Evaluation of Classification in More Than Two Classes

Daniel Volovici

1.03.2016

Abstract

Machine Learning is the most important part of Artificial Intelligence in the same sense as we cannot speak about intelligence without the capacity of learning. One of the basics type of learning is to learn to classify objects or putting labels on objects. If you are able to recognize that an object have the attributes of a class C or not (meaning that it is part of class non C), than you will be able to classify in more than one classes: with the strategy one-vs-all or with the strategy one-vs-one. Classification as a learning task imply training with examples of objects a priori labeled with the class which they belong. But if in data we do not have definitions of classes, splitting data into groups has the name of clustering. The idea behind clustering is that probably the data are produced by different processes or that they belong naturally to different groups. So, the best way to evaluate the quality of the clustering is to try to cluster data generated to be part of different classes.

The most used way for evaluation of classification and clustering methods is the confusion matrix defined for two classes. Starting from this matrix it is obtained the measures of Precision, Recall and the Fmeasure. Exist a generalization to n classes using a nxn matrix. But for the situation where exist a different number of clusters than the number of original classes we must use a nxm contingency matrix also named association matrix. And because the degree of association is measured by the dominance of the principal diagonal it is very important to use time efficient methods of manipulation of the lines and columns of matrixes.

Keywords list: classification, clustering, contingency matrix, association, matrix, precision, recall.

1 Introduction

A method used to centralize experimental data is to write them into a table, a two-dimensional array. An important decision is to choose what will be describe on lines and what on columns. When a test is performed repeatedly

a number of times or on many individuals it is preferred that each individual to be represented on a line of the table and the results of measurements for that individual to be put on the cells of that line according with the column designated for every measure (for each attribute).

Table 1: Centralization of experimental data

	Attribute 1	Attribute 2	...	Attribute j	...	Attribute m
Individual 1	a_{11}	a_{12}	...	a_{1j}	...	a_{1m}
Individual 2	a_{21}	a_{22}	...	a_{2j}	...	a_{2m}
...		.....				
Individual i	a_{i1}	a_{i2}	...	a_{ij}	...	a_{im}
...		.....				
Individual n	a_{n1}	a_{n2}	...	a_{nj}	...	a_{nm}

An important influence on this type of tables came from medical tests for diagnosis of different diseases. The set of measurements is related with a collection of diagnosis tests designed to signal the presence or the absence of a desease. The actual paradigm is to apply the test to a number of subjects and, after that, to centralize these data counting the number of those with positive and of that with negative results. Decision about the result of the test (to be consider positive or negative) will be taken accordingly the values of attributes. If it is used only the value of a single attribute to decide if the test is positive or negative, then usually we calculate a threshold to be compared with the attribute's value. In the attribute's value is lower than the threshold we decide that the test is on one side, for example positive (or vice versa); if the value is higher we decide that the test results is on the other side (in our example, negative).

It is possible to reorder the rows of the table in descending (or ascending) order of the attributes values so that be easier to visualize the distribution of those with pozitive/negative test result (Fig. 1). Going on and centralizing even more we will obtain the exact number of those with positive test and those with negative result (Fig. 1).

The most important problem impacting in the field of medical diagnosis is the revelation that tests are never perfect and, because of this situation, it is possible to appear *false positive* and *false negative* results.

The same situation appear in all the fields where we decide to classify or to cluster an object to a group. When the characteristics of the groups are known before testing the procedure is consider **classification** and when we do not know a priori the groups and sometimes not even the number of them, we speak about **clustering**.

In the literature dedicated to epidemiology the data related to the results

	Attribute
Individual 1	a_1
Individual 2	a_2
...	...
Individual P-1	a_{P-1}
Individual P	a_P = Threshold
Individual P+1	a_{P+1}
...	...
Individual $P+N$	a_{P+N}

	Number of experimental results
Individuals tested positive	P
Individuals tested negative	N

Figure 1: Centralization of tests with a single attribute

of a test is usually [FFF14] represented in the form of a table as the one in Table 2. In this area researchers have access to results of the test and usually they do not know for sure if the humans being tested are really healthy or have the disease. They need a so called *gold standard* to find the truth. This is also denominated as *criterion standard* or *reference standard* and could imply expensive and sometimes dangerous additional testings. Because of the fact that we have access only to the experimental results of the test it is natural to represent the possible outcomes in rows and let the columns to be assigned to the estimation of the presence of the disease.

Table 2: The results of a test for diagnosing a desease

		DISEASE	
		Present	Absent
TEST	Positive	tp	fp
	Negative	fn	tn

In many other situation, researchers have access to the true values of the items analysed and in these situations it will be more natural another arrangement of the data in the table. We may think of transposing the matrix from Table 2. One of the most important is the case of simulations when we intend to test different methods on data known to be part of a class. In all these cases it will be useful the confusion matrix.

2 Confusion matrix

The performance of a learning algorithm is visualised using *confusion matrix* (also named *error matrix*) which is a table where each row represents the instances in an actual class and the columns represents the instances in a

predicted class according with the algorithm. This representation is natural in classification, but also must be considered for testing the performances of clustering methods: we could generate data so that every point to be a realization of a class and to observe how well the clustering algorithm group the data in clusters more or less similar with the real (true but unknown) starting generation process. In the community of statisticians working in clustering the confusion matrix has the name *contingency matrix*.

Table 3: Confusion matrix for binary classification (for 2 classes)

		Estimated Cluster Predicted Class		
		Examples estimated as *positive*	**Examples estimated as*negative***	
True Class (real examples to be observed)	Positive examples	***tp***	***fn***	the number of positive examples ***tp+fn***
	Negative examples	***fp***	***tn***	the number of negative examples ***tn+fp***

In the subfield of Machine Learning specialised in problems with classification (supervised and unsupervised) are very important two measures of quality: *Precision* and *Recall*. In Information Retrieval [Rij79] and especially in Text Retrieval [CM12] the evaluation measures have a meaning easy to understand:

- $-Recall$, proportion of all true members of class retrieved by the algorithm;
- $-Precision$, proportion true members of class from the number of those considered positive.

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

For unsupervised classification (clustering) the problem is a little bit different: exist two different classes and we assume they have the same importance. This is the reason to consider more important the other two evaluation measures: *Accuracy* (also named *Success Rate*) and *Error Rate*.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

$$Error\,Rate = \frac{fp + fn}{tp + tn + fp + fn} \tag{4}$$

$$Accuracy = 1 - Error\,Rate$$

Accuracy is a measure [VBCM10] related with the association between the true sharing of examples in the two true classes on one side and the distribution of them in the two estimated clusters on the other side. And because statistics offers more powerful tools [Fle81] for estimating the degree of association [And73] we suggest to transform the problem of evaluation in one of association. For this purpose we will transform notations from Table 4 in those in Table 5. Starting from here it is possible to generalize to n classes like in Table 6.

3 Contingency matrix

The term of *contingency matrix* is used especially in statistics for representing in form of a tableau the frequency distribution of variables (very used for multivariate variables). It is also named *cross tabulation* it was introduced by Karl Pearson. In multivariate statistics it is vary important to discover dependencies between different variables. If some dependency exist we could determine a degree of association between variables.

In the present context, that of determination of the quality of clustering/classification, we are interested to measure the degree of association between real (true) classes and estimated clusters. So we will consider as variables the membership to classes/clusters. The membership will be consider a variable with m (number of classes) nominal values; on every line it will represent a class and on every column a cluster (or the estimated class). The contingency matrix could be consider also as an *association matrix* between real/true classes and clusters(estimated classes).

Table 4: The problem of splitting examples in one estimated class ***yes*** and ***no*** membership

		Estimated Class	
		yes	**no**
True Class	Positive examples	tp	fn
	Negative examples	fp	tn

In an ideal situation, the clustering method put examples in exact one correct class with no mistakes, no false positive and no false negative examples like in the example on Table 7. In this type of situation is no problem

Table 5: Transformation of the problem of one class into one with two classes

		Estimated Class	
		K_1	K_2
True Class	C_1	$a_{11} = tp$	$a_{12} = fn$
	C_2	$a_{21} = fp$	$a_{22} = tn$

Table 6: Generalization to n classes

		Estimated Class					
		K_1	K_2	$\cdots$	K_j	$\cdots$	K_n
True Class	C_1	a_{11}	a_{12}		a_{1j}		a_{1n}
	C_2	a_{21}	a_{22}		a_{2j}		a_{2n}
	$\cdots$						
	C_i	a_{i1}	a_{i2}		a_{ij}		a_{in}
	$\cdots$						
	C_n	a_{n1}	a_{n2}		a_{nj}		a_{nn}

to identify what cluster correspond with every class and to reorder lines and columns for obtaining a matrix with all nonzero elements on the principal diagonal.

Table 7: Example of an ideal clustering/classification

		Estimated Class (Cluster)			
		K_1	K_2	K_3	K_4
True Class	C_1	80	0	0	0
	C_2	0	50	0	0
	C_3	0	0	30	0
	C_4	0	0	0	20

For a possible situation more close to real situations like that from Table 8 we have some examles assigned to different other groups and we can consider them as false positive or false negative. It is important to be aware that are different types of false positives, one type for every class other than the true one; and different types of false negatives for every other cluster than the one assigned with the associated class.

Because we consider valid that association that maximize the *Accuracy* we want to maximize the sum of the cells corresponding to found associations. So we will rearrange the lines and the columns so that the sum of the cells on the principal diagonal to be maximum and we will obtain the correspondence: $C_1 - K_1$, $C_3 - K_2$, $C_2 - K_3$ and $C_4 - K_4$.

Table 8: Example of a not ideal clustering/classification

		Estimated Class (Cluster)			
		K_1	K_2	K_3	K_4
True Class	C_1	76	1	1	2
	C_2	0	0	30	0
	C_3	1	47	1	1
	C_4	3	2	0	15

4 Evaluation method

Table 9: A complex example of clustering

		Estimated Cluster			
		K_1	K_2	K_3	K_4
True Class	C_1	94	27	70	44
	C_2	69	56	10	4
	C_3	21	53	35	19
	C_4	0	33	1	3

Because in this more complex distribution of examples the great values in some of the cells are not very significant because it is possible to have many examples in one class and/or one cluster. To establish the importance of the value in one cell we could compare it with an uniform random distribution of the examples. The method used [WFH11] for this goal is to summarize the values on every line l_i and on every column n_j. The total number of examples is Sum.

$$l_i = \sum_{j=1}^{n} a_{ij}$$

$$n_j = \sum_{i=1}^{n} a_{ij}$$

$$Sum = \sum_{i=1}^{n} l_i = \sum_{j=1}^{n} n_j$$

Table 10: Working on the matrix

		Estimated Cluster				
		K_1	K_2	K_3	K_4	
True Class	C_1	94	27	70	44	$l_1 = 235$
	C_2	69	56	10	4	$l_2 = 139$
	C_3	21	53	35	19	$l_3 = 128$
	C_4	0	33	1	3	$l_4 = 37$
		$n_1 = 184$	$n_2 = 169$	$n_3 = 116$	$n_4 = 70$	$Sum = 539$

If all the examples were uniform random distributed according with the numbers belonging to classes and clusters a computed proportionally:

$$f_{ij} = \frac{l_i \cdot n_j}{Sum} \tag{5}$$

Table 11: Uniform random distribution of examples

		Estimated Cluster			
		K_1	K_2	K_3	K_4
True Class	C_1	80	74	51	31
	C_2	47	44	30	18
	C_3	44	40	28	17
	C_4	13	12	8	5

The majority of textbooks related with contingency matrix uses the square of the difference between the value a_{ij} and the uniform randomized value f_{ij} because they try to determine if it is true the hypothesis that exist an association between variables [JS11], [Fle81]. The squared values are used form obtaining χ^2 evaluation value or the *kappa* criterion. We use the differences normalized, but not squared, Δ_{ij} because we intend to find also the best possible association between classes and clusters according with the three criteria above mentioned.

$$\Delta_{ij} = \frac{a_{ij} - f_{ij}}{\sqrt{f_{ij}}} \tag{6}$$

For obtaining the most adequate assignment of clusters to true classes I propose three criteria:

1. each cluster is associated with a class and only with one;
2. assignment of the cluster with the class is better if the number of the related cell is greater;
3. maximizing the total sum of normalized differences Δ_{ij} on the cells of the association (rearranged on the principal diagonal of the matrix).

Table 12: Assignment of clusters to classes

		Estimated Cluster			
		K_1	K_2	K_3	K_4
True Class	C_1	1.53	-5.43	2.73	$2.44_{(3)}$
	C_2	$3.12_{(2)}$	1.88	-3.64	-3.3
	C_3	-3.43	2.03	$1.41_{(4)}$	0.58
	C_4	-3.55	$6.28_{(1)}$	-2.46	-0.82

5 Conclusions

In conclusion, we propose a method fos assigning the association (the correspondence) between the true class and the estimated cluster. Reordering the lines and the columns we obtain a matrix with the principal diagonal representing the values of guessed classes. Using that form of the matrix could calculate the *Acurracy* of the group. In the future we will try to make the procedure feasible in real time for great number of classes and to try to use for establishing the optimal number of clusters.

References

[And73] M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press, 1973.

[CM12] Radu Creţulescu and Daniel Morariu. *Text Mining: tehnici de clasificare şi clustering al documentelor*. Editura Albastră, Cluj-Napoca, 2012.

[FFF14] Robert H. Fletcher, Suzanne W. Fletcher, and Grant S. Fletcher. *Clinical epidemiology: the essentials*. Wolters Kluwer/Lippincott Williams & Wilkins Health, 5th ed edition, 2014.

[Fle81] J.L. Fleiss. *Statistical Methods for Rates and Proportions*. Wiley Series in Probability and Statistics. Wiley, 1981.

[JS11] Nathalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, New York, NY, USA, 2011.

[Rij79] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[VBCM10] Daniel Volovici, Macarie Breazu, Gabriel Dacian Curea, and Daniel Ionel Morariu. Statistical methods for performance evaluation of web document classification. *Studies in Informatics and Control*, 19(2):169, 2010.

[WFH11] Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.