Comparing Quality Measures for Contrast Pattern Classifiers

Milton García-Borroto[1], Octavio Loyola-Gonzalez[1,2], José Francisco Martínez-Trinidad[2], and Jesús Ariel Carrasco-Ochoa[2]

[1] Centro de Bioplantas. Carretera a Moron km 9, Ciego de Avila, Cuba
{mil,octavioloyola}@bioplantas.cu
[2] Instituto Nacional de Astrofísica, Óptica y Electrónica. Luis Enrique Erro No. 1, Sta. María Tonanzintla, Puebla, México,C.P. 72840
{ariel,fmartine}@ccc.inaoep.mx

Abstract. Contrast pattern miners and contrast pattern classifiers typically use a quality measure to evaluate the discriminative power of a pattern. Since many quality measures exist, it is important to perform comparative studies among them. Nevertheless, previous studies mostly compare measures based on how they impact the classification accuracy. In this paper, we introduce a comparative study of quality measures over different aspects: accuracy using the whole training set, accuracy using pattern subsets, and accuracy and compression for filtering patterns. Experiments over 10 quality measures in 25 repository databases show that there is a huge correlation among different quality measures and that the most accurate quality measures are not appropriate in contexts like pattern filtering.

Keywords: quality evaluation, contrast patterns, emerging patterns.

1 Introduction

A supervised classifier predicts the class of a query object based on the relationships it finds among its description and the descriptions of the objects in the training sample. An accurate prediction is an important component of the classifier behavior, but in some domains the classifier and its results should be also easily understandable by the user. In some cases, the lack of comprehensibility may cause a 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 [1].

Most understandable classifiers are based on patterns, which are expressions defined in some language that describe some properties of an object collection. An important family of understandable and accurate classifiers is based on *contrast patterns* [2]. A contrast pattern is a pattern that describes differences between two given datasets. Then, if a contrast pattern appears in a query object, it can be taken as an evidence towards the pattern class.

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 311–318, 2013.

Algorithms for finding contrast patterns or for classification based on contrast patterns usually employ a quality measure for evaluating the discriminative power of a pattern [3]. Since many authors have introduced different quality measures, it is important to perform both theoretical and experimental studies, in order to help users to select the appropriate one for a given task. Nevertheless, most published studies present comparisons mainly based on the classifier accuracy [4]. In this paper, we introduce a comparative study about a set of quality measures for contrast patterns over the following aspects:

- Accuracy using the whole training sample
- Accuracy using percentages of the best evaluated patterns
- Ability to be used to filter patterns:
  - Accuracy using a classifier based on aggregated support
  - Compression ratio

Additionally, we perform a correlation study that reveals that all the studied quality functions can be grouped into four groups. Functions in the same group have so similar behaviour that it is enough to use one of them in future comparative studies. This result is consistent with other results shown in this paper.

2 Quality Measures

In this section, we succinctly describe the quality measures used in this paper. More details can be found on each associated reference. Lets define the function count($pattern, set$) as the number of objects in *set* containing *pattern*, and the function

$$\text{support}(pattern, set) = \frac{\text{count}(pattern, set)}{|set|}$$

as the ratio of objects in *set* containing *pattern*. We also consider, for a given universe U, $|U| = N$, $|I| = \text{count}(I, U)$, $\neg I$ as the pattern negation, and $|\neg I| = \text{count}(\neg I, U) = N - |I|$.

A quality measure $q(I, D_p, D_n) \rightarrow R$ returns a value, which is larger while the pattern I better discriminates objects between the positive class D_p and the negative class D_n (both classes form a partition of the universe $U = D_p \cup D_N$,$D_p \cap D_N = \emptyset$). In this paper, we investigate the following quality measures:

Confidence. $Conf = \text{count}(I, D_p)/\text{count}(I, U)$, predictive ability of the pattern for the positive class [5].

Growth Rate. $GR = \text{support}(I, D_p)/\text{support}(I, D_n)$, ratio of the positive and negative class supports [6].

Support Difference. $SupDif = \text{support}(I, D_p) - \text{support}(I, D_n)$, support difference between positive and negative classes [7].

Odds Ratio. $Odds = \frac{\text{support}(I,D_p)/(1-\text{support}(I,D_p))}{\text{support}(I,D_n)/(1-\text{support}(I,D_n))}$, ratio of the pattern odds from D_p to D_n [8].

Gain. $Gain = \text{support}(I, D_p)(\log \frac{\text{support}(I,D_p)}{\text{support}(I,U)} - \log \frac{|D_p|}{|U|})$ [9].

Length. $Length = 1/|I|$, inverse of the number of items in $|I|$. We use the inverse because shorter patterns are more desirable for discrimination [10].

Chi-square. $\chi^2 = \sum_{X\in\{I,\neg I\}} \sum_{Y\in\{D_p,D_n\}} \frac{(\text{count}(X,Y)-E(X,Y))^2}{E(X,Y)}$, where $E(X,Y)$ is the expected frequency count of pattern X in class Y. This measure assesses how significantly different is a pattern support with respect to the universe support [7].

Mutual Information. Estimates how correlated is the pattern distribution with respect to the expected pattern frequencies per class [5].

$$MI = \sum_{X\in\{I,\neg I\}} \sum_{Y\in\{D_p,D_n\}} \frac{\text{count}(X,Y)}{N} \log \frac{\text{count}(X,Y)/N}{|X||Y|/N^2}$$

Weighted Relative Accuracy. $WRACC = \frac{|I|}{|D_p|+|D_n|}(\frac{\text{count}(I,D_p)}{|I|} - \frac{|D_p|}{N})$, used in the subgroup discovery field [11].

Strength. $Strength = \frac{\text{support}^2(I,D_P)}{\text{support}(I,D_p)\,\text{support}(I,D_n)}$ measures how strongly the pattern appearance indicates the class of the query instance containing it. [12]

Although most of these quality measures were defined for two-class problems, we use them in multi-class problems using the one-vs-rest approach [13]

3 Comparing Quality Functions

Accuracy Comparison. A good quality measure should assign higher evaluations to patterns that contribute more to the correct classification of query objects. That is why it is frequent to evaluate quality measures using the accuracy of a supervised classifier, which uses the measure information during the classification process. Nevertheless, in a contrast pattern classifier there are many parameters that impact the classifier accuracy like thresholds, aggregation scheme, and normalization procedures, among others. Then, using the classifier accuracy as an estimation of the behavior of the quality measure can be error prone.

To minimize the parameter influence in the classification accuracy, we perform the accuracy comparison using a simple classification algorithm. This algorithm assigns O to the class with the maximum aggregated support, calculated with the top-quality patterns from those contained in O. As this classifier is mostly based on the quality values, we expect that accuracy differences are mostly due to the specific quality measure behavior. In this way, we can safely estimate the behavior of the quality measure based on the accuracy. The pseudocode of the algorithm is the following:

Input: Set of patterns P, quality function q, query object O
Output: Class assigned to O

1. $S \longleftarrow$ patterns in P contained in O
2. $MaxQual \longleftarrow \text{argmax}_s(q(s))$
3. $S' \longleftarrow \{s \in S : q(s) = MaxQual\}$
4. **return** class with maximum aggregated support of patterns in S'

Accuracy Comparison Using a Pattern Subset. If we use a small subset of the pattern collection in a supervised classifier, the global classifier accuracy usually deteriorates. This behavior is mainly due to query objects that do not contain any pattern, causing classifier abstention to appear. If we select a percentage of the best patterns, using some quality measure, we expect the best quality measure to obtain the highest accuracy.

Filtering Patterns. Most pattern filtering methods iterate through the pattern collection, selecting those that fulfill some criterion. To obtain a subset with the best patterns, the pattern collection is sorted according to some quality measure. In this section, we evaluate the ability of quality measures to be used in a pattern filtering procedure. We use the following filtering algorithm:

Input: Set of patterns P, quality function q, training sample T
Output: Selected patterns R

1. $R \longleftarrow \emptyset$
2. **foreach** $o \in T$
   (a) Find $S =$ patterns in P contained in o
   (b) **if** $S \cap R = \emptyset$ **then** add to R the pattern in S with the highest q value
3. **return** R

The filtering algorithm uses a greedy heuristic to find the smallest pattern subset covering the full training sample, selecting the pattern with the highest quality evaluation. In this way, the best quality measure is expected to obtain the smallest and most accurate filtered subset.

4 Experimental Results

Experimental Setup. For mining patterns, we use LCMine miner [14], pruning each decision tree in order to obtain non pure patterns. LCMiner extracts patterns from a collection of decision trees, induced using a particular diversity generation procedure. For comparing the accuracy of a classifier based on emerging patterns, we perform a Friedman test with all the results [15]. Then, when we find significant differences, we perform the Bergmann-Hommel dynamic post-hoc, because it is more powerful than the classical Nemenyi and Holm procedures [16]. Post-hoc results are shown using *critical distance* (CD) diagrams, which present the order of the classifier accuracy, the magnitude of differences between them, and the significance of the observed differences in a compact form [15]. In a CD diagram, the rightmost classifier is the best classifier, while two classifiers sharing a thick line means they have statistically similar behavior. We use a 2 times 5 fold cross validation procedure averaging results, as suggested in [15].

All databases used for experiments were taken from the UCI repository of Machine Learning [17]. We selected small databases with balanced classes, because emerging pattern classifiers are very sensitive to class imbalance [18]. According to the feature type, there are pure numerical, pure categorical, and mixed

databases. The number of features ranges from 4 to 60. Databases are *breast-cancer*, *breast-w*, *cleveland*, *colic*, *credit-a*, *credit-g*, *crx*, *cylinder-bands*, *diabetes*, *haberman*, *heart-c*, *heart-h*, *heart-statlog*, *hepatitis*, *ionosphere*, *iris*, *labor*, *lung-cancer*, *sonar*, *tae*, *tic-tac-toe*, *vote*, *wdbc*, *wine*, and *wpbc*.

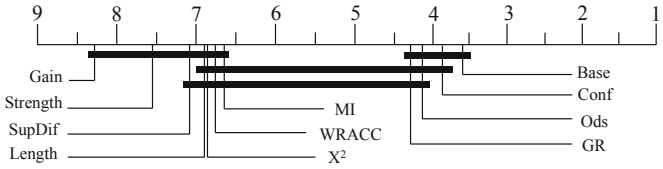

Fig. 1. CD diagram for accuracy comparisons

Accuracy Comparison. Results of the accuracy comparison (Figure 1) reveals that the quality measures Conf, Odds, and GR obtain the most accurate classifier. Their results are statistically similar to the base classifier, which uses the whole pattern collection to achieve classification. The good behavior of GR is not surprising, because it has been used as quality function in many papers. Additionally, it is used in the definition of *emerging patterns* [6], which are contrast patterns whose GR is above certain threshold.

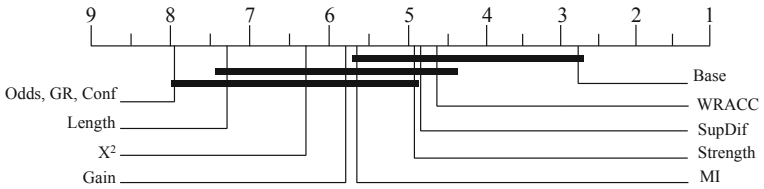

Fig. 2. CD diagram for accuracy comparisons taking the 10% of the best patterns

Accuracy Comparison Using a Pattern Subset. To compare accuracies using pattern subsets, we created pattern collections containing different percentages of the whole collection. We finally chose 10%, because it is the lowest value where accuracies are significantly similar to the unfiltered classifier, so the filtering procedure does not significantly deteriorates the classifier. Results shown in Figure 2 reveal an outcome not consistent with Figure 1, because the top accurate classifiers like Conf, GR and Odds had the worst behavior. To explain this behavior we must realize that Conf, GR and Odds returns the same value on all patterns having zero support on the negative class. In this way, a pattern with support $(1, 0)$ is considered as good as a pattern with support $(0.001, 0)$. On the other hand, the quality measures with the best results like WRACC, SupDif and Strength can easily differentiate among them, assigning the former a significant higher quality value.

Filtering Patterns. Figures 3 and 4 show the accuracy and compression ratio results from the pattern filtering experiment. Results are consistent with those

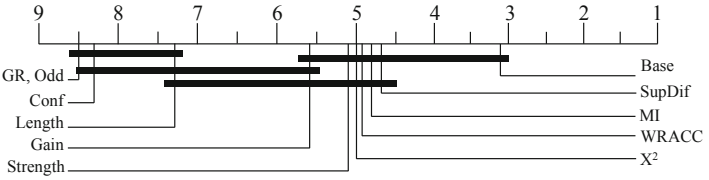

Fig. 3. CD diagram for accuracy comparisons with filtered pattern subset using each quality measure

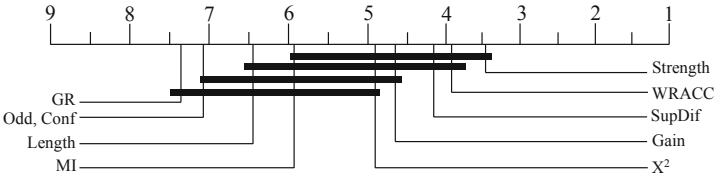

Fig. 4. CD diagram for compression ratio of the filtering procedure using each quality measure

previously shown, being the most accurate quality measures those that distinguish among single class patterns. As the compression ratio ranges from 0.01 to 0.10, with average 0.05, it looks promising to follow these ideas to obtain future pattern filters.

5 Correlation Study

According to their definitions, many quality measures seem to be very similar, being most of them variations of other measures. Additionally, during the analysis of the experiments, we also realize that many quality measures behave very similarly in all the experiments and databases. That is why we performed a two-way Pearson correlations analysis using the quality values extracted from all the emerging patterns per database. A Pearson correlation is a measure of association between two numerical variables. Pearson values range from -1 (perfect negative correlation) to 1 (perfect positive correlation). Since the results are highly consistent among all databases, we only show in Table 1 the results in *colic* database.

Correlation results allows us to cluster measures in four different groups, with very high inner correlations and very low outer correlations. These groups are completely consistent with other experimental results presented in this paper. The groups are the following:

Group 1. Conf, GR, Odds
Group 2. WRACC, Gain, SupDif, Strength, MI
Group 3. Length
Group 4. χ^2

Table 1. Non-trivial correlations among quality functions on *colic* database. An "X" appears where qualities have correlations above 0.75

	χ^2	Conf	Gain	GR	Length	MI	Odds	Strength	SupDif	WRACC
χ^2										
Conf				X			X			
Gain						X		X	X	X
GR	X						X			
Length										
MI		X						X	X	X
Ods	X			X						
Strength		X				X			X	X
SupDif		X				X		X		X
WRACC		X				X		X	X	

This clustering information can be useful in at least two tasks. First, we can simplify future researches on quality measures, using a single quality measure per cluster. Second, we can take a single quality measure per cluster to obtain a diverse measure collection, which can be used in some combination schemes.

6 Conclusions

In this paper, we present a comparative study about a set of quality measures for contrast patterns, which are used to evaluate the discriminative power of a pattern. We have addressed the necessity to provide theoretical and experimental comparisons to help users select among the existing measures, since previous studies lack comparisons over different relevant aspects.

After analysing experiments over 10 quality measures in 25 repository databases, we lead to the following conclusions:

- Many quality measures are strongly correlated, obtaining very similar results among them. Quality measures used in this paper can be grouped in four clusters: **Group1**={Conf, GR, Ods}, **Group2**={WRACC, Gain, Supdif, Strength, MI}, **Group3**{Length}, and **Group4**={χ^2}.
- On most databases, quality measures in Group1 are better estimations of the real pattern value for classification.
- Quality measures in Group1 can be very inaccurate in domains like pattern filtering, because they cannot distinguish among patterns supported by a single class

References

1. Martens, D., Baesens, B., Gestel, T.V., Vanthienen, J.: Comprehensible credit scoring models using rule extraction from support vector machines. European Journal of Operational Research 183(3), 1466–1476 (2007)

2. Dong, G.: Overview of Results on Contrast Mining and Applications. In: Dong, G., Bailey, J. (eds.) Contrast Data Mining: Concepts, Algorithms, and Applications, pp. 353–362. Chapman & Hall/CRC, United States of America (2012)
3. Fang, G., Wang, W., Oatley, B., Ness, B.V., Steinbach, M., Kumar, V.: Characterizing discriminative patterns. Computing Research Repository, abs/1102.4 (2011)
4. An, A., Cercone, N.: Rule quality measures for rule induction systems: Description and evaluation. Computational Intelligence 17(3), 409–424 (2001)
5. Bailey, J.: Statistical Measures for Contrast Patterns. In: Dong, G., Bailey, J. (eds.) Contrast Data Mining: Concepts, Algorithms, and Applications, pp. 13–20. Chapman & Hall/CRC, United States of America (2012)
6. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: ACM International Conference on Knowledge Discovery and Data Mining (KDD), pp. 43–52 (1999)
7. Bay, S.D., Pazzani, M.J.: Detecting change in categorical data: Mining contrast sets. In: ACM International Conference on Knowledge Discovery and Data Mining (KDD), pp. 302–306 (1999)
8. Li, J., Yang, Q.: Strong compound-risk factors: Efficient discovery through emerging patterns and contrast sets. IEEE Transactions on Information Technology in Biomedicine 11(5), 544–552 (2007)
9. Yin, X., Han, J.: CPAR: Classification based on predictive association rules. In: SIAM International Conference on Data Mining, SDM (2003)
10. Li, J., Li, H., Wong, L., Pei, J., Dong, G.: Minimum description length principle: Generators are preferable to closed patterns. In: 21st National Conf. on AI, pp. 409–414 (2006)
11. Lavrac, N., Kavsek, B., Flach, P.A., Todorovski, L.: Subgroup discovery with cn2-sd. Journal of Machine Learning Research with CN2-SD 5, 153–188 (2004)
12. Ramamohanarao, K., Fan, H.: Patterns based classifiers. World Wide Web 10, 71–83 (2007)
13. Abudawood, T., Flach, P.: Evaluation measures for multi-class subgroup discovery. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part I. LNCS (LNAI), vol. 5781, pp. 35–50. Springer, Heidelberg (2009)
14. García-Borroto, M., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Medina-Pérez, M.A., Ruiz-Shulcloper, J.: LCMine: An efficient algorithm for mining discriminative regularities and its application in supervised classification. Pattern Recognition 43(9), 3025–3034 (2010)
15. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)
16. García, S., Herrera, F., Shawe-Taylor, J.: An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. Journal of Machine Learning Research 9 (2008)
17. Merz, C.J., Murphy, P.M.: Uci repository of machine learning databases, Technical report, Department of Information and Computer Science, University of California at Irvine (1998)
18. Loyola-González, O., García-Borroto, M., Medina-Pérez, M.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., De Ita, G.: An Empirical Study of Oversampling and Undersampling Methods for LCMine an Emerging Pattern Based Classifier. In: Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F., Rodríguez, J.S., di Baja, G.S. (eds.) MCPR 2013. LNCS, vol. 7914, pp. 264–273. Springer, Heidelberg (2013)