Journal of Cybernetics and Informatics

published by

Slovak Society for Cybernetics and Informatics

Volume 14, 2014

http://www.kasr.elf.stuba.sk/sski/casopis/

ISSN: 1336-4774

Journal of Cybernetics and Informatics 14 (2014)
http://www.kasr.elf.stuba.sk/sski/casopis/

DATA DIMENSIONALITY REDUCTION BY GENETIC ALGORITHMS DATA MINING

M. Said. Abdel Moteleb and Nermin K. Abdel Wahab

Electronics Research Institute (ERI), Dokki, Cairo, Egypt

Abstract

Power quality monitors handle and store several gigabytes of data within a week and hence automatic detection, recognition and analysis of power disturbances require robust data mining techniques. Literature reveals that much work has been done to evolve several feature extraction and subsequent classification techniques for accurate power disturbance pattern recognition. However, the features extracted have been rarely evaluated for their usefulness. Classification fusion combines multiple classifications of data into a single classification solution of greater accuracy. Feature extraction aims to reduce the computational cost of feature measurement, increase classifier efficiency, and allow greater classification accuracy based on the process of deriving new features from the original features.

1 INTRODUCTION

There are many applications of data dimensionality reduction. Some of them are: Customer relationship management, text mining, image retrieval, protein classification, instruction detection, hand written digit recognition and face recognition. Automatic feature subset selection distinguishes the proposed Power Quality disturbance classification method from all other reported approaches. In particular, Genetic Algorithms (Gas) are employed to select features that encode important power quality information and improve classification performance. In [1] various methods for data dimensionality reduction were discussed, then a brief explanation for why GAs was employed and how it work was presented also. GAs belong to the class of randomized heuristic search techniques, offering an attractive approach to feature subset selection. Many of the applications of data selection can be included in the field of data mining; hence, in this paper we continue using GAs for data dimensionality reduction for data mining applications.

2 REVISION ON GAS

A. Terminology

The basic principles of GA were first proposed by Holland [2]. GA is inspired by the mechanism of natural selection, a biological process in which stronger individuals are likely be the winners in a competing environment, GA uses a direct analogy of such natural evolution. It presumes that the potential solution of a problem is an individual and can be represented by a set of parameters. These parameters are regarded as the genes of a chromosome and can be structured by a string of values. A positive value, generally known as fitness value, is used to reflect the degree of "goodness" of the chromosome for solving the problem, and this value is closely related to its objective value.

B. Genetic Operators and basic Principals

Population denotes a set of specific entities. In biological terms these entities are organisms of the same species. In the world of GAs the organisms are from the species "candidate solution" — living in an artificial environment defined by a given problem. The first generation of candidate solutions is a sample of possible parameter combinations for the problem.

Fitness is a term from evolution theory. It is a measure of the survival and the reproduction probability and fertility of an entity. The fitness function of the GA defines the environment for the artificial evolution. Each candidate solution in the population will be evaluated by the fitness function. A higher fitness results in a higher survival chance and a higher reproduction rate. A GA maximizes the fitness of a population.

Selection is the evolutionary term for "survival of the fittest", referring to the probability for an organism to survive and reproduce. Most Gas described in the literature have been "generational" — at each generation the new population consists entirely of off-springs formed by parents in the previous generation (Mitchell 1996). The parent generation is completely discarded. These GAs rely only on selection for reproduction.

Reproduction is another basic principle of evolution. In GAs the fertility of a candidate solution is determined by the relative fitness. A fitter solution will reproduce more often than a less fit entity.

Crossover is the recombination of the subsequences from two chromosomes to create two off-springs. The new chromosomes share the genes of both parents. From the evolutionary point of view, crossover secures the continuation of successful traits.

Mutation is a very important evolutionary aspect for GAs. While crossover can produce many new variants of existing solutions, mutation has the power to produce completely new solutions. It is randomly applied after crossover to mutate one or more genes in an offspring.

C. Implementation of a genetic algorithm

There is no golden rule to implement a GA [3]. The biological and evolutionary concepts of GA leave much room for interpretation. Many additional concepts can be introduced to optimize the GA for a specific problem.

3 FEATURE SELECTION/ EXTRACTION CONCEPT

A. Feature Selection

A process that chooses an optimal subset of features according to a objective function.

B. Feature Extraction

Feature reduction refers to the mapping of the original high-dimensional data onto a lower-dimensional space.

C. Feature Selection vs. Feature Extraction

In feature selection only a subset of the original features are selected, while in feature extraction all original features are used and the transformed features are linear combinations of the original features [4]. Objective of both is to reduce dimensionality and remove noise hence improve mining performance.

4 BACKGROUND ON USING GAS FOR FEATURE SELECTION/EXTRACTION

Genetic Algorithms (GAs) have been shown to be an effective tool to use in data analysis and pattern recognition [5], [6], [7], [8], [9], [10], [11]. An important aspect of GAs in a learning context is their use in pattern recognition. There are two different approaches to applying GA in pattern recognition:

1. Apply a GA directly as a classifier. Bandyopadhyay and Murthy in [12] applied GA to find the decision boundary in N dimensional feature space.

2. Use a GA as an optimization tool for resetting the parameters in other classifiers.

Most applications of GAs in pattern recognition optimize some parameters in the classification process. Many researchers have used GAs in feature selection [13], [14], [15], [16]. GAs have been applied to find an optimal set of feature weights that improve classification accuracy. First, a traditional feature extraction method such as Principal Component Analysis (PCA) is applied, and then a classifier such as k- NN is used to calculate the fitness function for GA [17], [18]. Combination of classifiers is another area that GAs have been used to optimize. Kuncheva and Jain in [19] used a GA for selecting the features as well as selecting the types of individual classifiers in their design of a Classifier Fusion System. GA is also used in selecting the prototypes in the case-based classification [20].

5 DATA MINING OVERVIEW

Recent tremendous technical advances in processing power, storage capacity, and interconnectivity is creating unprecedented quantities of digital data. Data mining, the science of extracting useful knowledge from such huge data repositories, has emerged as a young and interdisciplinary field in computer science. Data mining techniques have been widely applied to problems in industry, science, engineering and government, and it is widely believed that data mining will have profound impact on our society. The growing consensus that data mining can bring real value has led to an explosion in demand for novel data mining technologies and for students who are trained in data mining students who have an understanding of data mining techniques, can apply them

to real-life problems, and are trained for research and development of new data mining methods. Courses in data mining have started to sprawl all over the world [21].

6 REAL APPLICATION

A. Problem statement

Several web-based educational systems with different capabilities and approaches have been developed to deliver online education in an academic setting. In particular, Michigan State University (MSU) has pioneered some of these systems to provide an infrastructure for online instruction. The research presented here was performed on a part of the latest online educational system developed at MSU, the *Learning Online Network with Computer-Assisted Personalized Approach (LON-CAPA)*. LON-CAPA is involved with two kinds of large data sets: 1) educational resources such as web pages, demonstrations, simulations, and individualized problems designed for use on homework assignments, quizzes, and examinations; and 2) information about users who create, modify, assess, or use these resources. In other words, we have two ever-growing pools of data. [22] Investigated methods for extracting useful and interesting patterns from these large databases of students using online educational resources and their recorded paths within the system.

B. Map the Problem to Genetic Algorithm

Here the second approach of the GA pattern recognition (see Section 4) to optimize a combination of classifiers. Our objective is to predict the students' final grades based on their web-use features, which are extracted from the homework data.

Dataset and Class Labels. As test data we selected the student and course data of a LON-CAPA course, PHY183 (Physics for Scientists and Engineers I), which was held at MSU in spring semester 2002. This course integrated 12 homework sets including 184 problems, all of which are online. About 261 students used LON-CAPA for this course.

Extractable Features. An essential step in doing classification is selecting the features used for classification. Below a discussion of the features from LON-CAPA that were used, how they can be visualized (to help in selection) and why we normalize the data before classification.

The following features are stored by the LON-CAPA system:

1 Total number of correct answers.

2 Getting the problem right on the first try, vs. those with high number of tries.

3 Total number of tries for doing homework.

4 Time spent on the problem until solved.

5 Total time spent on the problem regardless of whether they got the correct answer or not.

6 Participating in the communication mechanisms, vs. those working alone. LONCAPA.

Classifiers. Six different classifiers using the LON-CAPA datasets are compared in this study. The classifiers used in this study include *Quadratic Bayesian classifier, 1-nearest neighbor (1-NN), k-nearest neighbor (k-NN), Parzen-window, multi-layer perceptron (MLP),* and *Decision Tree*. These classifiers are some of the common classifiers used in most practical classification problems. Finally, to improve performance, a combination of classifiers is presented.

C. Optimizing the CMC Using a GA

The feature vector for the predictors is the set of six variables for every student:

Success rate, Success at the first try, Number of attempts before correct answer is derived, the time at which the student got the problem correct relative to the due date, total time spent on the problem, and the number of online interactions of the student both with other students and with the instructor.

A population of six dimensional weight vectors with values between 0 and 1, are randomly initialized corresponding to the feature vector and experimented with different number of population sizes. We found good results using a population with 200 individuals. The GA Toolbox supports binary, integer, real-valued and floating-point chromosome representations. Real-valued populations may be initialized using the Toolbox function *crtrp*. For example, to create a random population of 6 individuals with 200 variables each: we define boundaries on the variables in *FieldD* which is a matrix containing the boundaries of each variable of an individual.

```
FieldD = [ 0 0 0 0 0 0; % lower bound
           1 1 1 1 1 1]; % upper bound
```

We create an initial population with Chrom = crtrp(200, FieldD), So we have for example:

```
Chrom =  0.23 0.17 0.95 0.38 0.06 0.26
         0.35 0.09 0.43 0.64 0.20 0.54
         0.50 0.10 0.09 0.65 0.68 0.46
         0.21 0.29 0.89 0.48 0.63 0.89
```

Here the simple genetic algorithm (SGA), which is described by Goldberg in [23] was used. The SGA uses common GA operators to find a population of solutions which optimize the fitness values.

Recombination "Stochastic Universal Sampling" [24] is used for selection. Using the equation:

$$F(x_i) = \frac{f(x_i)}{\sum_{i=1}^{N} f(x_i)}$$

Where $f(x_i)$ is the fitness of individual x_i and $F(x_i)$ is the probability of that individual being selected.

Crossover. The crossover operation is not necessarily performed on all strings in the population. Instead, it is applied with a probability Px where Px = 0.7. There are several functions to make crossover on real-valued matrices. One of them is *recint*, which performs intermediate recombination between pairs of individuals in the current population, OldChrom, and returns a new population after mating, NewChrom. Each row of OldChrom corresponds to one individual. *recint* is a function only applicable to populations of real-value variables. Intermediate recombination combines parent values using the following formula [25]:

$$Offspring = parent1 + Alpha \times (parent2 - parent1)$$

Alpha is a scaling factor chosen uniformly in the interval [-0.25, 1.25].

Mutation. There are several functions to make mutation on real-valued population. We used *mutbga*, which takes the real-valued population, OldChrom, mutates each variable with given probability and returns the population after mutation, *NewChrom = mutbga*(*OldChrom, FieldD, MutOpt*) takes the current population, stored in the matrix OldChrom and mutates each variable with probability by addition of small random values (size of the mutation step). We considered 1/600 as our mutation rate. The mutation of each variable is calculated as follows:

$$Mutated\ Var = Var + MutMx \times range \times MutOpt(2) \times delta$$

Where delta is an internal matrix which specifies the normalized mutation step size; MutMx is an internal mask table; and MutOpt specifies the mutation rate and its shrinkage during the run. The mutation operator *mutbga* is able to generate most points in the hypercube defined by the variables of the individual and the range of the mutation. However, it tests more often near the variable, that is, the probability of small step sizes is greater than that of larger step sizes.

Fitness Function. During the reproduction phase, each individual is assigned a fitness value derived from its raw performance measure given by the objective function. This value is used in the selection to bias towards more fit individuals. Highly fit individuals, relative to the whole population, have a high probability of being selected for mating whereas less fit individuals have a correspondingly low probability of being selected.

Experiment Results. In this real application the GAs improved the performance (the accuracy) of the classifiers in all cases.

Table I. Comparing the CMC performance on PHY183 dataset using GA and without GA in the cases of 2-Classes, 3-Classess, and 9-Classes [22].

	Performance %		
Classifier	2-Classes	3-Classes	9-Classes
CMC of 4 Classifiers without GA	83.87	61.86	49.74
GA Optimized CMC	94.09	72.13	62.25
Improvement	10.22	10.26	12.51

7 CONCLUSION

In this paper we reviewed the concept of GAs, feature selection and extraction, genetic operators, data mining and how to optimize it using GAs. A real application for data mining optimization using GAs in educational systems was explained. A comparison between using classifiers without GAs vs. the same classifiers with GAs was also presented. The combination of multiple classifiers with GA leads to a significant accuracy improvement in all cases.

REFERENCES

[1] M. Said.Abdel Moteleb, Nermin K. Abdel Wahab and Essam W. Helmy. ' Data Dimensionality Reduction by Genetic Algorithms'. Peit011 Conference. 2011

[2] J. H. Holland, Adaption in Natural and Artificial Systems. Cambridge, MA: MIT Press, 1975

[3] Nermin K. Abdel Wahab. ' Investigation of Genetic Algorithms - Based Face Detection and Recognition'. Peit011 Conference. 2011

[4] Lei Yu, Jieping Ye and Huan Liu. 'Dimensionality Reduction for Data Mining - Techniques, Applications and Trends'.

[5] Freitas, A.A.: A survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery, See: www.pgia.pucpr.br/~alex/papers. A chapter of: A. Ghosh and S. Tsutsui. (Eds.) Advances in Evolutionary Computation. Springer-Verlag, 2002

[6] Jain, A. K.; Zongker, D.: Feature Selection: Evaluation, Application, and Small Sample Performance, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 19, No. 2, February, 1997

[7] Falkenauer E.: Genetic Algorithms and Grouping Problems. John Wiley & Sons, 1998

[8] Pei, M., Goodman, E.D. and Punch, W.F.: Pattern Discovery from Data Using Genetic Algorithms, Proceeding of 1st Pacific-Asia Conference Knowledge Discovery & Data Mining (PAKDD-97) Feb. 1997

[9] Park Y and Song M.: A genetic algorithm for clustering problems. Genetic Programming 1998: Proceeding of 3rd Annual Conference, Morgan Kaufmann, 1998, 568–575.

[10] Michalewicz Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd Ed., Springer-Verlag, 1996

[11] De Jong K.A., Spears W.M. and Gordon D.F.: Using genetic algorithms for concept learning. Machine Learning 13, 1993, 161–188.

[12] Bandyopadhyay, S., and Muthy, C.A.: Pattern Classification Using Genetic Algorithms. Pattern Recognition Letters, Vol. 16, 1995, 801-808.
[13] Bala J., De Jong K., Huang J., Vafaie H., and Wechsler H.: Using learning to facilitate the evolution of features for recognizing visual concepts. Evolutionary Computation 4(3) - Special Issue on Evolution, Learning, and Instinct: 100 years of the Baldwin Effect, 1997
[14] Guerra-Salcedo C. and Whitley D.: Feature Selection mechanisms for ensemble creation: a genetic search perspective. In: Freitas AA (Ed.) Data Mining with Evolutionary Algorithms: Research Directions - Papers from the AAAI Workshop, 13-17. Technical Report WS-99-06. AAAI Press, 1999
[15] Vafaie, H. and De Jong, K.: Robust feature Selection algorithms. Proceeding of IEEE International Conference on Tools with AI, Boston, Mass., USA. Nov. 1993, 356-363.
[16] Martin-Bautista M.J., and Vila M.A.: A survey of genetic feature selection in mining issues. Proceeding Congress on Evolutionary Computation (CEC-99), Washington D.C., July 1999, 1314-1321.
[17] Pei, M., Goodman, E.D., and Punch, W.F.: Pattern Discovery from Data Using Genetic Algorithms. Proceeding of 1st Pacific-Asia Conference Knowledge Discovery & Data Mining (PAKDD-97), 1997
[18] Punch, W.F., Pei, M., Chia-Shun, L., Goodman, E.D., Hovland, P., and Enbody R.: Further research on Feature Selection and Classification Using Genetic Algorithms. In 5th International Conference on Genetic Algorithm, Champaign IL, 1993, 557-564.
[19] Kuncheva, L.I., and Jain, L.C.: Designing Classifier Fusion Systems by Genetic Algorithms. IEEE Transaction on Evolutionary Computation, Vol. 33, 2000, 351-373.
[20] Skalak D. B.: Using a Genetic Algorithm to Learn Prototypes for Case Retrieval an Classification. Proceeding of the AAAI-93 Case-Based Reasoning Workshop, Washigton, D.C. American Association for Artificial Intelligence, Menlo Park, CA, 1994, 64-69.
[21] Soumen Chakrabarti, Martin Ester, Usama Fayyad, Johannes Gehrke, Jiawei Han, Shinichi Morishita, Gregory Piatetsky-Shapiro and Wei Wang. ' Data Mining Curriculum: A Proposal (Version 0.91)'. Intensive Working Group of ACM SIGKDD Curriculum Committee. August 5, 2004
[22] Behrouz Minaei-Bidgoli and William F. Punch. 'Using Genetic Algorithms for Data Mining Optimization in an Educational Web-Based System'.
[23] Guerra-Salcedo C. and Whitley D.: Feature Selection mechanisms for ensemble creation: a genetic search perspective. In: Freitas AA (Ed.) Data Mining with Evolutionary Algorithms: Research Directions, Technical Report WS-99-06. AAAI Press, 1999
[24] Baker, J. E.: 'Reducing bias and inefficiency in the selection algorithm', Proceeding ICGA 2, Lawrence Erlbuam Associates, Publishers, 1987, 14–21.
[25] Muhlenbein and Schlierkamp-Voosen D.: Predictive Models for the Breeder Genetic Algorithm: I. Continuous Parameter Optimization, Evolutionary Computation, Vol. 1, No. 1, 1993, 25–49.