Expert Systems With Applications 83 (2017) 104–113

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Evaluation of quality measures for contrast patterns by using unseen objects

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

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

ARTICLE INFO

Article history:
Received 17 October 2016
Revised 18 January 2017
Accepted 18 April 2017
Available online 19 April 2017

Keywords:
Contrast patterns
Quality measures
Quality estimation
Meta-analysis

ABSTRACT

Contrast patterns, which lie in the core of most understandable classifiers, are frequently evaluated by quality measures. Since many different quality measures are available, they should be compared to select the most appropriate for each applications. This paper introduces a method to compare quality measures, using a set of mined patterns and a collection of objects not used for mining. The comparison is performed by correlating quality values with a quality estimation of the patterns. Additionally, a meta-learning study is performed to show that combining quality measures could be better than using the best single measures in isolation. The results of this paper can help researchers to create new quality measures or to find new combinations of quality measures to create better understandable classification systems.

1. Introduction

A supervised classifier predicts the class of a query object based on a model built using a training sample. Although an accurate prediction is an important component of the classifier quality, the lack of comprehensibility of classification results may cause reluctance to use certain classifiers. For example, when credit has been denied to a customer, the Equal Credit Opportunity Act of the US requires the financial institution to provide the reasons for rejecting the application; indefinite or vague reasons for denial are illegal (Martens, Baesens, Gestel, & Vanthienen, 2007).

A pattern is an expression defined in a language that describes a collection of objects. For example, [$Age < 6 \wedge Sex =$ "*Female*"] is a pattern, expressed in conjunctive form, that describes a set of girls. A pattern that appears significantly more in a group or class than in the remaining groups or classes, capturing existing contrasts, is named *contrast pattern*. For example, since the pattern [$Albumin > 3.85 \wedge Bilirubin \in [0.65, 2.55]$] appears in 57% hepatitis survivors and only in the 3% non-survivors, it can be considered a contrast pattern of class *Survivor*.[1] Contrast patterns lie in the core of most understandable classifiers (Dong, 2012). Contrast patterns contained in a query object can be used to find its class, and also provide an explanation of the classification in terms that are easy to understand by experts in real-world applications. Examples of contrast pattern-based classifiers are emerging patterns, decision trees, decision rules, and CARs.

To differentiate good from bad contrast patterns, or to rank contrast patterns, a form to numerically measure the goodness of a contrast pattern should be used. These measures received different names in different communities, like quality measures (Loyola-González, García-Borroto, Martínez-Trinidad, & Carrasco-Ochoa, 2014), interestingness measures (Geng & Hamilton, 2007), rule quality (Natarajan & Shekar, 2007), pattern discriminative ability (Bailey, 2012), and subgroup evaluation criteria (Herrera, Carmona, Gonzalez, & del Jesus, 2011).

Perhaps the two most used quality measures are *support* and *confidence*. Support (Agrawal, Imieliński, & Swami, 1993) is the probability of the pattern to appear on its class, so it can be used to filter between random patterns (which usually have low supports) and useful patterns. Confidence (Agrawal et al., 1993), on the other hand, measures the probability of the class once the pattern

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

[1] This pattern was extracted from database *hepatitis*, from the UCI repository (Bache & Lichman, 2013).

http://dx.doi.org/10.1016/j.eswa.2017.04.038
0957-4174/

is found. In this way, if a pattern with high confidence appears on the query object, it is a strong signal that the object should belong to the pattern's class.

There are different approaches to compare quality measures. One approach is to test the compliance of some theoretical properties, like symmetry and invariance to scale (Geng & Hamilton, 2007). Unfortunately, there is no general agreement about which properties are better for a given task, and no empirical relation between the properties and accuracy has been found. Another approach is to evaluate a quality measure by using it in some classification tasks (Loyola-González et al., 2014). In this way, the accuracy of the classifier can be used as an indirect estimation of the performance of the quality measure. Unfortunately, other factors that affect the accuracy, like minimal support thresholds or aggregation scheme, can also bias the results.

Some interesting questions arise here: How can we empirically evaluate the quality measures for supervised classification? How can we estimate the quality of a contrast pattern, in order to compare it with the evaluation provided by another quality measure? Are the result dependent on some database characteristics? Can synergistic combinations of quality measures outperform their individual measures? This paper presents answers to these questions by introducing a new method to estimate the quality of a contrast pattern based on the use of an unseen object collection. In order to empirically evaluate a quality measure, we propose correlating its results and the estimated quality. We also perform a study about how some database properties influence the correlation results. Additionally, we present a meta-learning analysis to show that combining quality measures could be better than using the best single measures in isolation.

This paper consists of five sections. Section 2 presents the related works. Section 3 contains materials and methods used throughout the paper. Section 4 presents the main results of this paper as well as a detailed discussion of the experimental results. Finally, conclusions are presented in Section 5.

2. Related work

Quality measures appear in diverse applications through different pattern recognition fields, including the following:

- Decision rules: Algorithms based on decision rules usually use quality measures to sort and filter mined rules (An & Cercone, 2001).
- Decision trees: Inducing decision trees is usually stopped if the pattern associated with the node gets a quality value above certain threshold, like χ^2 quality measure in C4.5 (Kuncheva, 2004). Additionally, confidence can be used to select the best tree in an ensemble (Guerrero-Enamorado & Ceballos-Gastell, 2016; Hai-Long, Yew-Kwong, & Wee-Keong, 2015).
- Emerging patterns: the growth rate is used to select which patterns should be used to build the model (Dong & Li, 1999), while χ^2 is used to define a particular type of high quality pattern (Ramamohanarao & Fan, 2007).
- Association rules: Quality measures help to reduce the number of mined association rules, to facilitate the user's comprehension (Czibula, Marian, & Czibula, 2015; Diatta, Ralambondrainy, & Totohasina, 2007).

2.1. Evaluating contrast patterns using quality measures

In this paper, a probabilistic notation for representing quality measures is used. The probability of finding an object with a given pattern P is denoted by $p(P)$, while the probability of not finding an object with a given pattern is denoted as $p(\neg P)$. With respect to a given class C, probabilities of finding an object of a given class or from a different class are denoted as $p(C)$ and $p(\neg C)$, respectively. Joint probabilities are then denoted as $p(PC)$, $p(P\neg C)$, and so on. The number of objects in the database will be denoted as N.

Estimating the importance and accuracy of a contrast pattern only by confidence ($p(C|P)$) and support ($p(PC)$) has several drawbacks. For example, mining contrast patterns based only on support and confidence can lead up to 95% of mined patterns to be useless (Berzal, Blanco, Sanchez, & Vila, 2002). The main drawback of the confidence is that it is unable to detect statistical independence between C and P. Other measures solve this drawback, like *Brins* ($\frac{p(P)p(\neg C)}{p(p\neg C)}$) (Brin, Motwani, Ullman, & Tsur, 1997b) and *Lift* ($\frac{p(PC)}{p(P)p(C)}$) (Piatetsky-Shapiro & Steingold, 2000). These measures return values close to 1 if there is independence, but since they are unbounded it is hard to compare their values from different patterns or to establish a threshold. A similar idea, but using difference instead of division is introduced in *leverage* ($p(C|P) - p(P)p(C)$) (Webb & Zhang, 2005). Another disadvantage of the confidence is that it cannot be used to compare patterns from different classes if the number of objects per class is very different. To solve this drawback, the *centered confidence* ($p(C|P) - p(C)$) (Lenca, Vaillant, Meyer, & Lallich, 2007) can be used.

The main drawback of the support is that it is mainly dependent on the sampling procedure used to collect the data, therefore many useless patterns might have higher support than many useful patterns. Another problem with the support is that it is very sensitive to class imbalance, because the probability of the majority class will usually be higher for any pattern. This problem can be solved using *coverage* ($p(P|C)$) (An & Cercone, 2001), but it does not penalizes the appearance of the pattern in the other classes. Mixing confidence and coverage tries to get the best of both measures, as it is done in the quality measure *cosine* ($\sqrt{p(C|P)p(P|C)}$) (Tan, Kumar, & Srivastava, 2004).

A different idea to evaluate patterns is to contrast the support of a pattern in a class with respect to the other classes. This idea is used by the quality measures: *growth rate* ($\frac{p(P|C)}{p(P|\neg C)}$) (Dong & Li, 1999), *support difference* ($p(P|C) - p(P|\neg C)$) (Bay & Pazzani, 1999), *Sebag–Shoenauer* ($\frac{p(PC)}{p(P\neg C)}$) (Sebag & Schoenauer, 1988), and *least contradiction* ($\frac{p(PC)-p(P\neg C)}{p(C)}$) (Azé & Kodratoff, 2002). Other quality measures are based on combinations of other qualities, like *strength*, based on growth rate, weighted sum of confidence and coverage, and product of confidence and coverage.

The collection of all quality measures used in this paper appears on Tables Table 9 and A.2, in Appendix A. Each quality measure is described by the symbol used, the names used by different authors (with proper reference), and their equation using the probabilistic notation.

2.2. Evaluating and comparing quality measures

To evaluate a quality measure, the fulfillment to some important criteria have been investigated. Geng and Hamilton (2007) reviewed 39 quality measures, evaluating 11 criteria. Criteria evaluated are as heterogeneous as the symmetry under variable permutation, the invariance to scaling, the monotonicity, and the easiness of expressing the semantic of the measure. In a similar way, Lenca et al. (2007) performed a similar study with 20 measures and a different set of properties. In that paper, measures were correlated by their behavior on a set of 10 databases. Unfortunately, there is no general agreement about which are the best criteria for evaluating quality measures and there are contradictions among authors about which are the most important ones (Geng & Hamilton, 2007).

A different approach for comparing quality measures is to evaluate them using some indirect measure. An example of this approach compares 36 quality measures on two different databases,