Cost-Sensitive Feature Selection using Particle Swarm Optimization: A Post-Processing Approach

Syed Imran Ali
Department of Computer Science & Engineering,
Kyung Hee University
Yongin-si, Korea
imran.ali@oslab.khu.ac.kr

Wajahat Ali Khan
Department of Computer Science & Engineering,
Kyung Hee University
Yongin-si, Korea
wajahat.alikhan@oslab.khu.ac.kr

Sungyoung Lee
Department of Computer Science & Engineering,
Kyung Hee University
Yongin-si, Korea
sylee@oslab.khu.ac.kr

Sang-Ho Lee
Department of Internal Medicine,
Kyung Hee University
Seoul, Korea
lshkidney@khu.ac.kr

Abstract— **Feature selection is one of the important tasks in machine learning. Feature selection task deals with selecting a subset of feature from an original feature set. An important consideration in feature selection is the usefulness of a feature i.e. a set of feature which is selected is neither irrelevant nor redundant. Most of the existing algorithms in the domain of feature selection are designed to optimize the aforementioned objective. In our research we have addressed a third dimension of usefulness i.e. cost of the feature. Cost-effectives of a solution is most apt in cases where there is an asymmetric cost of data acquisition such as medical diagnosis applications. In this regard, our research deals with enhancing the existing feature selection techniques with a post-processing stage in which cost consideration is also accounted for. The resultant solution is optimized for both important as well as cost-effective features. We have used particle swarm optimization with post processing over chronic kidney disease dataset for generating a feature subset set which is both salient and cost-effective.**

Keywords— Cost-sensitive Feature Selection, Particle Swarm Optimization, Post Processing, Data Classification

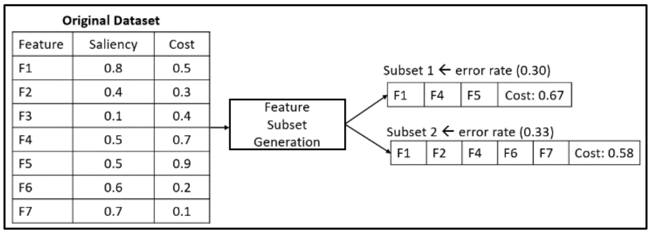
I. INTRODUCTION

A number of data driven approaches for decision making rely on machine learning models. In this regard, supervised learning models, arguably, are one of the most successful and widely used modeling approaches. In supervised learning framework, a model is learned i.e. inferred from data which is capable of producing a mapping between a set of independent features and a dependent feature [1].

Cardinality of the input data i.e. number of independent features, highly influence this resultant model built from data. It is report that if the original data contains a set of features which are irrelevant and redundant than the model may over-fit. It negatively impacts the model. Therefore, huge array of feature selection algorithms are designed to address the issue of 'curse of dimensionality' [2, 3, 4].

Traditionally an important assumption taken in building a classification model is the symmetric cost of data acquisition. This assumption is not valid in cases such as medical diagnosis where there cost of data acquisition is asymmetric e.g. variability of economic cost for different medical tests [5].

Furthermore, cost can be accounted for not only in economic terms but also in terms of the degree of availability of a certain medical test, risk of exposure to a patient, computational cost in terms of data acquisition, etc. [6] Hence, the notion of cost adds another important dimension to the overall usefulness of a solution i.e. how practical is the application? Figure 1. depicts a case in which two subsets, subset-1 and subset-2, are generated from an original dataset. Both the subsets have equal error rate while they have asymmetric cost associated with them. Subset-1 has fewer features then subet-2 but subset-2 has relatively less cost than the subset 1. As it can be seen, fewer features don't always result in a lower overall cost. This example illustrates a trade-off situation where no solution is strictly the best and hence both the subsets maybe made part of the final solution which is presented to the decision maker. Where a preference is pre-specified then a further refinement can be made to the final solution.


Original Dataset
Feature Subset Generation
Subset 1 ← error rate (0.30)
F1 | F4 | F5 | Cost: 0.67
Subset 2 ← error rate (0.33)
F1 | F2 | F4 | F6 | F7 | Cost: 0.58

Feature	Saliency	Cost
F1	0.8	0.5
F2	0.4	0.3
F3	0.1	0.4
F4	0.5	0.7
F5	0.5	0.9
F6	0.6	0.2
F7	0.7	0.1

Figure 1. A case for an asymmetric feature cost

In order to account for the cost factor, generally there are two approaches i.e. design a cost-sensitive classifier or design a cost-sensitive feature selection method. In this research, we have opted for the latter approach since the former is tightly bound to a specific class of classification model such as decision trees and forces the decision maker to do the same. In case of cost-sensitive feature selection, a decision maker can obtain a subset of features which are optimized for both label prediction and cost. Thereafter, any existing classification model may be used with the obtained feature subset.

In this research, we enhanced an existing feature selection algorithm with a post-processing step such that as it becomes a two step-process. In first step, a set of highly salient candidate features are obtained, and in the subsequent step through cost assignment a final solution is obtained. We have used particle swarm optimization (PSO) for generating a set of candidate solutions. In PSO, unlike non-evolutionary techniques, a set of solutions are simultaneously optimized for a given fitness function [7]. Nearest-neighbor classifier is used as a fitness function since it is non-parametric and makes no assumptions regarding the underlying data distribution.

II. PROPOSED METHODOLOGY

In this section, we elaborate our proposed two-stage methodology, Independent Multi-Swarm PSO with Post Processing (IMSPSO-PP), which leverages multi-swarm PSO for cost based feature subset selection and ensemble model generation as shown in Figure 2.

In the first-stage of the methodology, a set of unique solutions are generated through a multi-swarm induced PSO. Multi-swarm mechanism provides necessary diversity in the resultant solutions in order to process for cost-assignment and model creation. It is important to note that, solutions are optimized for accuracy i.e. objective 1, at this stage.

The second part of the methodology deals with evaluating evolved solutions for cost-assignment i.e. objective 2, through a post processing mechanism. Once, cost assignment takes place, then solution reordering initiates based on solutions that satisfy both objective 1 and objective 2. Such candidate solutions which are inferior than others are eliminated. As mentioned in preceding section, cost is a multidimensional notion that can relate to economical cost, data accusation cost and computational cost of acquiring data for a feature, among others. Hence, higher the cost of a candidate solution, lower the desirability of that solution.

In the proposed methodology, both stages can operate independent of each other i.e. any evolutionary algorithm can be used to generate a set of promising solutions and a post-processing mechanism can embed cost-factor afterwards. Furthermore, the choice of classification model is entirely of the decision maker. In this study, we have selected a three-classifier ensemble for both label prediction and knowledge consolidation tasks. Label prediction deals with providing a predicted label against a test instance, whereas, knowledge consolidation integrates rules from a set of interpretable models and enables a decision maker to get insights into the underlying problem.

The main strength of the proposed methodology lies in generating a set of multiple candidate solutions in parallel for cost-aware feature selection. In this regard, a labelled classification dataset is sampled into a set of multiple partitions with replacement. The decision for the number of partitions is based on the number of computation cores in the processing system.

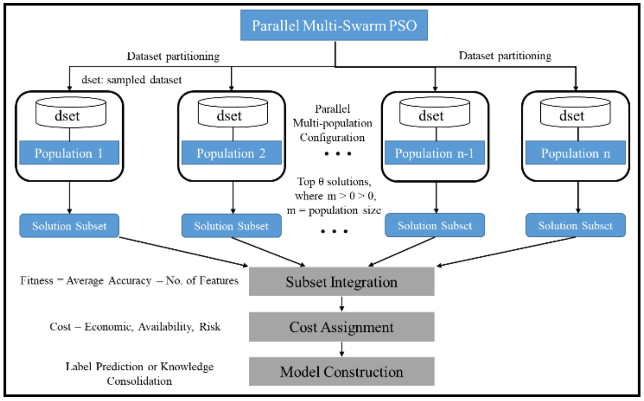


Figure 2. Proposed methodology for PSO-based cost-sensitive feature selection

PSO is a population based technique in which a set of candidate solutions are evolved in each iteration. In the proposed multi-swarm scheme, an original single-population swarm is divided into a set of non-overlapping sub-swarms. Each swarm is fed a sampled dataset which is horizontally partitioned. In this regard, each swarm evolves a set of solutions for the specific dataset partition. Since, a wrapper based FSS approach is adopted in this study therefore, it'd be relatively computationally expensive to evaluate a candidate solution as compared to evaluating a solution on a subset of data. Furthermore, in the proposed scheme, multiple-swarms can run in an independent manner, and at the end of the execution, solutions can be combined from multiple swarms, hence, providing a holistic view of the solution space while incurring computational cost of a single run of the PSO algorithm.

In this regard, there are two key objectives to evolve multiple swarms independently:

- Parallel execution using compute resources
- Solution diversity using multi-swarm configuration

Step-1 of the proposed methodology is detailed in Algorithm 1. As mentioned in the algorithm, a dataset is divided into a set of partitions, on each partition a PSO population will optimize solutions in an independent non-overlapping configuration.

Algorithm 1. Parallel Multi-Swarm PSO

```
Input:
    Dataset: Dset
Output:
    ConsolidateSolutions: List <particle, fitness>
Begin
  mSwarm ← Retrieve_Cores ()
  dSet ← Horizontal_Parition(mSwarm)
  Parallel |>
  Foreach swarm in mSwarm:
      Encoding (swarm) #Equation 2
      Initializing (swarm)
      Evaluating (swarm, dSet)
      gbest ← Locate-leader (swarm)
      While (termination != 'TRUE')
          Foreach particle in swarm:
              current = Position-Update (particle):
              Assign_Fitness (particle)
              IF current.fitness > pbest THEN:
                  Update (pbest)
              EndIF
          gbest ← Evaluate leader (swarm)
      EndWhile
  EndFor
  Parallel <|
  Solution <particle, fitness> ← SolutionConsolidate (mSwarm, k)
End
```

Step-2 of the proposed methodology is based on subset integration and cost assignment.

Subset Integration deals with combining all the unique feature subsets acquired from multiple swarms. Each swarm produces a set of solutions for a given dataset partition. Number of solutions are based on the size of a swarm. In this study, all the swarms carry equal number of particles.

Cost Assignment for a feature subset deals with averaging cost of each feature in a subset over the total cost of all the features, while cost computation of an individual feature is given in Eq. (1).

$$FC_i = \frac{Ecost_{fi}}{tot_{Mcost}} + \frac{Acost_{fi}}{tot_{Acost}} + \frac{Rcost_{fi}}{tot_{Rcost}} \qquad (1)$$

Where *Ecostfi*, *Acostfi*, and *Rcostfi* are the feature *i*'s economic cost, availability cost and risk cost, respectively. Normalized cost values are assigned to each feature. Each feature subset is associated with a cost factor which is between 0 and 1, inclusive. Higher the value, more would be the cost.

Model Construction is the last step of the methodology. It is based on decision maker's discretion to select a classifier which may serve for either label prediction or knowledge consolidation tasks. Figure 3. depicts one possible approach for model construction. Where "DT" refers to a decision tree model.

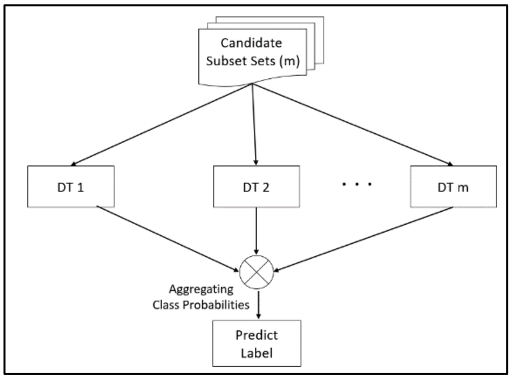


Figure 3. Ensemble based model construction for label prediction

An objective of a prediction model is to given insights regarding the underlying problem. In this case interpretable machine learning models are generally used. Decision tree is one of the popular interpretable machine learning models. An ensemble model based on decision trees can be constructed, where each decision tree is fed a reduced dataset. Each decision tree algorithm constructs a model which may be translated into a production rule of the form $IF\ \boldsymbol{\chi}\ THEN\ \boldsymbol{\Upsilon}$, where "$\chi$" represent a set of independent features, and "Υ" represent a dependent feature. Figure 4 depicts one of the configurations for creating an ensemble model for knowledge consolidation. In this knowledge consolidation approach a set of interpretable models are constructed such as shown in Figure 4. Each model is ranked based on its predictive accuracy score. Afterwards, rules are extracted. Traversing from the root node to one of its leaf node results in one conjunctive production rule. Once, rules are extracted from all the models, then contradictory and redundant rules are filtered out. Remaining rules may be re-evaluated based on a support-confidence threshold. The main intention is to elaborate on different approaches which may be utilized by the decision maker for model construction approach and a detailed discussion on knowledge consolidation process is not within the scope of this paper.

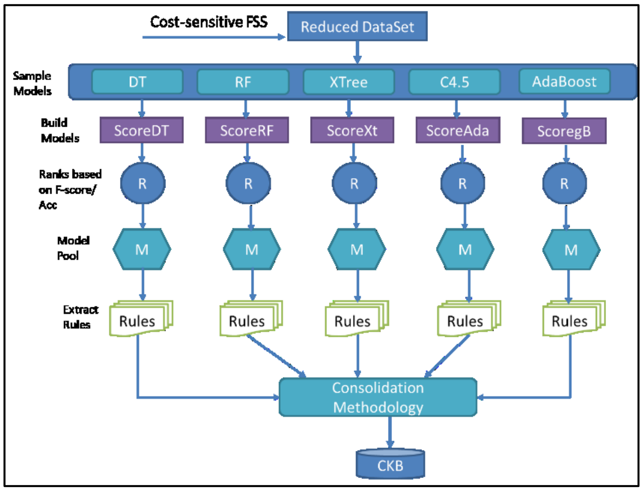


Figure 4. Ensemble model for knowledge consolidation

III. EXPERIMENTAION AND RESULTS

In this section, we will discuss about the dataset used for the experimentation of the proposed methodology. We have used chronic kidney disease (CKD) dataset, publicly available on UCI repository [8]. The dataset contains 400 patients record, where 250 patients are diagnosed with "ckd". Furthermore, dataset contains 25 features including the class label. Dataset is preprocessed for missing values and the continuous values are discretized using. Weka is used for executing PSO based feature selection method [9]. TABLE I provides details of the CKD dataset. Cost of feature is given by C-score using Eq. (1), and Feature score, F-Score, using Eq. (2). Symmetric uncertainty is used as a univariate measure for scoring each feature, this score is in turn used to obtain a percentile-rank of feature in order to show its importance in a given dataset.

$$FeatureScore_i = \frac{(\#Features - Rank_i)}{\#Features} \qquad (2)$$

Table 1. Chronic Kidney Disease dataset description

FID	Feature Name	Data Type	F- Score	C-Score
1	Age	Numeric	0.20	0.00
2	Blood Pressure (mm/Hg)	Number	0.06	0.01
3	Specific Gravity	Nominal	0.89	0.01
4	Albumin	Nominal	0.79	0.08
5	Sugar	Nominal	0.42	0.06
6	Red Blood Cells	Binary	0.42	0.11
7	Pus Cells	Binary	0.59	0.09
8	Pus Cells Clumps	Binary	0.41	0.09
9	Bacteria	Binary	0.24	0.13
10	Blood Glucose Random	Numeric	0.40	0.11
11	Blood Urea	Numeric	0.16	0.09
12	Serum Creatinine	Numeric	0.06	0.09
13	Sodium	Numeric	0.06	0.02
14	Potassium	Numeric	0.08	0.15
15	Hemoglobin	Numeric	0.73	0.01
16	Packed Cell Volume	Nominal	0.59	0.01
17	White Blood Cell Count	Numeric	0.16	0.09
18	Red Blood Cell Count	Numeric	0.87	0.09
19	Hypertension	Binary	0.88	0.01
20	Diabetes Mellitus	Binary	0.81	0.08
21	CAD	Binary	0.36	0.15
22	Appetite	Nominal	0.71	0.01
23	Pedal Edema	Nominal	0.65	0.01
24	Anemia	Nominal	0.54	0.09

In Figure 5, blue line represents error rate while orange line represents accumulated cost of a solution. For example, when population size is 5, then total number of features elected are 14, which incur error rate of 2.75 and total cost of 0.96. Based on preliminary experiments, as shown in Figure 5, we have selected population size of 15. Selected features are: 3,6,7,9,16,19,20,22,23,24.

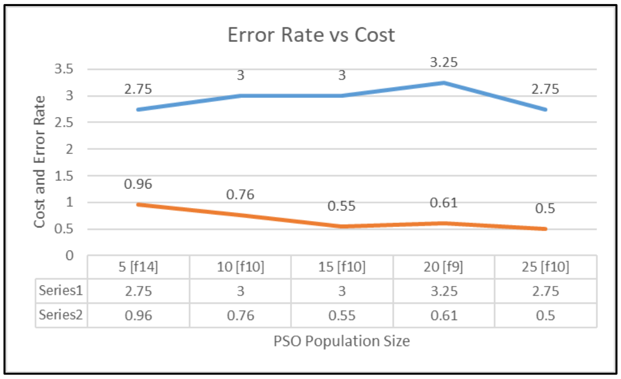


Figure 5. Effects of population size on error rate and solution cost

We have divided our original dataset into 4 sub-partitions with replacement. Each partition of 100 records is fed to a PSO population of 15 particles. Figure 6 depicts results of our reduced dataset. As it can be see that our proposed method has reduced the overall cost of the final solution from 2.6 to 0.55, while it also managed to reduce the overall error rate as well. Hence, it demonstrates that embedding a post-processing step for cost assignment selects features which are optimized not only for reducing the error rate but also yields a low-cost solution, hence, enhancing the overall applicability of the application as well.

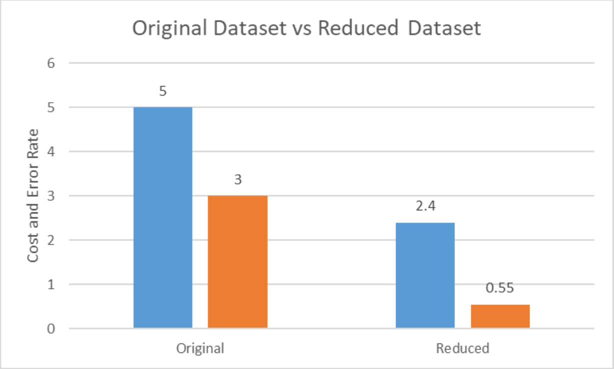


Figure 6. Original dataset vs reduced dataset

IV. CONCLUSION

Cost-sensitive feature selection has recently gained a lot of traction from the machine learning community. Since, in a number of application domains the traditional assumption of asymmetric cost is not valid. Therefore, enhanced feature selection techniques are proposed in order to account for the cost factor as well. In this research we have proposed a multi-swarm PSO methodology for cost-sensitive feature selection. This research demonstrates the effectiveness of our methodology in achieving solutions which are both efficient in terms of error rate and cost-effective.

This research can be advanced in a number of directions, such as the dominance relations among solutions can be extracted which would show the decision trade-off and therefore produce multiple solutions instead of a single solution. Another dimension of this research is to investigate in more detail the knowledge consolidation process in which multiple interpretable models are integrated in the form of production rules in order to extract non-trivial patterns from the dataset.

ACKNOWLEDGMENTS

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01629) supervised by the IITP(Institute for Information & communications Technology Promotion)"and IITP-2017-0-00655 and NRF-2016K1A3A7A03951968 and NRF-2019R1A2C2090504 and NRF-2019R1G1A1011296

REFERENCES

[1] Kodratoff, Yves. Introduction to machine learning. Elsevier, 2014.

[2] Bessa, M. A., et al. "A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality." Computer Methods in Applied Mechanics and Engineering 320 (2017): 633-667.

[3] Chen, Stephen, James Montgomery, and Antonio Bolufé-Röhler. "Measuring the curse of dimensionality and its effects on particle swarm optimization and differential evolution." Applied Intelligence 42.3 (2015): 514-526.

[4] Fong, Simon, Raymond Wong, and Athanasios V. Vasilakos. "Accelerated PSO swarm search feature selection for data stream mining big data." IEEE transactions on services computing 9.1 (2015): 33-45.

[5] Jiang, Liangxiao, Ganggang Kong, and Chaoqun Li. "Wrapper Framework for Test-Cost-Sensitive Feature Selection." IEEE Transactions on Systems, Man, and Cybernetics: Systems (2019).

[6] Zhang, Yong, et al. "Cost-sensitive feature selection using two-archive multi-objective artificial bee colony algorithm." Expert Systems with Applications 137 (2019): 46-58.

[7] García-Nieto, José, et al. "Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis." Information Processing Letters 109.16 (2009): 887-896.

[8] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[9] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.