Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 38 (2011) 4375-4379

Awakening to reality

Elixir ISSN: 2229-712X

Bagged ensemble of genetic algorithm for signature verification

M.Govindarajan and RM.Chandrasekaran

Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar-608002, Tamil Nadu, India.

ARTICLE INFO

Article history:

Received: 22 August 2011;
Received in revised form:
26 August 2011;
Accepted: 31 August 2011;

Keywords

Classification rate,
Data Mining,
Ensemble Method,
Genetic Algorithm,
Online Signature Verification.

ABSTRACT

Data Mining is the use of algorithms to extract the information and patterns derived by the knowledge discovery in databases process. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. The Verification of handwritten Signature, which is a behavioral biometric, can be classified into off-line and online signature verification methods. The feasibility and the benefits of the proposed approach are demonstrated by means of data mining problem: online Signature Verification. This paper addresses using ensemble approach of Genetic Algorithm for online Signature Verification. Online signature verification, in general, gives a higher verification rate than off-line verification methods, because of its use of both static and dynamic features of problem space in contrast to off-line which uses only the static features. We show that proposed ensemble of Genetic Algorithm is superior to individual approach for Signature Verification in terms of classification rate.

Introduction

Security is one of the major issues in today's world and most of us have to deal with some sort of passwords in our daily lives; but, these passwords have some problems of their own. If one picks an easy-to-remember password, then it is most likely that somebody else may guess it. On other hand, if one chooses too difficult a password, then he or she may have to write it somewhere (to avoid inconveniences due to forgotten passwords) which may again lead to security breaches. To prevent passwords being hacked, users are usually advised to keep changing their passwords frequently and are also asked not to keep them too trivial at the same time. All these inconveniences led to the birth of the biometric field. The verification of handwritten signature [6], which is a behavioral biometric, can be classified into off-line and online signature verification methods.

Online signature verification, in general, gives a higher verification rate than off-line verification methods, because of its use of both static and dynamic features of problem space in contrast to off-line which uses only the static features. Despite greater accuracy, online signature recognition is not that prevalent in comparison to other biometrics.

Data Mining [1] has become a very useful technique to reduce information overload and improve decision making by extracting and refining useful knowledge through a process of searching for relationships and patterns from the extensive data collected by organization. The extracted information is used to predict, classify, model, and summarize the data being mined. Data mining technologies, such as rule induction, neural networks, genetic algorithms, fuzzy logic, and rough sets are used for classification and pattern recognition in many industries. The primary objective of this paper is ensemble of Genetic Algorithm is superior to individual approach for signature verification in terms of classification rate.

Signature Verification Methods

The most commonly used protection mechanisms today are based on either what a person possesses (e.g. an ID card) or what the person remembers (like passwords and PIN numbers). However, there is always a risk of passwords being cracked by unauthenticated users and ID cards being stolen, in addition to shortcomings like forgotten passwords and lost ID cards.

To avoid such inconveniences, one may opt for the new methodology of Biometrics [7], which though expensive will be almost infallible as it uses some unique physiological and/or behavioral characteristics possessed by an individual for identity verification.

Examples include signature, iris, face, and fingerprint recognition based systems.

Forgeries can be classified into four types-random, simple, skilled and traced. Generally online signature verification methods display high accuracy rates (closer to 99%) than off-line methods (90-95%) in the case of all the forgeries. This is because in off-line verification methods, the forger has to copy only the shape of the signature [9].

On the other hand , in the case of online verification methods, since the hardware used captures the dynamic features of the signature as well, the forger has to not only copy the shape of the signature but also the temporal characteristics (pen tilt, pressure applied, signing velocity etc.) of the person whose signature is to be forged.

In addition, he has to simultaneously hide his own inherent style of writing the signature, thus making it extremely difficult to deceive the device in the case of online signature verification.

Online verification methods [10] can have an accuracy rate of as high as 99%.

The reason behind is its use of both static and dynamic (or temporal) features, in comparison to the off-line, which uses only the static features.

The major differences between off-line and online verification methods do not lie with only the feature extraction phases and accuracy rates, but also in the modes of data acquisition, preprocessing and verification/recognition phases, though the basic sequence of tasks in an online verification (or recognition) procedure is exactly the same that of the off-line.

Tele:
E-mail addresses: govind_aucse@yahoo.com

Genetic Algorithm (GA)

In this section we present a brief introduction about genetic algorithm. A more detailed introduction can be found in [8].

The genetic algorithm is a model of machine learning which derives its behaviour from a metaphor of some of the mechanisms of evolution in nature. This done by the creation within a machine of a population of individuals represented by chromosomes, in essence a set of character strings.

The individuals represent candidate solutions to the optimization problem being solved. In genetic algorithms, the individuals are typically represented by n-bit binary vectors. The resulting search space corresponds to an n–dimensional boolean space. It is assumed that the quality of each candidate solution can be evaluated using a fitness function.

Genetic algorithms use some form of fitness-dependent probabilistic selection of individuals from the current population to produce individuals for the next generation. The selected individuals are submitted to the action of genetic operators to obtain new individuals that constitute the next generation. Mutation and crossover are two of the most commonly used operators that are used with genetic algorithms that represent individuals as binary strings. Mutation operates on a single string and generally changes a bit at random while crossover operates on two parent strings to produce two off springs. Other genetic representations require the use of appropriate genetic operators.

The process of fitness-dependent selection and application of genetic operators to generate successive generations of individuals is repeated many times until a satisfactory solution is found. In practice, the performance of genetic algorithm depends on a number of factors including: the choice of genetic representation and operators, the fitness function, the details of the fitness-dependent selection procedure, and the various user-determined parameters such as population size, probability of application of different genetic operators, etc. The basic operation of the genetic algorithm is outlined as follows:

Procedure:

```
begin
t <- 0
initialize P(t)
while (not termination condition)
t <- t + 1
select P(t) from p(t - 1)
crossover P(t)
mutate P(t)
evaluate P(t)
  end
end
```

Since genetic algorithms were designed to efficiently search large spaces, they have been used for a number of different application areas such as camera calibration [11], signature verification [13], medical diagnosis [14], facial modeling [12] and handwritten recognition [15].

Bagging versus Genetic Algorithm Classifier Ensemble

Bagging and boosting [2] are two such techniques. They are examples of ensemble methods, or methods that use a combination of models.

Each combines a series of k learned models (classifiers or predictors), M_1, M_2,…,M_K, with the aim of creating an improved composite model, M*. Both bagging and boosting can be used for classification as well as prediction.

Bagging Classifiers

Bagging [3] is a "bootstrap" ensemble method that creates individuals for its ensemble by training each classifier on a random redistribution of the training set. Each classifier's training set is generated by randomly drawing, with replacement, t, N examples – where N is the size of the original training set; many of the original examples may be repeated in the resulting training set while others may be left out. Each individual classifier in the ensemble is generated with a different random sampling of the training set.

Given a set, D, of d tuples, bagging works as follows. For iteration I (i= 1,2,…k), a training set , D_i, of tuples is sampled with replacement from the original set of tuples, D. Note that the term bagging stands for bootstrap aggregation [4]. Each training set is a bootstrap sample. Because sampling with replacement is used, some of the original tuples of D may not be included in Di, whereas others occur more than once. A classifier model, M_i, is learned for each training set, D_i. To classify an unknown tuple, X, each classifier, M_i, returns its class prediction, which counts as one vote. The bagged classifier, M*, counts the votes and assigns the class with the most votes to X. Bagging can be applied to the prediction of continuous values by taking the average value of each prediction for a given test tuple.

The bagged classifier often has significantly greater accuracy than a single classifier derived from D, the original training data. It will not be considerably worse and is more robust to the effects of noisy data. The increased accuracy occurs because the composite model reduces the variance of the individual classifiers. For prediction, it was theoretically proven that a bagged predictor will always have improved accuracy over a single predictor derived from D.

Algorithm: Bagging

The bagging algorithm creates an ensemble of models (classifiers or predictors) for a learning scheme where each model gives an equally weighted prediction.

Input:

D, a set of d training tuples;

K, the number of models in the ensemble;

A learning scheme (eg., decision tree algorithm, back propagation, etc.)

Output: A composie model, M*.

Method:

```
(1) for i = 1 to k do // create k models:
(2) create bootstrap sample, $D_i$, by sampling D with replacement;
(3) use $D_i$ to derive a model, $M_i$;
(4) endfor
To use the composite model on a tuple, X:
(1) if classification then
(2) let each of the k models classify X and return the majority vote;
(3) if prediction then
(4) let each of the k models predict a value for X and return the average predicted value;
```

Several researchers have investigated the combination of different classifiers to from an ensemble classifier. An important advantage for combining redundant and complementary classifiers is to increase robustness, accuracy, and better overall generalization. In this approach we first constructed the Genetic Algorithm and applied 10-fold cross validation technique and evaluated error rate from the mean square error. Secondly, bagging is performed with Genetic Algorithm to obtain a very good generalization performance. We show that proposed ensemble of Genetic Algorithm is superior to individual approach for signature verification in terms of classification rate.

Methodology

Signature Verification dataset

The Source of the data is the raw measurement from a Nintendo Power Glove. It was interfaced through a Power Glove Serial Interface to a Silicon Graphics 4D/35G workstation. The glove definitely falls into the category of "cheap and nasty". Position information is calculated on the basis of ultrasound emissions from emitters the glove to a 3-microphone "L-Bar" that sits atop a monitor. There are two emitters on the glove: and three receivers. This allows the calculation of 4 pieces of information: x (left/right), y (up/down), z (backward/forward), and roll (is the palm pointing up or down?).x,y and z are measured with 8 bit accuracy. "x,y,z" should not be taken to be the normal 3-dimensional orthogonal basis. In particular, 1 unit in the z direction is not of similar distance to 1 unit in the x or y directions. These x,y,z positions are relative to a calibration point which is when the palm is resting on the seated signer's thigh. Roll 4 is 4 bits. The data is susceptible to occasional "spikes" caused by random ultrasound noise. Median filters have been found to be beneficial in solving this problem. Finger bend is generated by conductive bend sensors on the first four fingers. Values vary between 0 (Straight) and 3 (full bent). Accuracy is 2 bits. The gloves automatically apply a hysteresis filter on these bend sensors, At best, these measurements should be treated skeptically.

Representation and Operators

In this subsection we present the choice of a representation for encoding candidate solutions to be manipulated by the genetic algorithm.

Each individual in the population represents a candidate solution to the feature subset selection problem. Let *m* be the total number of features available to choose from to represent the patterns to be classifier.

The individual (chromosome) is represented by a binary vector of dimension *m*. If a bit is a 1, it means that the corresponding feature is selected, otherwise the feature is not selected. This is the simplest and most straightforward representation scheme [17].

As mentioned before, other genetic representations require the use of appropriate genetic operators.

Since we are representing a chromosome through a binary string, the operators mutation and crossover operates in the following way: Mutation operates on a single string and generally changes a bit at random.

Thus, a string 11010 may, as a consequence of random mutation get changed to 11110.

Crossover on two parent strings to produce two offsprings. With a randomly chosen crossover position 4, the two strings 01101 and 11000 yield the offspring 01100 and 11001 as a result of crossover.

Parameter Settings

Our experiments used the following parameter settings:

Population size: 100

Number of generation: 20

Probability of crossover: 0.9

Probability of mutation: 0.07

The parameter settings were based on results of several preliminary runs. They are comparable to the typical values reported in the literature [16].

Feature Ensemble Selection

The main idea of ensemble methodology is to combine a set of models, each of which solves the same original task, in order to obtain a better composite global model, with more accurate and reliable estimates or decisions that can be made from using a single model. Some of the drawbacks of the filters and wrappers can be solved by using ensemble [18]. As mentioned above filters perform less than wrappers. Due to the voting process, noisy results are filtered. Secondly, the drawback of wrappers which cost computing time is solved by operating bunch of filters.

Objective Function and Fitness Evaluation

The fitness evaluation is a mechanism used to determine the confidence level of the optimized solutions to the problem. Usually, there is a fitness value associated with each chromosome, e.g., in a minimization problem, a lower fitness value means that the chromosome or solution is more optimized to the problem while a higher value of fitness indicates a less optimized chromosome. Our problem consists of optimizing two objectives: Minimization of the error rate and there by maximizing the classification rate of the classifier.

Experiments

Experiments Using Genetic Algorithm

The data set (See Table I) described in section II is being used to test the performance of Genetic Algorithm. Mean square error (MSE) was evaluated using 10-fold cross validation as cross validation [5] is the best technique to get a reliable error estimate.

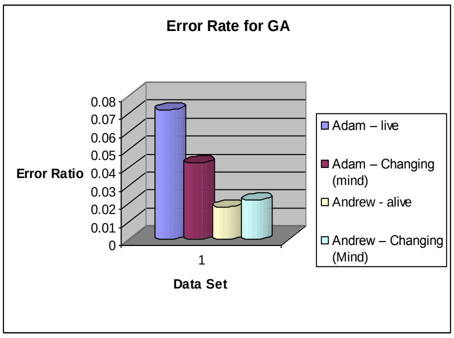

Figure 1: Error Rate for GA

Experiments Using Ensemble of Genetic Algorithm

The data set described in section II is being used to test the performance of bagging with Genetic Algorithm. Mean square error was evaluated using ensemble Method.

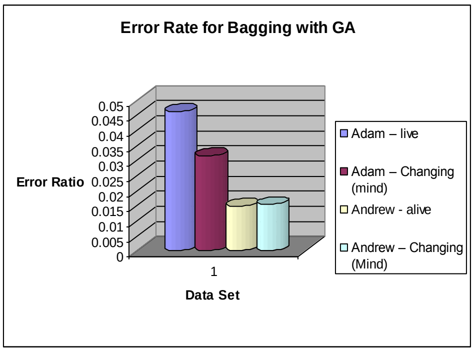

Figure 2: Error Rate for bagging with GA

Summary and Conclusions

In this article we suggested solution to some key problems of existing signature verification systems. Our research has clearly shown the importance of using ensemble approach for signature verification systems. An ensemble helps to indirectly combine the synergistic and complementary features of the different learning paradigms without any complex hybridization. Since all the considered performance measures could be optimized such systems could be helpful in several real world applications. We have achieved higher classification rate with respect to the lower error rate for the ensemble classifier (Figure 2) compared to that of single classifier (See Figure 1). We show that proposed ensemble of Genetic Algorithm is superior to individual approach for signature verification in terms of classification rate. We note however, that the difference in error rate figures tend to be very small and may not be statistically significant. More definitive conclusions can only be made after analyzing more comprehensive sets of signature verification data.

Acknowledgment

Authors gratefully acknowledge the authorities of Annamalai University for the facilities offered and encouragement to carry out this work. This part of work is supported in part by the first author got Career Award for Young Teachers (CAYT) grant from All India Council for Technical Education, New Delhi. They would also like to thank the reviewer's for their valuable remarks.

References

[1] Ian H.Witten and Eibe Frank, "Data Mining-Practical Machine Learning Tools and Techniques", Elsevier, 2005.

[2] Jiawei Han , Micheline Kamber " Data Mining – Concepts and Techniques" Elsevier, 2003.

[3] Margaret H.Dunham, "Data Mining-Introductory and Advanced Topics" Pearson Education, 2003.

[4] David Opitz, Richard Maclin, "Popular Ensemble Methods: An Empirical Study", Journal of Artificial Intelligence Research 11, 1999.

[5] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of International Joint Conference on Artificial Intelligence, pp. 1137–1143.

[6] C. Hook, J. Kempf, and G. Scharfenberg, "New pen device for biometrical 3d pressure analysis of handwritten charakters, words and signatures," in Proc. ACM Workshop Biometrics: Methods and Applications (WBMA), Berkeley, CA, 2003, pp. 38–44.

[7] O. Rohlík, P. Mautner, V. Matousek, and J. Kempf, "HMM based handwritten text recognition using biometrical data acquisition pen," in Proc.IEEE Int. Symp. Computational Intelligence in Robotics and Automation,vol. 2, Kobe, Japan, 2003, pp. 950–953.

[8] M.Mitchell. An introduction to genetic algorithms. MIT Press, Cambridge - MA, 1996.

[9] J. Ashbourn, Biometrics: Advanced Identity Verification: The Complete Guide. London, U.K.: Springer-Verlag, 2000.

[10] V. S. Nalwa, "Automatic on-line signature verification," in Biometrics: Personal Identification in Networked Society, A. K. Jain, R. Bolle, andS. Panhanti, Eds. Boston, MA: Kluwer, 1999.

[11] Q.Ji and Y.Zhang. Camera calibration with genetic algorithms. IEEE Trans. on Systems, Man, and Cybernetics, Part A:Systems and Humans, 31(2):120–130, 2001.

[12] S.Y.Ho and H.L.Huang. Facial modeling from an uncalibrated face image using a coarse-to-fine genetic algorithm. Pattern Recognition, 34(5):1015–1031, 2001.

[13] V.E.Ramesh and N.Murty. Off-line signature verification using genetically optimized weighted features. Pattern Recognition, 32(2):217–233, 1999.

[14] J.Yang and V.Honavar. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, 13(1):44–49, 1998.

[15] G.Kim and S.Kim. Feature selection using genetic algorithms for handwritten character recognition. In 7th IWFHR, pages 103–112, Amsterdam-Netherlands, 2000.

[16] A.E.Eiben, R.Hinterding, and Michalewicz. Parameter control in evolutionary algorithms. IEEE Trans.on Evolutionary Computation, 3(2):124–141, 1999.

[17] K.F.Man, K.S.Tang, and S.Kwong. Genetic Algorithms: Concepts and Designs. Springer-Verlag, London-UK, 1999.

[18] Rokach L., Chizi B., Maimon O. (2007). A Methodology for Improving the Performance of Non-ranker Feature Selection Filters, International Journal of Pattern Recognition and Artificial Intelligence, 21(5):809-830.

Dr.M.Govindarajan received the B.E and M.E and Ph.D Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 2001 and 2005 and 2010 respectively. He did his post-doctoral research in the Department of Computing, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, United Kingdom. He is currently an Assistant Professor at the Department of Computer Science and Engineering, Annamalai University, Tamil Nadu, India. He has presented and published more than 40 papers in Conferences and Journals. His current Research Interests include Data Mining and its applications, Algorithms, Text Mining, Neural Networks, genetic Algorithms, support vector machine, Radial Basis Function, ontology based Reasoning, Case Based Reasoning. He has conducted National Conference on Recent Trends in Data Mining and its Applications (March 11-12, 2006). He was the recipient of the Achievement Award for the field and to the Conference Bio-Engineering, Computer science, Knowledge Mining (2006), Prague, Czech Republic and All India Council for Technical Education "Career Award for Young Teachers (2006), New Delhi, India. He is Life Member of Computer Society of India, Indian Society for Technical Education and Session Member of Indian Science Congress Association, Associate member of Institute of Engineers

Dr. RM. Chandrasekaran received the B.E Degree in Electrical and Electronics Engineering from Madurai Kamaraj University in 1982 and the MBA (Systems) in 1995 from Annamalai University, M.E in Computer Science and Engineering from Anna University and PhD Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 1998 and 2006 respectively. He is currently working as a Professor at the Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India. From 1999 to 2001 he worked as a software consultant in Etiam, Inc, California, USA. He has conducted Workshops and Conferences in the Areas of Multimedia, Business Intelligence and Analysis of algorithms, Data Mining. He has presented and published more than 50 papers in conferences and journals and is the author of the book Numerical Methods with C++ Program (PHI, 2005). His Research interests include Data Mining, Algorithms, Networks, Software Engineering, Network Security, and Text Mining. He is Life member of Computer Society of India, Indian Society for Technical Education, Institute of Engineers and Indian Science Congress Association.

Table 1: Properties of Dataset

Signature Verification	Error Rate (MSE)
Adam – live	0.0719 %
Adam – Changing (mind)	0.0425 %
Andrew - alive	0.0180 %
Andrew– Changing (Mind)	0.0219 %

Table 2: Error Rate for GA

Signature Verification	Instances	Attributes
Adam – live	132	15
Adam – Changing (mind)	94	15
Andrew - alive	111	15
Andrew Changing (Mind)	137	15

Table 3: Error Rate for Bagging with Ga

Signature Verification	Error Rate (MSE)
Adam – live	0.0462 %
Adam – Changing (mind)	0.0316 %
Andrew - alive	0.0147 %
Andrew– Changing (Mind)	0.0153 %