17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

A Mixed Integer Optimisation Approach for Data Classification with Multiple Groups

Gang Xu, Nan Shao, Lazaros G. Papageorgiou

Centre for Process Systems Engineerng, Department of Chemical Engineering, University College London, Torrington Place, London WC1E 7JE, United Kingdom
Email: l.papageorgiou@ucl.ac.uk

Abstract

In this work, a mixed integer optimisation approach is proposed to solve the problem of data classification with multiple groups. An iterative solution procedure is developed to assign multiple boxes for each single class. The applicability of the proposed approach is demonstrated by two illustrative datasets. The computational results indicate that the optimisation-based framework is competitive in terms of prediction accuracy when compared with other standard classification models.

Keywords

Data Classification, Machine Learning, Hyper-box Representation, Mixed Integer Optimisation

1. Introduction

Data classification is one of the fundamental problems in machine learning and data mining. It involves the identification of patterns from training data and the membership prediction of newly sampled observation. Various classifiers have been used in many applications such as business aspects [1], flow regime identification [2,3] and fault diagnosis [4]. Initial approaches include linear discriminant analysis (LDA) [5] and k-Nearest Neighbor (k_NN) algorithm. Alternatively, neural networks have drawn more attentions because of their

ability to approximate nonlinear classification functions without any assumptions of training data distribution. A min-max neural network classifier was proposed [6]. N-dimensional fuzzy set hyper-boxes were determined by minimum and maximum points with a corresponding membership function. Moreover, support vector machines (SVM) [7] approach was applied to many practical data classification problems including flow regime identification and protein secondary structure prediction. SVM provides hyper-planes with the maximum separating margin to discriminate two classes of training samples. Kernel functions are incorporated to design nonlinear classification functions. Finally, classification models can be developed by mathematical programming (MP) techniques without knowing any assumption of group distribution. Discriminant function is initially generated as linear programming (LP) models [8, 9]. A mixed integer programming model (MIP) was then proposed to extend LP representations [10]. Binary variables were introduced to indicate whether training samples are correctly classified. The total number of correctly classified samples was maximised. Recently, Sueyoshi addressed a series of non-parametric discriminant analysis approaches for two-class and multi-class data classification problems [11]. Glen applied piecewise linear classifiers to approximate nonlinear discriminant functions [12]. Finally, Uney and Turkay [13] proposed a mixed integer linear programming (MILP) model using hyper-box representations.

In this paper, a mixed integer optimisation approach for the multi-class data classification problem is presented by generalising our previous work on process plant layout [14] to *M*-dimensions (where *M* is the number of attributes used for data classification). The proposed approach is also based on a hyper-box representation, which is similar to the one developed by Uney and Turkay [13]. In the next section, a brief description of the proposed approach is provided. An iterative solution algorithm is introduced in section 3 and a testing procedure is described in section 4. Two illustrative datasets are tested in section 5 to demonstrate the applicability of our methodology. Finally, some concluding remarks are made in section 6.

2. Model Description

Consider a multi-class data classification problem with *C* classes and *S* training samples. Each sample is characterised by *M* independent attributes. The class membership of each sample is known. The proposed approach is based on a MILP representation. Hyper-boxes with *M* dimensions are adopted to recognise the patterns hidden in the training data samples. Data enclosing constraints are applied to determine the optimal dimensions and locations of each hyper-box so as to cover the maximum number of correctly classified samples. Non-overlapping constraints are used to avoid hyper-boxes from different classes occupying the same location. The objective function used is the minimisation of the total number misclassified samples. It should be mentioned that the

proposed MILP representation assigns only one hyper-box to each class. Multi-boxes solution algorithms will be introduced in the next section to improve the training and testing accuracy.

3. An Iterative Solution Algorithm

In this section, an iterative solution procedure is proposed to assign N $(N \geq C)$ hyper-boxes to classify C groups of data samples. After allocating one hyper-box to each class by solving the single level MILP model described in section 2, new boxes are introduced to capture any misclassified samples during previous iterations and the modified MILP model with more hyper-boxes is then solved. The algorithm will terminate when the objective functions of two successive iterations have the same value. It should be noted that when a new box is added, the non-overlapping conditions are activated only for those boxes which belong to different classes. Therefore, potential overlapping happens between boxes that belong to the same class but not for boxes with different class memberships. Next, the following sets are defined for the description of the iterative algorithm:

Sets

H	Set of hyper boxes that belong to the same class
Δ	Set of misclassified samples
i_s	Hyper-box which sample s belongs to

The steps of the proposed approach are outlined below:

Step 1: Initialise $\Delta = \phi$, $H = \phi$, $N=C$.

Step 2: Solve the single level MILP.

Step 3: Identify samples outside hyper-boxes. Update Δ.

Step 4: Add one more box for each class to samples in Δ. Update N, H, i_s.

Step 5: Formulate new MILP problem with more added boxes. Non-overlapping constraints and variables are generated for i and j $\notin H$.

Step 6: Solve the modified MILP model using updated N boxes.

Step 7: If the objective function values of two successive iterations are the same, STOP; otherwise, go to STEP 3.

4. Testing Procedure

An important task for any classification method is its ability to perform a successful prediction based on the patterns captured through the training process. According to our hyper-box approach, the distances between the new

testing sample *s* to all established hyper-boxes are calculated. If sample *s* is within one of the hyper-boxes, its membership is identified directly as the class that is represented by the hyper-box enclosing the sample. If the sample is outside all existing hyper-boxes, sample *s* will be classified to the nearest one.

5. Computational Results

Two real datasets are used in this section to evaluate the applicability of the proposed methodology. The first example introduced by Sueyoshi [11] is associated with the bankruptcy of firms in US electric power industry. This dataset includes 61 non-default firms (group 1) and 22 default firms (group 2). The performance of each firm is determined by 13 independent financial ratios. The second dataset reflects the flow regime map of gas-liquid, two-phase flow in microsystems. This dataset collects 115 experimental data samples covering 5 flow regimes (Bubbly, Churn, Slug-Annular, Bubbly-Slug and Slug). The flow pattern of each sample is identified by measuring the superficial velocity of gas and liquid phases (this dataset shown in Figure 1 is provided by Dr. P. Angeli, UCL, through personal communication). The computational results from the iterative MILP approach are compared with five other standard classifiers including LDA, k_NN, NN and two MILP formulations for data classification with multiple groups proposed by Gelhrein [10] and Sueyoshi [11] (see Tables 1 and 2). The testing performances of all classification methods are compared through the following three themes:

Scenario A: 70% of the samples of each class are extracted randomly for training and the rest are used for testing.

Scenario B: 70% of the complete data samples are selected randomly for training and testing is applied to the remaining samples.

Scenario C: leave-one-out scheme. Each sample is dropped out for testing after training the remaining samples.

The proposed mixed integer optimisation approach is implemented in GAMS [15] using CPLEX mixed integer optimisation solver with 1% margin of optimality. LDA and k-NN are performed by MASS and class packages using the statistical computing language R (http://www.r-project.com). All neural network classification are applied using the weka open source machine learning software (http://www.cs.waikato.ac.nz/ml/weka/) with the following parameter settings: Model: Multi-layer Perceptron, Number of Hidden layers: 2, Learning Rule: Momentum (0.7), Step Size: 0.1, Maximum Number of Epochs: 10000, Weight Update Method: Batch Learning and Termination Method: Cap the number of epochs. Because of the random nature of scenarios A and B, both schemes are repeated 50 times and the mean prediction accuracies for all six classification methods are reported. The best testing performance in each scenario is indicated in bold.

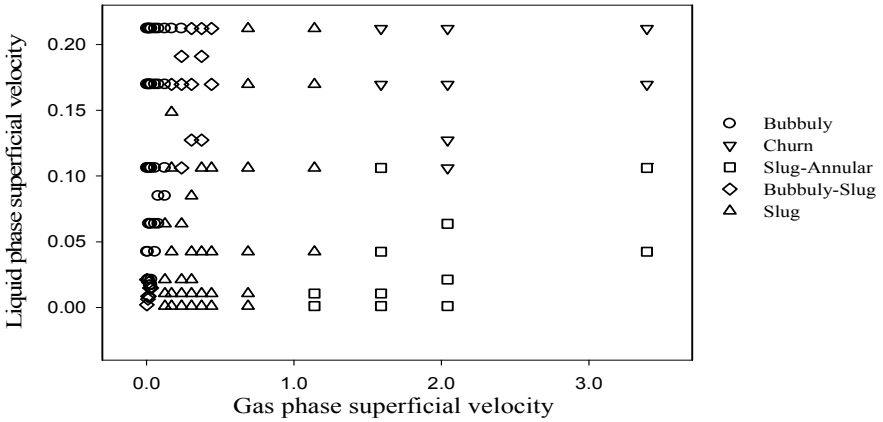

Figure 1. Gas-liquid two phase flow regime in microsystems

Table 1. Computational results for Example 1

Model	Scenario A	Scenario B	Scenario C
Our work	**92.67%**	**91.67%**	**91.57%**
Gelherin (1986)	86.75%	84.67%	81.93%
Sueyoshi (2006)	88.50%	89.25%	89.16%
k_NN	89.16%	89.44%	89.15%
LDA	89.68%	90.24%	90.36%
NN	91.25%	90.99%	91.56%

Table 2. Computational results for Example 2

Model	Scenario A	Scenario B	Scenario C
Our work	**80.29%**	**80.70%**	80.87%
Gelherin (1986)	80.17%	79.47%	78.26%
Sueyoshi (2006)	41.70%	39.65%	43.48%
k_NN	79.23%	79.05%	**81.74%**
LDA	67.53%	66.06%	75.65%
NN	72.84%	71.43%	71.30%

In the first dataset, our work outperforms other classification models in terms of all three different scenarios; achieving prediction accuracy of above 91%. NN approach also shows its ability to achieve good prediction accuracy because of its adoption of nonlinear discriminant functions (see Table 1).

The computational results of the second dataset for all presented methods indicate that the flow pattern of an experimental sample in microsystems can be successfully predicted by our approach with more than 80% accuracy. In most cases, our method still gets the best prediction accuracy among all six classifiers in terms of three testing scenarios (see Table 2).

6. Conclusions

An efficient mixed integer optimisation approach has been proposed to solve the classification problem with multiple groups. Hyper-boxes are used to enclose training samples which belong to the same class. In order to improve the training and testing accuracy, an iterative solution procedure has been presented to assign multiple boxes for each class. The memberships of new samples have been identified by calculating the distances between testing samples to all established hyper-boxes. Finally, the applicability of the proposed methodology has been demonstrated through two illustrative datasets. The prediction performance of our approach has been compared with five other standard classifiers over three different scenarios. The computational results indicate that our approach is competitive in terms of prediction accuracy when compared with other alternative classification methodologies.

Acknowledgements

The authors gratefully acknowledge Dr. P. Angeli for providing the flow regime dataset. GX acknowledges support from the Centre for Process Systems Engineering.

References

1. T. Sueyoshi, Euro. J. Oper. Res., 152 (2004) 45.
2. L.A. Tarca, B.P.A. Grandjean and F. Larachi, Chem. Eng. Sci., 59 (2004) 3303.
3. T.B. Trafalis, O. Oladunni and D.V. Papavassiliou, Ind. Eng. Chem. Res., 44 (2005) 4414.
4. L.H. Chiang, M.E. Kotanchek and A.K. Kordon, Comput. Chem. Eng., 28 (2004) 1389.
5. R. Fisher, Ann. Eugenics., 7 (1936) 179.
6. P.K. Simpson, IEEE. T. Neural. Networ., 3 (1992) 776.
7. C. Cortes and V. Vapnik, Mach. Learn., 20 (1995) 273.
8. N. Freed and F. Glover, Eur. J. Oper. Res., 7 (1981) 44.
9. N. Freed and F. Glover, Decis. Sci., 12 (1981) 68.
10. W.V. Gehrlein, Oper. Res. Let., 5 (1986) 299.
11. T. Sueyoshi, Eur. J. Oper. Res., 169 (2006) 247.
12. J.J. Glen, J. Oper. Res. Soc., 56 (2005) 331.
13. F. Uney and M. Turkay, Eur. J. Oper. Res., 173 (2006) 910.
14. L.G. Papageorgiou, and G.E. Rotstein, Ind. Eng. Chem. Res., 37 (1998) 3631.
15. A. Brooke, D. Kendrick, A. Meeraus and R. Raman, GAMS: A user's guide GAMS development Corp. Washington, DC (1998).