Data Modelling and Specific Rule Generation via Data Mining Techniques

Plamena Andreeva

Abstract: *Data Mining techniques are useful for analyzing data from many different dimensions and for identifying relationships. Non-parametric data models are explored and a heuristic approach is proposed for specific rule generation in practical cases. The most suited algorithm for a specific application is presented and learning methods evaluation is given. A problem of feature extraction and specific rule inferring from heart diseases data set is considered and experimental results are presented and discussed.*

Key words**:** *Machine Learning, Data Mining, Diagnostic Inference, Rule generation, Modelling.*

INTRODUCTION

Information technology development over the last years grows rapidly and alters from single use centralized systems to distributed, multi purpose systems. In such systems a useful tool for processing information and analyzing feature relationships is needed. Data mining (DM) technique has become an established method for improving statistical tools to predict future trends [3, 8]. There are a huge variety of learning methods and algorithms for rule extraction and prediction. Data mining (or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information.

The aim is to achieve fast and simple learning models that result in small rule bases, which can be interpreted easily. In this particular study different data models are explored and evaluated by the test accuracy. For training the model non-parametric density estimation is used for improving the initial accuracy. First the unsupervised learning is conducted, and then a heuristic from experts is applied for specific rule generation. In the last section visual results from the experiments are presented and discussed.

PROBLEM STATEMENT

Detecting a disease from several factors or symptoms is a many-layered problem that also may lead to false assumptions with unpredictable effects. Therefore, the attempt of using the knowledge and experience of many specialists collected in databases to support the diagnosis process is needed. The goal is to obtain simple intuitive models for interpretation and prediction. The advantage of combining such simple learning density functions and feature selection mechanism is that the resulting relational model is easy to understand and interpret [2]. Preliminary testing shows that knowledge extracted from heart diseases data can be efficiently used for classification of diagnosis.

If we make the rules more general, a greater number of the cases can be matched by one or more of the rules. To minimize their number some of the features are removed. The specific rule generation is based on pruned decision tree, where the most expressive attribute is increasingly weighted. The determination of the number of clusters is a central problem in data analysis.

In the conducted experiments the collected data records are preprocessed (scaled, cleaned) and classified. Each measurement is presented as a pixel in multidimensional space and data points are mapped by means of a *Gaussian kernel* to a high dimensional feature space, where the minimal enclosing sphere can be calculated. When mapped back to data space this sphere can be separated into several components, each enclosing a cluster of points. Separating the classes with a large margin minimizes the bound on the expected generalization error. In the case of non-separable classes, it minimises the number of misclassifications whilst maximizing the margin with respect to the correctly classified examples. Unlike other algorithms, it makes no assumptions about the

relationships between a set of features (attributes) in a feature space. This allows us to identify and determine the most relevant features used in a model and the model's feature dependencies. As a result, non-linear modelling is done very accurately and classifiers are automatically generated. ML tuning methodology does not make any assumptions about correlation between features, as opposed to techniques that assume statistical independence.

USEFULNESS OF THE MODEL

If the goal is not just to represent the data set but also to *make inferences* about its structure, it is essential to analyze whether the data set exhibits a clustering tendency, as sated in [6]. The results of the cluster analysis need to be validated. A potential problem is that the choice of the number of clusters may be critical. Good initialisation of the cluster centroids may also be crucial; some clusters may even be left empty if their centroids lie initially far from the distribution of data. The Bayesian rule is the optimal classification rule [7] but only if the underlying distribution of the data is known.

We have included into our DM analysis frequently used algorithms of estimating parameters of non-supervised classifiers as well as methods of empirical segmentation and heuristic rule extraction [1]. One of the most important data mining tools is visualization of the available information, especially of multidimensional data. The visualization of several attributes in one computer screen is implemented for the visual heuristic analysis of correspondence between estimated parameters class value. Here we use standard methods of 2d and 3d graphics embedded in WEKA shell [8]. The visual class relations for the first 4 attributes of the heart example dateset are shown on fig.1.

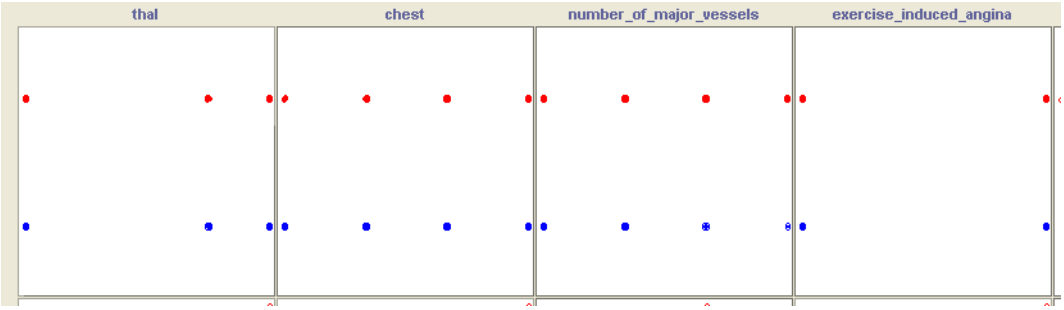

Figure 1. Representation of „thal", „chest", „n_major_vessel" and „ex_angina" attributes in relation to Class (on Y axis) for the Heart dataset.

Standard methods used in data mining are principal component analysis and Kohonen' self organizing maps (SOM) [4, 5]. However, the component analysis is a linear projection method not always well representing the structure of multidimensional data. SOM is not suitable to visualize large sets of multidimensional data.

Parametric techniques rely on knowledge of the probability density function of each class. On the contrary, non-parametric classification does not need the probability density function and is based on the geometrical arrangement of the points in the input space. We apply a non-parametric technique, *k-nearest neighbors* to verify the discriminability of the different feature spaces. Since non-parametric techniques have high computational cost, we make use from some expert's assumptions that lead to dimensionality reduction. The estimation of the local probability density at each point in the feature space is first calculated and then a minimal risk based optimisation is conducted. The density estimate group contains: k-nearest neighbour; radial basis functions; Naive Bayes; Polytrees; SOM; LVQ; and the kernel density method. After the optimal model is selected, the test set is run and compared. The accuracy and precision are calculated and results are given in table1.

When using non-linear RBF model the correctly classified cases are 84.07%. This outperformed the linear model, which did with an average accuracy of 75.4%. Compared against a Naïve Bayes, which achieved an average test accuracy of 78.6%, the kernel

Table 1. Model accuracy comparisons for the examined heart dataset.

Model	Test accuracy	Precision	T Positive Rate	Expert refinemnt
PART C4.5	75.738 %	0.757	0.767	81.28 %
Naïve Bayes	78.563 %	0.795	0.800	84.24 %
Decision Table	82.4348 %	0.841	0. 877	84.33 %
Neural nets	82.773 %	0.840	0.840	N/A
Voted perceptron	83.704 %	0.844	0.793	83.74 %
SMO	84.074 %	0.845	0.873	N/A
RBF Gaussian	84.074 %	0.845	0.873	85.31 %
Repeated Inc Pruning	84.3576 %	0.823	0.813	81.33 %
Kernel density	**84.4444 %**	**0.880**	**0.800**	**87.67 %**

density algorithm is the optimal non-linear model selected on the training set (with Density (precision of 0.88) achieved test accuracy of 84.44% which is the best result in the experiments. This is at least partially due to the use of 10-fold cross validation and to a model that generalizes well. The auto-training approach for selecting the optimal model requires finding the optimal combination of all parameters.

The *decision-tree method* like the *nearest-neighbours* method, exploits clustering regularities for the purposes of classifying new examples. It constructs a decision-tree representation of the data and provides a hierarchical description of the statistical structure of the data. It shows implicitly which variables are more significant with respect to classification decisions. Most clustering methods based on heuristic are approximate estimation for particular probability models.

LEARNING MODELS

The basis of the model consists in viewing a numeric value, i.e. measure as being dependent on a set of attributes, dimensions. Each classifier uses its own representation of the input pattern and operates in different measurement systems. A well-known approach is the weighted sum, where the weights are determined through a Bayesian decision rule. Regression is the oldest and most well known statistical technique (for continuous quantitative) that the DM community utilizes. For categorical data (like colour, name or gender) DM technique is successfully used [9]. This technique is much easier to interpret by human.

If the resulting attribute distribution is broad and flat we know that the partial observation does not contain sufficient relevant information to predict this attribute. If the distribution has a sharp single peak we can predict the attribute value with confidence.

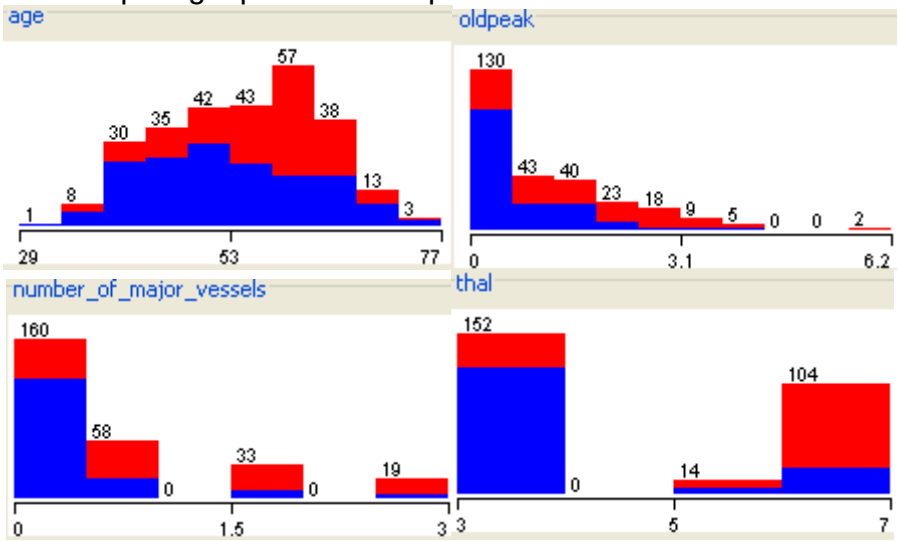

Figure 2. The most relevant attribute distribution („thal") is used for diagnosis prediction.

The distribution's visualization for the first 4 important attributes is given on figure 2. The most relevant attribute used for diagnostic prediction is „thal", obtained from experts. The effects of noise and deviation from the normal distribution in the data pose natural limitations to both methods' prediction capabilities. Most clustering methods based on heuristic are approximate estimation for particular probability models. The goal of the described data mining techniques is to aid the development of a reliable model.

SPECIFIC RULE EXTRACTION

The default rule relies only on knowledge of the prior probabilities, and clearly the decision rule that has the greatest chance of success is to allocate every new observation to the most frequent class. However, if some classification errors are more serious than others we adopt the minimum risk (least expected cost) rule and the class C_k is that with the least expected cost.

A rule-set set is formed from C4.5 decision tree algorithm by identifying each root-to-leaf path with a rule. Each rule is simplified by successively dropping conditions (attribute-tests). The difference lies in the sophistication of criteria used for retracting a trial generalisation when it is found to result in inclusion of cases not belonging to the rule's decision class. In the noise-free taxonomy problem a single „false positive" was taken to bar dropping the given condition. After that we reveal which rule explains the presence of disease most accurately. The final predictions are based on the most accurate rule. All the records where the predicted value fits the actual value are explained by the specific generated rules. The proportion between the success rate of the positive and negative predictions is the result of the proportion between the price of a miss and the price of a false alarm. The specific rule is: *If (thal >= **4.5**) and (chest >= **4**) => class is **„Yes"**.*

Class distributions	thal <= 4.5	„No"	„YES"
		0.7828947	0.217105
	thal > 4.5	„No"	„YES
		0.2627118	0.737288

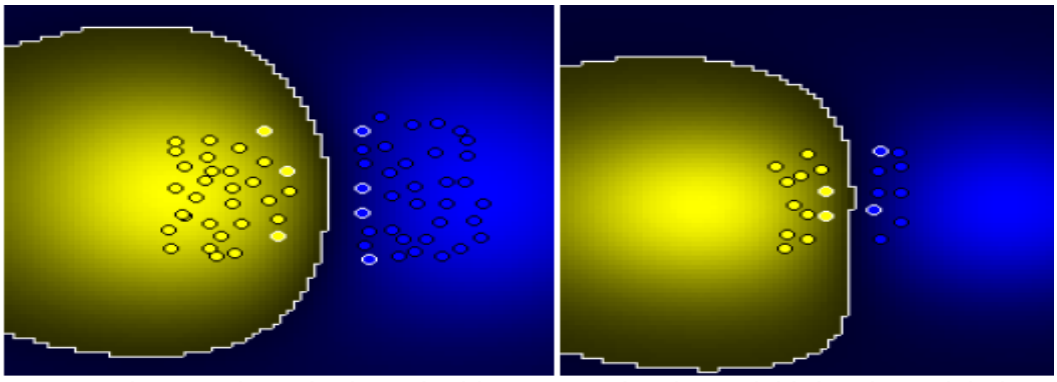

Figure 3. The frontiers designed with a Gaussian kernel (right picture) is based only on the selected support vectors instead of a real class distribution (on the left picture)

As illustrated in figure 3 on a very simple problem, the frontiers designed with a Gaussian kernel confirm that it tends to draw unreliable separation frontiers in the input data space (based only on the selected support vectors instead of a real class distribution). In our approach we assume that we have to estimate the *n* dimensional density function $f_x(p)$ of an unknown distribution. Then, the probability, *P* that a vector *x* will fall in a region *R* is:

$$P = \int_R f(x)\,dx \qquad (1)$$

Suppose that *n* observations are drawn independently according to f_x . Then we can approach *P* by *k/n* where *k* is the number of these *n* observations falling in *R*. The

estimation for f_x is an average function of x and samples x_i. In general this estimation is:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i, \lambda_n) \text{ , where } K(x, x_i, \lambda_n) \text{ are kernel functions.} \qquad (2)$$

If λ_n is very large, the kernel function changes very slowly with x, resulting in a very smooth estimate for f_x . On the other hand, if λ_n is small then $\hat{f}(x)$ is the superposition of n sharp normal distributions with small variances centred at the samples producing a very erratic estimate of f_x. The expected value of the estimate is an averaged value of the unknown densities. For skewed distribution kernel width is proportional to the neighbor distance. When applying these algorithms to concrete tasks we have to consider which learning algorithm is best suited for which problem. A satisfactory answer requires certain know-how of this area, which can be acquired only with experience.

EXPERIMENTAL RESULTS

In diagnosis applications the outcome may be the prediction of disease *vs.* normal or in prognosis applications. The input features may include clinical variables from medical examinations, laboratory test results, or other measurements. The objectives of feature selection are: reducing the cost of production of the predictor, increasing its speed, improving its prediction performance and/or providing an interpretable model.

The purpose of this experimental dataset is to predict the presence or absence of heart disease given the results of various medical tests carried out on a patient. This dataset contains 13 attributes, which have been extracted from a larger set of 75. There are two classes: presence and absence (of heart disease). RBF Gaussian model and SMO performed well on the heart dataset. This may reflect the careful selection of attributes by the doctors. After expert refinement Kernel density performed the best. The achieved result from **87.67 %** gives good perspectives especially when lognormal or skewed distributions are estimated. The leading correlation coefficient (that gives a measure of predictability) is 0.7384 and as such is not very high. Therefore the discriminating power of the linear discriminant is only moderate.

Despite being one of the fastest methods for learning support vector machines, SMO (sequential minimal optimization) is often slow to converge to a solution—particularly when the data is not linearly separable in the space spanned by the non-linear mapping.

The optimal model is then picked based on the highest accuracy value and then the whole training dataset is retrained with the optimization parameters of the selected model to produce a new optimized model. The user can create a model by choosing the type of model, for example linear or non-linear, as well as the parameters for that type of model. It is clear that if we choose the model (and hence the class) to maximise the accuracy value, then we will choose the correct class each time. We note that an optimal diagnosis assumes all costs to be expressed on a single numerical scale (need not correspond to economic cost).

Non-parametric density estimation usually requires a large amount of training data to provide a good estimate of the true distribution of a data set. Because of this property and the small size of the heart data set, the high testing accuracy we achieved was unexpected. The most important factor is how well the training set represents the actual distribution of the data. Due to the accuracy of our classifiers, it appears that the patients with the higher „thal" attribute are highly related to the positive class. The density estimates could be improved by finding more accurate estimates of the a priori probabilities by sampling the patient population. Traditionally model selection and parameterization is difficult for new data sets, even for experienced users. We generated models by: manually specifying which type of model and parameters to use, performing a Search across various model types and parameters, and by doing an DM analysis.

CONCLUSIONS AND FUTURE WORK

This exploration tries to aid automatic classification of diagnosis from heart diseases data. Although oriented to a specific problem, knowledge extraction from medical datasets and generating rules for predicting outcomes, the examined models can be easily tuned to other diagnostic problems based on data analysis and visual representations.

The practical application of the DM model selection needs explicit expert knowledge and more experimental collections. At the same time, more clinical studies are necessary to conclude the important problems and real advantages of diagnosis. The knowledge concerning which algorithm is applicable can be summarised in the form of rules, which can be constructed via ML methods. This concept has strong implications for the geometric interpretation of the shape of the feature map. Preliminary results obtained with our approach are promising. The study of the data distribution through the detection of the models seems to be robust. There are a lot of still open problems and questions in data modelling and specific rule representation analysis and up to our opinion, physicians and computer researchers should work hard together in order to achieve really valuable computer-assisted tools for precise diagnosis and therapy. With the proposed DM analysis we try to extract knowledge and rules. We are currently implementing this approach in Active Vessel computer-aided workstation for cardiology diagnosis and treatment with partners from Spain.

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

[2]. Andreeva P., M. Dimitrova and A. Gegov, (2006) Information Representation in Cardiological Knowledge Based System, *SAER'06, 23-25 Sept*. (submitted)

[3]. Data Mining Software, www.chel.com.ru/~rav/data_mining_software.html.

[4]. Kohonen T., (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, pp.59–69.

[5]. Müller, J.-A., F. Lemke, *Self Organising Data Mining,* Libri, Hamburg 2000, ISBN 3-89811-861-4, www.knowledge miner.net.

[6]. Russell, S. and Norvig, P. (2003) *Artificial Intelligence: a modern approach*. Prentice-Hall, 2nd edition.

[7]. Stutz J., P. Cheeseman, (1996) Bayesian classification (autoclass): Theory and results. *In Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

[8]. University of Waikato, *ML Program WEKA*, www.cs.waikato.ac.nz/ml/weka.

[9]. Witten I. H. (1998). Generating Accurate Rule Sets Without Global Optimization, in Shavlik, J., ed., *Machine Learning: Proc. of 15 Int. Conf.*, Morgan Kaufmann.

ABOUT THE AUTHOR

Res. Assoc. Plamena Andreeva (MSc. in Elecronics), Department of Knowledge Based Control Systems at the Institute of Control and System Research – BAS, 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.