# An investigation of genetic algorithm-based feature selection techniques applied to keystroke dynamics biometrics

**Tuany Mariah Lima do Nascimento, Andrelyne Vitória Monteiro de Oliveira, Márjory Da Costa Abreu, Laura Emmanuella Alves dos Santos Santana**

[1]Universidade Federal do Rio Grande do Norte
marjory@dimap.ufrn.br, lauraemmanuella@gmail.com

***Abstract.*** *Due to the continuous use of social networks, users can be vulnerable to online situations such as paedophilia treats. One of the ways to do the investigation of an alleged pedophile is to verify the legitimacy of the genre that it claims. One possible technique to adopt is keystroke dynamics analysis. However, this technique can extract many attributes, causing a negative impact on the accuracy of the classifier due to the presence of redundant and irrelevant attributes. Thus, this work using the wrapper approach in features selection using genetic algorithms and as KNN, SVM and Naive Bayes classifiers. Bringing as best result the SVM classifier with 90% accuracy, identifying what is most suitable for both bases.*

## 1. Introduction

The use of social networks grew enormously, serving to connect and share information among different groups of people. However, users can be vulnerable to risky situations, paedophilia is one of them. One of the ways to do the investigation of an alleged pedophile is to validate whether the individual is, in fact, the genre that it is said to be. For this, a possible technique to be adopted is the keystroke dynamics [Fairhurst and Da Costa-Abreu 2011]. The keystroke dynamics is a behavioral biometric which consists of analyzing the way the user types in a terminal, monitoring the keyboard to identify the user based on his/her usual typing rhythm patterns. Keystroke dynamics can be extracted through fixed text or free texts, that is, texts that have been predetermined for all individuals under observation and texts that have periodic monitoring of the keystrokes, respectively . This technique is accessible since external hardware is not required to collect data from the keyboard only [Darabseh and Namin 2015].

Studies have shown that keystroke dynamics is a viable technique for gender recognition such as [Tsimperidis et al. 2018, Antal and Nemes 2016], however, this technique can extract many attributes and dealing with a large amount of attributes can have negative impact. In an intuitive way, the greater the number of attributes in a database, the easier it is to extract knowledge models, however, in practice this may not to be true, since there may be irrelevant or redundant attributes and the presence of noise it can confuse the classifier as well as impair its accuracy, the computational cost increases exponentially with a very large number of attributes, making it difficult to construct the model [Faceli et al. 2011].

Features selection is considered one of the most important steps in the process of data mining, machine learning and pattern recognition, since it aims to reduce dimensionality by selecting a subset features relevant to the build of a model, because the presence of

irrelevant or redundant features can affect the performance of the classification algorithm [Santana 2012]. Feature selection can be done using a filter or wrapper-based approach which will be explored more later in this work.

Thus, the present work aims to perform a comparative analysis between features selection techniques for the gender recognition on the databases of keystroke dynamics using a wrapper approach with genetic algorithms.

## 2. Background

In this session we will address some basic concepts in order to establish a theoretical basis on the present work.

### 2.1. Genetic Algorithms

Genetic algorithms are a search technique and optimization based on the Darwinian principle of natural selection proposed in the book, The Origin of Species in 1859. This algorithm can be used to features selection of a database from a wrapper approach and was used in this work. Its development is based on biological mechanisms as heredity and evolution. The algorithm is a population based method and a chromosome is used to represent each solution to the given problem [Santana and Canuto 2014].

The operations of the genetic algorithm work by randomly generating an initial population, which will be the first generation of all, then each individual of this population is evaluated using an objective function or fitness function, which is used to define how close the individual is of the optimal solution. From the evaluation carried out, the fittest individuals are selected to continue in the genetic process, they are placed in a temporary population which can be called parents and they are responsible for the next generation. This loop involving evaluation, selection and genetic operators is repeated until reach ending condition [Santana and Canuto 2014]. Thus, the initial population was 30 individuals and randomly chosen. Parent selection was done through binary tournament and the genetic operators used were mutation and single point crossing using rate equal to 0.5 and 0.9. Moreover, elitism, which is the choice of the best individual to be passed on to the next generation, was also used in this work. The parameters of the genetic algorithm were chosen empirically.

## 3. Related Work

Gender recognition can be applied in biometric identification systems based on the recognition of speech, face, iris. In some cases, gender classification across the face is performed earlier.

In the work of [Antal and Nemes 2016], two databases were used in their experiments. For the features selection performed in this work, the authors used the Weka data mining software [Hall et al. 2009] making use of the filter approach, as correlation evaluation method was used. As classifiers, Random Forest was used to evaluate the data sets. The authors used two types of cross-validation, 10-fold common cross-validation, and leave-one-user-out cross validation, respectively, to evaluate the accuracy of the methods. The classification measures were performed in both databases and then after using the subset with five remaining attributes during the selection process using the filter approach. The results were promising with cross-validation in both sets of data.

In works such as [Tsimperidis et al. 2018], the authors preferred in their work the use of free text due to better integration with the volunteers' regular typing activities and is less intrusive. About 10,000 known and unknown attributes were collected. For the select features performed in the work, the authors used the filter approach, using information measures. Several experiments were performed using five classifiers, they are: SVM, MLP, NB, RF and RBFN The RBFN model correctly predicts the gender of the unknown user, reaching an accuracy of 95.6%.

## 4. Methodology

Since the papers presented previously point to the improvement in the performance of classification systems from the features selection, and that it is feasible to recognition gender from databases of keystroke dynamics [Kawamura and Chakraborty 2017], this work will allow the identification of which features selection techniques may result in smaller subsets of attributes with increased system accuracy and be reliable to be used in security applications.

The experiment was conducted in three stages using the databases, Brazilian hand-based behavioral biometrics [Da Silva et al. 2016] and GREYC [Giot et al. 2009] Table 1. The first step was the features selection with the wrapper approach using genetic algorithms for the KNN, SVM and Naive Bayes classifiers. All experiments were performed 30 times, for the third and final step to calculate the mean and standard deviation in order to perform statistical tests with T-test.

| | Volunteers | Males | Females | Features | Instances |
|---|---|---|---|---|---|
| [Da Silva et al. 2016] | 77 | 20 | 50 | 42 | 231 |
| [Giot et al. 2009] | 133 | 98 | 35 | 60 | 7555 |

**Table 1. Database Details**

In order to perform a comparative analysis between the attribute selection techniques, we used three classifiers, which were from the standard implementation of the Weka toolbox [Hall et al. 2009]. These are $K$-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Naive Bases, as well as the genetic algorithm, the choice of machine learning algorithms parameters, were chosen after several tests were performed. They were, for KNN (K = 11, K = 3 and Euclidean distance, for Brazilian hand-based behavioral biometrics and GREYC database, respectively), SVM (C was set to 10 and 100 and Kernel was Puk, for base data Brazilian hand-based behavioral biometrics and GREYC, respectively) and finally the default configuration of Naive Bayes was used. These classifiers were chosen because they belong to different learning paradigms and can bring different aspects to the comparative analysis to be performed.

## 5. Results and Conclusion

As mentioned previously, our experiments were conducted using two databases, Brazilian hand-based behavioral biometrics and GREYC, having 43 and 60 attributes and 231 and 7555 instances, respectively. In both databases, data were separated by 66% for training and 33% for testing with the KNN, SVM and Naive classifiers.

According to the results presented in Table 2, it can be observed that Naive presented the least accuracy for the two databases, which can be justified by a possible correlation between the attributes of the base since this model assumes the independence

| | No Selection | | |
|---|---|---|---|
| **Database** | **KNN** | **SVM** | **Naive Bayes** |
| [Da Silva et al. 2016] | 72.73 $\pm 2.74(42)$ | 73.55 $\pm 2.63(42)$ | 55.54$\pm 3.54(42)$ |
| [Giot et al. 2009] | 86.72 $\pm 0.58(60)$ | 78.28 $\pm 0.67(60)$ | 69.33 $\pm 0.67(60)$ |
| | **Wrapper Approach** | | |
| [Da Silva et al. 2016] | 87.85$\pm 0.84(15)$, | 88.10 $\pm 0.90(31)$ | 81.64 $\pm 0.97(10)$ |
| [Giot et al. 2009] | 88.86 $\pm 0.23(42)$ | 90.05$\pm 0.41(45)$ | 77.44 $\pm 0.27(12)$ |

**Table 2. Accuracy (Acc) rates in percentage and total of selected features (SF)**

between attributes. The KNN obtained higher accuracy for both bases, and for the Brazilian hand-based behavioral biometrics base the SVM got a bit better adjusted. It can be observed that there was an increase in accuracy in both bases for all classifiers when compared to the results with the complete base. The SVM classifier was the most adequate in both bases with an increase in accuracy of 14.55% and 11.77% even though it did not remove many attributes such as Naive Bayes.

The process of selecting attributes, is one of the most important steps in the process of data mining, pattern recognition and machine learning and with it databases containing many attributes, may require a high computational power. Thus, the selection of a subset of features that contains relevance to the class may increase the accuracy of the classification. The use of genetic algorithms seems to be better, although it has an increase in computational cost, but may show us better results in terms of accuracy as in a better subset. Where our results showed that the wrapper approach is better compared to no selection and statistically proven. The best classifier for both databases that resulted in greater accuracy to classify a user's gender was the SVM with 90% accuracy

## References


Antal, M. and Nemes, G. (2016). Gender recognition from mobile biometric data. In *2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 243–248.

Da Silva, V. R., De Araujo Silva, J. C. G., and Da Costa-Abreu, M. (2016). A new brazilian hand-based behavioural biometrics database: data collection and analysis. In *7th International Conference on Imaging for Crime Detection and Prevention (ICDP 2016)*, pages 1–6.

Darabseh, A. and Namin, A. S. (2015). Effective user authentications using keystroke dynamics based on feature selections. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 307–312.

Faceli, K., Lorena, A. C., J.Gama, and Carvalho, A. C. P. L. F. (2011). *Artificial Intelligence: A Machine Learning Approach*. LTC, Rio de Janeiro.

Fairhurst, M. and Da Costa-Abreu, M. (2011). Using keystroke dynamics for gender identification in social network environment. In *4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*, pages 1–6.

Giot, R., El-Abed, M., and Rosenberger, C. (2009). Greyc keystroke: A benchmark for keystroke dynamics biometric systems. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Kawamura, A. and Chakraborty, B. (2017). A hybrid approach for optimal feature subset selection with evolutionary algorithms. In *2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, pages 564–568.

Santana, L. E. A. S. (2012). *Optimization classifiers committees: An approach based on filter for selecting subsets of attributes*. PhD thesis, Universidade Federal do Rio Grande do Norte, Natal.

Santana, L. E. A. S. and Canuto, A. M. P. (2014). Filter-based optimization techniques for selection of feature subsets in ensemble systems. *Expert Systems with Applications*, 41(4, Part 2):1622 – 1631.

Tsimperidis, I., Arampatzis, A., and Karakos, A. (2018). Keystroke dynamics features for gender recognition. *Digital Investigation*, 24:4 – 10.