ORIGINAL RESEARCH

An adaptive methodology to discretize and select features

Miguel Ángel Álvarez de la Concepción[1], Luis González Abril[2], Luis Miguel Soria Morillo[1], Juan Antonio Ortega Ramírez[1]

1. Computer Languages and Systems Department, University of Seville, Spain. 2. Department of Applied Economics I, University of Seville, Spain

Correspondence: Miguel Ángel Álvarez de la Concepción. Address: Computer Languages and Systems Department, University of Seville, Spain. Email: maalvarez@us.es.

Received: June 20, 2012 **Accepted:** December 15, 2012 **Online Published:** February 18, 2013
DOI: 10.5430/air.v2n2p77 **URL:** http://dx.doi.org/10.5430/air.v2n2p77

Abstract

A lot of significant data describing the behavior or/and actions of systems can be collected in several domains. These data define some aspects, called features, that can be clustered in several classes. A qualitative or quantitative value for each feature is stored from measurements or observations. In this paper, the problem of finding independent features for getting the best accuracy on classification problems is considered. Obtaining these features is the main objective of this work, where an automatic method to select features is proposed. The method extends the functionality of Ameva coefficient to use it in other tasks of machine learning where it has not been defined.

Key words

Ameva, Feature selection, Discretization, Entropy

1 Introduction

The problem to obtain the best accuracy, sensitivity, specificity, etc. in classification is one of the main problems in a lot of research areas like analysis and pattern recognition. It requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of features. One of them is in medical area when a doctor needs to know if a patient has cancer or not through thousands of gen values. In this sense, there are a lot of classifiers in the bibliography that process data sets to get the best results. Also, it is a central problem in machine learning. For example, there are classifiers based on SVM [1], Naive Bayesian [2], C4.5 [3], etc. that have been developed in the last years.

One of the most important preprocess in classification is the discretization because it allows algorithms to run very fast. This process establishes a relationship between continuous variables and their discrete transformation through developed functions. Therefore, it is possible to qualitatively model a series of continuous values if a label is assigned to them. Some studies [4] have shown that executing a prior process to discretize continuous features is more efficient than working directly with the continuous values. The discretization process reduces the computation memory usage and time in the application that develops classification algorithms. Also, it is used to manage the values of a feature more easily. As same as classifiers, there are a lot of discretization methods like GUDA-CCC [5], EDISC [6], CD [7] and others [8], [9]. Also, Ameva [10] is the discretization method that it is used in this paper.

It has been confirmed as one of the most promising correlation algorithms due to its reduced execution time and the small number of intervals provided. This behavior is outstanding when the data set has a large number of classes, although it has a slight reduction in the capacity of identification [11].

The other important problem in the classification process is the selection of features [12]. Usually, the obtained experimental data is not filtered about relevant features in systems. A lot of techniques for feature selection [13]-[15] have been developed. Some of these techniques are based on SVM [16] or Naïve Bayes [17].

The Ameva discretization algorithm [10] performs the discretization process effectively and quickly, so the set of values of a feature is greatly reduced. Because Ameva uses the statistic $\mathcal{X}^2$ to determine the relationship between features and classes, it is possible to use this algorithm to determine the relationship between features.

In this paper, a methodology based on Ameva is developed in order to select the main features of a data set. This method exploits the advantages of Ameva in runtime and brings a different approach which was developed on.

The rest of this paper is organized as follows: first, the definition of the problem is presented in Section 2. Also, the Ameva discretization algorithm and the entropy coefficient are presented. Section 3 presents the methodology to determine the best feature selection using the Ameva and the entropy coefficients. Section 4 reports the obtained results of applying the methodology in an example. The paper is finally concluded with a summary of the most important points.

2 Discretization

Let $X = \{x_1, x_2, \dots, x_N\}$ be a data set of a continuous attribute $\mathcal{X}$ of mixed-mode data such that each example x_i belongs to only one of ℓ classes of the variable denoted by

$$\mathcal{C} = \{C_1, C_2, \dots, C_\ell\}, \ell \geq 2$$

A continuous attribute discretization is a function $\mathcal{D}: \mathcal{X} \longrightarrow \mathcal{C}$ which assigns a class $C_i \in \mathcal{C}$ to each value $x \in \mathcal{X}$ in the domain of the property that is being discretized.

Let us consider a discretization $\mathcal{D}$ which discretizes $\mathcal{X}$ into k intervals:

$$\mathcal{L}(k; \mathcal{X}; \mathcal{C}) = \{L_1, L_2, \dots L_k\}$$

where L_1 is the interval $[d_0, d_1]$ and L_j is the interval $(d_{j-1}, d_j], j = 2,3, \dots, k$. Thus, a discretization variable is defined as $\mathcal{L}(k) = \mathcal{L}(k; \mathcal{X}; \mathcal{C})$ which verifies that, for all $x_i \in X$, a unique L_j exists such that $x_i \in L_j$ for $i = 1,2, \dots, N$ and $j = 1,2, \dots, k$. The discretization variable $\mathcal{L}(k)$ of attribute $\mathcal{X}$ and the class variable $\mathcal{C}$ are treated from a descriptive point of view. Having two discrete attributes, a two dimensional frequency table (called contingency table) as show in the Table 1 can be built.

In Table 1, n_{ij} denotes the total number of continuous values belonging to the C_i class that are within the interval L_j. $n_{i\cdot}$ is the total number of instances belonging to the class C_i, and $n_{\cdot j}$ is the total number of instances that belong to the interval L_j, for $i = 1,2, \dots, \ell$ and $j = 1,2, \dots, k$. So that:

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}\, , n_{\cdot j} = \sum_{i=1}^{\ell} n_{ij}\, , N = \sum_{i=1}^{\ell} \sum_{j=1}^{k} n_{ij} \tag{1}$$

Table 1. Contingency table

$C_i\|L_j$	L_1	$\cdots$	L_j	$\cdots$	L_k	$n_{i\cdot}$
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ik}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_ℓ	$n_{\ell 1}$	$\cdots$	$n_{\ell j}$	$\cdots$	$n_{\ell k}$	$n_{\ell\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot k}$	N

2.1 The Ameva discretization

Given Table 1, a discretization criterion based on the Contingency Coefficient ($\mathcal{X}^2$) is defined which measures the independency between class variable $\mathcal{C}$ and discretization variable $\mathcal{L}(k)$.

It is well known in Statistics that two given discrete attributes $\mathcal{C}$ and $\mathcal{L}(k)$ are (statistically) independent if, for all $C_i \epsilon \mathcal{C}$ and $L_j \epsilon \mathcal{L}(k)$,

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}, i = 1, \dots, \ell, j = 1, \dots, k$$

that is, no association exists between the two attributes.

Therefore, one way to measure the association (or independency) between class variable $\mathcal{C}$ and discretization variable $\mathcal{L}(k)$ is to analyse the value

$$\sum_{i=1}^{\ell} \sum_{j=1}^{k} \left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{N}\right)^2$$

Nevertheless, it is better to consider a relative measure denoted by $\mathcal{X}^2(k) \overset{\text{def}}{=} \mathcal{X}^2(\mathcal{L}(k), \mathcal{C}|\mathcal{X})$:

$$\mathcal{X}^2(k) = \sum_{i=1}^{\ell} \sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{N}\right)^2}{\frac{n_{i\cdot} n_{\cdot j}}{N}}$$

By using (1), it is not difficult to prove that:

$$\mathcal{X}^2(k) = N\left(-1 + \sum_{i=1}^{\ell} \sum_{j=1}^{k} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}}\right) \tag{2}$$

and

$$\max_{X, \mathcal{L}(k), C} \mathcal{X}^2(k) = N\left(\min\{\ell, k\} - 1\right) \tag{3}$$

In order to compare this coefficient against several discretization variables $\mathcal{L}(k)$ for $k \geq 2$, the Ameva coefficient, $Ameva(k) \overset{\text{def}}{=} Ameva(\mathcal{L}(k), \mathcal{C}|\mathcal{X})$, is defined as follows:

$$Ameva(k) = \frac{\mathcal{X}^2(k)}{k(\ell - 1)}$$

For $k, \ell \geq 2$. The Ameva criterion has the following properties:

- The minimum value of $Ameva(k)$ is 0 and when this value is achieved then both discrete attributes $\mathcal{C}$ and $\mathcal{L}(k)$ are statistically independent and viceversa.
- The maximum value of $Ameva(k)$ indicates the best correlation between class labels and discrete intervals. If $k \geq \ell$ then, for all $x \in C_i$ a unique $j0$ exists such that $x \in L_{j0}$ (remaining intervals $(k - \ell)$ have no elements); and if $k < \ell$ then, for all $x \in L_j$, a unique $i0$ exists such that $x \in C_{i0}$ (remaining classes have no elements) i.e. the highest value of the Ameva coefficient is achieved when all values within a particular interval belong to the same associated class for each interval.
- The aggregated value is divided by the number of intervals k, hence the criterion favors discretization schemes with the lowest number of intervals.
- From (3), it is followed that $Ameva_{max}(k) \overset{\text{def}}{=} \max_{\mathcal{X},\mathcal{L}(k),\mathcal{C}} Ameva(k) = \frac{N(k-1)}{k(\ell-1)}$ if $k < \ell$ and $\frac{N}{k}$ otherwise. Hence, $Ameva_{max}(k)$ is an increasing function of k if $k \leq \ell$, and a decreasing function of k if $k > \ell$. Therefore, $\max_{k\geq 2} Ameva_{max}(k) = Ameva_{max}(\ell)$ i.e. the maximum of the Ameva coefficient is achieved in the optimal situation, it is to say, when all values of C_i are in a unique interval L_j and viceversa.

Therefore, the aim of the Ameva method is to maximize the dependence relationship between the class labels $\mathcal{C}$ and the continuous-values attribute $\mathcal{L}(k)$, and at the same time to minimize the number of discrete intervals k.

2.2 The entropy

If $\ell = 1$ or $k = 1$ then it is not possible to use the Ameva method (Note 1). Let us see these two cases (see Table 2 and Table 3).

Equation (2) can not be calculated using Table 2 because it is not possible to divide by 0. Nevertheless, all the instances belong to the same class, therefore can be concluded that the dependence is maximum. In this case, let us indicate that $A^*(1) = 1$.

Table 2. Contingency table at first case ($\ell = 1$).

$C_i\|L_j$	L_1	$\cdots$	L_j	$\cdots$	L_k	$n_{i\cdot}$
C_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	N
$n_{\cdot j}$	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1k}	N

Regarding to Table 3, Ameva method can not be used because $\mathcal{X}^2(k) = 0$ and the Ameva coefficient does not give any information about the dependence. However, the dependence is not minimum and a new coefficient is necessary. By taking into account that if all instances are distributed equally in all classes, the dependence is minimum, and if exists i such that $n_{i1} = N$, the dependence is maximum. Hence the following coefficient, called Entropy, is considered:

$$A(1) = 1 + \frac{1}{N \ln \ell} \sum_{i=1}^{\ell} n_{i1} \ln\left(\frac{n_{i1}}{N}\right)$$

It holds that $0 \leq A(1) \leq 1$, and:

- If $A(1) = 0$, then $n_{i1} = \frac{N}{\ell}$ (minimum dependence).

- If $A(1) = 1$, then a unique n_{i1} exists that $n_{i1} = N$ (maximum dependence).

Table 3. Contingency table at second case $(k = 1)$.

$C_i\|L_j$	L_1	$n_{i\cdot}$
C_1	n_{11}	n_{11}
$\vdots$	$\vdots$	$\vdots$
C_i	n_{i1}	n_{i1}
$\vdots$	$\vdots$	$\vdots$
C_ℓ	$n_{\ell 1}$	$n_{\ell 1}$
$n_{\cdot j}$	N	N

3 The methodology

Given an attribute X_i where $i = 1,2,\dots,s$, the Ameva discretization algorithm is applied to this attribute so obtained intervals are considered as a new set of classes. This set of classes is denotes as follows:

$$\mathcal{C}^i = \{C_1^i, C_2^i, \dots, C_{\ell_i}^i\} \tag{4}$$

Let us consider $X^p \subset X$ as the data subset that belongs to the class $C_p \in \mathcal{C}^i$ where $p = 1,2,\dots,\ell$. From (4), for each attribute X_j with $j = 1,2,\dots,s$, a g_{ijp} value is obtained from $\mathcal{C}^i$ as follows:

- If the X^p data subset all belong to the same class C^i then $g_{ijp} = A^*(1) = 1$.
- If the subset of data belongs to different classes, then:
    - If values of the attribute X_j are always in the same interval, then $g_{ijp} = A(1)$.
    - If values of the attribute X_j are not always in the same interval, then $g_{ijp} = Ameva_N(\ell_i)$, where $Ameva_N(\ell_i)$ is defined as follows (Note 2):

$$Ameva_N(\ell_i) = \frac{\ell_i'}{N_p} Ameva(\ell_i)$$

provided that N_p is the number of instances of the class X^p and ℓ_i' is the number of intervals of the attribute X_i for which there is at least one value in the data subset.

Given $i,j = 1,2,\dots,s$, a g_{ijp} value can be obtained applying this methodology for all class $C_p \in \mathcal{C}$ $(p = 1,2,\dots,\ell)$, and by considering different statistics as follows:

$$g_{ij}^{min} = \min_p g_{ijp}$$

$$g_{ij}^{geo} = \sqrt[\ell]{\prod_{p=1}^{\ell} g_{ijp}}$$

$$g_{ij}^{ari} = \frac{1}{\ell}\sum_{p=1}^{\ell} g_{ijp}$$

$$g_{ij}^{max} = \max_{p} g_{ijp}$$

It is well-known that the following relationship is holded:

$$g_{ij}^{min} \leq g_{ij}^{geo} \leq g_{ij}^{ari} \leq g_{ij}^{max}$$

The main properties of the matrix $G = (g_{ij})$, that is,

$$G = \begin{pmatrix} 1 & g_{12} & \cdots & g_{1s} \\ g_{21} & 1 & \cdots & g_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ g_{s1} & g_{s2} & \cdots & g_{ss} \end{pmatrix}$$

are the following: i) it is squared but non symmetric matrix; ii) the values of the main diagonal are 1; iii) $0 \leq g_{ij}, g_{ji} \leq 1$.

From the G matrix, a method of generating rules of dependence between attributes can be defined. For example, a possible rule is the next: given a threshold value, U, if max $\{g_{ij}, g_{ji}\} > U$ and $i < j$ where $i, j = 1,2, \ldots, s$ and $i \neq j$, then the X_j variable is eliminated. Let us illustrate it with an example in the next section.

4 Two experiments

Let us consider the Iris Plant Database (Note 3) from UCI Repository which is perhaps the best known database to be found in the pattern recognition literature. This data set is considered due to its simplicity since this methodology is not completely defined yet.

The data set contains four attributes (sepal length, sepal width, petal length and petal width) and three classes (Setosa, Versicolor and Virginica) of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from each other.

The matrices generated by the presented methodology in this paper are:

$$G_{Iris}^{min} = \begin{pmatrix} 1 & 0.4898 & 0.6667 & 0.0998 \\ 0.3265 & 1 & 0.0350 & 0.0930 \\ 0.0280 & 0.0586 & 1 & 0.0303 \\ 0.0545 & 0.0998 & 0.0836 & 1 \end{pmatrix} \tag{5}$$

$$G_{Iris}^{geo} = \begin{pmatrix} 1 & 0.7883 & 0.8736 & 0.4638 \\ 0.6886 & 1 & 0.3271 & 0.4530 \\ 0.1727 & 0.2674 & 1 & 0.1293 \\ 0.1573 & 0.3222 & 0.2244 & 1 \end{pmatrix} \tag{6}$$

$$G_{Iris}^{ari} = \begin{pmatrix} 1 & 0.8299 & 0.8889 & 0.6999 \\ 0.7755 & 1 & 0.6783 & 0.6977 \\ 0.4039 & 0.4617 & 1 & 0.3672 \\ 0.3753 & 0.4783 & 0.4063 & 1 \end{pmatrix} \tag{7}$$

$$G_{Iris}^{max} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad (8)$$

This result shows that it is possible to determine the dependence of attributes of a data set from the Ameva discretization algorithm and the adjustments to resolve the inconsistencies outlined above with the entropy.

The coefficients in the minimum matrix (5) determine the lowest coefficients of dependence between two attributes. These coefficients provide information about there is a class for which the two attributes have less dependency. If these values are high, it is possible to conclude that the dependence between two attributes is high. Therefore, these coefficients are a minimum threshold for each pair of attributes.

A similar conclusion can be obtained from the maximum matrix (8). The coefficients provide information about there is a class for which the two attributes have a high dependence. In this case, these coefficients are the maximum threshold values for each pair of attributes.

Given a data set, the best result is achieved when the maximum and minimum matrix are the same. In this case, all the attributes are the same dependence with other regardless of the original class. Thus, there is only one matrix for generate the discrimination rules.

The geometric mean matrix (6) and the arithmetic mean (7) represent a global value of dependency. While the geometric mean matrix rewards the worst situations about a class, leading to a low value on the global coefficient, the arithmetic mean matrix balances the values of the coefficients.

A possible interpretation to determine which attributes are dependent of each other is to establish a threshold value. From this limit, two attributes are dependent if the average of the coefficients g_{ij} and g_{ji} of the arithmetic mean matrix is greater than or equal to this value.

In this case, the threshold value of 0.75 is established to check which attributes are dependents. The pair g_{ij}, g_{ji} that reaches this threshold is g_{12}, g_{21} because the arithmetic mean of g_{12} and g_{21} is greater than 0.75. It is necessary indicate that the sepal length and the sepal width features are the first and second attributes in the experiment.

Thus, in order to carry out a classification problem can be declared that the X_1 and X_2 features are similar. Let us see this affirmation by using as classification algorithm the Support Vector Machine (SVM) [18].

A performance for the 1-v-r SVM, in the form of accuracy rate, has been evaluated on models using the Gaussian kernel with $\sigma = 1$, and $C = 1$. The criteria employed to estimate the generalized accuracy is the 5-fold cross-validation on the whole set of training data. This procedure is repeated 120 times in order to ensure good statistical behavior. The obtained results are:

- With all features, the accuracy is 0.9320.
- Without the sepal length feature, the accuracy rate is 0.9341.
- Without the sepal width feature, the accuracy rate is 0.9667.

Furthermore to check that the accuracy rate is not less when a feature is eliminated, the methodology has discovered that these features introduce noise in the classification problem when both are used at the same time because the results are improved without the second feature.

Now, consider the Glass Identification (Note 4) Dataset, also from UCI Repository, to prove the methodology with another classification method.

The data set contains nine attributes (refractive index, Sodium, Magnesium, Aluminum, Silicom, Potassium, Calcium, Barium, Iron) (Note 5) and seven classes (building windows float processed, building windows non float processed, vehicle windows float processed, vehicle windows non float processed, containers, tableware and headlamps).

The generated matrices are:

$$G_{Glass}^{min} = \begin{pmatrix} 1 & 0.0864 & 0.1621 & 0.1526 & 0.1828 & 0.0937 & 0.2566 & 0.0200 & 0.0210 \\ 0.0607 & 1 & 0.1244 & 0.2184 & 0.0952 & 0.3224 & 0.0602 & 0.0064 & 0.0502 \\ 0.0968 & 0.1680 & 1 & 0.1577 & 0.1382 & 0.2324 & 0.1359 & 0.0676 & 0.0227 \\ 0.1344 & 0.0496 & 0.1062 & 1 & 0.0680 & 0.2534 & 0.0719 & 0.1239 & 0.0496 \\ 0.0236 & 0.2464 & 0.1361 & 0.0876 & 1 & 0.0972 & 0.0284 & 0.0045 & 0.0372 \\ 0.0416 & 0.3599 & 0.2023 & 0.1897 & 0.1244 & 1 & 0.0422 & 0.0060 & 0.0351 \\ 0.2505 & 0.1293 & 0.2047 & 0.1408 & 0.1755 & 0.3704 & 1 & 0.1247 & 0.0303 \\ 0.1080 & 0.1603 & 0.2230 & 0.3297 & 0.1016 & 0.2199 & 0.0936 & 1 & 0.0450 \\ 0.0606 & 0.0399 & 0.0428 & 0.0463 & 0.0643 & 0.0912 & 0.0493 & 0.0502 & 1 \end{pmatrix}$$

$$G_{Glass}^{geo} = \begin{pmatrix} 1 & 0.1105 & 0.1859 & 0.2392 & 0.2651 & 0.1680 & 0.2925 & 0.0762 & 0.0333 \\ 0.0809 & 1 & 0.2780 & 0.3359 & 0.1195 & 0.4461 & 0.0890 & 0.2031 & 0.0614 \\ 0.1433 & 0.2830 & 1 & 0.3181 & 0.1525 & 0.3355 & 0.2408 & 0.2626 & 0.0320 \\ 0.1755 & 0.2344 & 0.3042 & 1 & 0.0973 & 0.3202 & 0.1173 & 0.4599 & 0.0565 \\ 0.1952 & 0.2976 & 0.2422 & 0.1401 & 1 & 0.2040 & 0.1471 & 0.0593 & 0.0422 \\ 0.0682 & 0.4720 & 0.3455 & 0.2298 & 0.1515 & 1 & 0.0670 & 0.1598 & 0.0588 \\ 0.3566 & 0.1677 & 0.2744 & 0.2813 & 0.2178 & 0.4770 & 1 & 0.1541 & 0.0361 \\ 0.1506 & 0.3618 & 0.3457 & 0.5367 & 0.1609 & 0.2939 & 0.1249 & 1 & 0.0547 \\ 0.2040 & 0.0631 & 0.1552 & 0.1749 & 0.1652 & 0.2299 & 0.3136 & 0.3342 & 1 \end{pmatrix}$$

$$G_{Glass}^{ari} = \begin{pmatrix} 1 & 0.1131 & 0.1873 & 0.2449 & 0.2696 & 0.1757 & 0.2956 & 0.0945 & 0.0346 \\ 0.0832 & 1 & 0.2968 & 0.3421 & 0.1212 & 0.4515 & 0.0934 & 0.3401 & 0.0620 \\ 0.1486 & 0.2924 & 1 & 0.3426 & 0.1540 & 0.3449 & 0.2617 & 0.3011 & 0.0355 \\ 0.1828 & 0.2755 & 0.3369 & 1 & 0.1035 & 0.3263 & 0.1229 & 0.5214 & 0.0567 \\ 0.2630 & 0.3102 & 0.2499 & 0.1439 & 1 & 0.2318 & 0.1779 & 0.0839 & 0.0424 \\ 0.0745 & 0.4759 & 0.3567 & 0.2324 & 0.1534 & 1 & 0.0739 & 0.2609 & 0.0615 \\ 0.3634 & 0.1702 & 0.2865 & 0.3002 & 0.2215 & 0.4807 & 1 & 0.1554 & 0.0379 \\ 0.1621 & 0.3824 & 0.3549 & 0.5497 & 0.2124 & 0.2992 & 0.1269 & 1 & 0.0556 \\ 0.2686 & 0.0662 & 0.2454 & 0.2861 & 0.2110 & 0.2585 & 0.3897 & 0.4178 & 1 \end{pmatrix}$$

$$G_{Glass}^{max} = \begin{pmatrix} 1 & 0.1681 & 0.2279 & 0.3083 & 0.3234 & 0.2379 & 0.3757 & 0.2191 & 0.0494 \\ 0.1135 & 1 & 0.4177 & 0.3998 & 0.1497 & 0.5440 & 0.1529 & 0.4686 & 0.0730 \\ 0.2003 & 0.3868 & 1 & 0.5766 & 0.2017 & 0.4854 & 0.4861 & 0.4013 & 0.0751 \\ 0.3050 & 0.3512 & 0.5092 & 1 & 0.1585 & 0.4457 & 0.1927 & 0.7000 & 0.0642 \\ 0.4974 & 0.5004 & 0.3048 & 0.1836 & 1 & 0.5007 & 0.2904 & 0.1116 & 0.0494 \\ 0.1227 & 0.5554 & 0.4854 & 0.2824 & 0.1910 & 1 & 0.1486 & 0.4153 & 0.0917 \\ 0.4336 & 0.2137 & 0.4890 & 0.4973 & 0.2979 & 0.5689 & 1 & 0.1834 & 0.0692 \\ 0.2842 & 0.4686 & 0.4730 & 0.6650 & 0.6630 & 0.4153 & 0.1603 & 1 & 0.0695 \\ 0.5107 & 0.1041 & 0.5004 & 0.5153 & 0.5034 & 0.5013 & 0.5058 & 0.5287 & 1 \end{pmatrix}$$

As can be seen, the coefficients are lower than the matrices in the previous example. In this case, the threshold value of 0.4 is established to check which attributes are dependents in the arithmetic matrix. The pairs g_{ij}, g_{ji} that reaches this threshold is g_{26}, g_{62} and g_{48}, g_{84} because the arithmetic mean of g_{26} and g_{62}, and g_{48} and g_{84} are greater than 0.4. The Sodium, Aluminum, Potassium and Barium are the second, the fourth, the sixth and the eighth attributes in the experiment.

Thus, the X_2 and X_6 features and X_4 and X_8 features are similar. To prove this affirmation, a K-Nearest Neighbor classification algorithm is used with $k = 3$. The criteria employed to estimate the generalized accuracy is the 10-fold cross-validation on the whole set of training data in this case. The procedure is repeated 120 times. The obtained results are:

- With all features, the accuracy is 0.7152.
- Without the Sodium feature, the accuracy rate is 0.7284.
- Without the Potassium feature, the accuracy rate is 0.7058.
- Without the Aluminum feature, the accuracy rate is 0.6877.
- Without the Barium feature, the accuracy rate is 0.7149.
- Without the Sodium and the Aluminum features, the accuracy rate is 0.6762.
- Without the Sodium and the Barium features, the accuracy rate is 0.7286.

Without the Potassium and the Aluminum features, the accuracy rate is 0.6719.

Without the Potassium and the Barium features, the accuracy rate is 0.7156.

Once more, the methodology has discovered that these features introduce noise in the classification problem when both pairs are used at the same time.

5 Conclusions

We have studied a method of discretization, Ameva, whose objective is to maximize the dependence between the intervals that divide the values of an attribute and the classes to which they belong, providing at the same time the minimum number of intervals.

After that, we have developed a methodology to reduce the number of features of a data set based on the dependence between them. To the best of knowledge, there are not existing researches that directly address the problem to reduce the number of features using an approach similar to ours.

This development is based on taking advantage of Ameva discretization algorithm. Thus, a new coefficient has been developed to determine the dependence between features. Hence, we have reduced the number of values of features and the number of features from a qualitative reasoning.

To test the development of the methodology, it has been applied to two well-known data sets to obtain the dependent relationship between their features. Nevertheless, we think that this approach can be satisfactorily applied in this area when the data set has a lot of instances and features, and one of these features determines the class which each instance belongs to. Another data sets must fulfill these characteristics.

Finally, after applying the discrimination of features obtained in the methodology, the modified data sets have been carried out for the classification tests to verify the effectiveness of the methodology.

The next step to complement this development is the design of an automatic method of creation of feature discrimination rules. Subsequently, we must define some improvements in this methodology to automatically know the dependence between features without setting manually a threshold value.

Acknowledgement

This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness ARTEMISA (TIN2009-14378-C02-01) and Simon (TIC-8052) of the Andalusian Regional Ministry of Economy, Innovation and Science.

References

[1] Ghoggali, N., Melgani, F. and Bazi, Y. “A multiobjective genetic SVM approach for classification problems with limited training samples”. IEEE Transactions on Geoscience and Remote Sensing. 2009; 47 (6): 1707-1718. http://dx.doi.org/10.1109/TGRS.2008.2007128

[2] Stojadinovic, A., Potter, B.K., and et al. Development of a prognostic naive bayesian classifier for successful treatment of nonunions. The Journal of Bone and Joint Surgery (American). 2011; 93 (2): 187-194. http://dx.doi.org/10.2106/JBJS.I.01649

[3] Jiang, S.Y. and Yu, W. A combination classification algorithm based on outlier detection and C4.5. Advanced Data Mining and Applications. 2009; 5678: 504-511. http://dx.doi.org/10.1007/978-3-642-03348-3_50

[4] Entezari-Maleki, R., Iranmanesh, S.M. and Minaei-Bidgoli, B. An experimental investigation of the effect of discrete attributes on the precision of classification methods. World Applied Sciences Journal. 2009; 216-223. http://dx.doi.org/10.1109/ICICT.2009.5267189

[5] Zeng, A., Gao, Q. and Pan, D. A global unsupervised data discretization algorithm based on collective correlation coefficient. Modern Approaches in Applied Intelligence. 2011; 6703: 146-155. http://dx.doi.org/10.1007/978-3-642-21822-4_16

[6] Shehzad, K. EDISC: a class-tailored discretization technique for rule-based classification. IEEE Transactions on Knowledge and Data Engineering. 2012; 24 (8): 1435-1447. http://dx.doi.org/10.1109/TKDE.2011.101

[7] Wang, C., Wang, M., She, Z. and Cao, L. CD: a coupled discretization algorithm. Advances in Knowledge Discovery and Data Mining. 2012; 7302: 407-418. http://dx.doi.org/10.1007/978-3-642-30220-6_34

[8] Jiang, F., Zhao, Z. and Ge, Y. A supervised and multivariate discretization algorithm for rough sets. Rough Set and Knowledge Technology. 2010; 6401: 596-603. http://dx.doi.org/10.1007/978-3-642-16248-0_81

[9] Zhao, J., Han, C.Z., Wei, B. and Han, D.Q. A UMDA-based discretization method for continuous attributes. Advanced Materials Research. 2012; 403-408: 1834-1838. http://dx.doi.org/10.4028/www.scientific.net/AMR.403-408.1834

[10] González, L., Cuberos, F.J., Velasco, F. and Ortega, J.A. Ameva: an autonomous discretization algorithm. Expert Systems with Applications. 2009; 36 (3): 5327-5332. http://dx.doi.org/10.1016/j.eswa.2008.06.063

[11] González, L., Cuberos, F.J., Velasco, F. and Ortega, J.A. A new approach to qualitative learning in time series. Expert Systems with Applications. 2009; 36 (6): 9924-9927. http://dx.doi.org/10.1016/j.eswa.2009.01.066

[12] Saeys, Y., Inza, I. and Larrañaga, P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23 (19): 2507-2517. http://dx.doi.org/10.1093/bioinformatics/btm344

[13] Hua, J., Tembe, W.D. and Dougherty, E.R. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition. 2009; 42 (3): 409-424. http://dx.doi.org/10.1016/j.patcog.2008.08.001

[14] Witten, D.M. and Tibshirani, R. A framework for feature selection in clustering. Journal of the American Statistical Association. 2010; 105 (490): 713-726. http://dx.doi.org/10.1198/jasa.2010.tm09415

[15] Ma, Y. and Zhan, L. Research on the evaluation of feature selection based on SVM. Informatics in Control, Automation and Robotics. 2012; 133: 407-414. http://dx.doi.org/10.1007/978-3-642-25992-0_57

[16] Maldonado, S. and Weber, R. A wrapper method for feature selection using Support Vector Machines. Information Sciences: an International Journal. 2009; 179 (13): 2208-2217. http://dx.doi.org/10.1016/j.ins.2009.02.014

[17] Chen, J., Huang, H., Tian, S. and Qu, Y. Feature selection for text classification with Naive Bayes. Expert Systems with Applications. 2009; 36 (3): 5432-5435. http://dx.doi.org/10.1016/j.eswa.2008.06.054

[18] González, L., Angulo, C., Velasco, F. and Català, A. Dual unification of bi-class support vector machine formulations. Pattern recognition. 2006; 39 (7): 1325-1332. http://dx.doi.org/10.1016/j.patcog.2006.01.007