Information Sciences 374 (2016) 179–192

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Effect of class imbalance on quality measures for contrast patterns: An experimental study[☆]

Octavio Loyola-González[a,b,*], José Fco. Martínez-Trinidad[a], Jesús Ariel Carrasco-Ochoa[a], Milton García-Borroto[c]

[a] *Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Enrique Erro No. 1, Sta. María Tonanzintla, Puebla, C.P. 72840, México*
[b] *Centro de Bioplantas, Universidad de Ciego de Ávila, Carretera a Morón km 9, Ciego de Ávila, C.P. 69450, Cuba*
[c] *Instituto Superior Politécnico José Antonio Echeverría, Calle 114 No. 11901, Marianao, La Habana, C.P. 19390, Cuba*

ARTICLE INFO

Article history:
Received 4 March 2016
Revised 13 September 2016
Accepted 16 September 2016
Available online 19 September 2016

Keywords:
Contrast patterns
Quality measures
Imbalanced databases

ABSTRACT

Contrast pattern-based classifiers rely on the discriminative power of contrast patterns. For this reason, many quality measures have been proposed to evaluate the quality of a contrast pattern. These measures allow to distinguish among contrast patterns with low and high discriminative ability for classification. In the literature, many comparative studies among quality measures for contrast patterns have been proposed but all of them were performed without taking into account the class imbalance level. However, in many class imbalance problems, those patterns extracted from the minority class have low support, which could negatively affect their discriminative ability. Therefore, in this paper, we present an experimental study of the effect of class imbalance on quality measures for contrast patterns. This study determines which quality measures for contrast patterns are the best for class imbalance problems; both regarding and disregarding the class imbalance level. Also, for the best quality measures we performed a pairwise comparison to determine which other quality measures have statistically similar behavior to them. This will help to simplify future research since it can be used only one quality measure among those with similar performance.

1. Introduction

A *pattern* is an expression written in a language that covers a set of objects for a specific problem. In classification problems, a *contrast pattern* (CP) is a pattern that covers significantly more objects belonging to a class than objects belonging to other classes [11].

In several classification problems, contrast pattern-based classifiers are used to create models that are easy to understand by an expert in the application domain [11,36,41]. However, most of these algorithms often extract a lot of contrast patterns for each class of the problem [13,14,16]. Hence, it is important to select just those contrast patterns that will provide a model easy to understand with high classification accuracy.

[☆] This work was partly supported by National Council of Science and Technology of Mexico under the scholarship grant 370272.

[*] Corresponding author.

E-mail addresses: octavioloyola@inaoep.mx, octavioloyola@gmail.com (O. Loyola-González), fmartine@inaoep.mx (J.Fco. Martínez-Trinidad), ariel@inaoep.mx (J.A. Carrasco-Ochoa), mgarciab@ceis.cujae.edu.cu (M. García-Borroto).

http://dx.doi.org/10.1016/j.ins.2016.09.040
0020-0255/© 2016 Elsevier Inc. All rights reserved.

In supervised classification, a *quality measure* (QM) assigns a higher value to a pattern when it discriminates better objects of a class from objects in the other classes. Consequently, a quality measure allows generating a pattern ranking based on the discriminative power of the patterns, which can be used for filtering and selecting the best contrast patterns [15,24,34,41]. Thus, in this paper, we will say that a quality measure Q_1 has *better behavior* than another quality measure Q_2 if, at the classification stage, the patterns selected from the ranking of Q_1 provide better accuracy than those coming from the ranking of Q_2.

In the literature, several quality measures for contrast patterns have been proposed [1,2,4–8,12,20,25–27,29,44,46,49,50]. Also, since these quality measures are based on different approaches issued by experts in application domains, many comparative studies among quality measures for contrast patterns have been reported [5,11,15,18,24,29,34,41]. All these studies have been performed without taking into account the class imbalance level, i.e., these studies were done over datasets with low imbalance level. However, in some classification problems, such as online banking fraud detection, prediction of protein sequences, detection of diabetic retinopathy and maculopathy, and face recognition; the objects are not equally distributed into the problem classes. Commonly, there is a class (known as minority class) with significantly less objects than the others classes of the problem. These problems are known as class imbalance problems or problems with imbalanced databases [36].

The main difficulties that arise when mining contrast patterns from class imbalance problems are the following:

(i) Contrast patterns extracted from the minority class are commonly fewer (sometimes there are no patterns at all) than the contrast patterns extracted from other classes.
(ii) Contrast patterns from the minority class have lower supports than the patterns from the other classes.

Such difficulties might negatively affect the minority class at the classification stage, while the remaining classes of the problem tend to be favored [31,35].

Additionally, as we have already commented, all the studies of quality measures for contrast patterns have been proposed without taking into account the impact of the class imbalance level over the quality measure results. Therefore, the following interesting questions arise: Do quality measures have the same behavior for contrast patterns extracted from imbalanced databases? Is there any relation between the behavior of quality measures and the different levels of class imbalance?

To the best of our knowledge, this is the first experimental study of the effect of class imbalance on quality measures for contrast patterns. In this paper, we first identify a set of quality measures for contrast patterns that obtain significantly better behavior in our experiments; regarding and disregarding the class imbalance level. After, we detect a set of quality measures which are highly affected by class imbalance. Then, we determine which quality measures have statistically similar behavior to the best quality measures, with the aim of simplifying future research since it is possible to use only one quality measure among those with similar performance. Finally, we contrast the results of this study against the results presented by Loyola-González et al. [34] where a similar study was performed without taking into account the class imbalance level.

The rest of the paper is organized as follows: Section 2 provides some basic concepts and a revision of the state-of-the-art on quality measures for contrast patterns. Section 3 contains materials and methods used throughout the paper. Section 4 presents the main results obtained in this study as well as a detailed discussion of them. Finally, conclusions and future work are presented in Section 5.

2. Quality measures for contrast patterns

A quality measure aims to determine the discriminative power of patterns discovered through a data mining process. However, this is a subjective and complex task that currently is an active and important research field in data mining [11,15,18,19,34,37].

Based on many studies [18,19,37], quality measures for contrast patterns can be categorized in two groups:

- *Objective*, which are based on probabilities or statistics. The aim is to evaluate the ability of a pattern for discriminating objects of a class from objects in other classes [37,38].
- *Subjective*, which are based on a criterion issued by an expert in the application domain [30,43].

Objective measures are the most used for experimental studies because they don't take into account neither the context of the application domain nor the goals and background knowledge of experts [19]. Since subjective measures are based on a specific criterion issued by an expert in the application domain, we do not include these measures in this study.

An objective quality measure can be defined as a function $q(P, C, \bar{C}) \to R$, which assigns a higher value to a pattern P when it discriminates better objects between a class C and the remaining problem classes $\bar{C}$. (The classes form a partition of the universe $U = C \cup \bar{C}, C \cap \bar{C} = \emptyset$) [15,34].

Despite their apparent differences, some quality measures are based on a combination of other quality measures. For example, *Cosine* ($\sqrt{p(C|P)p(P|C)}$) [49] is a combination of *Support* ($p(P|C)$) [1] and *Confidence* ($p(C|P)$) [1]. On the other hand, some quality measures expressed through apparently different equations are equivalent [19,34,41].

In the literature, there are many studies on quality measures for contrast patterns. In 1991, Piatetsky-Shapiro [45] proposed three principles that should be complied by an objective measure. Nevertheless, some researchers [19,48] consider that one of the proposed principles may seem too rigid for some quality measures. Later, in 2001, An and Cercone [5] performed a meta-analysis over several quality measures. The authors found rules with the aim to offer an insight about which