Performance of Algorithms for Interval Discretization of Biomedical Signals

L.M. Soria Morillo[1], L. Gonzalez-Abril[2], and J.A. Ortega Ramírez[1]

[1] Computer Languages and Systems Dept., University of Seville, 41012 Seville, Spain
[2] Applied Economics I Dept., University of Seville, 41018 Seville, Spain

Abstract— **A methodology to quantify the dependence between features using the Ameva discretization algorithm and the advantages of qualitative models is presented in this paper. This approach will be applied over medical data sets. A comparison among Ameva and other related works has been done. The results, as will be depth explained in this paper, show that Ameva-based methodology can be used to determine the dependence between features in a fast and understandable way from data sets with a high number of attributes and low number of instances. This is a quite important feature in genomic environments among others. This methodology has been applied to some well-known medical data sets and the results obtained shown that is a good alternative to other established algorithms in terms of clarity and computational cost.**

Keywords— **Discretization algorithm, binning, data mining, machine learning, e-health systems**

I. Introduction

The problem of classification is one of the main problems in data analysis and pattern recognition that requires the construction of a classifier, that is, a function that assigns a class label to instances described by a set of features. The induction of classifiers from data sets of classified instances is a central problem in machine learning. For that purpose, a large number of methodologies based on SVM [1], Naive Bayesian [2], C5.0 [3], etc. have been developed.

Discretization is an important preprocess in classification. This process establishes a relationship between continuous variables and their discrete transformation through functions. Some studies [4] have shown that it is more efficient to execute a prior process of discretization of continuous features. This process reduces the computation time and memory us~~age in the application of classificat~~ion algorithms and it is used to manage more effectively the set of values of a feature. Some relevant discretization methods are Ameva [5], Chi2 [6], Khiops [7], CAIM [8] and others [9].

The Ameva discretization method has been confirmed as one of the most promising algorithms resulting in faster execution and providing a smaller number of intervals. This behavior is outstanding when the data set have a large number of classes, although it has a slight reduction in the capacity of identification [5, 10].

Another problem in the process of classification is the existence of irrelevant features [11]. When data is obtained experimentally, is not considered what features are relevant to the study system. Several techniques [12, 13, 14] have been developed to reduce the number of features and determine which are relevant to the system. Some of these techniques are based on principal components analysis [15] or factorial analysis [16].

The Ameva discretization algorithm [10] performs the discretization process effectively and quickly, so the set of values of a feature is greatly reduced, but not reduce the number of features. Because Ameva uses the statistic χ^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 new method that quantitatively relates the dependence of features by using the Ameva discretization algorithm and the advantages of a qualitative model has been developed. This method uses Ameva exploiting its advantages 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 for establish the notation of the rest of the paper. Also, the Ameva discretization algorithm and the Entropy coefficient are presented. Section 3 presents the new methodology for determined the dependence between features using the Ameva algorithm and the entropy coefficient. Section 4 reports the obtained results of applying the methodology over different medical datasets. The paper is finally concluded with a summary of the most important points and future works.

II. Discretization

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

$$\mathscr{C} = \{C_1, C_2, \ldots, C_\ell\}, \quad \ell \geq 2 \tag{1}$$

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

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

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

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

$C_i \vert L_j$	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

Table 1: Contingency table

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 $i = 1, 2, \ldots, \ell$ and $j = 1, 2, \ldots, k$. So that:

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

A. The Ameva discretization

Given discrete attributes $\mathscr{C}$ and $\mathscr{L}(k)$, the contingency coefficient, denoted by $\chi^2(k) \overset{def}{=} \chi^2(\mathscr{L}(k), \mathscr{C}|X)$, defined as

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

is considered. It is straightforward to prove that

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

Hence, the Ameva coefficient, $Ameva(k) \overset{def}{=} Ameva(\mathscr{L}(k), \mathscr{C}|X)$, is defined as follows:

$$Ameva(k) = \frac{\chi^2(k)}{k(\ell - 1)} \tag{4}$$

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 $\mathscr{C}$ and $\mathscr{L}(k)$ are statistically independent and viceversa.
- The maximum value of $Ameva(k)$ indicates the best correlation between the class labels and the discrete intervals. If $k \geq \ell$ then, for all $x \in C_i$ a unique j_0 exists such that $x \in L_{j0}$ (the remaining intervals $(k - \ell)$ have no elements); and if $k < \ell$ then, for all $x \in L_j$, a unique i_0 exists such that $x \in C_{i0}$ (the 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 $Ameva_{max}(k) \overset{def}{=} \max_{X, \mathscr{L}(k), \mathscr{C}} Ameva(k) = \frac{N(k-1)}{k(\ell-1)}$ 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 (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 dependency relationship between the class labels $\mathscr{C}$ and the continuous-values attribute $\mathscr{L}(k)$, and at the same time to minimize the number of discrete intervals k.

B. The entropy

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

$C_i \vert L_j$	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

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

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

$C_i \vert L_j$	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

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

With respect to Table 3, Ameva method is not possible to use because $\chi^2(k) = 0$ 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 the instances are distributed equally in all classes, the dependence is minimum, and if exists i such that $n_{i1} = N$, the dependence is maximum, then 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 $0 \le A(1) \le 1$, 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).

Let us indicate these pathologic cases do not happen in a standard discretization, but it is necessary taking into account in the presented methodology in the next section.

III. The methodology

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

$$\mathscr{C}^i = \{C_1^i, C_2^i, \ldots, C_{\ell_i}^i\} \qquad (5)$$

Let us consider $X^p \subset X$ as the data subset that belongs to the class $C_p \in \mathscr{C}$ where $p = 1, 2, \ldots, \ell$. From (5), for each attribute X_j with $j = 1, 2, \ldots, s$, a G_{ijp} value is obtained from $\mathscr{C}^i$ 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 belonging to different classes, then:
  · If the values of the attribute X_j are always in the same interval, then $G_{ijp} = A(1)$.
  · If the 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:

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

provide 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.

Note 1. *This new Ameva coefficient is chosen in order to obtain a normalized value* $0 \le Ameva_N(\ell_i) \le 1$.

Furthermore, it is straightforward to prove that if $i = j$ *for* $i = 1, 2, \cdots, s$*, then* $G_{iip} = 1$*, for all* $p = 1, 2, \cdots, \ell$.

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

- The minimum $G_{ij}^{min} = \min_p G_{ijp}$.
- The geometric mean $G_{ij}^{geo} = \sqrt[\ell]{\prod_{p=1}^{\ell} G_{ijp}}$.
- The arithmetic mean $G_{ij}^{arit} = \frac{1}{\ell} \sum_{p=1}^{\ell} G_{ijp}$.
- The maximum $G_{ij}^{max} = \max_p G_{ijp}$.

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

$$G_{ij}^{min} \le G_{ij}^{geo} \le G_{ij}^{arit} \le G_{ij}^{max}$$

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

$$G = \begin{pmatrix} 1 & G_{12} & \cdots & G_{1n} \\ G_{21} & 1 & \cdots & G_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ G_{n1} & G_{n2} & \cdots & 1 \end{pmatrix}$$

are the following: i) it is square and symmetric matrix; ii) the values of the main diagonal are 1; and iii) $0 \le G_{ij}, G_{ji} \le 1$.

From the G matrix, a method of generating rules of dependence between attributes can be defined. Thus, if a threshold value is set to decide whether two attributes are dependent, the elimination of features can be made. Let us illustrate it with an example in the next section.

IV. Application to medical data

Let us consider three different medical datasets to be used in this application:

- EEG Eye State dataset ([1]). This dataset was obtained using Emotiv EEG Neuroheadset during 117 seconds. During this time, eye states (open or closed) were detected. This process was carried out via a camera during the EEG measurement and added later manually to the file after analysing the video frames. The dataset is composed of 14980 instances and 15 attributes.
- Diabetes 130-US hospitals for years 1999-2008 [17] ([2]). The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc. This dataset is composed of 100000 instances grouped in 55 attributes. Must be noted that in this case some missing values have been detected.
- Multi-Class Cancer Diagnosis Using Tumor Gene Expression Signatures [18] ([3]). In this dataset have been subjected 218 tumor samples, spanning 14 common tumor types, achieved purely by molecular classification. In this case, the tumor type has been classified using discretization technique based on Ameva. The last dataset contains 144 instances with different cancer types and 16063 attributes for each one of these instances.

In this section, first dataset processing using the presented methodology will be shown. The reason for select this dataset is the number of attributes. In order to show the different matrices generated by our algorithm, the lower the number of attributes, the greater the clarity of the explanation. The summary of the remainder dataset results will be shown later.

The matrices generated by the presented methodology in this paper from the EEG Eye State dataset are:

This result shows that it is possible to determine the dependence of attributes of a dataset from the Ameva discretization algorithm and the adjustments to resolve the inconsistencies outlined above with the entropy. The coefficients in the minimum matrix for the labels of the data set under study (Table 4) determine the lowest coefficients of dependence between all attributes. As can be identified from this analysis, attribute 1 is low correlated with attributes 4, 6 and 12. However, it's high correlated with attributes 3, 11, 13 and 14. These coefficients provide information about what is minimum correlation level for the both attributes compared in a concrete position of the table. 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, so it's not possible to determine, if the level is low, if both attributes are not correlated.

A similar conclusion can be obtained from the maximum correlation matrix (Table 7). The coefficients provide information about what is the maximum correlation value, regardless of the class, for each pair of attributes compared. In this case, these coefficients are the maximum threshold values for each pair of attributes. By hence, in this case it's possible to conclude, if the correlation level is low enough, that two attributes are not correlated.

The most accurate result is achieved when the maximum and minimum matrix are similar. In this case, the pair of attributes under comparison, have the same dependence each other regardless of the original class. Thus, it would be possible to choose any matrix for generate the discrimination rules.

The arithmetic mean (Table 6) and the geometric mean matrices (Table 5) 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_{25}, G_{52} because the arithmetic mean of G_{25} and G_{52} is greater than 0.75. This conclusion indicates that the features 2 and 5 are high correlated for most of original labels and therefore, one of them can be removed when the classification algorithm is executed.

Thus, in order to carried out a classification problem can be declared that the X_2 and X_5 features are similar. This conclusion can be demonstrated using a classification algorithm. In this case, in order to show the final results, Support Vector Machine (SVM) [1] will be used.

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 10-folds cross-validation on the

[1] https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State

[2] https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

[3] http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view\&paper_id=61

1,000	0,685	0,718	0,278	0,650	0,363	0,683	0,641	0,674	0,707	0,742	0,440	0,761	0,862
	1,000	0,548	0,173	0,872	0,339	0,848	0,660	0,839	0,783	0,799	0,202	0,691	0,615
		1,000	0,487	0,517	0,523	0,553	0,585	0,547	0,594	0,605	0,634	0,635	0,726
			1,000	0,072	0,612	0,102	0,283	0,114	0,169	0,179	0,793	0,244	0,330
				1,000	0,291	0,886	0,709	0,895	0,826	0,800	0,171	0,702	0,592
					1,000	0,313	0,482	0,334	0,380	0,352	0,585	0,367	0,414
						1,000	0,759	0,869	0,762	0,805	0,256	0,745	0,639
							1,000	0,802	0,742	0,725	0,440	0,691	0,658
								1,000	0,852	0,817	0,272	0,749	0,641
		Sym.							1,000	0,831	0,336	0,780	0,696
										1,000	0,335	0,849	0,735
											1,000	0,410	0,496
												1,000	0,775
													1,000

Table 4: G_{EEG}^{min}

1,000	0,745	0,733	0,350	0,713	0,450	0,729	0,666	0,728	0,759	0,793	0,471	0,809	0,877
	1,000	0,593	0,184	0,900	0,399	0,851	0,671	0,841	0,791	0,816	0,257	0,752	0,678
		1,000	0,545	0,544	0,573	0,561	0,611	0,572	0,611	0,634	0,637	0,666	0,732
			1,000	0,101	0,663	0,149	0,315	0,154	0,216	0,214	0,810	0,295	0,396
				1,000	0,370	0,923	0,737	0,918	0,855	0,839	0,199	0,756	0,655
					1,000	0,421	0,548	0,421	0,461	0,432	0,641	0,457	0,486
						1,000	0,761	0,901	0,816	0,837	0,268	0,775	0,678
							1,000	0,803	0,764	0,753	0,445	0,718	0,666
								1,000	0,880	0,854	0,277	0,787	0,688
		Sym.							1,000	0,866	0,337	0,811	0,742
										1,000	0,341	0,876	0,775
											1,000	0,430	0,529
												1,000	0,804
													1,000

Table 5: G_{EEG}^{geo}

1,000	0,748	0,733	0,360	0,716	0,460	0,731	0,667	0,731	0,761	0,794	0,472	0,810	0,877
	1,000	0,595	0,184	0,901	0,404	0,851	0,671	0,841	0,791	0,816	0,264	0,755	0,681
		1,000	0,549	0,545	0,576	0,561	0,612	0,572	0,611	0,634	0,637	0,667	0,732
			1,000	0,106	0,665	0,160	0,317	0,161	0,223	0,218	0,811	0,301	0,403
				1,000	0,380	0,924	0,738	0,919	0,855	0,840	0,202	0,758	0,658
					1,000	0,439	0,552	0,433	0,470	0,441	0,643	0,468	0,492
						1,000	0,761	0,902	0,818	0,838	0,268	0,775	0,679
							1,000	0,803	0,764	0,753	0,445	0,718	0,666
								1,000	0,881	0,854	0,277	0,789	0,690
		Sym.							1,000	0,866	0,337	0,811	0,743
										1,000	0,341	0,876	0,776
											1,000	0,430	0,530
												1,000	0,804
													1,000

Table 6: G_{EEG}^{arit}

1,000	0,811	0,748	0,441	0,782	0,557	0,779	0,692	0,787	0,815	0,846	0,503	0,859	0,892
	1,000	0,643	0,195	0,930	0,468	0,853	0,682	0,843	0,800	0,833	0,326	0,818	0,747
		1,000	0,610	0,573	0,628	0,569	0,638	0,598	0,628	0,664	0,639	0,699	0,737
			1,000	0,141	0,718	0,217	0,352	0,208	0,276	0,257	0,828	0,358	0,475
				1,000	0,470	0,961	0,766	0,942	0,885	0,879	0,232	0,814	0,725
					1,000	0,566	0,622	0,531	0,560	0,529	0,702	0,568	0,570
						1,000	0,764	0,935	0,874	0,870	0,281	0,806	0,718
							1,000	0,804	0,786	0,782	0,450	0,746	0,674
								1,000	0,910	0,892	0,282	0,828	0,739
		Sym.							1,000	0,902	0,339	0,843	0,790
										1,000	0,346	0,904	0,816
											1,000	0,451	0,564
												1,000	0,834
													1,000

Table 7: G_{EEG}^{max}

training set. This procedure is repeated 120 times in order to ensure good statistical behavior and to reduce the risk of false positives and wrong conclusions. The obtained results are shown below:

- With all features, the accuracy rate is 0.9184.
- Removing the second feature, the accuracy rate is 0.9113.
- Removing the fifth feature, the accuracy rate is 0.9109.
- Removing the fourth feature, the accuracy rate is 0.8791.

This example shows that the same algorithm executed from different configurations of the original dataset, obtain different results to be be studied. In first place, the accuracy of the algorithm using all features is taken as gold standard for the rest of configurations. If second feature is removed, the accuracy is quite similar to that obtained when fifth feature is removed. This is due to both features are high correlated. Instead, when fourth feature is removed, the difference in terms of accuracy between the las two configurations and the last one is evident. Furthermore, besides check that the accuracy rate is not less when a feature is removed, the methodology has discovered that these features introduce noise in the classification system. Therefore, under some conditions, it could be profitable to remove these correlated feature in order to improve the overall accuracy of used algorithms.

V. Conclusions and future work

We have studied a method of discretization, Ameva, which objective is to maximize the dependence between the intervals on that divide the values of an attribute and the classes to which they belong. Ameva algorithm provides at the same time the minimum number of intervals, that is high recommended to improve the classification speed and reduce the energy consumption when these algorithms are executed under critical power systems, such as mobile phones for example.

Later, a methodology to reduce the number of feature set based on dependence criteria was presented. In this vein, there are not existing researches that directly address the features number reduction problem using a similar approach. This technique is based on the correlation between labels an class intervals for each pair of features and is based on Ameva discretization algorithm. The discretization algorithm selected has been used in a field that was not previously defined. Also, a new coefficient has been developed to determine the dependence between features when Ameva failed.

Finally, the development of the methodology has been tested. The process has been applied to some medical data set to obtain the dependent between their features. These kind of information has a peculiar feature, so the number of instances is very low with respect to the number of features. Compared to other not medial data sets, this is a characteristic under the Ameva algorithm has not been previously executed. Nevertheless, once made the testing over different data sets, it can be demonstrated that the advantages of this approach are clear when it is used with several instances and features. Furthermore, if one (or more) of these features determines the label each instance belongs to, results are even better.

Regarding to future works that complements this research, the design of an automatic method for creation of feature discrimination rules is under development. This system will allow to define some improvements in this methodology to automatically setting of threshold values. Currently, these values are set manually depending on the case and the profile of each attribute.

Acknowledgments

This research is partially supported by the projects of the Spanish Ministry of Economy and Competitiveness HERMES (TIN2013-46801-C4-1-r) and Simon(TIC-8052) of the Andalusian Regional Ministry of Economy, Innovation and Science.

References

1. González Luis, Angulo Cecilio, Velasco Francisco, Catala Andreu. Dual unification of bi-class support vector machine formulations *Pattern recognition.* 2006;39:1325–1332.
2. Wang Qiong, Garrity George M, Tiedje James M, Cole James R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy *Applied and environmental microbiology.* 2007;73:5261–5267.
3. Govindarajan M. Text mining technique for data mining application in *Proceedings of world academy of science, engineering and technology*;26:544–549Citeseer 2007.
4. Entezari-Maleki Reza, Iranmanesh Seyyed Mehdi, Minaei-Bidgoli Behrouz. An experimental investigation of the effect of discrete attributes on the precision of classification methods in *Information and Communication Technologies, 2009. ICICT'09. International Conference on*:215–220IEEE 2009.
5. Gonzalez-Abril L, Cuberos Francisco Javier, Velasco Francisco, Ortega Juan Antonio. Ameva: An autonomous discretization algorithm *Expert Systems with Applications.* 2009;36:5327–5332.
6. Liu Huan, Setiono Rudy. Chi2: Feature selection and discretization of numeric attributes in *tai*:388IEEE 1995.
7. Boulle Marc. Khiops: A statistical discretization method of continuous attributes *Machine learning.* 2004;55:53–69.
8. Kurgan Lukasz, Cios Krzysztof J, others . CAIM discretization algorithm *Knowledge and Data Engineering, IEEE Transactions on.* 2004;16:145–153.
9. Yepes Alejandro G, Freijedo Francisco D, Doval-Gandoy Jesus, Lopez Oscar, Malvar Jano, Fernandez-Comesa Pablo. Effects of discretization methods on the performance of resonant controllers *Power Electronics, IEEE Transactions on.* 2010;25:1692–1712.

10. Gonzalez-Abril L, Velasco Francisco, Ortega Juan Antonio, Cuberos Francisco Javier. A new approach to qualitative learning in time series *Expert Systems with Applications.* 2009;36:9924–9927.

11. Guyon Isabelle, Elisseeff André. An introduction to variable and feature selection *The Journal of Machine Learning Research.* 2003;3:1157–1182.

12. John George H, Kohavi Ron, Pfleger Karl, others . Irrelevant features and the subset selection problem in *Machine Learning: Proceedings of the Eleventh International Conference*:121–129 1994.

13. Yang Jihoon, Honavar Vasant. Feature subset selection using a genetic algorithm in *Feature extraction, construction and selection*:117–136Springer 1998.

14. Faraoun KM, Rabhi A. Data dimensionality reduction based on genetic selection of feature subsets *INFOCOMP Journal of Computer Science.* 2007;6:36–46.

15. Rocchi L, Chiari L, Cappello A. Feature selection of stabilometric parameters based on principal component analysis *Medical and Biological Engineering and Computing.* 2004;42:71–79.

16. Khosla Nitin. *Dimensionality Reduction Using Factor Analysis*. PhD thesisGriffith University, Australia 2004.

17. Strack Beata, DeShazo Jonathan P, Gennings Chris, et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records *BioMed research international.* 2014;2014.

18. Ramaswamy Sridhar, Tamayo Pablo, Rifkin Ryan, et al. Multiclass cancer diagnosis using tumor gene expression signatures *Proceedings of the National Academy of Sciences.* 2001;98:15149–15154.