Data Mining: The Mushroom Database

Hemendra Pal Singh*
**Seth Gyaniram Bansidhar Podar College, Nawalgarh, Rajasthan*
hemendrapalsingh1983@gmail. com

Abstract: This study will focus on the use of Data Mining techniques on previously analyzed data sets. The data mining tool Weka will be used. Weka stands for Waikato environment for knowledge analysis, and "is a popular suite of machine learning software written in Java, developed at the University of Waikato. WEKA is free software available under the GNU General Public License". The purpose of the study is to extend previous studies by running new data sets of stylometry, keystroke capture, and mouse movement data through Weka using various data mining algorithms. The study will also extend previous research at Pace University into the uses of a human- machine interface to increase the accuracy of machine learning. To this end, the study will use a nominal data set, the Mushroom Database.

Keyword-*Data Mining, Mouse Movement, Keystroke Capture, Mushroom Database, Stylometry.*

I. INTRODUCTION

Raw data is useless without techniques to extract information from it. According to I. H. Witten and E. Frank, "Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic advantage. The data is invariably present in substantial quantities"[8]. Different types of learning techniques can be used, including classification, association rules, clustering, attribute selection, normalization, instance based measures and decision trees. Selection of a learning technique is a difficult task that depends on the database and the types of desired results.

This project analyzes the stylometry, keystroke capture, and mouse movement data sets collected at Pace University in the fall of 2007. A Mushroom database compiled in the Audubon Society Field Guide to North American Mushrooms will also be analyzed [4]. In addition, an application will be developed to demonstrate a technique for creating a human machine interactive, web-enabled client-side text-based classification tool. In the Interactive Visual System for flower identification, the selection of the final result is a human decision whereas in the Mushroom database application, the human interaction determines only the details related to an instance. The actual final prediction is made based on machine learning. Analysis was done using the Mushroom database, which contains 8124 instances and 23 attributes. The attributes for the Mushroom Database are listed in the table in the appendix.

II. BACKGROUND

In order to explain the use of various algorithms in this study, the algorithms will be discussed. Naïve Bayes and Apriori will be used against the Stylometry data set. IBk will be used against the Keystroke Capture and Mouse Movement data sets. J48 will be used with the Mushroom Database. The choices of these techniques and their implementation will be discussed in detail in the methodologies section. First, some background information on these algorithms is given below.

According to Witten and Frank in Data Mining, the Naïve Bayes method is, "based on Bayes's rule and 'Naïvely' assumed independence — it is only valid to multiply probabilities when the events are independent. The assumption that attributes are independent (given the class) in real life certainly is a simplistic one"[8]. However, Naïve Bayes shows very high accuracy with

certain data sets especially when redundant attributes are eliminated.

In order to discuss the Apriori algorithm, two key terms must be defined. The *support* is the accuracy required. The *confidence* is a measure of the correctness of the rule and can be determined through counting the number of transactions that fulfill the rule. Any association rule without enough *support* is rejected. In Apriori, x1, x2, . . . , xn predict y, where y is a classification which may in some cases represent useful information such as the likelihood of a future purchase. Apriori accepts only nominal data although numeric data can be discretized to nominal form.

IBk is the k-nearest-neighbor technique, where k represents the number of clusters, and the 'I' and 'B' stand for Instance Based. According to Witten and Frank, IBk "uses the instances themselves to represent what is learned rather than inferring a rule set or decision tree[8]". Instance based learning uses a distance measure to determine the classification of instances based on their proximity to a new unknown instance. Often Euclidean distance is used; the equation for calculating Euclidean distance is provided below.

Given two points, (x_1, y_1) and (x_2, y_2), Euclidean distance can be calculated as:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

It is also important to note that when k is set equal to 1 in IBk, the result will be the same as if the algorithm IB1 had been used.

Weka's implementation of C4. 5 revision 8 is J48. C4. 5, in turn, is the result of improvements of ID3. All of these algorithms utilize what Witten and Frank call the "divide-and-conquer approach to decision tree induction," or, the "top-down induction of decision trees". This approach was created by J. Ross Quinlan of the University of Sydney, Australia. Improvements to ID3, which resulted in C4. 5 "include methods for dealing with numeric attributes, missing values, noisy data, and generating rules from trees. . . . "

III. FOCUS OF STUDY

The purpose of Data Mining, as defined by I. Witten & E. Frank, is, "the process of discovering patterns in data, the patterns discovered must be meaningful in that they lead to some advantage"[9].

The focus of this study is to run algorithms in WEKA against the Mushroom Database, as well as the mouse movement, keystroke capture, and stylometry data sets that have been collected in Fall 2007. In addition this study will cross reference its findings with those of previous studies and discuss the results. Each data set will be analyzed based on a focused methodology literature survey and will be analyzed accordingly. The following sections will discuss the prior studies from the literature survey.

A. Mushroom Database

Schlimmer's dissertation will act as a starting point for further analysis of the Mushroom database[6]. Shimmer sets forth that his rule set of four rules should be treated as a landmark. Therefore, this study will compare its resulting decision tree to Schlimmer's rule sets in terms of accuracy and ease of use. The study will demonstrate the effectiveness of the resulting decision tree through the development of an application with a graphical user interface wherein a user may determine the edibility or poisonousness of a mushroom sample based on their answers to pertinent questions.

The application will be a text-based, web-enabled human machine interactive client-side application that is extensible. The purpose of making the application web-based is that it will be accessible to anyone for use. This significantly increases the value of the application itself, as, unlike previous applications developed at Pace University, it will

not require special equipment. One such previous application, "Interactive Visual System", was built as an application for use on a desktop or a Sharp Zaurus SL-5500 handheld computer[2]. Unlike the "Interactive Visual System" the Mushroom database application will be accessible and usable from all manner of Web browser accessible devices. While the "interactive visual system" enabled the classification of photographs taken of different flowers, this system presumes that the user has first hand knowledge of, or access to, their mushroom sample. One of the contributions made by this study will be the extension and reuse of a technique proven to be effective by the "interactive visual system", human-machine interaction. The Mushroom database application will demonstrate not only the proven effectiveness of a human-machine interactive visual system, but will also show the capacity for other types of systems that use data mining techniques along side human interaction.

B. Stylometry, Keystroke Capture, and Mouse Movement

Previous studies conducted by students and/or faculty of Pace University have documented attempts at classification of the stylometry, keystroke capture, and mouse movement data sets. Often these studies show excellent accuracy, however in some cases they provide a starting point rather than a definitive result. Therefore, many of the techniques used in this study will draw upon the conclusions of previous studies. The purpose of running these data sets through the algorithms offered by WEKA is to improve the accuracy of classifications.

IV. RESEARCH METHODOLOGY

Several different methodologies will be used to analyze the various data sets. First, classifiers that do not generate rules will be used on the Mushroom database such as PRISM and will be compared to the accuracy of an unpruned tree. In this instance, the purpose of using such disparate approaches is to analyze the accuracy of the Mushroom database application. The application will use an unpruned decision tree to evaluate input by a user. The Stylometry data set will be analyzed using classifiers in an attempt to discover rules for author identification. The mouse movement and keystroke capture data sets will be analyzed using a nearest neighbor approach in an effort to extend previous studies in these areas to the new data sets. These approaches and the reasons for these approaches will now be described in more detail.

A. Mushroom Database

As a primary approach to the Mushroom database, a J48 unpruned tree will be used in the design and implementation of the Mushroom database application. As a secondary approach, classifiers that do not generate rules such as IBk and the Voted Perceptron will be compared to the results of the J48 unpruned tree. As a tertiary approach, the PRISM classifier, which does generate a rule set, will be run on the Mushroom database. All of these approaches will then be compared to an optimal rule set described in the dissertation, "Mushroom database" by Jeff Schlimmer[6]. Schlimmer found optimal rules for the Mushroom database using back propagation methods in 1987 to 95% accuracy. However, if it is possible that a greater accuracy can be found, and used to advise users of the Mushroom database application then these results will be useful in analyzing similar data sets wherein the question of whether or not a fungus or plant is edible could be answered with great accuracy.

B. Stylometry

The Stylometry data set will be analyzed using Naïve Bayes as a frequency analysis and a discretized Apriori as a classifier for the feature vector data. The study, "The Use of Stylometry for Email Author Identification: A Feasibility Study" by Goodman, Hahn, Marella, Ojar, and Westcott, attempts to analyze Stylometry data on the basis of a keystroke capture approach[3].

Some of the approaches therein had moderate success; however, these successes could be improved upon, perhaps through the use of Apriori. Furthermore, as an experimental means to step away from analyzing email authorship through keystroke capture, Naïve Bayes will be used to help generate a rule set that will focus more on stylistic idiosyncrasies for email author identification. As mentioned in the Feasibility Study, historical examples of authorship analysis focused on "lexical, syntactic, content, and complexity features". The Feasibility Study goes on to state that "This was a long, painstaking, manual process," but with the use of Naïve Bayes, perhaps that process can be expedited.

C. Keystroke Capture

The analysis of the keystroke capture data set will further the approach used in "keystroke biometric recognition on Long-Text Input: A Feasibility Study" by Curtin, Tappert, Villani, Ngo, Simone, Fort, and Cha [1]. The thesis of the Feasibility Study is continued and applied in more detail in "Keystroke Biometric Recognition Studies on Long-Text Input under Ideal and Application-Oriented Conditions" by Curtin, Tappert, Villani, Ngo, Simone, Fort, and Cha [7]. This analysis of the new Keystroke Capture data collected in Fall of 2007 will apply the k-nearest neighbor classifier, IBk, to feature vector data. While this analytical approach is not new, perhaps it will reinforce the findings of these previous studies though applying them to newer data.

D. Mouse Movement

As per the recommendations made for further study in, "Mouse Movements Biometric Identification: A Feasibility Study" by Weiss, Ramapanicker, Shah, Noble, and Immohr, the Mouse Movement data will be analyzed using the IBk algorithm, a k-nearest neighbor classifier [7]. While this same approach was used in the feasibility study, it has not been extended to the new data set collected in the Fall of 2007. Furthermore, through choosing different values for k in the k-nearest neighbor approach, different levels of accuracy may be reached.

V. IMPLICATIONS OF INITIAL STUDY

A. Mushroom Database

Six separate data sets were used in evaluating the Mushroom Database. The client provided three data sets that were taken from Schlimmer's compilation of the Audubon Society's Mushroom data. These sets, referred to herein as training, test 1, and test 2, each contained 1000 instances, for a total of 3000 instances. Schlimmer's data set contained 8124 instances and was split into three sets, referred to herein as strain, stest 1, and stest 2, each with 2708 records.

Preliminary results for the Mushroom database show an extremely high level of accuracy with classifiers that generate rules, classifiers that do not generate rules, and with the J48 unpruned tree. Indeed, the lowest level of accuracy reached was with the use of the voted perceptron. The percentages of correctly classified instances were 99.5%, 95.3%, and 98.4% on the training, test 1, and test 2 data sets, respectively. At the next level of accuracy, the PRISM classifier, which does generate rules, was able to reflect the following percentages on training, test 1, and test 2 data: 100%, 99.4%, and 99.7%. However, in order to reach this level of accuracy, one attribute had to be removed. Attribute 11, Stalk-root, which had a number of missing values, was not included in these PRISM analyses. At the next level of accuracy, IBk is the lazy classifier, which had 100%, 99.5%, and 100% accuracy on the training, test 1, and test 2 sets. Perhaps surprisingly, the most accurate results were found using the J48 unpruned tree, which had 100%, 99.6%, and 100% accuracy on the training, test 1, and test 2 data. Furthermore, the unpruned tree itself presents a much more accurate means by which to evaluate new Mushroom data. This is because the PRISM rules are disjunctive and therefore, will not as accurately catch the classification of

every instance within a human-machine interactive application. The reason for this is that in programming, disjunctive if statements often leave certain possibilities unaccounted for. While there is the potential for this in the unpruned tree, it can more easily be countered through the use of nested if statements. The J48 unpruned tree, although more accurate and cohesive than the PRISM rule sets, has a larger number of possibilities that must be evaluated in order to reach a conclusion than does the optimal rule set described by Jeff Schlimmer in his dissertation, "The Mushroom database. "

Schlimmer's rule set could also be used for a human-machine interactive application that would help to identify mushrooms based on their classifications. However, Schlimmer achieved a 95% accuracy in his evaluations versus a 99.6% accuracy at the lowest in the J48 unpruned tree analysis. Therefore, as one of the objectives in creating a human-machine interactive mushroom identification application is to warn users about potentially poisonous mushrooms, the higher accuracy result will be used in generating the application.

The training set, strain, evaluated to 100% accuracy in correct classifications of instances using the same J48 unpruned tree algorithm. However, this training set, strain, resulted in a less accurate unpruned tree. This is illustrated through the results of running stest 1 and stest 2, wherein no higher than 92.8% accuracy was reached. For this reason, the initial J48 unpruned tree results achieved through running the training set provided by the client were used in developing the Mushroom database application. It is interesting to note that when the entire Schlimmer data set of 8124 records was run through Weka using the J48 unpruned tree algorithm, a somewhat similar decision tree as that of the client's training results was achieved.

B. Mouse Movement Data Set

Preliminary results for the mouse movement data set show a high level of accuracy with certain techniques used with IBk, the k-nearest-neighbor classifier. The highest accuracy, 100% in correctly classified instances can be reached using IBk, with KNN set to 1 and a percentage split of 80% with cross validation. The same 100% of correctly classified instances can also be reached using IBk, with KNN set to 3 and a percentage split of 80% with cross validation. Other approaches did not have the same percentage of correctly classified instances. For example, using best first attribute selection to find the four best attributes, which were average curve speed, average curve time, average click duration, and SD click duration, it is possible to narrow the number of attributes included in the analysis. In order to do so, preprocessing removed all but the best four attributes and the class. The training and test sets contained only the best four attributes. The training set had 20 instances, whereas the test set had 5 instances. Then, KNN was set to 1 and the training set was used with and without cross validation. In both instances, 100% of instances were correctly classified in the training data while 80% of correctly classified instances were obtained on the test set. Clearly, the 80% percentage split approach was the most accurate.

VI. DESCRIPTION OF MUSHROOM DATABASE APPLICATION

The Mushroom database application is a web-enabled, text-based, human machine interface that displays the correct classification of an instance based on the observations submitted by the human user. The application is written entirely on the client side in HTML and JavaScript and therefore is extensible. The application is accessible from any web-enabled device. The prototype can be used with other types of decision trees that are based on other types of databases. The application makes use of the J48 unpruned tree results as they

were found to be more accurate than the rule set discovered by Schlimmer. The application uses form input to determine the values of an attribute. The application has the capacity to be reconfigured for use with a server-side script that accesses a database. This could be of use as there is a great potential for developing a similar application that would work with the results of various unpruned trees stored in a database. In such an application, the user would specify what type of instance they wish to classify initially. To this end, the first array in the code is written as the second level of classification.

VII. LIMITATIONS AND OPPORTUNITIES FOR RESEARCH

As it was stipulated in the research methodology section of the present technical paper, the studies are covering areas of mushroom database, stylometry, keystroke capture, and mouse movement. data mining techniques are used in the classification process, based on the feature data provided by the other teams, considered front-end systems. A multitude of classifiers/ methods can be found during the course of the project: J48 unpruned tree, IBk, voted perceptron, PRISM, naïve bayes, k-nearest neighbor. All of these classification/learning methods have differences in applicability, meaning there is no best method but only optimum ones, depending on the particular project's data set. That is the maneuvering space, and studies can be continued with more or less datasets but most important, can be improved by finding the optimum classifier and/or choosing the right parameters for the process of classification.

It is important to understand the fact that most of the algorithms of classification/authentication involved in the project are not producing a 'clean cut' result. They depend on user's decisions, like what should be the threshold for an optimum classification. To emphasize the importance of the chosen method, the associated parameters, and the limitations of it, a chosen example is Naïve Bayes method. Although good results were obtained using sophisticated learning methods on many datasets, Naïve Bayes does just as well, or even better. Of course, there are many datasets for which the method does not do so well; "— Attributes are treated as though they were completely independent, the addition of redundant ones skews the learning process."[8] Relatively, the same arguments can be brought in the case of IBk, where the accuracy of the methods can be improved by selecting an optimum value for k.

In conclusion, the quality of the project depends on the fine balance between datasets and chosen methods. On the same idea, a comparative analysis of different classification/learning methods, based on the same feature data, had been examined. Most likely, even if the studies are involving the best-chosen methods, they will be improved by future students/customers, who upon analyzing the projects and their results will fill in the gaps.

VIII. CONCLUSION

While it has not yet been possible to run the stylometry and keystroke capture data sets collected in Fall 2007, this study will draw conclusions based on the Mushroom Database and Mouse Movement data set results.

The highest accuracy result from running the training, test 1 and test 2 data sets of the Mushroom database through Weka was achieved with the J48 unpruned tree. This unpruned tree was then used to generate a web-enabled human-machine interactive mushroom identification application. This application further increases the accuracy of individual mushroom classifications through human interaction. Furthermore, the usability of the application increases its usefulness. That is, as this application is web-based and does not require any special skills besides basic knowledge of the Internet, it will be accessible to casual users and academics alike.

The highest accuracy result from running the mouse movement data set through Weka was achieved with 80% percentage split using either KNN equal to 1 or KNN equal to 3. The percentage of correctly classified instances obtained through this methodology was 100%.

REFERENCES

[1] M. Curtin, C. Tappert, M. Villani, G. Ngo, J. Simone, H. St. Fort, and S. -H. Cha. "Keystroke Biometric Recognition on Long-Text Input: A Feasibility Study," *Proc. Student/Faculty Research Day, CSIS*, Pace University, White Plains, NY, May 2006, pp. 1-5.

[2] A. Evans, J. Sikorski, and P. Thomas, "Interactive Visual System," *Proc. Student/Faculty Research Day*, CSIS, Pace University, White Plains, NY, May 2003, pp. 1-6.

[3] R. Goodman, M. Hahn, M. Marella, C. Ojar, and S. Westcott, "The Use of Stylometry for Email Author Identification: A Feasibility Study," *Proc. Student/ Faculty Research Day, CSIS*, Pace University, White Plains, NY, May 2007, pp. 1-7.

[4] G. H. Lincoff (Pres.), *The Audubon Society Field Guide to North American Mushrooms*, New York, Alfred A. Knopf, 1981.

[5] M. Ritzmann and L. Weinrich, "Strategies for Managing Missing or Incomplete Data in Biometric and Business Applications," *Proc. Student/Faculty Research Day, CSIS*, Pace University, White Plains, NY, May 2007, pp. 1-6.

[6] M. Villani, C. Tappert, G. Ngo, J. Simone, H. St. Fort, and S. -H. Cha, "Keystroke Biometric Recognition Studies on Long-Text Input under Ideal and Application-Oriented Conditions," *Proc. Student/ Faculty Research Day, CSIS*, Pace University, White Plains, NY, May 2006, pp. 1-8.

[7] A. Weiss, A. Ramapanicker, P. Shah, S. Noble and L. Immohr, "Mouse Movements Biometric Identification: A Feasibility Study," *Proc. Student/ Faculty Research Day, CSIS*, Pace University, White Plains, NY, May 2007, pp. 1-8.

[8] Witten, I. H. and Frank, E. , *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition, Morgan Kaufman Publishers, San Francisco, 2005.