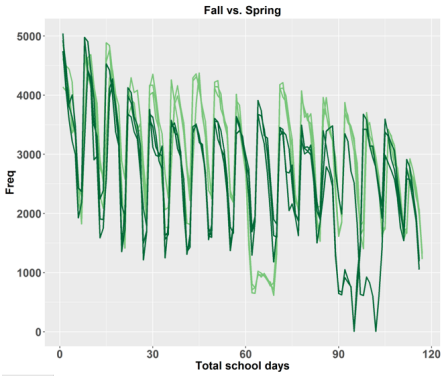

Fig. 4. Overall trend Fall/Spring

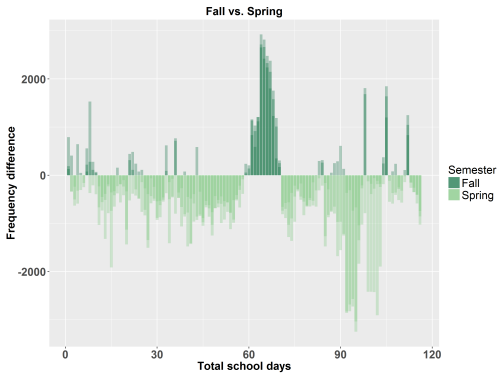

Fig. 5. Frequency differences Fall/Spring

observation is that during finals week (the last few days of each semester) the frequency is relatively low compared to other days. We could therefore conclude that school events such as vacations (IsVacations) and finals week (IsFinals) have negative effects on student workout trends and these two variables should be considered as input features in the model prediction process. Similarly, we should also put all university holidays (Labor Day, Martin Luther King Day (MLK), Presidents Day and Veterans Day) into consideration when building the predictive model (IsHolidays).

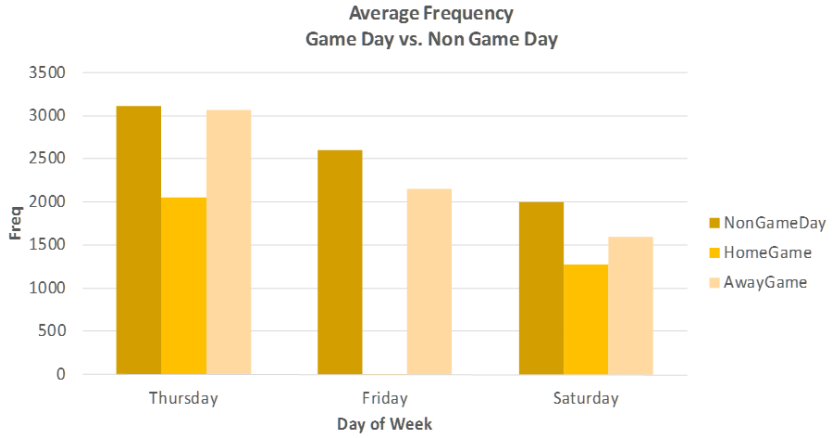

Fig. 6. Average frequency on game days vs. non game days

Figure 5 plots the frequency differences between Fall and Spring semesters. An interesting phenomenon we could obtain from it is that the overall frequency seems higher in Spring than in Fall. One reason could be the impact of weather.

Many psychology studies have showed that climate can influence people's behaviors [11,14], thus it is possible that in Spring semesters, warmer climate leads to more active behaviors. Another explanation could be the impact from a significant school event, the football game, as they are always held over Fall semesters, not Spring. We assume that if people are going to watch the game, it is very likely for them to miss their normal workout routine. To show the effect of game days, we did plots to compare that under the same external conditions (same semester and day of week), how do the number of people visiting the Rec on a football game day differ from a normal day. Figure 6 showed that students tend to miss their workout on a game day. Especially the games on Friday or Saturday, the influence is significant. Moreover, students seem to get more excited about watching a home game live than an away game from TV; the frequency drops much more on home game days than away game days. Because of this finding, we are adding two more features to our predictive model (IsHomeGame and IsAwayGame).

5 Data Modeling

The second step of this work is to learn a model to predict the crowdedness at the Rec. This section describes our decision tree model and presents preliminary results.

5.1 Model Selection

The main challenge of any learning task is how to select a suitable model. Model selection includes algorithm selection, feature selection, and parameter selection. In this work, we decided that the Decision Tree (DT) algorithm is sufficient for our dataset because: 1) our data set contains continuous features (date and time) that need to be discretized, and DT handles this process automatically; 2) Decision tree is a non-parametric algorithm and does not make assumptions on the probability distribution of our data [3]; 3) DT can be viewed not only as a machine learning algorithm but also as a data mining approach. It has been used to explore data in a variety of areas such as decision-making support for business [4]. Since our work aims to help students in making decisions on fitness timing, it is a data-driven decision-making application that can benefit from the use of DT.

Many approaches such as greedy search and ensemble methods have been developed and commonly used in feature selection process [9]. In this work, however, the potential features were selected based on the EDA results because of the lack of features in the original data set (only contains time stamps), we have to identify possible relevant factors from the real-world based on observations of our data. The model was trained in Weka and evaluated with 10-fold cross-validation method. We were able to achieve reasonable accuracy without any parameter selection or tuning. Section 5.2 describes detailed settings of our experiments and presents the preliminary results

5.2 Preliminary Result

This work aims to learn a predictive model to suggest when the Rec will be more or less busy so that students can pick more desirable times to workout. In practice, students may care more about the relative crowdedness, instead of the exact number of people at the Rec. Thus, we turn the regression problem into a classification problem. In particular, instead of predicting the number of people at the Rec for a given time interval, we predict whether this number is high or low. If the crowding level is high, students could avoid going to the Rec, or vice versa. Through EDA, we were able to find out that within a one hour time window, the maximum number of people who visited the Rec was 597. Therefore, we discretized the number into six crowdedness classes: low (0-100), med-low (101-200), med (201-300), med-high (301-400), high (401-500), very-high (501-600) and attempted to solve a classification problem.

In section 4, we were able to extract possible features for model learning tasks through data mining. Table 1 concluded the seven features: time of the open hours of Rec, day of the week, and binary representation of whether it is finals week, holidays, vacations, or football game days (home or away). We used Weka [24] to learn the tree model, in particular, with C4.5 algorithm, and evaluated it with 10-fold cross-validation approach. Most of the parameters were kept as default in Weke, we only changed two parameters: pruned to unprunded tree, and minimum number of instances from 2 to 1. The accuracy of the learned tree model was only 57%.

We then worked on improving this accuracy by adding two new features, date and semester labels (Fall and Spring). Surprisingly, the accuracy of the new learned tree model increased to 77.5%. Although no significant relations on how weather influences students' workout habits were observed through EDA, it is still worth looking into the statistical correlation between them. We then tried to add weather information to see if it would help with improving the accuracy of the predicted output. We integrated temperature data obtained from [1,10,16] with the student activity data. Specifically, we transformed the temperature data from numeric to nominal (from low to high) to avoid over fitting the training data. We found that the accuracy of the new learned model could achieve as high as 82.9%. Therefore, we can confirm that weather has an impact on student workout trends and more climate data should be collected in our future work.

Time	DayofWeek	IsFinals	IsHolidays	IsVacations	IsHomeGame	IsAwayGame
5:00:00	Friday	0	1	0	1	0
...	...	...	...	...	...	...

date	semester	temperature
9/30/2014	Fall	med
...	...	...

Table 1. Feature set: the seven initial features on top, accuracy was only 57%. After adding the date and semester features, accuracy improved to 77.5%. After adding the temperature feature, the final ten features achieved 82.9%

```
=== Confusion Matrix ===

     a     b     c     d     e     f   <-- classified as
 10109   808    19     3     0     0 |     a = low
  1079  6226   427    39     0     0 |     b = med-low
    26   454  1952   378    46     3 |     c = med
     1    26   412   841    86     6 |     d = med-high
     0     0    39   103   148    10 |     e = high
     0     0     0     4    20    14 |     f = very-high
```

Fig. 7. Confusion Matrix of the learned Decision Tree model using Weka

The confusion matrix of the final learned tree model is shown in Figure 7. It shows the total number of correctly or incorrectly classified instances for each class. It is worth to note that for those instances that were labeled low, none of them were classified as high or very-high. And for those instances that were labeled high or very-high, they were never classified as low or med-low, which was exactly what we want.

Algorithms	7 features	9 features	10 features	time
Decision Tree	57%	77.5%	82.9%	0.92 sec
Naïve Bayes	-	-	67.1%	0.04 sec
SVM	-	-	64.2%	63.49 sec

Table 2. Accuracy and running time (in seconds) comparison of Decision Tree, Naïve Bayes, and SVM

To compare with Decision Tree algorithm, we also tried *Naïve Bayes* [19] and *Support Vector Machine* (SVM) [22] (with Weka default RBF kernal) classifiers on the same data set with 10 features. They both underperformed Decision Tree. In addition, SVM took much longer to learn the model compared against the other two classifiers. Table 2 summarizes our results.

6 Conclusions and Future Work

In this work, we explored and visualized interesting patterns of student workout activities in terms of different factors (time, day, semester, etc.). We also successfully learned a Decision Tree model to predict crowdedness at the Rec for a given time interval. By comparison, we showed the advantages of a Decision Tree model in this real-world application over Naïve Bayes and Support Vector Machine classifiers.

Our long term goal is to make students more (quantitatively) satisfied with their experience at the Rec and/or (quantitatively) increase the number of times they visit the Rec to exercise. Future work will be focused on the development of agent in the manner of reinforcement learning. This data-driven agent will be able to take information that was extracted from data mining process, perform intelligent reasoning and prompt recommendations to users on desirable times to go to the Rec. Upon prompting, the user will give feedback to the agent based on their behaviors: if the user took the suggestion and visited the Rec, CougCard record at the Rec entry will be sent back to the agent as a positive reward for further analysis; if the user ignored the recommendation, the agent will receive a negative reward and should intelligently adjust it's recommending strategies for more precise future recommendations. Moreover, our agent system will be able to assist Rec managers with shift scheduling based on usage patterns and fitness event planning according to students' fitness demands.

In addition, more analysis can be performed on the existing model to achieve a more robust result. The main drawback of our current work is that we do not have measure of when do students leave the Rec therefore do not know the exact time of their stay. One solution to this problem is to set up a swipe-out system at the Rec's exit gate — we will be able to calculate the exact occupancy at the building for some time interval by comparing the number of swipe-ins and swipe-outs. In terms of improving the accuracy, we could discover and add more features to our current model such as weather conditions (e.g., rainy, sunny, etc.), or do regression instead of classification to predict the absolute number of people at the Rec for a given time interval. A web-page or mobile app of this agent could be built so that students can access it 24/7. We also expect to increase the number of areas the agent monitors and predicts crowdedness (e.g., the campus food court) and it will be interesting to look into how the crowdedness at different places related to each other. For example, if there are more people at food court there might be fewer people at the Rec, or if there are fewer people at the Rec during finals week, the library might be more crowded. There is also a potential of extending our agent to different fields by collecting and analyzing a different sets of data, such as predicting the crowdedness at different stores (clothes, electronic, sports, etc.) at a shopping mall during holidays with customer entries data. This idea raises the problem that unlike student card data, which has very low noise, data collected from multiple sources could be noisy and need more pre-processing and cleaning. However, our agent will be able to provide a good benchmark on how to process the data more efficiently such that better mining results could be produced.

Acknowledgments. The authors would like to thank Terry Quinn, Craig Howard, Joanne Greene, and Ryan Savage for their generous help during the data collecting process and Bei Peng for her contributions to experiment design. The authors would also like to thank Dr. Janardhan Rao Doppa, Dr. Assefaw Gebremedhin, Chris Cain, and Viresh Duvvuri for their encouragement and comments that helped improve this work. This research has taken place in the In-

telligent Robot Learning (IRL) Lab, Washington State University. IRL research is support in part by grants from AFRL FA8750-14-1- 0069, AFRL FA8750-14-1-0070, NSF IIS-1149917, NSF IIS-1319412, and USDA 2014-67021-22174. This research has been approved by IRB Certification of Exemption: #14565.

References

1. Meso west (2015), http://mesowest.utah.edu/, Accessed: 2015-12-10
2. Abadi, A., Rajabioun, T., Ioannou, P., et al.: Traffic flow prediction for road transportation networks with limited traffic data. Intelligent Transportation Systems, IEEE Transactions on 16(2), 653–662 (2015)
3. Alpaydin, E.: Introduction to machine learning. MIT press (2014)
4. Berson, A., Smith, S., Thearling, K.: An overview of data mining techniques. Building Data Mining Application for CRM (2004)
5. Cao, L., Gorodetsky, V., Mitkas, P., et al.: Agent mining: The synergy of agents and data mining. Intelligent Systems, IEEE 24(3), 64–72 (2009)
6. Cao, L., Weiss, G., Philip, S.Y.: A brief introduction to agent mining. Autonomous Agents and Multi-Agent Systems 25(3), 419–424 (2012)
7. Gehrke, J.D., Wojtusiak, J.: Traffic prediction for agent route planning. In: Computational Science–ICCS 2008, pp. 692–701. Springer (2008)
8. Google Maps: Popular times (2016), https://support.google.com/business/answer/6263531?hl=en, Accessed Jan 30, 2016
9. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182 (2003)
10. Horel, J., Splitt, M., Dunn, L., Pechmann, J., et al.: Mesowest: Cooperative mesonets in the western united states. Bulletin of the American Meteorological Society 83(2), 211 (2002)
11. Hsiang, S.M., Burke, M., Miguel, E.: Quantifying the influence of climate on human conflict. Science 341(6151), 1235367 (2013)
12. Huang, W., Song, G., Hong, H., Xie, K.: Deep architecture for traffic flow prediction: Deep belief networks with multitask learning (2014)
13. InsiderGuide Inc.: Disney World Wait Times, Touring Plans Free by Undercover Tourist (2015), https://www.undercovertourist.com/apps/, Accessed Jan 30, 2016
14. Klimstra, T.A., Frijns, T., Keijsers, L., Denissen, J.J., Raaijmakers, Q.A., van Aken, M.A., Koot, H.M., van Lier, P.A., Meeus, W.H.: Come rain or come shine: Individual differences in how weather affects mood. Emotion 11(6), 1495 (2011)
15. Klusch, M., Lodi, S., Moro, G.: Agent-based distributed data mining: The kdec scheme. In: Intelligent information agents, pp. 104–122. Springer (2003)
16. Lammers, M., Horel, J.D.: Mesowest 2.0: Providing real-time and archival surface observations with a modern and interactive website (2015), https://ams.confex.com/ams/95Annual/videogateway.cgi/id/29995?recordingid=29995, 31st Conference on Environmental Information Processing Technologies
17. Lv, Y., Duan, Y., Kang, W., Li, Z., Wang, F.Y.: Traffic flow prediction with big data: A deep learning approach (2015)
18. NoWait Inc.: NoWait - Get in Line At Restaurant Without Reservation (2015), http://nowait.com/, Accessed Jan 30, 2016
19. Rish, I.: An empirical study of the naive bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. vol. 3, pp. 41–46. IBM New York (2001)

20. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics 21(3), 660–674 (1991)
21. Tukey, J.W.: Exploratory data analysis (1977)
22. Vapnik, V.N., Vapnik, V.: Statistical learning theory, vol. 1. Wiley New York (1998)
23. Waze Inc.: Waze - GPS, Maps and Social Traffic (2016), `https://www.waze.com`, Accessed Jan 30, 2016
24. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, 2nd edn. (2005)