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

A Data Mining Study of Mouse Movement, Stylometry, and Keystroke Biometric Data

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 extends earlier studies by running previously and newly obtained Mouse Movement, Stylometry, and Keystroke Capture data through various data mining algorithms. All the data sets were analyzed using the k-nearest-neighbor classifier while the Stylometry data set was also analyzed using decison rule and k-means clustering techniques. Various preprocessing techniques and other manipulations were also applied to the data. High identification and authentication classification accuracies were achieved in many cases.

1. Introduction

Raw data is useless without techniques to extract information from it. According to Witten and 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" [9]. 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.

2. Background

The Weka data mining tool was used in this study and this section describes the algorithms and tools used. Decision rule (PredictiveApriori), k-means clustering (simpleKmeans) and k-nearest-neighbor (IBk) algorithms were used against the Stylometry data set (Weka names in parentheses). The k-nearest-neighbor algorithm was also used against the Keystroke Capture and Mouse Movement data sets. Although the choices of these techniques and their implementations are discussed in the methodologies section, some background information on these algorithms is given below.

The k-nearest-neighbor 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.

In order to discuss the PredictiveApriori algorithm, it is necessary first to define two key terms related to the Apriori algorithm. 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 PredictiveApriori, 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. According to the information documents contained in Weka, "Predictive Apriori ... combines the standard confidence and support statistics into a single measure." [5] PredictiveApriori accepts only nominal data although numeric data can be discretized to nominal form.

In the simpleKmeans algorithm, the parameter k represents the number of clusters desired. The classes are not necessarily known before the clustering algorithm is run, and even the overall number of classes need not be known. The output of the clustering algorithm is k clusters, which should correspond to any known classes in terms of instance distribution, but will not be labeled as such. The mapping of classes to clusters must be determined later if desired. Seeds randomize the initial assignment of instances to clusters.

Two tools are also described - percentage split and cross validation. A percentage split uses a percentage of the total records as training data and the remaining

percentage as test data. The user determines the percentage. The percentage split does not affect the number of attributes used. Rather a random selection of a percentage of the full records including all values for those records for all attributes is used as training and the rest are used as test.

Cross validation is a process by which machine learning models can be verified and strengthened as they are built. The number of folds in cross validation is determined by the user. The records are divided into the given number of folds. These folds, or equal partitions of the data, are each in turn used for testing as the rest of the folds are used for training. Given three folds, first folds one and two will be used for training while fold three is used for test. Then the process is repeated with folds one and three as training with fold two for test. Finally, folds two and three are used for training with fold one as test.

The leave-one-out procedure is a special form of cross validation, wherein one record rather than a full fold or subset of records is left out of the training process and later classified during the testing process.

3. Focus of Study

This study extends previous studies by running Weka algorithms against the Mouse Movement, Keystroke Capture, and Stylometry data sets collected during the fall of 2007, and against the Keystroke Capture data set collected prior to fall 2007.

Previous studies conducted by students and faculty of Pace University documented user identification classification of the Mouse Movement, Stylometry, and Keystroke Capture 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 verify the earlier results and to improve the accuracy of classification results. Furthermore, since new data has also been collected, some of these results are on larger data sets.

4. Research Methodology

Authentication and identification experiments were run on the Mouse Movement, Stylometry, and Keystroke Capture data sets. For the Keystroke Capture data set, longitudinal authentication and identification experiments were run on the new data while authentication and identification experiments were run on the old data. Both the authentication and the identification experiments used k-nearest neighbor. The authentication experiments used dichotomy-model data sets, wherein the records used in the identification experiments were preprocessed to classify instance pairs as belonging to either the same or different classes [2, 10].

Additional experiments were run on each of the biometric data sets. Several different methodologies have been used in these additional tests to analyze the biometric data sets. The Stylometry data set has been analyzed using classifiers in an attempt to discover rules and other means of author identification. The Keystroke Capture data set has been analyzed using a nearest neighbor approach with cross validation and percentage splits.

4.1. Mouse Movement

Although the k-nearest neighbor classifier was used in an earlier identification study [8], it has been extended to authentication and idenfitication experiments using larger data sets data collected in the fall of 2007.

4.2. Stylometry

An earlier identification study of email authorship focused on "lexical, syntactic, content, and complexity features" [6]. Here, a larger Stylometry data set is analyzed using the k-nearest neighbor classifier for authentication and identification experiments. Additional tests were run using a discretized PredictiveApriori, a normalized simpleKmeans and a normalized IBk with cross validation and percentage splits.

4.3. Keystroke Capture

The analysis of the Keystroke Capture data set furthers the k-nearest-neighbor approach previously used [7]. Both longitudinal authentication and longitudinal identification tests were run on the new Keystroke Capture data collected in the fall of 2007. Authentication and identification experiments were also run on the old Keystroke Capture data set, collected previous to the fall of 2007, for 36 subjects who each submitted records for the copy task and free text task on a desktop and a laptop; these were the four quadrants of the previous experiment. The new Keystroke Capture data sets also used these four quadrants; however, in this study, these data sets were analyzed using longitudinal authentication and identification experiments.

5. Experimental Results

5.1. Mouse Movement Data Set

The Mouse Movement data set contains 205 records, 30 records from each of five subjects, 15 records for one subject, and 10 records for each of four subjects. These data were collected in the fall of 2007 and tested in identification experiments [1].

Authentication tests were run on the Mouse Movement dichotomy data [2] using IBk with k=1. Table 1 shows the results of the Mouse Movement authentication experiments on the dichotomy data.

Train	Test	Accuracy
First 5 Subjects 1000 records	**Last 5 Subjects 1000 records**	56.5%
Last 5 Subjects 1000 records	**First 5 Subjects 1000 records**	56.5%
First 5 Subjects 6555 records	**Last 5 Subjects 4005 records**	66.74%
Last 5 Subjects 4005 records	**First 5 Subjects 6555 records**	68.62%

Table 1: Results of Authentication Experiments on the Mouse Movement data.

In addition to the authentication tests on the dichotomy data, an identification test was also run on the new Mouse Movement normalized data set of 205 records.

Table 2 presents the results of the identification experiment, which was run using IBk with k=1. The test used a full data set for training and a full data set for testing. Table 2 shows that 93% accuracy was achieved on the full data set.

Train	Test	Accuracy
Full (205 samples from 10 subjects)	**Full (205 samples from 10 subjects)**	93%

Table 2: Results of Identification Experiment on the Mouse Movement data.

5.2. Stylometry Data Set

The Stylometry data set contains 120 records, ten records from each of twelve subjects [4]. Authentication and identification experiments were run on the data.

The dichotomy data for the Stylometry authentication experiments contains 1770 records for each subset of six subjects [2]. Each subset was run against the other yielding the results in Table 3.

Train	Test	Accuracy
First 6 Subjects (1770 records)	**Last 6 Subjects (1770 records)**	76.89%
Last 6 Subjects (1770 records)	**First 6 Subjects (1770 records)**	66.89%

Table 3: Results of Authentication Experiments on the Stylometry data.

Table 4 presents the results of the identification experiments, which used IBk with k=1. The first test used a full data set for training and a full data set for testing. The second test used the first five records from each of twelve subjects for a total of sixty records as training and the last five records from each of twelve subjects for a total of sixty records as test. The third test used the last five records from each of twelve subjects for a total of sixty records as training and the first five records from each of twelve subjects for a total of sixty records as test.

Train	Test	Accuracy
Full leave one out (10 samples from each of 12 subjects)	**Full leave one out (10 samples from each of 12 subjects)**	26.67%
Full (10 samples from each of 12 subjects)	**Full (10 samples from each of 12 subjects)**	97.5%
First 5 (5 samples from each of 12 subjects)	**Last 5 (5 samples from each of 12 subjects)**	21.67%
Last 5 (5 samples from each of 12 subjects)	**First 5 (5 samples from each of 12 subjects)**	23.33%

Table 4: Results of Identification Experiments on the Stylometry data.

Additional tests were also run on the Stylometry data. Two additional algorithms were used to analyze the Stylometry data. PredictiveApriori and SimpleKmeans were used in addition to IBk. For these tests, various preprocessing techniques and various manipulations of the settings specific to each algorithm were applied.

5.2.1. PredictiveApriori. The PredictiveApriori algorithm is geared towards finding association rules. For these runs, the numeric attributes were normalized and discretized while all nominal attributes were removed with the exception of author name, which was selected as the predicted class. The other nominal attributes were removed because their values in some cases were too specific to a particular class and therefore, the rules that resulted from running the data with all nominal attributes included were over fit. All 120 instances from this data set were used.

The rules generated used class association rules (car) and therefore predicted the user name class such that user name was on the right hand side of every rule.

The accuracy of the first four rules was the same 89% for all four. The accuracy of rules 5 though 10 ranged between 86% and 81%.

PredictiveApriori made an interesting experiment because of the potential for using the predictive rules that were found to determine the identity of the author. However, PredictiveApriori rules are predictive rather than conclusive. The algorithm analyzes the current data to find rules that predict future values.

5.2.2. SimpleKmeans. The clustering algorithm, simpleKmeans, was used. As there were seven subjects, or authors, in this Stylometry data set, with ten records per subject, all clustering tests used simpleKmeans with 12 clusters and the classes to clusters evaluation. The random seed was varied from test to test for ten tests.

In analyzing the results of simpleKmeans, a discussion of the highest and lowest accuracies is irrelevant. Rather, the average accuracy achieved is a more accurate prediction of the potential for this algorithm to correctly classify instances. As the seeds are used to randomize the initial assignment of instances to clusters, they do not generate an overall pattern. That is, it is not apparent what level of accuracy any particular seed will yield. Ten tests were run with random seeds. The tests resulted in an average accuracy of 32.17%.

5.3. Keystroke Capture Data Set

The three new Keystroke Capture data sets each contain five records from each of four subjects and were collected at two week intervals (259 attrributes). The three new Keystroke Capture data sets were each collected two weeks apart from the same subjects.

The old data sets contained approximately 5 records from each of 36 subjects (256 attributes).

Each of the data sets (new and old) contained four subsets of data: copy task data on a desktop, copy task data on a laptop, free text data on a desktop, and free text data on a laptop. Table 5 shows the number of instances per subset of the old Keystroke Capture data.

Data Set	TASK	Computer Type	Instance Count
Identification	Copy (36 subjects)	Desktop	180
		Laptop	180
	Free (36 subjects)	Desktop	176
		Laptop	180
Authentication	Copy (36 subjects)	Desktop	860
		Laptop	860
	Free (36 subjects)	Desktop	841
		Laptop	860

Table 5. Specifications for old Keystroke Capture data sets

Eight authentication experiments were run on the three new Keystroke Capture dichotomy-model data sets using IBk with k=1. Each data set contained 190 instances and the original 256 attributes [2]. In these experiments, the full first data set was used for training and the full second and third data sets recorded at different collection times were used for testing. Table 6 shows the accuracy results, indicating that accuracy is maintained for data input at later times, at least for later times of two and four weeks for this small sample of subjects.

Train	Test	Type	Accuracy
1 (5 samples from each of 4 subjects)	**2 (5 samples from each of 4 subjects)**	Copy Desk	95.79%
		Free Desk	96.32%
		Copy Lap	91.58%
		Free Lap	92.11%
1 (5 samples from each of 4 subjects)	**3 (5 samples from each of 4 subjects)**	Copy Desk	88.95%
		Free Desk	98.42%
		Copy Lap	100.00%
		Free Lap	93.68%

Table 6: Results of Longitudinal Authentication Experiments on the Keystroke Capture data.

The new data sets were also compared in longitudinal identification experiments, using IBk with k=1. As with the authentication experiments, the full first data set was used for training and the full second and third data sets were used for testing. Each of the four subsets was tested individually against its counterpart in the training set. Table 7 shows the accuracy results, again indicating that accuracy is maintained for data input at later times.

Train	Test	Type	Accuracy
1 (5 samples from each of 4 subjects)	**2 (5 samples from each of 4 subjects)**	Copy Desk	95%
		Free Desk	100%
		Copy Lap	100%
		Free Lap	85%
1 (5 samples from each of 4 subjects)	**3 (5 samples from each of 4 subjects)**	Copy Desk	80%
		Free Desk	100%
		Copy Lap	100%
		Free Lap	100%

Table 7: Results of Longitudinal Identification Experiments on the Keystroke Capture data.

Sixteen authentication experiments were run on the old Keystroke Capture dichotomy data using IBk with k=1. These data sets came from a four quadrant experiment in which 36 subjects submitted 5 records each per quadrant. These tests are arranged according to the previous identification analysis of the data [7]. Table 8 shows the results of these authentication experiments, and in many cases high accuracies were achieved.

Experiment	Train	Test	Accuracy
Train (18 subjects) **Test** (18 subjects)	Desk Copy	Desk Copy	87.94%
	Desk Free	Desk Free	90.24%
	Lap Copy	Lap Copy	91.03%
	Lap Free	Lap Free	92.06%
Copy Task (36 subjects)	Desktop	Laptop	93.55%
	Laptop	Desktop	87.21%
Free Text (36 subjects)	Desktop	Laptop	77.44%
	Laptop	Desktop	91.62%
Desktop (36 subjects)	Copy	Free Text	88.08%
	Free Text	Copy	88.49%
Laptop (36 subjects)	Copy	Free Text	72.33%
	Free Text	Copy	95.81%
Different Mode/ Keyboard (36 subjects)	Desk Copy	Lap Free	81.40%
	Lap Free	Desk Copy	92.33%
Different Keyboard/ Mode (36 subjects)	Lap Copy	Desk Free	83.59%
	Desk Free	Lap Copy	91.51%

Table 8: Results of Authentication Experiments on the old Keystroke Capture data.

Similarly, twelve identification experiments were also run on the old Keystroke Capture data set using IBk with k=1.

Table 9 presents the results of these experiments showing a wide range of recognition accuracies.

Experiment	Train	Test	Accuracy
Copy Task (36 subjects)	Desktop	Laptop	83.34%
	Laptop	Desktop	51.67%
Free Text (36 subjects)	Desktop	Laptop	40.56%
	Laptop	Desktop	52.84%
Desktop (36 subjects)	Copy	Free Text	47.72%
	Free Text	Copy	51.11%
Laptop (36 subjects)	Copy	Free Text	18.89%
	Free Text	Copy	57.78%
Different Mode/ Keyboard (36 subjects)	Desk Copy	Lap Free	31.67%
	Lap Free	Desk Copy	55.56%
Different Keyboard/ Mode (36 subjects)	Lap Copy	Desk Free	38.07%
	Desk Free	Lap Copy	54.45%

Table 9: Results of Identification Experiments on the old Keystroke Capture data.

Eight additional identification tests were run on the old data, again using IBk with k=1, but using the processing techniques of an 80% split and cross validation. Other values for Knn were also attempted, but without an improvement.

Each of the following tests was run against each of the four subsets of the additional data set: cross validation with 180 folds (leave-one-out procedure) and an 80% split of training and test data.

Table 10 details the findings of the eight tests. The cross validation and 80% split tests achieved very high accuracy results. In the case of the 80% split, this may be partly because so much of the data is used as training data. However, in the case of cross validation, the classification accuracy increases because the classifier is strengthened by the cross validation technique as the classification model is being built.

Task	Method	Data set	Accuracy
Copy Task (36 Subjects)	Cross Val 180	Desktop	97.78%
	80% Split	Desktop	100.00%
	Cross Val 180	Laptop	96.11%
	80% Split	Laptop	100.00%
Free Text (36 Subjects)	Cross Val 180	Desktop	94.89%
	80% Split	Desktop	94.44%
	Cross Val 180	Laptop	98.33%
	80% Split	Laptop	97.22%

Table 10. Results of Additional Idenfitication Experiments on the old Keystroke Capture data.

6. Limitations and Opportunities for Future Research

This study analyzes Mouse Movement, Stylometry, and Keystroke Capture data sets using data mining techniques. Numerous algorithms and methodologies have been used during the course of the project. All data sets were analyzed using IBk, while the Stylometry data set was also analyzed using PredictiveApriori and simpleKmeans. All of these machine-learning methods have differences in applicability, meaning there is no one best method; rather, there are only optimal methods, depending on the particular data set.

It is important to note that most of the algorithms involved in the project do not produce 100% accuracy. For example, although highly accurate results were obtained using sophisticated learning methods on many datasets, some approaches were more successful than others.

The most successful approaches have been shown in detail and future researchers may find that they can improve the results found in this study using similar techniques.

Future researchers may be particularly interested to try a different approach to the authentication experiments. In these experiments, a community of subjects was authenticated against another community of subjects. However, for use of the biometric information to identify an individual, it would be more efficient to attempt authentication based solely on the subject in question. Experiments that would lead to an adequate system for identifying idividuals would require splitting the data set into separate data sets that hold only the within and between class records pertaining to each of the subjects.

7. Conclusion

Authentication and identification tests were run on Mouse Movement, Stylometry, and Keystroke Biometric data. In most cases, the identification tests resulted in higher accuracies than the authentication tests. Recommendations have been made for legitimately increasing the classification accuracy of the authentication experiments.

Additional tests were also run on all of the data sets. The k-nearest neighbor approach was used with cross validation and 80% splits on the Stylometry and Keystroke Capture data sets, which showed high accuracy results for non-training data. Decision rule and k-means clustering were used on the Stylometry data set and made for interesting experiments, in that the decision rules may be useful to future researchers and in that the clustering algorithm shows an average accuracy that is similar to the accuracies achieved using k-nearest neighbor with cross validation on the modified Stylometry data.

This study has extended previous studies by running additional experiments on the Mouse Movement, Stylometry, and Keystroke Biometric data, new and previously obtained, using the data mining tool Weka. The data mining algorithms with which the experiments were conducted are widely used and provide an entry point for future researchers into the use of data mining with biometric data sets.

8. References

[1] N. Ajufor, A. Amalraj, R. Diaz, M. Islam, M. Lampe, "Refinement of a Mouse Movement Biometric System," Proc. CSIS Research Day, Pace Univ., May 2008.

[2] S. Bharati, R. Haseem, R. Khan, M. Ritzmann and A. Wong, "Biometric Authentication System using the Dichotomy Model," Proc. CSIS Research Day, Pace Univ., May 2008.

[3] T. Buch, A. Cotoranu, E. Jeskey, F. Tihon, and M. Villani, "An Enhanced Keystroke Biometric System and Associated Studies," Proc. Seidenberg Research Day, Pace Univ., May 2008.

[4] K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, and S. Westcott, "Stylometry for E-mail Author Identification and Authentication," Proc. CSIS Research Day, Pace Univ., May 2008.

[5] E. Frank, M. Hall, G. Holmes, R. Kirkby, F. Pfahringer, I. H. Wittne, L. Trigg, "WEKA A Machine Learning Workbench for Data Mining", http://www.cs.waikato.ac.nz/~eibe/pubs/weka_dmh.ps.gz, Accessed November 2007.

[6] 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.

[7] M. Villani, C.C. Tappert, G. Ngo, J. Simone, H. St. Fort, and S. Cha, "Keystroke Biometric Recognition Studies on Long-Text Input under Ideal and Application-Oriented Conditions," Proc. *CVPR 2006 Workshop on Biometrics*, New York, NY, June 2006.

[8] 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.

[9] Witten, I. H. and Frank, E., Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufman Publishers, San Francisco, 2005, p.5.

[10] S.Yoon, S. - S. Choi, S. -Y. Cha, Yillbyung Lee, C. C. Tappert, "On the Individuality of the Iris Biometric", ICGST-GVIP Journal, Volume (5), Issue (5), May 2005, pp.63-70.