A quantitative methodology to identify related features in data sets

Alvarez, M.A.[a]; Gonzalez-Abril, L.[b]; Ortega, J.A.[a]; Soria, L.M.[a]; Cuberos, F.J.[a]
[a]Department of Computer Science, University of Seville
[b]Department of Applied Economics I, University of Seville
Seville, Spain
{maalvarez, luisgon, jortega, lsoria, fjcuberos}@us.es

Abstract

In this paper, a methodology which quantifies the dependence between features in a data set is developed. This methodology uses the Ameva discretization algorithm. In particular, it uses the Ameva coefficient to quantify the dependence. Furthermore, a new coefficient called entropy has been proposed for cases where it is not possible to apply the Ameva discretization algorithm. Thus, different matrices of inter-dependence are built providing a grade of dependence between two features. Finally, to verify the qualities of this methodology, a simple method to discard features based on it is applied to a well-known data set in a classification process and promising results for the carried out system are obtained.

1 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.

One of the most important preprocess in classification is the discretization. This process establishes a relationship between continuous variables and their discrete transformation through functions. Therefore, it is possible to model qualitatively a series of continuous values if a label is assigned to them. Some studies [4] have shown that execute a prior process to discretize continuous features is more efficient than work directly with the continuous values. This process reduces the computation time and memory usage in the application of classification algorithms and it is used to manage the set of values of a feature more effectively. Some relevant discretization methods are Ameva [5], Khiops [6], CAIM [7] and others [8; 9].

The Ameva discretization method has been confirmed as one of the most promising algorithms due to its reduced execution time and the smaller 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 [5; 10].

Another problem in the classification process is the existence of irrelevant features [11]. When data is obtained experimentally, is not considered what features are relevant for the studied system. Several techniques [12; 13; 14] have been developed to reduce the number of features and to determine which are relevant for the system. Some of these techniques are based on principals 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 do 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 methodology based on Ameva algorithm is developed in order to reduce the number of 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 to 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 to determine the dependence between features using the Ameva algorithm and the entropy coefficient. Section 4 reports the obtained results of applying the methodology in two datasets. The paper is finally concluded with a summary of the most important points and future works.

2 Discretization

Let $X = \{x_1, x_2, \ldots, x_N\}$ 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

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

A continuous attribute discretization is a function $\mathcal{D} : \mathcal{X} \rightarrow \mathcal{C}$ 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 discrete intervals:

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

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 $\mathcal{L}(k) = \mathcal{L}(k;X;\mathcal{C})$ which verifies that, for all $x_i \in X$, a unique L_j exists such that $x_i \in L_j$ for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, k$. 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 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:

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

2.1 The Ameva discretization

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

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

Hence, the Ameva coefficient, $Ameva(k) \overset{def}{=} Ameva(\mathcal{L}(k), \mathcal{C}|X)$, 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 $\mathcal{C}$ 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 j_0 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 i_0 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 $Ameva_{max}(k) \overset{def}{=} \max_{X,\mathcal{L}(k),\mathcal{C}} Ameva(k) = \frac{N(k-1)}{k(\ell-1)}$ 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. 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 $A^*(1) = 1$.

$C_i \vert L_j$	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 ($\ell = 1$)

$C_i \vert L_j$	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 ($k = 1$)

Regarding to Table 3, Ameva method can not be used 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 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).

Note 2.1 *Let us indicate these pathologic cases do not happen in a standard discretization, but it will be necessary taking into account in the presented methodology in the next section.*

3 The methodology

Given an attribute X_i where $i = 1, 2, \ldots, s$, 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 denoted as follows:

$$\mathcal{C}^i = \{C_1^i, C_2^i, \ldots, C_{\ell_i}^i\} \tag{5}$$

Also, a new matrix that contains the Ameva coefficients for all attributes can be built.

Let us consider $X^p \subset X$ as the data subset that belongs to the class $C_p \in \mathcal{C}$ 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 $\mathcal{C}^i$ 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:

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

    provide that N_p is the number of instances of the class X^p and $\ell_i^{'}$ is the number of intervals of the attribute X_i for which there is at least one value in the data subset.

Note 3.1 *This new Ameva coefficient is chosen in order to obtain a normalized value $0 \leq Ameva_N(\ell_i) \leq 1$ as same as $A(1)$.*

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 applying this methodology for all class $C_p \in \mathcal{C}$ $(p = 1, 2, \cdots, \ell)$, and by considering one statistic, the arithmetic mean, $g_{ij} = \frac{1}{\ell}\sum_{p=1}^{\ell} g_{ijp}$.

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

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

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

The matrix described above does not indicate the degree of goodness of dependencies, so it is necessary to introduce an adjustment factor. This factor is the normalized Ameva coefficient, so the above matrix is adjusted as follows:

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

where $g_{ij}^* = \frac{g_{ij} Ameva_N(X_i) + g_{ji} Ameva_N(X_j)}{Ameva_N(X_i) + Ameva_N(X_j)}$.

Now, the main properties of the matrix $G^* = (g_{ij}^*)$ are the following: i) it is square and symmetric matrix; ii) the values of the main diagonal are 1; and iii) $0 \leq g_{ij}^*, g_{ji}^* \leq 1$.

Finally, a threshold value, δ, is set for establish the dependence between features for the two matrix. While exists $g_{ij} > \delta$, then X_i or X_j are dependents:

- If $g_i > g_j$, then g_i is discarded.
- If $g_i < g_j$, then g_j is discarded.

where $g_i = \sum_{m=1}^{s} g_{im}$ and $g_j = \sum_{m=1}^{s} g_{mj}$. The reasoning is the same with g_{ij}^*. Let us illustrate it with examples in the next section.

4 Experimentation

Let us consider the Glass Identification Dataset[1] and the Image Segmentation Dataset[2] from UCI Repository. These data sets are considered due to their simplicity.

The first data set contains 9 attributes, 6 classes and 214 instances. The second data set contains 19 attributes, 7 classes and 210 instances.

The results of applying this methodology are shown in the following tables for a classification test with the following characteristics:

- 10 loops.
- Cross Validation with 10 folders.
- K-Nearest Neighbor with $k = 3$.
- Three cases of the datasets with $\delta \in [0, 1]$:
  - Original.
  - Typified.
  - Typified and discretized.

[1] Available at http://archive.ics.uci.edu/ml/datasets/Glass+Identificati

[2] Available at http://archive.ics.uci.edu/ml/datasets/Image+Segmentati

δ	s		Original (%)		Typified (%)		Typified and discretized (%)	
	G	G^*	G	G^*	G	G^*	G	G^*
0	1	1	43.69	42.80	43.13	42.94	**39.72**	39.72
0.1	2	1	42.01	43.36	41.78	43.60	**39.72**	39.72
0.2	3	2	53.08	42.52	51.59	42.38	**40.05**	39.72
0.3	5	4	65.37	63.22	67.10	58.50	**64.11**	46.21
0.4	7	7	69.63	68.13	**70.93**	66.78	**65.65**	65.56
0.5	8	8	69.02	68.88	70.89	67.94	**65.65**	65.09
0.6	9	9	69.77	69.11	71.64	70.98	65.65	65.14
0.7	9	9	70.14	68.93	71.45	70.79	65.42	64.95
0.8	9	9	69.67	69.39	70.93	71.59	66.07	65.14
0.9	9	9	69.49	69.30	71.50	72.20	66.03	65.98
1	9	9	**69.72**	**69.72**	71.03	71.12	65.19	65.70

Table 4: Accuracy percent of Glass Identification Dataset

δ	s		Original (%)		Typified (%)		Typified and discretized (%)	
	G	G^*	G	G^*	G	G^*	G	G^*
0	1	1	28.86	14.29	27.86.13	14.29	14.29	**14.29**
0.1	2	4	42.62	31.38	51.67	27.67	32.86	**17.57**
0.2	2	4	42.76	29.48	51.38	28.57	32.76	**17.71**
0.3	2	6	42.67	68.29	50.14	72.62	32.48	70.43
0.4	2	7	41.86	68.67	50.57	74.86	32.86	71.33
0.5	3	7	62.29	68.19	68.57	75.33	46.24	71.14
0.6	5	10	74.05	72.52	85.48	84.29	64.00	**85.86**
0.7	8	13	74.71	72.76	87.29	84.10	82.90	**85.76**
0.8	8	14	75.19	73.33	87.05	84.90	82.52	**85.10**
0.9	8	14	74.90	73.29	**87.33**	84.29	82.86	**85.29**
1	19	19	**76.19**	**76.29**	85.57	85.67	84.05	84.05

Table 5: Accuracy percent of Image Segmentation Dataset

The table of the accuracy percent of Glass Identification Dataset and Image Segmentation Dataset are shown in the Table 4 and 5, respectively.

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 relationship of dependency between features can be seen in the column "Typified and discretized" for the matrix G in the Table 4. For threshold values of 0.1 to 0.5 the accuracy percentage is almost the same and the number of features is different.

The best accuracy percentage is 70.93 and it has been obtained with a threshold value of 0.4, the matrix G, the typified dataset and only with 7 features. The original dataset has a 69.72 accuracy percentage with 9 features.

Also, it can be seen in the same column for the matrix G^* in the Table 5. For threshold values of 0 to 0.2 and 0.6 and 0.9, the accuracy percentage is almost the same and the number of features is different.

The best accuracy percentage is 87.33 and it has been obtained with a threshold value of 0.9, the matrix G, the typified dataset and only with 8 features. The original dataset has a 76.19 accuracy percentage with 19 features.

5 Conclusions and future work

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 a similar approach 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 quantitative reasoning.

To test the development of the methodology, it has been applied to two well-known data set for obtain the dependent relationship between their features. Nevertheless, we think that this approach can be satisfactorily apply 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. Another data sets must fulfill these characteristics.

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

Acknowledgments

This research is supported by the Spanish Ministry of Science and Innovation R&D project ARTEMISA (TIN2009-14378-C02-01)..

References

[1] L. González, C. Angulo, F. Velasco, and A. Catala. Dual unification of bi-class support vector machine formulations. *Pattern recognition*, 39(7):1325–1332, 2006.

[2] Q. Wang, G.M. Garrity, J.M. Tiedje, and J.R. Cole. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73(16):5261–5267, 2007.

[3] M. Govindarajan. Text Mining Technique for Data Mining Application. *Proceedings of World Academy of Science, Engineering and Technology*, 26:544–549, 2007.

[4] R. Entezari-Maleki, S.M. Iranmanesh, and B. Minaei-Bidgoli. An Experimental Investigation of the Effect of Discrete Attributes on the Precision of classification Methods. *World Applied Sciences Journal*, 7:216–223, 2009.

[5] L. Gonzalez-Abril, FJ Cuberos, F. Velasco, and JA Ortega. Ameva: An autonomous discretization algorithm. *Expert Systems with Applications*, 36(3):5327–5332, 2009.

[6] M. Boulle. Khiops: A statistical discretization method of continuous attributes. *Machine Learning*, 55(1):53–69, 2004.

[7] L.A. Kurgan and K.J. Cios. CAIM discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16(2):145–153, 2004.

[8] F.J. Ruiz, C. Angulo, N. Agell, X. Rovira, M. Sánchez, and F. Prats. A Discretization Process in Accordance with a Qualitative Ordered Output. *Proceeding of the 2005 conference on Artificial Intelligence Research and Development*, 131:273–280, 2005.

[9] R.P. Li and Z.O. Wang. An entropy-based discretization method for classification rules with inconsistency checking. 1:243–246, 2002.

[10] L. Gonzalez-Abril, F. Velasco, JA Ortega, and FJ Cuberos. A new approach to qualitative learning in time series. *Expert Systems with Applications*, 36(6):9924–9927, 2009.

[11] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[12] G. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. *Machine Learning Conference Proceedings*, pages 121–129, 1994.

[13] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. *Intelligent Systems and Their Applications, IEEE*, 13(2):44–49, 1998.

[14] KM Faraoun and A. Rabhi. Data dimensionality reduction based on genetic selection of feature subsets. *INFOCOM - Journal of Computer Science*, 6(2):9–19, 2007.

[15] L. Rocchi, L. Chiari, and A. Cappello. Feature selection of stabilometric parameters based on principal component analysis. *Medical and Biological Engineering and Computing*, 42(1):71–79, 2004.

[16] Nitin Khosla. *Dimensionality Reduction Using Factor Analysis*. PhD thesis, 2006.