Feature discovery for data mining

Manuel del Valle
Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain

Beatriz Sánchez
Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain
and
Telefónica Investigación y Desarrollo
Emilio Vargas 6, 28043 Madrid, Spain

Luis F. Lago-Fernández
Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain
and
Cognodata Consulting
Caracas 23, 28010 Madrid, Spain

Fernando J. Corbacho
Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain
and
Cognodata Consulting
Caracas 23, 28010 Madrid, Spain

Resumen

In most problems of data mining the human analyst previously constructs a new set of features, derived from the initial problem input attributes, based on a priori knowledge of the problem structure. These different features are constructed from different transformations which must be selected by the analyst. This paper provides a first step towards a methodology that allows the search for near-optimal representations in classification problems by allowing the automatic selection and composition of feature transformations from an initial set of basis functions. In many cases, the original representation for the problem data is not the most appropriate, and the search for a new representation space that is closer to the structure of the problem to be solved is critical for the successful solution of the problem. On the other hand, once this optimal representation is found, most of the problems may be solved by a linear classification method. As a proof of concept we present a classification problem where the class distributions have a very intricate overlap on the space of original attributes. For this problem, the proposed methodology is able to construct representations based on function compositions from the trigonometric and polynomial bases that provide a solution where some of the classical learning methods, e.g. multilayer perceptrons and decision trees, fail. The methodology consists of a discrete search within the space of compositions of the basis functions and a linear mapping performed by a Fisher discriminant. We play special emphasis on the first part. Finding the optimal composition of basis functions is a difficult problem because of its nongradient nature and the large number of possible combinations. We rely on the global search capabilities of a genetic algorithm to scan the space of function compositions.

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.107-113
ISBN: 84-9732-449-8

1. Introduction

Data mining has become an increasingly important field of research due to the large potential to be tapped from many commercial, scientific and industrial databases. Nevertheless, in most cases the knowledge discovery processes are still quite costly due to the number of iterations that the human analysts have to perform over the discovery loop [3]. To reduce this cost a number of tasks within the knowledge discovery loop can be partially automated. This is the case for feature selection and feature construction [4, 5, 6, 7, 8, 11, 13, 14, 17]. In most problems of knowledge discovery the human analyst previously constructs a new set of features, derived from the initial problem input attributes, based on a priori knowledge of the problem structure. These different features are constructed from different transformations which must be selected by the analyst.

For each new feature, a subset of input attributes must be selected (attribute selection) and a transformation to be applied to those attributes must be also selected (transformation selection). Both processes can be viewed as a search process and hence both can be automated to some degree by heuristic search. In this regard, domain knowledge can be introduced by choosing a set of bases that include transformations closer to the problem structure and heuristics that guide/bias the search process. The methodology described in this paper intertwines attribute selection and transformation selection in an overall search process implemented by a genetic algorithm. We have introduced the bias by means of the set of basis functions included. The different bases provide with different transformation properties, for instance the trigonometric basis introduces periodicity in an explicit manner. Furthermore, the architecture presented in this paper allows for basis function composition. Function composition enriches the expressive power by allowing the construction of features that have combined properties from the selected bases while giving rise to more compact representations. This is so since a basis closer to the problem structure gives rise to a more compact representation of the problem solution.

Classical methods for pattern classification are based on the existence of statistical differences among the distributions of the different classes. The best possible situation is perfect knowledge of these distributions. In such a case, Bayes classification rule gives the recipe to obtain the best possible solution. In real problems, however, class distributions are rarely available because the number of patterns is generally small compared with the dimensionality of the feature space. To tackle this problem many techniques of density estimation have been developed, both parametric and non-parametric [2]. When density estimation becomes too difficult, there is a variety of supervised learning algorithms, such as neural networks [1] or support vector machines [16], that try to find a non-linear projection of the original attribute space on to a new space where a simple linear discriminant is able to find an acceptable solution.

Let us assume a particular classification problem in which, when looking at the original attribute space, we observe an almost complete overlap among the class distributions. Following the Bayes rule, we see that for any point in this attribute space, the probabilities of belonging to any of the classes are all equal. We could be tempted to conclude that there is no solution to the problem better than choosing the class randomly. However, it could be that the overlapping is due to a bad representation of input data, and that there exists a transformation that separates the classes. We hypothesize that if such a transformation exists, there must exist a suitable basis in which it has a simple and compact expression. So solving such a problem can be reduced to finding the most appropriate basis or representation for the input data (with respect to the classification target). Once this representation is found, a linear discriminant will suffice to find a simple and compact solution. We propose an expansion of the work in [15] that incorporates other bases apart from the polynomial one. We use a genetic algorithm to perform both variable selection and search in the

transformation space, and a Fisher discriminant that performs the final linear projection. We show that this approach is able to solve problems where other methods fail to find a solution, even when the overlap is so large that there are no apparent statistical differences among the classes. This overlap may be due simply to the fact that the original representation of data is not well suited to the problem. Actually, it is well known that many classification problems are solved only after the application of some "intelligent" transformations provided by a domain "expert". Here we want to go a step closer into the automatic selection of these intelligent transformations, by allowing the algorithm to search for the optimal basis.

2. Methodology

When facing a two-class classification problem, our starting point is the assumption that there exists a non-linear function that projects the input data onto a unidimensional space where a linear separator is able to discriminate among the two classes. This function must have a simple and compact form in some basis, so finding an appropriate set of basis functions will strongly contribute to the simplification of the problem: the final projection may be constructed as a linear combination of these non-linear transformations. Here we propose to explore jointly the Taylor and Fourier bases, as well as compositions of both. We use a genetic algorithm (GA) to construct the non-linear transformations that operate on the raw input data, and a Fisher discriminant to perform the linear projection on the transformed attributes. In this regard our approach follows on the work developed by [15] with the EFLN algorithm, introducing two main differences: first, we do not limit the transformations to polynomials, but we expand the representation capabilities by adding trigonometric functions; and second, the linear projection is performed by a Fisher discriminant, instead of a linear neural network.

The proposed algorithm includes feature construction as well as feature selection. For the first task, it combines different bases of transformation (e.g. polynomial and trigonometric) to generate the input for the linear classifier. Feature selection is performed by the application of the genetic algorithm, which selects the best subsets of transformed variables by using the linear classifier error rate as the fitness criterion. Consequently our algorithm can be viewed as a wrapper method [7].

The general form for the input transformations operating in a n-dimensional feature space is given by the expression:

$$F(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} y_i^{a_i} T_i(b_i \pi y_i) \qquad (1)$$

where the x_i represent the original input variables, a_i and b_i are integer coefficients, and T_i is a trigonometric function (a sine or a cosine). Each y_i is either equal to x_i or to a new $F(x_1, x_2, ..., x_n)$. In this way compositions of polynomials and trigonometric functions can be constructed.

Our algorithm starts by generating K different function sets, each one composed of m functions as that of equation 1:

$$S_i = \{F_1^i, F_2^i, ..., F_m^i\}, i = 1, 2, ..., K \qquad (2)$$

Each of these sets S_i corresponds to an individual in the initial population which will be evolved by the genetic algorithm. The fitness of the individual S_i is calculated as the classification error of a Fisher linear discriminant operating on the transformed attributes $F_1^i(\mathbf{x}), F_2^i(\mathbf{x}), ..., F_m^i(\mathbf{x})$. Note that an exhaustive search over the space of input transformations would be computationally too expensive and would not scale properly on the number of input variables. This fact, together with the absence of gradient information, makes the use of an evolutionary approach very appropriate. In figure 1 we show a scheme of the algorithm, which is briefly described below:

1. *Initialize the first population of individuals randomly, and set the parameters for the GA, such as the number of iterations*

and the mutation probability.

2. *For each evolution iteration:*
   a) *For each individual:*
      1) *Generate the new features applying the input transformations to the original attributes.*
      2) *Evaluate its fitness value as the classification error of the Fisher Linear Discriminant applied to the transformed features on the training and validation data sets.*
   b) *Select the lowest error individuals for the next iteration.*
   c) *Generate a new population applying genetic operators and the individuals selected in (b).*
3. *Evaluate the most accurate individual on the test data set.*

3. Test case

We have applied the previous methodology to a synthetic data set that consists of two classes, A and B, in a two-dimensional input space, given by the attributes x and y. The problem presents the following properties: (i) there exists an appropriate non-linear transformation that is able to separate the classes with no error; and (ii) in the original input space the classes present a very high overlap and, given the number of examples, seem to follow the same distribution. This last fact makes the problem particularly difficult to solve. We present the results of our algorithm in comparison with the results obtained with other classification methods, namely multilayer perceptrons trained with backpropagation, decision trees trained with the C4.5 algorithm, and evolutionary FLNs that use the polynomial basis. The backpropagation algorithm was tested using networks of one single hidden layer, with different number of hidden units (ranging from 3 to 10) with a sigmoidal activation function. Different values for the learning rate between 0,01 and 0,3 were tried. For the decision trees, we used Quinlan's C4.5 algorithm [12] with probabilistic thresholds for continuous attributes, windowing, a gain ratio criterion to select tests and an iterative mode with ten trials. Finally, the evolutionary FLN was trained as described in [15], with polynomials of up to degree 3.

Specifically, let us consider the following problem. It consists of 2000 patterns in a two-dimensional input space, defined in the interval $[0 \leq x \leq 100, 0 \leq y \leq 100]$. We select 1000 patterns of each class. The patterns of class A are defined as:

$$(x, y) \in A \longleftrightarrow mod(int(x^2 y^2), 2) = mod(int(y), 2) \qquad (3)$$

where $int(x)$ is the integer part of x and $mod(x, 2)$ is the remainder of $x/2$. Class B patterns are those that do not satisfy the equality in eq. 3. Note that there exists a non-linear transformation that solves this problem with 0 error. However, in spite of the deterministic nature of the problem, the small number of patterns makes it appear that the two classes fully overlap, which makes the problem particularly difficult to most classification methods. To ilustrate this, we show in figure 2 the patterns of each class in the original input space.

We applied the previous three traditional methods to this problem, obtaining the results shown in table 1. Classification error rates are in all cases close to 50 %, which indicates that no improvement with respect to random class selection is achieved. This means that they are not performing much better than selecting the class randomly. The difficulty these traditional methods are confronting is due to the high overlap between the two classes. Note that for an absolute class overlap, even the best (Bayes) class estimator fails. However Bayesian decision theory assumes perfect knowledge of class distributions, which is not the present case. In fact, we know that below the apparent class mixing there is a hidden structure that the tested methods are not able to discover when just focusing on the original input space.

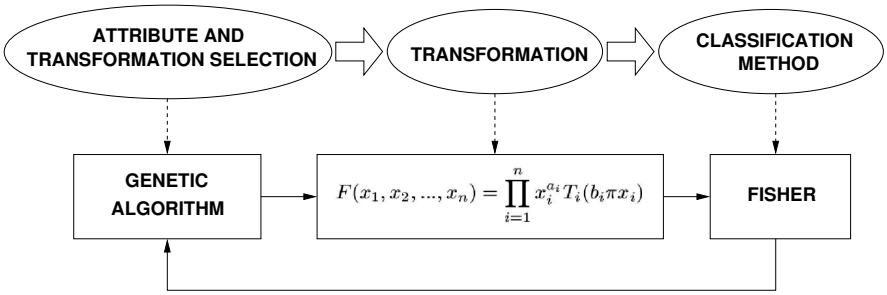

Figura 1: Schematics of the overall methodology. The genetic algorithm evolves individuals consisting of different sets of transformations that operate on the input data. The transformed attributes are then fed into a Fisher discriminant whose error rate determines the fitness of the individual, used by the genetic algorithm to compute the next generation of transformation sets.

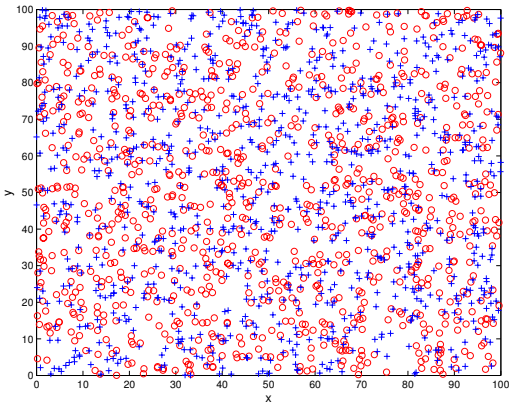

Figura 2: Input patterns for the test case consisting of two classes, A and B, and two attributes, x and y. The problem data consist of 1000 patterns of class A (circles) and 1000 patterns of class B (crosses).

Algorithm	Train Error %	Test Error %
Backprop	50,4	49,5
C4.5	46,4	50,2
EFLN	44,6	46,4

Cuadro 1: Comparison of performances of various classification methods on the problem of test case 2.

Finally, to test our algorithm we used a genetic algorithm with populations of up to 50 individuals, each one consisting of a set of $m = 6$ input transformations. The optimization was performed using a standard GA package [10]. All the trials we ran converged fastly to the optimal solution, the outcome of one of them is shown below:

$$\begin{pmatrix} \sin(\pi x^2 y^2)\sin(\pi y) \\ 0 \\ x^2 \sin(3\pi x^2 y^2) y \sin(2\pi x^2 y^2 \sin(3\pi x)\sin(\pi y)) \\ 0 \\ 0 \\ \cos(3\pi x^2 y^2) \end{pmatrix}$$

The corresponding Fisher projection is given by:

$$(-9{,}53 \quad 0{,}06 \quad 0 \quad 0 \quad 0 \quad 0)$$

Which produces the final transformation $-9{,}53 \sin(\pi x^2 y^2) \sin(\pi y)$ that separates the two classes with no error (see figure 3).

4. Conclusions

This paper presents a proof of concept for the construction of near-optimal problem representations in classification problems, based on the combination of functions selected from an initial family of transformations. The selection of an appropriate transformation allows the solution of complex nonlinear problems by a simple linear discriminant in the newly transformed space of attributes.

Work on progress includes the introduction of a more extensive family of basis functions that will allow for the construction of a wider repertoire of problem representations. Additionally, mechanisms to control the combinatorial explosion in the space of representations and the complexity of solutions will be analyzed. Additional work in progress also includes information/statistical measures that allow to uncover the structural/statistical properties of the input attributes and this in turn provides additional heuristics over which transformations to select.

Other advantages of the proposed method are that a closer, more compact problem representation usually allows for easier model interpretation [15], and, hence, a deeper understanding of the structure and mechanisms underlaying the problem under study. Related work on the extraction of hidden causes [9], which provide the generative alphabet, will be farther explored.

5. Acknowledgments

We want to thank P. Hoelgaard, M. Sánchez-Montañés and A. Sierra for very interesting comments and discussions, and Ministerio de Ciencia y Tecnología for financial support (BFI2003-07276).

Referencias

[1] Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford Univ. Press (1995)

[2] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley and Sons (2001) 84–214

[3] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining, U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.). Menlo Park, CA, AAAI Press (1996) 1–34

[4] Flach, P.A., Lavrac, N.: The role of Feature Construction in Inductive Rule Learning. ICML (2000) 1–11

[5] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Machine Learning (2003) 1157–1182

[6] Kramer, S.: Demand-Driven Construction of Structural Features in ILP. ILP **2157** (2001) 132–141

[7] Kohavi, R., John, G.H.: Wrappers for Feature Subset Selection. Artificial Intelligence **97** (1–2) (1997) 273–324

[8] Kudenko, D., Hirsh, H.: Feature Generation for Sequence Categorization. American Association for Artificial Intelligence (1998) 733–738

[9] Lago-Fernández, L.F., Corbacho, F.J.: Optimal Extraction of Hidden Causes. LNCS **2415** (2002) 631–636

[10] Levine, D.: Users Guide to the PGA-Pack Parallel Genetic Algorithm Library. T.R.ANL-95/18 (1996)

[11] Pagallo, G.: Boolean Feature Discovery in Empirical Learning. Machine Learning **5** (1) (1990) 71–99

[12] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc. (1992)

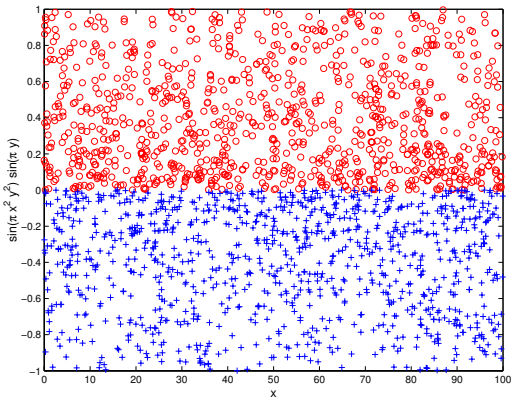

Figura 3: Plot of $sin(\pi x^2 y^2) sin(\pi y)$ vs x for the patterns of the test case. The final transformation discovered by the proposed algorithm allows for a linear separation of the two classes.

[13] Ragavan, H., Rendell, M.: Lookahead Feature Construction for Learning Hard Concepts. ICML (1993) 252–259

[14] Rennie, J.D.M., Jaakkola, T.: Automatic Feature Induction for Text Classification. MIT Artificial Intelligence Laboratory Abstract Book (2002)

[15] Sierra, A., Macías, J.A., Corbacho, F.: Evolution of Functional Link Networks. IEEE Trans. Evol. Comp. **5** (1) (2001) 54–65

[16] Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons (1998)

[17] Zucker, J.D., Ganascia, J.G.: Representation Changes for Efficient Learning in Structural Domains. ICML (1996) 543–551