

CSE 471 Project Writeup

Team 2

David Ganey, Kerry Martin, Evan Stoll, Michael Theut, and Ben Roos
 School of Computing, Informatics and Decision Systems Engineering
 Arizona State University
 CSE 471

Executive Summary—The CEO of CactusCard has commissioned a team to solve several problems facing the up-and-coming financial company. The company is in the process of releasing a new card, called the CactusCardPlus. In order for this new product to be a success, the CEO recognizes that CactusCard Credit must accomplish three distinct goals. First, the company needs to successfully market the new card, through both traditional strategies and in particular to users of social media. In addition to a powerful marketing campaign, the company must ensure it makes good decisions about which applicants are deserving of a line of credit. Failing to do this would put the company at a disadvantage, relative to the numerous credit agencies who do use data to make credit decisions. Finally, the success of the product hinges on its ability to protect users (and the company) from credit card fraud. As such, the team must be prepared to counter fraudulent users through a variety of strategies.

These tasks will be undertaken by a team of computer science students with knowledge of artificial intelligence. Using search algorithms, the team will maximize the marketing potential of the card. Using machine learning techniques, the team will utilize data to accurately predict which applicants will best utilize their line of credit. Using game theory, the team will develop successful counter-fraud strategies which can be used to protect the company.

Paragraph(s) about the marketing part

After devising a successful marketing strategy, the team turned its focus to the determination of credit worthiness. As seen below in Section II, the team used a well-known machine learning library called Weka to run various machine learning algorithms on the data set provided. This data set, which contained 15 anonymized credit attributes plus whether or not that individual was given a card, was used to train learning machines which were then evaluated on their ability to accurately predict the issuance decision. The team found that a machine running the Multilayer Perceptron algorithm was the most successful at predicting credit decisions and would serve the company well in the future.

Paragraph(s) about the game theory part

I. PART 1

II. PART 2

Part 2 of the project describes another task given by the CEO of the CactusCard Credit Company, wherein a program should be developed to provide “a *machine learning based approach* to identify individuals who should/should not be given the CactusCardPlus.” This section will formulate this problem and describe the steps taken to solve it.

A. Problem Importance

The decision whether or not to grant a line of credit (and, as a corollary, how large that line of credit should be) is one of the most important decisions made by a credit agency. In this age of big data, agencies which fail to accurately utilize existing data to make these decisions are likely to fall behind agencies who are doing so. In this problem, CactusCard Credit is fortunate to have a dataset which tracks many credit attributes for a large number of users.

B. Problem Formulation

The problem itself is a machine-learning problem, which at its core simply means a program designed to recognize patterns. Machine learning can encompass a wide variety of problems, which can be divided into “regression” problems and “classification” problems. In this case, the consumer research division has given the team a dataset which contains information about customers *for whom the decision to award or not to award the CactusCard has already been made*. This is important information – because customers can either be given the card or not (a binary *classification*), this falls under the realm of *classification* machine learning problems. Additionally, the fact that the classification program will be trained with data means this is a *supervised* learning problem.

As a supervised classification machine, the program must be able to read the data set and, with that information, draw conclusions about which credit variables

have the largest impact on whether or not the customer was given the card. To be an effective tool for the CactusCard corporation, the dataset supplied by the consumer research division should only include proper decisions – cases where the card was awarded to the correct individual and denied to individuals who would have abused the line of credit. Otherwise, the machine will learn “bad habits” and will draw conclusions from the data which are not useful to the company.

The dataset itself is comprised of 15 credit attributes for the individuals, as well as an indication of whether the individual received a card. The credit attributes are not specified, though the problem specification does include the datatypes for each attribute. (For example, attribute 1 can be either “b” or “a”, and therefore represents some binary information about the customer such as whether they are male or female). These attributes, therefore, are all weighted equally, which may present a disadvantage. Further discussion of this issue can be found in Section IV. Additionally, it is important to note that the dataset is approximately balanced — approximately half of the individuals were granted cards, while half were not. This is helpful in correctly training the classifier.

The goal of this component of the project is to determine an algorithm which will effectively use the dataset to learn how the credit attributes affect the decision to award the card, and from there use that knowledge to predict card decisions within a certain degree of accuracy.

C. Solution Proposals

Weka, a software tool developed at the University of Waikato, is a software package with implementations of a large number of machine learning algorithms. Weka can be used in multiple ways. One option is to simply use the Weka API to access the algorithms from Java code. This is advantageous when one wishes to write a full machine learning system, or integrate machine learning algorithms with an existing Java system. Another option is to use the Weka Explorer, which is a wrapper around the algorithms. This GUI tool allows the user to load a data set and run the algorithms on it, then view the results in various formats.

It is the latter method which is proposed to solve the task given by the CEO. Using the Weka Explorer, we have a straightforward method for gaining insight into the effect of the 15 credit attributes. We can use the Explorer to test various machine learning algorithms, and based on their success rates, determine which should be used in the future to make the actual issue decision. An additional advantage of using this method is that Weka is tested, proven, open-source code. The Github repository

has nearly 8,000 commits, and therefore represents much more stable code than the team could have produced in the time allotted to this study [?]. This provides greater value for the company.

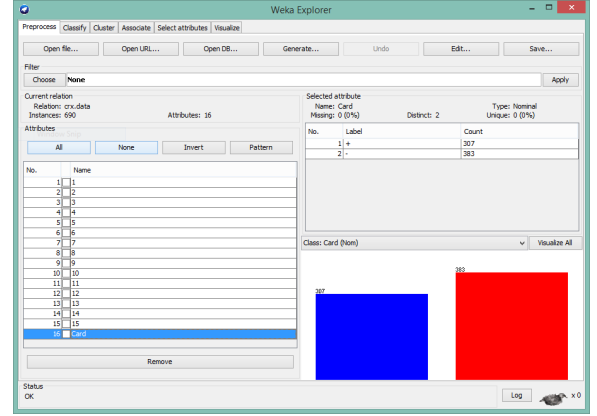


Fig. 1. The Weka Explorer interface

D. Test Plan

The team will evaluate several popular machine learning algorithms in the Weka Explorer. The success of each algorithm will be determined by a number of factors. The team must consider runtime, as the algorithm cannot take an extreme amount of time to complete or it will be useless. The most important factor, however, is the accuracy of the classification. Using random chance to classify card issuance would theoretically have an accuracy rate of approximately 50%, so the team must find an algorithm with a higher percentage than that. To ensure CactusCard’s continued financial success, the algorithm must predict card issuance correctly at least 80% of the time. The team proposes this number as it represents a realistic target for algorithms which make mistakes, yet still shows a tremendous potential advantage for CactusCard Credit.

E. Solution 1

Support Vector Machines (SVM) are, according to Russell and Norvig, “the most popular approach for ‘off-the-shelf’ supervised learning” [?]. SVMs do a good job of classification because they separate groups as much as possible (called a “maximum margin separator”). Additionally, SVMs can work in multiple dimensions (creating a “hyperplane”) which will be important with the high number of credit attributes in the data set. Finally, SVMs are considered a nonparametric method, because they store some of the training examples, and yet like parametric models, they often avoid overfitting the data [?].

To use an SVM on the data set provided by the CEO, Weka supports the addition of a library called libsvm. This Java library contains an implementation of the SVM algorithm which hooks into the Weka Explorer interface. The team will set the Explorer to use varying numbers of crossfold validation passes, in order to evaluate the classifier. This will split the dataset, train the SVM with a subset of the dataset, and then evaluate the SVM's effectiveness at predicting the card issuance decision by comparing the SVM's predictions to the actual data. The results from running the libsvm algorithm on the dataset are shown in the table below:

Crossfold validations	5	10	15
Time taken to build model (s)	.19	.24	.19
Correctly classified (%)	382 (55.36%)	385 (55.80%)	384 (55.65%)
Incorrectly classified (%)	308 (44.64%)	305 (44.20%)	306 (44.35%)

TABLE I. A TABLE SHOWING THE RESULTS OF THE LIBSVM APPLIED TO THE DATA SET

As we can see from the table, no significant difference exists when adjusting the number of crossfold validation passes. All of the attempts fall well within the expected runtime, and predict with approximately 55% accuracy whether someone was or was not given a CactusCard. While this is consistently better than randomly guessing, it only exceeds that percentage by approximately 5%. While SVMs are good classifiers, in this instance this algorithm is not a suitable choice for the company, as it fails the 80% criteria established in Section II-D.

F. Solution 2

Another type of machine learning algorithm is called a "multilayer perceptron." This is a type of neural network, which maps from inputs (in this case, the credit attributes) to outputs (whether someone was given a card). The MLP is suitable for this task because it uses "backpropagation," which is a supervised learning technique. TODO – EXPAND MORE ON THIS HERE

The multilayer perceptron is significantly more complex than the SVM, but conveniently Weka does include an implementation in the Explorer. Using the same technique as above, the team will evaluate this algorithm with various levels of crossfold validation to determine the optimal setting. The results are seen below:

Crossfold validations	5	10	15
Time taken to build model (s)	8.46	8.47	8.51
Correctly classified (%)	584 (84.64%)	572 (82.90%)	572 (82.90%)
Incorrectly classified (%)	106 (15.36%)	118 (17.10%)	118 (17.10%)

TABLE II. A TABLE SHOWING THE RESULTS OF THE MULTILAYER PERCEPTRON APPLIED TO THE DATA SET

Once again, it is clear that adjusting the number of crossfold validations has little to no effect on the accuracy of the classifier. As a result, the recommendation of the team is to use 5 validations on data sets of this size, so the algorithm does not take too long to run. While the time taken to build the data model was always nearly the same, the total runtime for all passes of the validation increased dramatically with larger numbers of cross fold validations. Since they appear to have no impact on the effectiveness of the algorithm, there is no need for the program to take that additional time. It is likely that on future datasets with larger numbers of individuals, the runtime would increase further still, reinforcing the decision that 5 crossfold validations is the correct choice.

Although no difference between the columns in Table II is visible, we can see immediately that the multilayer perceptron is a good choice for the CactusCard company. In all cases, the success of the classifier was above the criteria set above in Section II-D. The MLP algorithm averaged approximately 83% correctness, which is significantly higher than random chance. This algorithm would be a good choice for future data sets, where it could use the information it learned from this training set to choose who receives a card.

III. PART 3

IV. CONCLUSION

The conclusion goes here. Discuss issues with part 2 dataset – all attributes weighted equally

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

- [1] H. Kopka and P. W. Daly, *A Guide to L^AT_EX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.