

## A Refinement Model with Information Granulation Focused on Difficult to Distinguish Cases

Plamena Andreeva, Plamen Andreev, Maya Dimitrova, Petia Radeva

**Abstract:** *The paper proposes a different approach to data modeling. Analogous to the rejection method, where the misclassifications are removed and manually evaluated, we focus here on difficult to distinguish cases for binary classification. Such cases are further explored and information granulation is conducted based on the idea to stretch out the interesting intervals. This refinement model adopts the concept from database theory where the  $n \times n$  relations are resolved through additional internal attribute connections. We introduce an integration of target functions for such cases. The achieved experimental results from a test dataset are described and future work for user assistance in knowledge modelling is presented.*

**Key words:** *Information Granulation, Modelling, Decision Boundary, Classification, Machine Learning.*

### INTRODUCTION

The problem of learning from data has attracted increasing amount of interest recently [1, 3, 8, 9]. Here we mainly consider binary classification to model the behaviour of a system with its input and output variables. The purpose is the automatic detection of one specific type of outcomes. The new work on classification exploits more flexible classes of models, many of which attempt to provide an estimate of the joint distribution of the features within each class, which can in turn provide a classification rule. Statistical approaches are generally characterised by having an explicit underlying probability model, which provides a probability of being in each class rather than simply a classification. Other techniques, such as genetic algorithms and inductive logic procedures [4], are currently under active development and allow dealing with more general types of data, including cases where the number and type of attributes may vary, and where additional layers of learning are superimposed, with hierarchical structure of attributes and classes and so on.

The aim is to use not just cases with attributes matching exactly those of the given example, but near them in some sense. The minimum error decision rule would be to allocate to the most frequent class among these matching examples. Partitioning algorithms and decision trees in particular [7], divide up attribute space into regions of self-similarity: all data within a given box are treated as similar, and posterior class probabilities are constant within the box. For discrete variables the range of the smoothness parameter is the interval [0,1]. One extreme leads to the uniform distribution and the other to a one-point distribution.

### PROBLEM STATEMENT

When the classes of “positive” and “negative” diagnosis are largely imbalanced, operational curves are used to optimize the trade-off between false positive and false negative rates. Misclassifications can be drastically reduced at the expense of a small fraction of missed. This implies an imbalanced or a cost-sensitive problem [2]. In such problems, even a small error rate results in an unacceptably large number of false positive classifications. We propose a different approach to data modeling. Analogous to the rejection method [6], where the misclassifications are removed from the model and are given to experts for manual evaluation, we focus here on difficult to distinguish cases for binary classification. Such cases are further explored and information granulation is conducted based on the idea to stretch out the neighbourhood intervals of interest. This refinement model adopts the concept from database theory, where the  $n$  by  $n$  relations are resolved through additional internal attribute connections. We introduce an integration of target functions for such cases, which combine the local and global learning. The aim is to

help the expert by identifying cases in the diagnostic system, which contain “positive” with high probability thereby greatly reducing the inspection time. The proposed model gives an insight to the ambiguous cases.

**SUPPORT VECTORS**

The support vector machine (SVM) is a widely used tool for classification. Many efficient implementations exist for fitting a two-class SVM model [8, 9]. The user has to supply values for the tuning parameters: the regularisation cost parameter, and the kernel parameters. It seems a common practice to use a default value for the cost parameter, often leading to the least restrictive model.

For a set of  $n$  training pairs  $x_i, y_i$ , where  $x_i \in \mathfrak{R}^n$  is a vector of real-valued attributes for the  $i^{th}$  observation, and  $y_i \in \{-1, +1\}$  codes its binary response. The goal is to estimate a linear decision function in a 2 dimensional plane as illustrated on fig.1.

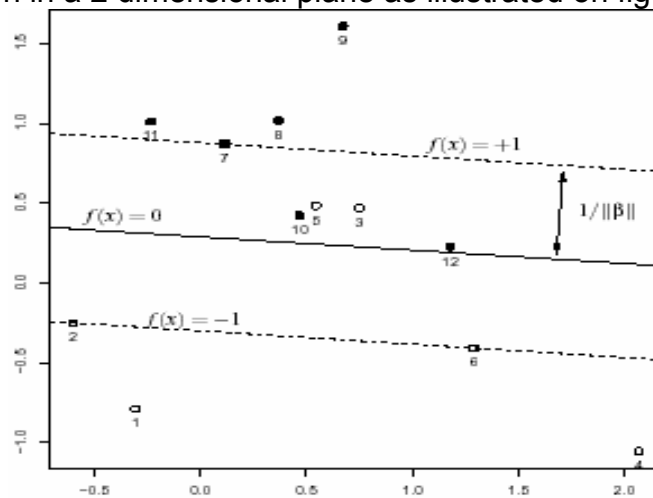


Figure 1. SVM model with parameter  $C=2$ , and the width of the margin is  $2/||\beta|| = 0.587$ . There are 2 misclassifications at point {3} and {5}, while three points are exactly on the margin {2}, {6}, {7}.

There are many ways to fit such a linear classifier, including linear regression, Fisher’s linear discriminant analysis and logistic regression. It is shown that  $1 / ||b||$  ( $b_0 + x^T_i \beta$ ) is the signed distance from  $x_i$  to the decision boundary. When the data are not separable, this criterion is modified. The cost parameter  $C$  controls the amount of overlap. If the data are not separable, as  $C$  gets large, the solution approaches the minimum overlap solution with largest margin, which is attained for some finite value of  $C$ .

The error here is defined as the difference between the predicted value  $f(x)$  and the actual value  $y$  of the outcome for any point  $(x; y)$ . It could be the squared residual or absolute deviation depending on the error metric employed. The optimal separation hyperplane is the one that has the largest margin  $M$  on given dataset. A separating hyperplane with the largest margin (defined by  $M = 2 / ||w||$ ) specifies support vectors, (training data points closest to it), which satisfy

$$y_i \left[ w^T x_i + b \right] = 1, \quad \text{for } j = 1, \dots, N_{SV} \tag{1}$$

where  $b$  is a parameter and  $N_{SV}$  is the number of support vectors.

**REFINEMENT MODEL**

The construction of classifiers is different. While bagging builds the individual classifiers separately, boosting builds them sequentially such that each new classifier is influenced by the performance of those built previously. Our method differs from these traditional classifiers in the management of the original training data. We propose an iterative design method, where the most informative attribute (in our case it is the 13<sup>th</sup> from heart dataset as shown on fig.2) is further stretched, so the interval is magnified. This

is opposite to the normalization, where all the attributes values are proportionally represented (scaled) to the same range.

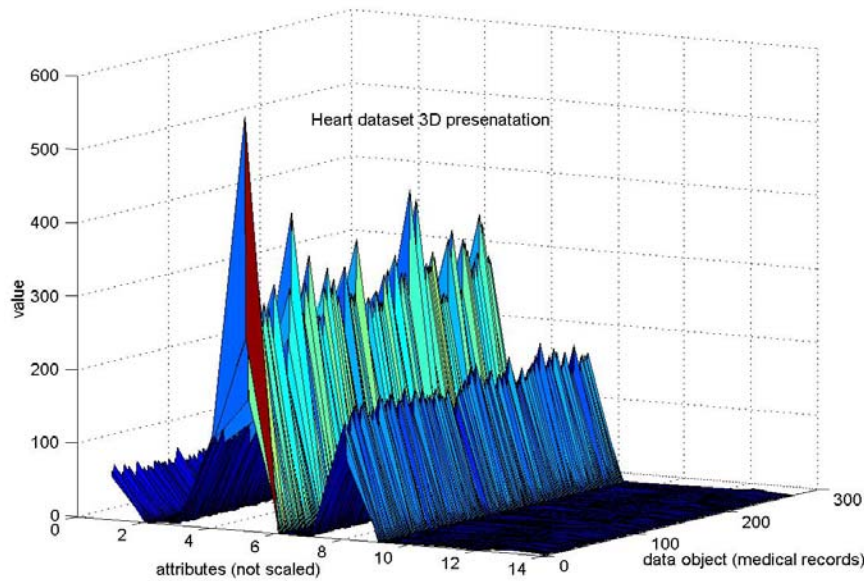


Figure 2. The attributes values distribution for the experimented dataset.

After the refinement, in the second pass of data modelling we fuse the learning criterion for the target function and so we combine the local and global functional  $J(f)$ . For example, setting  $J(f)$  to be the minimising the mean-squared error :

$$J(f) = \min_{w,b,\xi} \sum_i C_i \xi_i + \frac{1}{2} \|w\|^2 \quad (2)$$

with slack variables  $\xi$  for the number  $i=1, \dots, N$  of the whole training dataset.

Now we obtain the fused (mixed)  $J^*(f)$  as:

$$J^*(f) = \begin{cases} \min_{w,b,\xi} \sum_i C_i \xi_i + \frac{1}{2} \|w\|^2, & \text{if } I \text{ pass SVM} \\ \max_j [(1-f(x_j))^2 - d_j^2], & \text{if refinement on attributes} \end{cases} \quad (3)$$

where  $d_j$  is the distance of the data point  $x_j$  and the neighbours. This gives us optimal solutions without solving the optimization problems many times for each value. After the 1<sup>st</sup> pass of the modelling process we only concentrate on the specific attributes classification. This method generalizes the results of Hastie [4] to the case of asymmetric cost functions.

The refinement model is intuitively following the empirical diagnosis, when the doctor first focuses on the most difficult symptoms (in our study they are the attributes), then is asking more specific questions about the detailed scope of this attribute.

### INFORMATION GRANULATION ON DIFFICULT CASES

In supervised classification algorithms, a classifier is used to predict the class labels of unseen test data. These algorithms can be called global learning algorithms since the classifier is built based on the whole training dataset. In contrast, in local learning algorithms for a given test data point, a classifier is trained only with its neighbouring training data, and then the given test data point is classified by this locally learned model. It has been reported that local learning algorithms often outperform global ones [3] since the local models are trained only with the points that are related to the particular test data.

In our refinement model we use this approach but apply a mixed regularization based on the specific attribute domain. So we import the influence of its value. We are interested here only in the universe of discourse for the selected attribute. Our method keeps both the size of the original data and the attributes' values unchanged throughout the whole process. As a result in (3) our rules will always reflect precisely the nature of the original data. The results are based on two measures: test error numbers—the number of misclassifications on independent test samples and the error numbers of 10-fold cross validation. Support vector machines, though accurate, are not preferred in applications requiring high classification speed, due to the large number of support vectors.

## EXPERIMENTS AND RESULTS

For further comparison, we also used SVM and 3- nearest neighbours to conduct the same 10-fold cross validation. SVM achieved 100% accuracy, but SVM used all the 13 input features together with 40 support vectors and kernel evaluations in its decisions. It is difficult to derive understandable explanations of any diagnostic decision made by this system. In contrast, our method used only 5 trees and less than 17 rules. The 3-nearest neighbours classifier made 15 mistakes (5 in heart attack and 10 in normal). If only the top 2, 5, or 10 ranked features are used, the support vector machines did not achieve the perfect 100% accuracy; our method did not achieve the perfect 100% accuracy either.

The strategy is to start with a relatively large value of  $M$  and find all models of size  $M_L$  and less. In order to decrease  $M$  solutions that minimise the size and give a refinement of the difficult cases are found. The starting parameter values for the numerical search in each model are the solution values for the  $M$  most important terms of the previous pass. The stored results of the previous mining serve as background knowledge for consecutive data mining tasks.

To generate the training set, the total of the 180 records are sampled without overlap, so that tiling them together in the correct order will recover the entire given diagnose. The main software package used in the experiments is the free ML environment Weka version 3.4 available from [5].

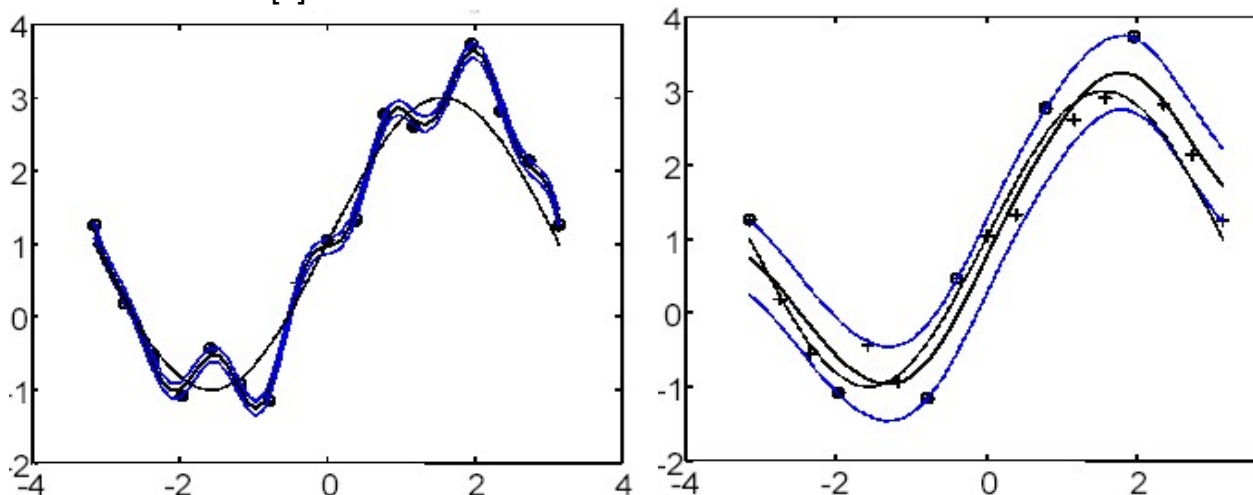


Figure 3. The experimental data given by circles, and the learning function is the solid line. On the right is the classification model after the 1<sup>st</sup> pass, and on the left is the refinement model.

The supervised learning algorithms can also result in good performance. In fact, local learning regularization is a flexible framework, which offers a possible way to adapt various supervised learning techniques to classification problems as long as the resulting local model can be expressed in the form of (3).

From the results it can be seen that performance based on the local model compares favourably to the other regularizers. Correspondingly the SVM has good performance where the data are close to a low-dimensional manifold embedded into high-dimensional space. These observations suggest that the property of the local model used by each regularizer can strongly affect the final classification performance.

Two heuristics can be used that can be derived from a SVM classifier trained on points with known labels. The “largest positive” heuristic selects the point that has the largest classification score among all examples still unlabeled. The “near boundary” heuristic selects the point, where the classification score has the smallest absolute value. Although the semantics of these two heuristics differ, in one of the cases we are trying to explore the space of positive examples as fast as possible, whereas in the other case the effort is focused on learning the boundary – in both cases the SVM has to be re-trained after each selection. In the original application of [7] the data samples were relatively small, therefore the SVM can be re-trained from scratch after addition of new points. Consequently, a better way to proceed is by applying refinement learning as presented in this paper.

Active learning is another powerful technique by means of which learning can be efficiently carried out in large data sets with limited availability of labels. *Active* means that instead of having all data labelled beforehand, an algorithm “*actively*” chooses examples for which labels must be assigned by a user. Various applications of the algorithm, such as in accident discovery, online monitoring of industrial devices, and surveillance of network traffic, can be predicted.

## CONCLUSIONS AND FUTURE WORK

In this study we propose an approach for data modelling which is suitable for difficult to distinguish cases and combines the local and global learning through integrating the granulated information. This approach intuitively mimics the expert’s decision making where the human is going into details when suspicious measurement are available. The refinement model requires that the label of each data point should be well estimated based on its neighbours. Comparison with related approaches illustrates that local learning regularization is a flexible framework, under which we can examine some current regularizers for classification. It also allows us to adapt various learning algorithms.

The study of the data refinement through the stretching out the interesting intervals and combining the target functions seems to be robust. The proposed model is better because it picks a better representation within a desired complexity since it is using a learning-related criterion to select an interesting representation instead of a merely geometric one as in SVM.

It is interesting to compare our approach of learning to recently proposed alternative approaches to online learning. There are a lot of still open problems and questions in data modelling and we plan to investigate further some diagnostics of unexpected failure in engineering and economics as well. A pragmatic area for our research will be applications to real-world problems.

## ACKNOWLEDGEMENT

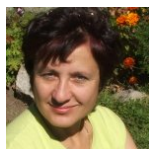
This work is partially supported by the NSRF of the Bulgarian Ministry of Education and Science as part of the Research Project № MI - 1509/2005 “*Multimodal User and Sensor Interface in a Computer System for Cardiological Diagnosis and Intervention*”.

## REFERENCES

[1]. Andreeva, P., M. Dimitrova. Methods for rule extraction in knowledge based information systems using learning models. *Int. Conf. Automatic and Informatics'2004*, 6-8 Oct., 2004, 199-203.

- [2]. Andreeva P., M. Dimitrova and A. Gegov. Information Representation in Cardiological Knowledge Based System, *SAER'06*, 23-25 Sept., 2006, 266-272.
- [3]. Joachims T. Transductive learning via spectral graph partitioning. In T. Fawcett and N. Mishra, eds, *Proc. 20th International Conference on Machine Learning*, AAAI Press 2003, 290–297.
- [4]. Hastie T., S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5. 2005, 1391–1415
- [5]. University of Waikato, *ML Program WEKA*, [www.cs.waikato.ac.nz/ml/weka](http://www.cs.waikato.ac.nz/ml/weka). 2007
- [6]. Vilarino F., L. Kuncheva, P. Radeva, ROC curves and Video Analysis Optimization in Intestinal Capsule Endoscopy, *Pattern Recognition Letters*, 27, 2000, 875-881.
- [7]. Witten I. H. Generating Accurate Rule Sets Without Global Optimization, in Shavlik, J., ed., *Machine Learning: Proc. of 15 Int. Conf.*, Morgan Kaufmann. 1998.
- [8]. Wu M., B. Schölkopf. Transductive Classification via Local Learning Regularization, *Proc. 23th Int. Conf. on Machine Learning*, 912–919. AAAI Press, 2006.
- [9]. Zhou D., O. Bousquet, T. N. Lal, J. Weston, B. Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf, eds, *Advances in Neural Information Processing Systems*. MIT Press, 2004.

### ABOUT THE AUTHORS



Res. Assoc. Plamena Andreeva (MSc. in Electronics) is with the Department of Knowledge Based Control Systems at the Institute of Control and System Research – Bulgarian Academy of Sciences, She is interested in DM, classification and clustering algorithms for practical application of these techniques. Sofia, Phone: +359 2 870 03 37, E-mail: [plamena@icsr.bas.bg](mailto:plamena@icsr.bas.bg), [andreeva.plamena@gmail.com](mailto:andreeva.plamena@gmail.com)



Eng. Plamen D. Andreev, technical expert at Air Traffic Service Authority, Direction Operating Systems, Department of Communication, Airport Sofia. He is mainly interested in simulating and modeling of traffic data, fault prediction and diagnostic. Sofia, Phone +359 2 937 43 43, E-mail: [pol\\_andreev@yahoo.com](mailto:pol_andreev@yahoo.com)



Dr. Maya Dimitrova is a research fellow at the Department of Hybrid Systems and Management of the Institute of Control and System Research – Bulgarian Academy of Sciences. Her main research interests include user-adaptive web design, web genre visualization, cognitive and adaptive human-computer interfaces, neural networks and information retrieval. Sofia, Phone +359 2 979 20 54, E-mail: [dimitrova@icsr.bas.bg](mailto:dimitrova@icsr.bas.bg)



Assoc. Prof. Petia Radeva (Dr. in Informatics). She is with the Computer Vision Center, Autonomous University Barcelona, since 1991 as lecturer and researcher. Her research interests include image recognition, clustering and classification. Spain, Barcelona, Phone +34 9 358 118 62, E-mail: [petia@cvc.uab.es](mailto:petia@cvc.uab.es)