Proceedings of Student-Faculty Research Day, CSIS, Pace University, May 2nd, 2008

Data Mining on a Mushroom Database

Clara Eusebi, Cosmin Gliga, Deepa John, Andre Maisonave
Seidenberg School of CSIS, Pace University, White Plains, NY, 10606, USA
{clara.eusebi, a.maisonave1982}@gmail.com, {cg26013w, dj89833w}@Pace.edu

Abstract

This study focuses on the use of data mining techniques to analyze a previously obtained data set. 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, and the data mining tool Weka. Various data mining algorithms are used against the Mushroom Database, including an unpruned decision tree, a voted perceptron algorithm, a covering algorithm that generates only correct rules, and the nearest neighbor classifier. Finally, an unpruned tree is used to develop a human-machine interactive application.

1. Introduction

A Mushroom Database is analyzed. Two data sets were used, one consisting of all records from a database compiled by J.S. Schlimmer [9] and the other consisting of a subset of records provided by the client, Dr. S. - H. Cha. In addition, an application was developed to demonstrate a technique for creating a human-machine interactive, web-enabled client-side text-based classification tool. In the article, "Interactive Visual System" by A. Evans et al [5], the flower identification tool requires the selection of the final result by a human whereas the Mushroom Database application developed in this study requires human interaction during the classification process. However, the actual final prediction is made based on machine learning. The Mushroom Database compiled by Schlimmer contains 8124 instances and 23 attributes. The subset provided by the client contains 3000 instances and the same 23 attributes.

2. Background

Several key concepts, algorithms and techniques, will be discussed because they are used in this study. First algorithms and techniques will be discussed including an unpruned decision tree (J48), a voted perceptron algorithm (VotedPerceptron), a covering algorithm that generates only correct rules (PRISM), the nearest neighbor classifier (IBk) and the best first search (BestFirst) for attribute selection technique (Weka names in parentheses). Then, Jeff Schlimmer's dissertation, "Concept Acquisition Through Representational Adjustment" [9], will be described and discussed as both of the data sets used in this study were drawn from the database compiled by Schlimmer as a part of his dissertation. Finally, the confusion matrix will be discussed because it is a key concept utilized in the analyses accomplished in this study.

2.1. Algorithms and Techniques

The k-nearest-neighbor (IBk) technique uses the majority class of the nearest k neighbors to determine the class of a new instance. The IB stands for "Instance Based," and instance-based learning uses a distance measure, often Euclidean distance, to determine the classification of instances based on their proximity to a new unknown instance.

A covering algorithm (PRISM) tests the rule that is being created in order to create a maximum accuracy rule. PRISM aims for maximum correctness of a rule and does not consider a rule to be valid unless it has 100% accuracy. Whenever a new rule is created, the full set of given data is used and tests are run until only conclusions relating to the target class are included in the rule. PRISM rules are therefore disjunctive.

The voted perceptron algorithm (VotedPerceptron) is a linear classifier and a precursor to neural networks. According to Witten and Frank, "The algorithm iterates until a perfect solution has been found, but it

will only work properly if a separating hyperplane exists, that is, if the data is linearly separable. Each iteration goes through all the training instances." Once a misclassified instance is found, the definition of the hyperplane is modified, such that the miscategorized instance can be reclassified. Since updates are made in order of the iterations through training instances, if a change is made to the definition of the hyperplane for a misclassified instance encountered after other misclassified instances, the previous changes may be subverted in that process. If the data is linearly separable, there will be a finite number of iterations. Otherwise, the algorithm will get stuck in an infinite loop and therefore, a maximum number of iterations must be specified. The voted perceptron is a modification of the perceptron algorithm wherein all the weight vectors encountered during the learning process vote on a prediction. A measure of correctness of a weight vector, based upon the number of successive trials in which it correctly classified instances, can be used as the number of votes given to the weight vector. This is the voted perceptron.

The unpruned decision tree (J48) was used in this study to find classifications for the Mushroom Database. Weka's implementation of C4.5 revision 8 is J48, which generates decision tree classification models. C4.5, in turn, is the result of improvements to 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"[12]. 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...."

Best first attribute selection (BestFirst) is a technique for finding the top attribute or attributes in the data. Best first records a list of all attribute subsets evaluated in order of their performance so that an earlier configuration can be selected if no better performance selections are found later on. An upper bound is usually set in best first attribute selection for the number of attributes that will be accepted. This prevents the search from continuing to evaluate the entire set of permutations and combinations for every number of attributes.

2.2. Schlimmer's Dissertation

In his dissertation, Jeff Schlimmer used a database and applied mining techniques with the purpose of obtaining good classifiers that can be used on similar databases. He compiled a mushroom database drawing the records from The Audubon Society Field Guide to North American Mushrooms [8]. The Audubon Society's past and current mission is to conserve and restore natural ecosystems. The data set compiled by Schlimmer comprised samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota families. Each species was identified as belonging to one of three classes, definitely edible, definitely poisonous, or unknown edibility and thus, not recommended. Because of the potential dangers associated with unknown edibility, this class was combined with the poisonous one, resulting in a bigger 'definitely poisonous' class. Because the Audubon Society's Field Guide states that there is no simple rule for determining the edibility of a mushroom, Schlimmer saw a chance to find rules that can sort and classify the mushrooms. This was done by implementing a supervised learning system based on decision trees. The maximum obtained accuracy was 95%. This is a relatively good accuracy, but with current improved techniques the accuracy can be even higher.

2.3. Confusion Matrix

The confusion matrix is used in the case of supervised learning; for unsupervised learning there is a matching matrix. In a confusion matrix, each column represents the instances in a predicted class, while each row represents the instances in an actual class. The information can be read horizontally and vertically. The matrix is used because it is easy to see if the system is mislabeling or confusing the classes and in this way it can be found out what the accuracy of a classification is. For a better understanding, below is an example.

```
   a   b   <-- classified as
 500   0 |   a = e (edible)
   5 495 |   b = p (poisonous)
```

In this example, the confusion matrix has two columns and two rows. When read horizontally, to classify 'a', there are 500 instances that belong to 'a' and 0 instances that belong to 'b'. That means that there are 500 instances that are correctly classified as being in the 'a' class, and 0 incorrectly classified instances. This relates to 100% accuracy on classifying 'a'. The second row has 495 instances correctly classified as being in the 'b' class and 5 instances incorrectly classified as being in 'a' class. This shows 99% accuracy in classifying 'b'. Overall, the classifier shows 99.5% accuracy. In conclusion, the confusion matrix should be read diagonally; the accuracy of the

classifier is higher as the number of incorrectly classified instances on each row is lower.

3. Literature Review

3.1. Data Mining Overview

In the article, "An Overview of Data Mining Techniques" by Alex Berson et al [3], the authors state that there is little difference between data mining and statistical techniques. According to the authors, "Today data mining has been defined independently of statistics though 'mining data' for patterns and predictions is really what statistics is all about." They go on to describe some of the major concepts in statistics and data mining including histograms, linear regression, nearest neighbor, clustering, hierarchical clustering, including agglomerative and divisive clustering techniques. Other topics discussed include what the authors call next generation techniques, such as decision trees, neural networks, and rule induction. The authors conclude that deciding what techniques to implement can be a difficult choice, but that some of the criteria for determining the optimal techniques can be "determined by trial and error."

In the article, "Understanding Data Mining: It's All in the Interaction" by Kurt Thearling [11], the author states that since the output of data mining is unpredictable, visualization and interaction are needed to gain user trust. The user needs to gain a better understanding of the output of data mining so that data mining tools can be used for different business solutions. This can be achieved through the visual presentation of data mining output in a meaningful way and through allowing user interaction along with visualization. Simple questions increase the interaction of the user who will be better able to understand data mining tools. Data mining application developers, according to the author, should develop data mining tools by taking into consideration the "understandability" and interactivity of their applications.

While Thearling makes suggestions for improving the client experience of data mining through interactivity and visualization, the Berson et al give an overview of statistical and data mining techniques rather than an overview of business concerns where data mining is involved. However, both articles take business concerns into consideration. Yet business concerns are not the main focus of Berson et al, who instead focus on helping the reader to understand the key concepts in statistics and data mining.

3.2. Decision Trees

The article entitled "Measuring Data Quality: Predictive Accuracy vs. Similarity of Decision Trees" [7] presents findings about the accuracy of some techniques that have been proposed to protect confidentiality of individual information. These techniques surfaced due to the inevitable separation and exchange of information between entities that collect the information and those that are extracting useful information from it; these collector entities are described as organizations, companies, medical service providers, government departments/agencies. The discussed techniques were implemented due to the growing public concern about privacy and laws/regulations that deal with privacy/confidentiality; they are called perturbation techniques, with the purpose of providing security in statistical databases and data mining.

Attribute perturbation techniques were used to produce several perturbed datasets from the original dataset. Decision trees and neural networks were used as classifiers on the original and perturbed datasets to predict the classes of the test, original and perturbed datasets. A goal was to obtain a good control technique that can preserve the quality of datasets. Another goal was to show the effect of perturbed data on predictive accuracy. The main work was based on comparison of the measure of data quality to the similarity of decision trees obtained from the original and perturbed datasets. The conclusion stated that more research is needed because the experimental results showed that measuring data quality by predictive accuracy of the decision trees is inconsistent with measuring data quality by similarity of the decision trees built on original and perturbed datasets. The conclusion also stated that random noise is not likely to cause the recognition of new patterns by the classifier; rather, it is more likely to strengthen or weaken preexisting patterns.

The work presented in the article entitled "Effect of the χ^2 test on construction of ID3 decision trees" [10] is about evaluation of the effect of using the $\chi 2$ statistic as a tool for identifying noise during ID3 tree construction. ID3 is a machine-learning system that was developed with the purpose of producing knowledge from a limited training set; it builds a decision tree through inductive steps. When using ID3 there is a chance that data contains noise, thus leading to construction of an inaccurate decision tree; this is a known weakness associated with decision trees. To alleviate this problem, the paper mentions two approaches. In the first one, inconsistencies are corrected by controlling the errors externally (a

separately codified system checks the validity of each concept description); this process is tedious, and still does not guarantee that all errors have been found. The second approach is based on an error-handling mechanism, which is integrated into the inductive system; this is a preferred approach, because it is easier to automate and it is operationally less complex. Based on the second approach, the work presented in the paper is about evaluation of the effects of $\chi 2$ test when applied to the development of ID3 decision trees. This is done by comparing the results from trees constructed with and without the $\chi 2$ test. The conclusion of the paper was that, based on experimental results, the $\chi 2$ test significantly affects the construction of ID3 trees, in the sense that the trees are less complex, and, in some cases, performance was poor. The paper also concluded that the most probable conclusion approach, wherein the most frequently occurring classification at a node is used as the node class, when used in conjunction with the $\chi 2$ test, performed consistently well in the presence of increasing noise.

These articles discuss the challenges related to noise in data when building and constructing decision trees. In the first article, the noise is deliberately induced in the data with the purpose of obfuscating the data from unwelcome attention; the second article evaluates the accuracy obtained by working with the $\chi 2$ test in conjunction with conclusive accuracy criteria for decision trees.

3.3. Visual Classification and Human-Machine Interaction

Data mining and machine learning strive in directions that haven't yet been explored; it is an underestimated science with an immense amount of possibilities.

Two articles that further this point are "*Machine assisted visual grading of rare collectibles over the Internet*" by Richard Bassett of Western Connecticut State University [2] and "*Visual Classification: An Interactive Approach to Decision Tree Construction*" by Mihael Ankerst [1], Christian Elsen, Martin Ester, and Hans-Peter Kriegel from the Institute of Computer Science, University of Munich.

Bassett's article is concerned with the visual grading of collectible coins. Bassett describes how it takes years and much expertise for any human to properly and professionally grade valuable coins; Bassett states, "This study examines the human visual recognition process in the grading of rare collectibles and addresses a number of issues, limitations and constraints inherent to human visual recognition as a way of explaining grading variation". Bassett continues in further detail to describe how a human must over time and with experience develop both long-term and short-term memory in order to be taken seriously as a coin grader. Coin graders must pay attention to categories such as condition, authenticity, age, and historical significance when grading. In regard to human grading, Bassett concludes that humans have limitations, and he puts the cognitive and visual recognition abilities of humans to the test, stating, "Extensive documented research in the field of cognitive psychology supports the claim that humans have great difficulty processing more than 5-9 chunks of visual input data at one time. Thus by our very own cognitive nature humans lose, drop or fail to consider anywhere from 5 to 15 detailed technical features when grading coins". As human beings, we put our bodies through physical, mental, and environmental factors that can assist in losing our cognitive abilities. No one human is built the same way as another and therefore, results cannot be exact but only similar. Limitations are caused by speculation and differences of opinion.

Machine learning has both advantages and flaws. Bassett describes two companies, the Professional Coin Grading Service (PCGS) and CompuGrade systems, each of which attempted to automate this process. In coin grading, there are both technical and non-technical considerations. However, machine learning alone has failed because a machine could not properly simulate the non-technical aspects of grading coins. Despite the high technical quality of the gradings accomplished by the two companies, these companies were unpopular and folded due to increasing software development costs and lack of popularity.

Bassett summarizes, "Both PCGS and CompuGrade attempted to build systems that they anticipated would become commercially viable and profitable. They soon discovered that the development of software could be a long and expensive process". All of the experiments for this study were accomplished over the Internet. Bassett employed various experiments and tests.

The very nature of the experiments that Bassett used shows the validity of a human-machine interface. His analysis of human coin grading alone and machine coin grading alone leads to the conclusion that neither approach is sufficient when taken separately. A human-machine interface, such as the online coin grading system that Bassett developed, holds more potential.

The second article, from the Institute of Computer Science located in the University of Munich, does not offer a new concept in machine learning and data

mining, but rather offers an improved system for human-machine interaction and visualization. The abstract states, "... we introduce a fully interactive method based on a multidimensional visualization technique and appropriate interaction capabilities". The visual human-machine interactive system was developed with classification as a main goal. The article states, "Classification is one of the major tasks of data mining. The goal of classification is to assign a new object to a class from a given set of classes based on the attribute values of this object". Throughout the course of this study, attributes and attribute values were mapped to colors to aid in visual classification. Their method known as 'Circle Segments', "maps d-dimensional objects to a circle which is portioned into d segments representing 1 attribute each". Once the training is completed, a "Data Interaction Window" presents the user with the option of reevaluating the data classifications based on the visual display.

Human-machine interactive and visual classification methodologies improve classification accuracy and the confidence of human users in data mining systems.

In the article, "Interactive Visual System" by A. Evans et al, a visual flower classification system is implemented on a Sharp Zaurus SL-5500 hand held computer [5]. This implementation of the Interactive Visual System illustrates that human-machine interaction greatly improves the accuracy of identification systems. Some preliminary effectiveness tests were run comparing the speed and accuracy of classification using just a machine versus an interactive application on a hand-held device versus on a desktop. Their results are presented as observations as they do not draw from a large enough database or user sampling, and therefore, cannot be considered statistically valid. Other limitations encountered include the limited size of the flower species database. The authors of this study anticipate that as the number of instances grows the accuracy of the identifications will decrease. Classification was accomplished by using the k-nearest-neighbor algorithm. Features are extracted interactively; the user can choose up to five features for extraction. All features are normalized and unweighted, and k-nearest-neighbor is applied with a value of k=1.

For this system, a graphical user interface was developed that allowed the user to narrow the field of analysis in the image through selecting the center of the flower with a stylus. Further interaction allowed users to select the petals with the stylus to further aid the device in feature extraction. Furthermore, the final identification decision is made by the user who selects from three images that were classified as the top three most likely matches by the device. Once a positive identification has been made by the user, the data can be added to the database. If no positive identification has been made, the data can be added as a new species.

In the article, "Interactive Flag Identification" by Hart, Tappert and Cha [6], an interactive system for the identification of flags based on photographs of actual flags in varying circumstances is proposed. The challenge in classifying such flags is that flags may be furled. Distortions caused by background imagery, such as the setting in front of which the flag stands, cause noise, which can greatly increase the difficulty of feature extraction. Therefore, users crop the image to the flag. Subsequently, users choose the correct flag based on an optimal list of matches identified by the machine. Feature extraction for flag identification is based on the concept of dividing the image into small blocks of pixels before extracting color and texture. The classifier used was k-nearest-neighbor. Classifications are based on 186 flag classes each containing 36 samples.

The article, "Forged Handwriting Detection" by Hung-Chun Chen [4], hypothesizes that forged handwritings are written more slowly than authentic handwriting and are therefore, less smooth than authentic handwriting. A system was developed for the purpose of identifying forged handwriting. The study uses the IBM Trans note, a pen-enabled notebook computer, and an associated SDK, which provides x and y coordinates as a function of time. This information yielded speed and acceleration of the handwriting samples. The researcher also evaluated the wrinkliness of the handwriting based on digital scans of handwriting samples. For the digital scans, the wrinkliness of the writing was extracted based on a machine count of the number of pixels on the boundary of the handwriting at both 300 and 600 dpi. Finally, the article concludes that there is a 93.5% probability that forged handwriting will be wrinklier than authentic handwriting.

The Interactive Visual System and the Interactive Flag Identification system are similar in terms of the methods used for human-machine interaction. In both cases, the user aids in feature extraction and makes the final identification decision based on optimal suggestions made by the machine or device. The Forged Handwriting Detection study uses feature extraction based on a different methodology and does statistical analysis on samples collected rather than using data mining techniques for identification.

4. Focus of Study

The purpose of data mining, as defined by Witten and Frank, is, “the process of discovering patterns in data … The patterns discovered must be meaningful in that they lead to some advantage” [12].

The focus of this study is to run algorithms in Weka against the Mushroom Database. In addition this study will cross reference its findings with those of previous studies and discuss the results.

Finally, a human-machine interactive application, the Mushroom Database application, which was developed during this study, is discussed.

Schlimmer's dissertation, "Concept Acquisition Through Representational Adjustment" [9] will act as a starting point for further analysis of the Mushroom Database. Schlimmer sets forth that his rule set of four rules should be treated as a landmark. Therefore, this study will compare an unpruned 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 is 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 is 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 [5]. Unlike the "Interactive Visual System" the Mushroom Database application is 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 is 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 system, but will also show the capacity for other types of systems that use Data mining techniques along side human interaction.

5. Research Methodology

Classifiers that generate rules and classifiers that do not generate rules were used on the Mushroom database and are compared to one another in terms of accuracy. The purpose of using such disparate approaches is to analyze their accuracy against one another for the Mushroom database.

As a primary approach to the Mushroom database, a J48 unpruned tree was 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 VotedPerceptron were compared to the results of the J48 unpruned tree. As a tertiary approach, the PRISM classifier, which does generate a rule set, was run on the Mushroom Database. All of these approaches are then compared to an optimal rule set described by Schlimmer. Schlimmer found optimal rules for the Mushroom Database 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.

6. Results of Study

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 VotedPerceptron. At the next level of accuracy, the PRISM classifier, which does generate rules, was also able to reflect a high accuracy. 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 very high accuracy on the data 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 minimal rule set described by Schlimmer.

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 95% accuracy in his evaluations on the 8124 records, whereas this study achieved 99.6% accuracy at the lowest in the J48 unpruned tree analysis on the subset of 3000 records. Therefore, as the objective in creating this human-machine interactive mushroom identification application is to warn users about potentially poisonous mushrooms, the higher accuracy result was used in programming the 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 set results was generated.

Data	#	Algorithm	Type	Accuracy
Client Data Sets 3000 rcds 23 attribs	1	IBk	train	100%
	2	IBk	test1	99.5%
	3	IBk	test2	100%
	4	PRISM 22 attribs	train	100%
	5	PRISM 22 attribs	test1	100%
	6	PRISM 22 attribs	test2	97.7%
	7	VP	train	99.5%
	8	VP	test1	95.3%
	9	VP	test2	98.4%
	10	J48	train	100%
	11	J48	test1	99.6%
	12	J48	test2	100%
Schlimmer Data Sets 8124 rcds 23 attribs	1	IBk	strain	100%
	2	IBk	stest1	90.7%
	3	IBk	stest2	38.3%
	4	PRISM 22 attribs	strain	100%
	5	PRISM 22 attribs	stest1	75.6%
	6	PRISM 22 attribs	stest2	52.1%
	7	VP	strain	99.8%
	8	VP	stest1	87.9%
	9	VP	stest2	31.6%
	10	J48	full	100%

Table 1. Summary of Results for Mushroom data sets, VP=VotedPerceptron

Table 1 shows the results of the same tests when run on the client data sets versus when they were run on the Schlimmer data sets. The only tests, which are not comparable, are the J48 tests, which have already been discussed. From this table it is clear that the second test set of the Schlimmer data, containing 2708 records, is either dissimilar to strain and stest1 or it contains a high level of noise. Acceptable levels of accuracy were reached on stest1 in the VotedPerceptron (VP) and PRISM tests. However, the highest level of accuracy reached on the stest1 data was achieved with IBk. While test 10 under the Schlimmer data sets in Table 1 has 100% accuracy, it is not comparable to the other tests because it used Schlimmer's full 8124 records as training data. That test was run in order to illustrate the differing yet somewhat similar J48 unpruned tree that can be achieved from Schlimmer's entire data set.

7. 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 main purpose of the Mushroom Database application is to inform users as to whether their mushroom is edible or poisonous. The application assumes that the observations submitted by the human user are accurate.

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 facilitate actualization of 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.

8. Limitations and Opportunities for Research

A multitude of algorithms and methodologies were used during this study: J48 unpruned tree, IBk, VotedPerceptron, and PRISM.

The 1000 record data sets provided by the client were incompatible with the 2708 records sets from Schlimmer's database. Even though the data sets contained the same numbers of attributes and were drawn from the same source, due to time limitations it was not feasible to pick through the data for minor incompatibilities, which made comparison tests between the client's data and Schlimmer's data impossible. This may also have been a formatting issue. For example, Weka rejected the client data set when it was used for testing if the Schlimmer data set had been loaded for training. Therefore, no comparative tests were run. Future researchers may want to look into the differences between the data sets, especially in light of the higher accuracy achieved on the client data set.

For the PRISM tests, the attribute stalk-root was removed due to a large amount of missing data for that attribute. This caused some inconsistency when it came to comparing accuracy rates from runs that used all 23 attributes. However, stalk-root was not removed from all runs because in some cases it was unnecessary as an acceptable level of accuracy was reached regardless of the missing data. Future researchers might want to look into the reactions of various algorithms and methodologies to missing data. Reviews of studies on noisy data have been included in this study; however, in depth research on the topic of missing and noisy data was outside of the scope of this study.

9. Conclusion

Both high and acceptable levels of accuracy were reached on the first test set from Schlimmer's data for all algorithms. However, the second test set of Schlimmer's data had low accuracy for all algorithms. It may be that the first two thirds of Schlimmer's full database are more similar to one another than they are to the last third.

The highest accuracy that resulted from running the subsets provided by the client 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. Higher accuracy in classifications can be achieved because human observations are used in the classification process. 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 is accessible to casual users and academics alike.

10. References

[1] M. Ankerst, C. Elsen, M. Ester, H. -P. Kriegel, "Visual Classification: An Interactive Approach to Decision Tree Construction", ACM, KDD-99, San Diego, CA, 1999, pp.392-396.

[2] R. Bassett, "Machine assisted visual grading of rare collectibles over the Internet", Proc. Student Research Day, CSIS, Pace University, May 9th, 2003, pp.-12.

[3] A. Berson, S. Smith and K. Thearling, "An Overview of Data Mining Techniques", http://www.thearling.com/text/dmtechniques/dmtechniques.htm accessed November 2007, pp.1-54.

[4] H. - C. Chen, "Forged Handwriting Detection", Proceedings of Student Research Day, CSIS, Pace University, May 9th, 2003, 9.1-9.6.

[5] 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.

[6] E. Hart, S. - H. Cha, and C. C. Tappert, "Interactive Flag Identification", CSIS Technical Reports, Pace University, White Plains, NY, 2004, pp. 1-8.

[7] Md. Z. Islam, P. M. Barnaghi and L. Brankovic. " Measuring Data Quality: Predictive Accuracy vs. Similarity of Decision Trees". Proceedings of the 6th International Conference On Computer And Information Technology (ICCIT2003), December 2003, Dhaka, Bangladesh, 2003, pp. 1-7.

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

[9] J. S. Schlimmer. (1987). Concept Acquisition Through Representational Adjustment (Technical Report 87-19). Doctoral dissertation, Department of Information and Computer Science, University of California, Irvine.
[10] M. Thakore and D. C. St. Clair, "Effect of the χ^2 Test on Construction of ID3 Decision Trees", ACM-SAC '93/2/93/IN, USA, 1993, pp.1-8.
[11] K. Thearling, "Understanding Data Mining: It's All in the Interaction", DS*, December 9, 1997, pp. 1-3.
[12] Witten, I. H. and Frank, E., Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufman Publishers, San Francisco, 2005, p.5.