Data mining & warehouse (Survey Report 3)

# Co-evolutionary algorithm for data mining in big data


**Contributed by(** *Students of BSIT 7A***):**

Hifsa Basharat 45903
Haris Anwer 45902
M Laraib Kiayani 45916

**Instructor**

Sir. Imran Memon

Bahria University
Discovering Knowledge

SURVEY PAPER

# Co-evolutionary algorithm for data mining in big data

By Hifsa Basharat, Haris Anwer, and Muhammad Laraib Kiayani

Correspondence:
*Hifsabasharat689@gmail.com;*
*Harisanwer125@gmail.com*
*skiayani404@hotmail.com*

*Hifsa, Haris and Laraib contributed equally in this work For the purpose of Submition of Semester assignment*

## Abstract

Cooperative Coevolution is a technique in the area of Evolutionary Computation. It has been applied to many combinatorial problems with great success. This contribution proposes a Cooperative Coevolution model for simultaneous performing some data reduction processes in classification with nearest neighbors methods through feature and instance selection. In order to check its performance, we have compared the proposal with other evolutionary approaches for performing data reduction.

Results have been analyzed and contrasted by using non-parametric statistical tests, finally showing that the proposed model outperforms the noncooperative evolutionary techniques. One of the major challenges in data mining is the extraction of comprehensible knowledge from recorded data. In this paper, a coevolutionary-based classification technique, namely co-evolutionary Rule Extractor (CORE), is proposed to discover classification rules in data mining. Unlike existing approaches where candidate rules and rule sets are evolved at different stages in the classification process, the proposed CORE coevolves rules and rule sets concurrently in two cooperative populations to confine the search space and to produce good rule sets that are comprehensive.

The proposed coevolutionary classification technique is extensively validated upon seven datasets obtained from the University of California, Irvine (UCI) machine learning repository, which are representative artificial and real-world data from various domains. Comparison results show that the proposed CORE produces comprehensive and good classification rules for most datasets, which are competitive as compared with existing classifiers in literature. Simulation results obtained from box plots also unveil that CORE is relatively robust and invariant to random partition of datasets.

## Introduction

One main process in data mining is the one known as data reduction. In classification, it aims to reduce the size of the training set mainly to increase the efficiency of the training phase (by removing redundant instances) and even to reduce the classification error rate (by removing noisy instances).

Instance Selection (IS) and Feature Selection (FS) are two of the most known data reduction techniques. In data mining. Both are really effective not only to reduce the size of the train set, but also to filtrate and clean noisy data, thus helping classifiers to improve its accuracy. Evolutionary Algorithms (EAs) are general purpose search algorithms that use principles inspired by nature to evolve solutions to problems.

EAs have been successfully used in data mining problems. Their capacity of tackling IS and FS as combinatorial problems is especially useful. Coevolution is a specialized trend of EAs. It tries to simultaneously manage two or more populations (also called species), to evolve them and to allow interactions among individuals of any population.

Unlike traditional gradient-guided data mining techniques, evolutionary computation techniques intelligently search the solution space by evaluating performances of multiple candidate solutions simultaneously and approach the global optimum in a non-deterministic manner. Although Evolutionary Computation (EC) techniques play an important role in several areas of data mining domain, they have achieved more popularity for rule based classification (rule induction), for the reason that sets of IF-THEN rules can easily be represented by choosing an encoding of rules that allocates specific substrings for each rule precondition and postcondition (Mitchell 1997).

The goal is to improve results achieved from each population separately. The Coevolution model has shown some interesting characteristics in the last years.

Also, it has been successfully applied in other problems, like function optimization. Our proposal combines Evolutionary IS and FS with Coevolution techniques, in order to improve the effectiveness of Evolutionary IS and FS applied to nearest neighbors classifiers in terms of accuracy. We have named our proposed model CoCHC (Cooperative Coevolution model using CHC algorithm). A wide range of classification data sets will be used to compare it with other non-coevolutionary models, in order to highlight the benefits of the use of Coevolution.

## Previous Work

This section shows the main topics of the background in which our contribution is based. First Section describes some evolutionary techniques applied to IS and FS problems. Section shows the EAs in which our model is based. Finally, second Section highlights the main characteristics of Cooperative Coevolution.

### *Evolutionary Instance and Feature Selection*

EAs have proved to be good mechanisms for data reduction in data mining. They have been widely used to tackle the FS and IS problems. The FS problem can be defined as a search process of P features from an initial set of M variables, with P <= M. It aims to eliminate irrelevant and/or redundant features and to obtain a simpler classification system. Also, this reduction can improve the accuracy of the model in classification. The IS problem can also be defined as a search process, where a reduced set S of instances is selected from the training set.

By choosing the most suitable points in the data set as instances for the training data, the classification process can gets greatly increased both its efficiency and accuracy In is proposed a hybridization of a genetic algorithm with local search operators for FS.

In, a complete study of the use of EAs in IS done, highlighting four EAs to complete this task: Generational Genetic Algorithm (GGA), Steady-State Genetic Algorithm (SGA), CHC Adaptive Search Algorithm(CHC) and Population-Based Incremental Learning (PBIL).

They concluded that EAs outperform classical algorithms both in reduction rates and classification accuracy. They also concluded that CHC is the most appropriate EA to make this task, according to the algorithms they compared. Beyond these applications, it is important to point out that both techniques can be applied simultaneously.

Despite the most natural way to combine these techniques is to use one first (i.e. IS), to get its results and to apply them to the second technique (i.e. FS), some authors have already tried to get some profit from the joint use of both approaches.

***CHC Algorithm***

As it is exposed in the previous section, CHC is a good example of EA which can be used in IS and FS. We have studied its main characteristics to select it as the baseline EA which will guide the search process of our model. During each generation, the CHC algorithm develops the following steps:

1. It uses a parent population of size R to generate an intermediate population of R

2. individuals, which are randomly paired and used to generate R potential offspring.

3. Then, a survival competition is held where the best R chromosomes from the parent and offspring populations are selected to form the next generation.

CHC also implements HUX recombination operator. HUX exchanges half of the bits that differ between parents, where the bit position to be exchanged is randomly determined. It also employs a method of incest prevention: Before applying HUX to two parents, the Hamming distance between them is measured.

Only those parents who differ from each other by some number of bits (mating threshold) are mated. If no offspring is inserted into the new population then the threshold is reduced.

No mutation is applied during the recombination phase. Instead, when the search stops making progress the population is reinitialized to introduce new diversity. The chromosome representing the best solution found is used as a template to re-seed the population, randomly changing 35% of the bits in the template chromosome to form each of the other chromosomes in the population.

We have selected CHC because it has been widely studied, being now a well-known algorithm on evolutionary computation. Furthermore, previous studies like support the fact that it can perform well on data reduction problems.

## Existing Problems and its Solutions

***Cooperative Coevolutive Model Based Using CHC***

CoCHC employs three populations which simultaneously coexist. They cooperate to get the best possible solution through the evolutionary search procedure.

Each population is focused on one reduction data task:

– The first population performs an instance selection.

– The second population performs a feature selection.

– The third population performs both instance and feature selections.

Algorithm 1 shows a basic pseudocode of the model proposed.

**Algorithm 1.** CoCHC algorithm basic structure

```
1  Generate ISPopulation,FSPopulation and IFSPopulation Randomly;
2  Select initial bestISArray, bestFSArray and bestIFSArray;
3  Evaluate all populations in the multiclassifier;
4  Select bestISArray, bestFSArray and bestIFSArray from each population;
5  while evaluations < max_evaluations do
6  |   Select best classifier in last generation;
7  |   Generate simple classifier output from best individuals of the last
   |   generation;
8  |   Do a CHC Generation on every population;
9  |   Update bestISArray, bestFSArray and bestIFSArray if a better global
   |   solution has been found;
10 end
   Output: bestISArray, bestFSArray and bestIFSArray
```

Instruction 1 generates the initial random populations. Instruction 2 evaluates all chromosomes by using simple classifiers and selects the best individual of each population. Instruction 3 evaluates all chromosomes by using the complete metaclassifier, and instruction 4 selects the new best individual of each population. In instruction 5 the evolutionary process starts. Instruction 6 selects the best classifier of the last generation (the best simple classifier in accuracy). This will help to break ties in the fitness function evaluation. In Instruction 7, the outputs of the simple classifiers from the best individuals of the last generation are saved.

Instruction 8 performs a CHC generation on each population. Instruction 9 updates the best global solution if a better solution (concerning one chromosome from each population)have been found. When a fixed number of evaluations run out, the evolutionary process is finished. Then, best global solution founded is returned. At this point, we have to describe three important issues to completely describe CoCHC:

The specification of the representation of the chromosomes, the structure of the multi classifier defined by a full solution and the definition of the fitness function.

***Experimental Framework***

To check the performance of CoCHC algorithm, we have used 18 data sets taken from the UCI Machine Learning Database Repository. Table 1 shows their main characteristics. For each data set, it is shown the number of examples, attributes and classes of the problem described.

**Table 1.** UCI Data sets used in our experiments

| Data set | Examples | Attributes | Classes | Data set | Examples | Attributes | Classes |
|---|---|---|---|---|---|---|---|
| Aut | 205 | 25 | 6 | Housevotes | 435 | 16 | 2 |
| Bal | 625 | 4 | 3 | Iris | 150 | 4 | 3 |
| Bands | 539 | 19 | 2 | Mammogr | 961 | 5 | 2 |
| Bupa | 345 | 6 | 2 | Pima | 768 | 8 | 2 |
| Car | 1728 | 6 | 4 | Sonar | 208 | 60 | 2 |
| Cleveland | 303 | 13 | 5 | Tic-tac-toe | 958 | 9 | 2 |
| Dermat | 366 | 34 | 6 | Vehicle | 846 | 18 | 4 |
| German | 1000 | 20 | 2 | Wisconsin | 699 | 9 | 2 |
| Glass | 214 | 9 | 7 | Zoo | 101 | 16 | 7 |

The data sets considered are partitioned by using the tenfold cross-validation (10-fcv) procedure. The parameters of CoCHC are: Population size = 50 (for each population), Number of evaluations = 10000, $\alpha = 0.6$, $\beta = 0.98$. The alpha parameter value was taken from the value used on the experiments of, but slightly increased because the simultaneous use of a FS component. The beta parameter value is near to 1 because in the FS component our model has to remove irrelevant attributes without provoke sudden changes which could decrease the overall accuracy.

Our proposal will be compared with three evolutionary algorithms based on the CHC model, for performing IS, FS and simultaneous IS-FS, respectively. The first one will be denoted by IS-CHC, the second one, FS-CHC; and the last one IFS-CHC.

Nearest neighbour rule (1-NN) is also used as a baseline algorithm. The parameters used for each EA involved in the experimental study are the same as the used by our approach.

## Future Directions

The proposed CORE has been examined on seven datasets obtained from UCI machine learning repository and has produced good classification results as compared to many existing classifiers. Most of the comparisons were performed statistically using box plots to show the robustness of the proposed classifier.

Extensive simulation results show that CORE outperformed two other rule-based classifiers (C4.5 and PART) in almost all test problems except for the Sick dataset and is very competitive as compared to statistical based techniques, such as Naïve Bayes (e.g. Naïve Bayes achieved better results only on the Heart- C problem).

One reason for CORE being less efficient than the C4.5 and PART in terms of generalization ability for Sick dataset could be due to the low number of rules used.

While CORE uses 3.49 (on average) number of rules, C4.5 uses 21.61 number of rules and PART uses 17.05 number of rules, these figures are about 600% and 500% more compared to the number of rules CORE uses. In order to improve the performance of CORE for this problem, the upper limit of allowed rules can be increased but at the expense of higher computational cost.

The performance comparisons to other evolutionary based classifiers (GP-Co, GGP, GBML, GPCE, GA-based FKIF, XCS and XCSR) are mainly restricted by the availability of data, e.g. not all the datasets used in our experiments were tested in other publications.

comparing it to some established methods widely used in the literature. The paired t-tests with Bonferroni correction made for multiple comparisons have been performed between CORE and three widely used classifiers by taking each dataset as independent observation.

These p-values suggested that there is not much significant difference between the means as opposed to the results presented in previous tables; this could be due to the number of independent observations taken being too small.

The results reported by CORE shown in previous tables show the lowest standard deviation in all

Table 2. Performance comparisons for the sick problem

| Algorithm | Reference | # Rules | Avg accuracy | Best accuracy | Standard deviation |
|---|---|---|---|---|---|
| CORE | – | 3.49 | 95.17% | 98.05% | 1.53% |
| C4.5 | (Quinlan 1993) | 21.61 | 98.59% | 99.53% | 0.12% |
| PART | (Frank and Witten 1998) | 17.05 | 98.27% | 99.06% | 0.16% |
| NaïveBayes | (John and Langley 1995) | – | 92.59% | 94.93% | 1.28% |

Since there have been so many classifiers proposed in literature over the years, it is very difficult, if not impossible, to include every one of them in the comparisons. Therefore the comparisons are not meant to be exhaustive, but to assess the reliability and robustness of CORE by

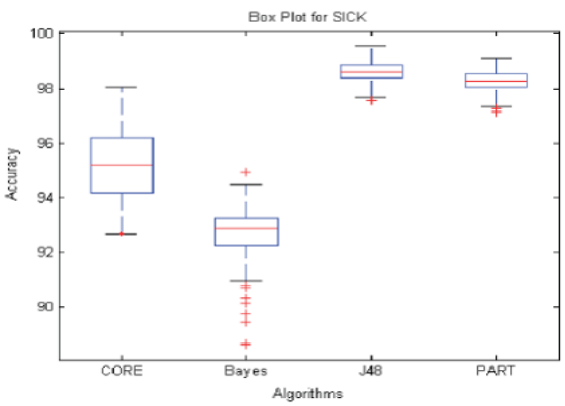


Figure 1. Box plot for the sick problem

the problems except Sick, compared to all other algorithms. The box plots show that CORE has a relatively lesser number of outliers as compared to traditional rule-based classifiers, which indicates that CORE is relatively more robust and less affected by the random partition of learning and testing sets.

## Discussion

Table 2 shows the average results obtained in test data in terms of accuracy. It also shows the reduction rate achieved in training data. The best results achieved in accuracy for each data set are remarked in bold.

Observing Table 2, we can make the following analysis:

- CoCHC achieves the best average result on accuracy.
- CoCHC outperforms all the remaining algorithms in 6 of 18 data sets.
- The loss in reduction rate achieved by CoCHC (compared with IS-CHC and IFS-CHC) is not too critical. It increases the average accuracy in 2% with respect to both them and keeps a good reduction rate.

## Conclusion

The purpose of this contribution is to present a cooperative coevolutionary model developed to tackle data reduction tasks to improve the classification based on the nearest neighbors technique. The proposal combines processes of evolutionary instance selection and feature selection techniques. The results show that the use of cooperative coevolution in data reduction based on feature and instance selection can obtain promising results to optimize the performance of nearest neighbor classification. This paper has proposed a cooperative coevolutionary algorithm (CORE) for rule extraction and classification in data mining applications. CORE utilizes the evolutionary algorithm principles which possess global search ability to search for rules and rule sets.

**Table 2.** Accuracy obtained in test data

| Algorithm | CoCHC | | IS-CHC | | FS-CHC | | IFS-CHC | | 1-NN |
|---|---|---|---|---|---|---|---|---|---|
| Data set | %Acc. | %Red. | %Acc. | %Red. | %Acc. | %Red. | %Acc. | %Red. | %Acc. |
| Automobile | 79.66 | 88.72 | 70.42 | 91.27 | **80.18** | 68.00 | 70.53 | 98.92 | 77.43 |
| Bal | 88.16 | 89.28 | 89.29 | 98.74 | 79.04 | 0.00 | **89.43** | 98.56 | 79.04 |
| bands | 71.81 | 82.47 | 70.14 | 97.30 | 71.07 | 49.47 | 68.28 | 99.66 | **74.04** |
| Bupa | 68.96 | 81.96 | 61.94 | 96.14 | 61.93 | 38.33 | **68.98** | 99.24 | 61.08 |
| Car | 89.18 | 70.88 | 86.69 | 98.37 | **89.58** | 18.33 | 88.66 | 98.27 | 85.65 |
| Cleveland | 56.76 | 84.52 | **57.81** | 97.47 | 50.83 | 46.15 | 57.10 | 99.46 | 53.14 |
| Dermatology | 95.37 | 85.60 | **97.55** | 96.36 | 95.11 | 54.71 | 96.44 | 99.18 | 95.35 |
| German | **72.20** | 82.80 | 71.70 | 98.69 | 69.30 | 38.50 | 71.90 | 99.89 | 70.50 |
| Glass | 69.99 | 78.08 | 69.11 | 93.14 | 71.37 | 43.33 | 67.06 | 97.83 | **73.61** |
| Housevotes | **95.14** | 89.46 | 93.32 | 98.24 | 94.47 | 65.00 | 93.54 | 99.87 | 92.16 |
| Iris | **95.33** | 81.89 | **95.33** | 95.56 | **95.33** | 45.00 | 94.67 | 98.11 | 93.33 |
| Mammographic | **82.00** | 97.39 | 80.23 | 99.17 | 72.94 | 56.00 | 81.59 | 99.90 | 74.72 |
| Pima | 72.01 | 87.73 | **76.07** | 98.50 | 68.62 | 50.00 | 74.11 | 99.75 | 70.33 |
| Sonar | 85.55 | 86.13 | 76.83 | 93.75 | **86.45** | 58.50 | 79.24 | 99.76 | 85.55 |
| Tic-tac-toe | **83.81** | 73.29 | 73.69 | 97.91 | 82.78 | 22.22 | 75.89 | 98.91 | 73.07 |
| Vehicle | **71.99** | 82.15 | 63.36 | 96.44 | 71.52 | 45.56 | 68.09 | 99.10 | 70.10 |
| Wisconsin | **96.28** | 91.62 | 96.27 | 99.32 | 95.14 | 47.78 | 95.28 | 99.78 | 95.57 |
| Zoo | 95.58 | 86.56 | **97.00** | 86.24 | 94.75 | 55.63 | 87.97 | 95.76 | 92.81 |
| Average | **81.66** | 84.47 | 79.26 | 96.25 | 79.47 | 44.58 | 79.37 | 99.00 | 78.75 |

**Table 3.** Results of Wilcoxon Signed-Ranks Test

| $\alpha = 0.1$ | IS-CHC | FS-CHC | IFS-CHC | 1-NN |
|---|---|---|---|---|
| CoCHC | +(.059) | +(.012) | +(.018) | +(.004) |

In addition to Table 2, we have performed a two-tailed Wilcoxon Signed-Ranks Test [4], to statistically analyze the results obtained in the experiment. Table 3 shows the p-values obtained by Wilcoxon test. The results offered by the test indicate us that the proposed model outperforms FS-CHC and IFS-CHC with a level of significance $\alpha = 0.05$, and it is better than IS-CHC considering a level of significance $\alpha = 0.1$.

These solutions are presented in high level linguistic rule sets that are easily comprehensible for humans. Unlike existing evolutionary approaches, the proposed coevolutionary classifier coevolves the rules and rule sets concurrently in two cooperative populations. The main population encodes a population of rules using Michigan encoding which are syntactically shorter thus making the search for good candidate solutions faster.

## References


Baluja, S.: Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning. Technical report, Pittsburgh, PA, USA (1994)

Cano, J.R., Herrera, F., Lozano, M.: Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study. IEEE Transactions on Evolutionary Computation 7, 561–575 (2003)

Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13, 21–27 (1967)

Demˇsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30 (2006)

Eiben, A.E., Smith, J.E.: Introduction to Evolutionary Computing. Springer, Heidelberg (2003)

Eshelman, L.J.: The CHC adaptative search algorithm: How to safe search when engaging in nontraditional genetic recombination. In: Foundations of Genetic Algorithms, pp. 265–283 (1990)Fragoudis, D., Meretakis, D., Likothanassis,

Integrating feature and instance selection for text classification. In: 8th ACM SIGKDD international conference on KDD, pp. 501–506 (2002)

Freitas, A.A.: Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer, New York (2002)

Ghosh, A., Jain, L.C.: Evolutionary Computation in Data Mining. Springer, Berlin (2005)

Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-

Wesley, Reading (1989)

Au, C., Leung, H.: Guided Mutations in Cooperative Coevolutionary Algorithms for Function Optimization. In: IEEE International Conference on Tools with Artificial Intelligence, pp. 407–414 (2007)

Liu, H., Motoda, H.: Instance Selection and Construction for Data Mining. The Springer International Series in Engineering and Computer Science (2001)

Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman & Hall/Crc Data Mining and Knowledge Discovery Series (2007)

Oh, I., Lee, J., Moon, B.: Hybrid Genetic Algorithms for Feature Selection. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 1424–1437 (2004)

Panait, L., Wiegand, R.P., Luke, S.: Improving coevolutionary search for optimal multiagent behaviors. In: International Joint Conferences on Artificial Intelligence, pp. 653–658 (2003)

Panait, L., Luke, S., Harrison, J.F.: Archive-Based Cooperative Cooevolutionary Algorithms. In: Genetic and Evolutionary Computation Conference, pp. 345–352 (2006)

Potter, M.A., De Jong, K.A.: Cooperative coevolution: an architecture for evolving coadapted subcomponents. Evolutionary Computation 8, 1–29 (2000)

# PLAGIARISM REPORT

10% similarity has been found in this content.

turnitin

## Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

| | |
|---|---|
| Submission author: | HIFSA BASHARAT |
| Assignment title: | Fyp 3 |
| Submission title: | Survey Report Assignment3 |
| File name: | SurveyReport3.docx |
| File size: | 411.92K |
| Page count: | 15 |
| Word count: | 4,651 |
| Character count: | 25,651 |
| Submission date: | 20-Dec-2019 12:54AM (UTC+0500) |
| Submission ID: | 1234507572 |


SURVEY PAPER

Co-evolutionary algorithm for data mining in big data

Abstract

Assignment Inbox: Fyp 3

| Assignment Title | Info | Dates | Similarity | Actions |
|---|---|---|---|---|
| Fyp 3 | ⓘ | Start 13-Dec-2019 11:04PM<br>Due 25-Dec-2019 11:59PM<br>Post 21-Dec-2019 12:00AM | 10% | Resubmit View |