A Study on Interestingness Measures for Associative Classifiers

Mojdeh Jalali-Heravi
University of Alberta, Canada
jalalihe@cs.ualberta.ca

Osmar R. Zaïane
University of Alberta, Canada
zaiane@cs.ualberta.ca

ABSTRACT

Associative classification is a rule-based approach to classify data relying on association rule mining by discovering associations between a set of features and a class label. Support and confidence are the de-facto "interestingness measures" used for discovering relevant association rules. The support-confidence framework has also been used in most, if not all, associative classifiers. Although support and confidence are appropriate measures for building a strong model in many cases, they are still not the ideal measures and other measures could be better suited.

There are many other rule interestingness measures already used in machine learning, data mining and statistics. This work focuses on using 53 different objective measures for associative classification rules. A wide range of UCI datasets are used to study the impact of different "interestingness measures" on different phases of associative classifiers based on the number of rules generated and the accuracy obtained. The results show that there are interestingness measures that can significantly reduce the number of rules for almost all datasets while the accuracy of the model is hardly jeopardized or even improved. However, no single measure can be introduced as an obvious winner.

Categories and Subject Descriptors

H.2.8 [**Database management**]: Data Mining

Keywords

Interestingness measures, Associative classifiers

1. INTRODUCTION

Associative classification [4, 23, 24] is a rule-based approach recently proposed to classify data by discovering associations between a set of features and a class label. To build an associative classification model, association rules whose consequent is a class label are generated using an association rule mining technique. Research shows promising results for associative classification and its potential for improvement to a more powerful classification paradigm.

Support and confidence are the default "interestingness measures" universally used for discovering relevant association rules. The support-confidence framework is similarly the most common framework used for mining and selecting rules of associative classifiers. Although these two measures are widely used, they are still not necessarily the ideal measures. This is because in many situations a huge set of rules is generated which could hinder the effectiveness in some cases for which other measures could be better suited. Yet, no systematic study has been done to identify a better framework or the most appropriate measure.

1.1 Background and Problem Definition

The associative classifier is a classifier that uses association rule mining in the training phase in order to generate classification rules. To use this classifier, datasets have to be transformed in a transactional format. Considering each attribute-value pair in a dataset as an item results in a transactional dataset in which a row of data looks like a transaction of items. Among items of each transaction, one is the class label of the related object. Using an association rule mining technique (e.g., Apriori [3], Eclat [36] or FP-growths [14]) on the resulting transactional data, frequent itemsets are mined and the ones of the form $\{A, c\}$ are extracted where A is a set of features and c is a class label (A and c are disjoint subsets of items). Among these frequent itemsets, the confident ones are chosen to build classification rules of the form $A \rightarrow c$. Then, these rules are used to predict class labels for objects with an unknown class.

As mentioned above, the support-confidence framework is the standard framework in association rule mining and inherited by associative classification. For a rule $A \rightarrow c$, support is the fraction of data samples having A and c together (i.e., $P(Ac)$). A rule is frequent if its support is greater than a minimum support threshold. Confidence is the conditional probability that a record is of class c given that it includes A (i.e., $P(c|A)$). A rule is confident if its confidence is above a minimum confidence threshold. To build an associative classifier, only strong rules, i.e., the rules that are both frequent and confident are used. Even with these two constraints, still a huge set of rules may be generated. Different approaches are used to prune the rules in the second phase of associative classifiers [9]. Finally, to classify an object O, two different approaches are typically used. The first way is to take into account only the best rule by choosing the rule that applies to O with the "highest rank" based on a defined ordering. The other way is to consider all rules that apply

SAC'10 March 22-26, 2010, Sierre, Switzerland.

to O by calculating the "average" value of the measure used in the defined ordering for the matching rules for each class label and choose for O the label with the highest average as prediction.

1.2 Disadvantages of Support and Confidence

While powerful in pruning the search space due to the antimonotonicity of support, the support-confidence framework has been criticized in the context of association rule mining by many authors [7, 2, 1]. For instance, it is difficult to tune. Choosing a large minimum support may lead to having only rules that contain obvious knowledge and missing exceptional cases that are interesting. On the other hand, assigning a low minimum support yields a huge number of rules which could be redundant or noisy.

Similarly, confidence is not a perfect measure as it considers nothing beyond the conditional probability of rules which may lead to confident associations but between statistically independent items.

Brin et al. show that with high support and confidence, a rule can even have negative correlation between its antecedent and consequent [7].

1.3 Approach

There are many rule interestingness measures already used in machine learning, data mining and statistics. Many different measures are introduced in the field of association rule mining as filters or rankers to weed-out the least relevant rules. All those measures can be directly applied to associative classifiers as well, although never tested or reported in the literature. This work focuses on probability-based objective rule interestingness measures for associative classification. Using these interestingness measures, there are two questions that should be answered:

First, can "interestingness measures" have any effect on the associative classifiers on its three different phases: rule generation, pruning and selection, so that the mining algorithm improves both in terms of increasing classification accuracy and decreasing the number of rules?

Second, if there are any improvements, is it possible to probe the best measure or measures which can beat the other measures for improving the results base on either the accuracy or the number of rules in all cases? There is a possibility that no one measure can be found to be effective in all circumstances. In this case, are there any relevant dataset characteristics or measure properties that can help build a classifier in order to predict an effective measure for a dataset?

To attempt to answer the above questions 20 different UCI datasets are used with 53 different measures to study the impact of "interestingness measures" on associative classifiers.

Section 2 describes the interestingness measures and their properties and introduces 53 different probability-based objective measures reportedly used in association rule mining. In Section 3 some related works studying interestingness measures are highlighted. The methodology of using interestingness measures in the three different phases of an associative classifier is discussed in Section 4. Experimental results, comparing the impact of interestingness measures on classification accuracy and the number of generated classification rules, are illustrated in Section 5.

2. INTERESTINGNESS MEASURES

Generating rules in association rule mining or with associative classifiers can lead to a very large set of rules which make it impossible, for even domain specialists, to study. Sifting through thousands or even millions of rules, inevitably containing irrelevant ones and noise, is impractical. To solve this problem, interestingness measures can be used for filtering or ranking association rules.

There are many different rule interestingness measures widely used in machine learning, data mining and statistics. In a study of 38 different measures, Geng and Hamilton [12] classify the interestingness measures in 3 main categories: *objective*, *subjective* and *semantics-based* measures. Objective measures are those that are not application-specific or user-specific and depend only on raw data. Subjective measures are those that consider users' background knowledge as well as data. As a special type of subjective measures, semantic-based measures take into account the explanation and the semantic of a pattern which are, like subjective measures, domain specific. For simplicity, our work only focuses on objective measures.

2.1 Objective Interestingness Measures

There is a large number of objective interestingness measures available in the literature. The 53 probability-based objective rule interestingness measures that we could find in all related literature are shown in Table 1.

To be able to analyze the objective measures, some properties are proposed for these measures in the literature. Four sets of properties are considered for objective interestingness measures. Piatetsky-Shapiro [29] has proposed the three main properties which are desirable for any objective interestingness measures. There are also other properties introduced by Major and Mangano [25], Tan et al. [32], Lence et al. [21, 22], and Geng and Hamilton [12], in total 16.

All the properties were introduced in the context of association rules. They can be used for finding similar measures or to find the appropriate measure for a problem domain if the required measure properties for that domain are known.

Using these properties, we clustered all the 53 measures in Table 1 with an agglomerative hierarchical clustering algorithm using average linkage. Having each measure as a vector of properties, the distance of two measures is based on a Hamming distance. Figure 1 shows different levels of this clustering till the maximum distance among measures in each cluster is 0.25.

The clustering can be compared with the work done by Tan et al. who clustered 18 measures with 8 properties[33], or Lenca et al. who clustered 20 measures using 6 properties [22]. Extensive comparison is reported in [16]. Here we only convey that groupings of interestingness measures are very similar with an Adjusted Rand Index [35] of 0.67 and an F-1 measure of 0.88. Discrepancies due to the fact that we are using a larger space with more measures and more properties, are outlined in [16].

3. RELATED WORK

Interestingness measures are used in different aspects of data mining. McGarry et al. [27] have used these measures to evaluate the worth of rules extracted from neural networks to discover their internal operation. Buntine [8] took advantage of these measures in probabilistic graphical model. Romao et al. [30] have used interestingness mea-

No.	Measure	Abbr	Ref
1	1-way support	1waySup	[12]
2	2-way support	2waySup	[12]
3	2-way support variation	2waySupVar	[12]
4	Accuracy	Acc	[12]
5	Added value	AddVal	[12]
6	Certainty factor	CerFac	[12]
7	Chi-square	Chi2	[31]
8	Class correlation ratio	CCR	[34]
9	Collective Strength	CollStr	[12]
10	Complement class support	CCS	[5]
11	Confidence	Conf	[12]
12	Confidence causal	ConfC	[17]
13	Confirm causal	CnfrmC	[17]
14	Confirm descriptive	CnfrmD	[17]
15	Confirmed-confidence causal	CCC	[17]
16	Confirmed-confidence descriptive	CCD	[17]
17	Conviction	Conv	[12]
18	Correlation coefficient	Corr	[12]
19	Cosine/IS	Cos	[12]
20	Dilated chi-square	D-Chi2	[20]
21	Example and counterexample rate	Ex&Cex	[12]
22	F-measure	FM	[28]
23	Ganascia	Gan	[19]
24	Gini index	Gini	[12]
25	Goodman-Kruskal	GK	[12]
26	Hyper confidence	HConf	[13]
27	Hyper lift	HLift	[13]
28	Implication index	ImpInd	[22]
29	Information gain	InfoGain	[12]
30	Intensity of implication	IntImp	[20]
31	Interestingness Weighting Dependency	IWD	[12]
32	Jaccard	Jacc	[12]
33	J-measure	JM	[12]
34	Kappa	Kappa	[32]
35	Klosgen	Klos	[12]
36	K-measure	KM	[28]
37	Laplace correlation	Lap	[12]
38	Least contradiction	LC	[12]
39	Leverage	Lev	[12]
40	Lift/interest	Lift	[12]
41	Loevinger	Loe	[12]
42	Mutual information	MutInfo	[12]
43	Odd multiplier	OddMul	[12]
44	Odds ratio	OddR	[12]
45	Piatetsky-Shapiro	PS	[12]
46	Recall/local support	LocSup	[12]
47	Relative risk	RelRisk	[12]
48	Sebag-Schoenauer	SS	[12]
49	Specificity	Spec	[12]
50	Support/global support	GlbSup	[12]
51	Yule's Q	YulQ	[12]
52	Yule's Y	YulY	[12]
53	Zhang	Zhang	[12]

Table 1: A list of objective rule interestingness measures, their abbreviations and references.

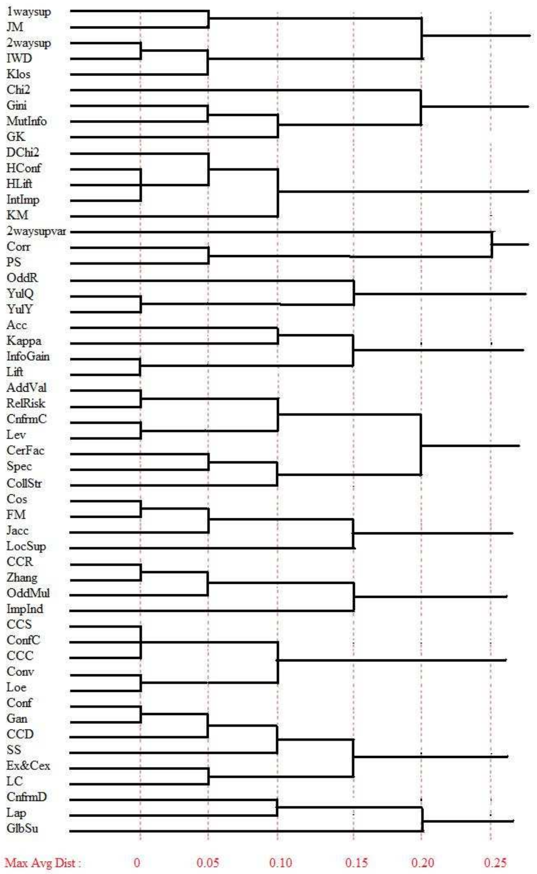

Figure 1: An agglomerative hierarchical clustering of measures based on their properties

sures in a genetic algorithm that optimizes expert beliefs to rank the interestingness of fuzzy prediction rules. Hilderman et al. [15] compared the various diversity measures used for ranking data summaries. Kononenko [18] discovered the properties of measures used in decision trees. And Gavrilov et al. [11] and Zhao et al. [37] compared objective functions used in clustering approaches.

The focus of this work is on interestingness measures (IM) for association and classification rules, hence, only the works related to these two areas are introduced in this section.

3.1 IM in association rule mining

Tan et al. [32, 33] have introduced five key properties that should be considered for selecting the right interestingness measure for a specific application. To study these properties, 21 different objective rule interestingness measures have been used. Using these five properties as well as three properties introduced by Piatetsky-Shapiro [29], Tan et al. have grouped some of these measures based on the correlation between their property vectors.

The work of Tan et al. also addresses finding the best measure for specific application domains using experts. Having a set of patterns, different measures can be used to rank these then the rankings compared using Pearson correlation to a manual ranking done by experts.

Lenca et al. [21] use a different way to find the best measure for a domain. They rank the measures based on their properties rather than using a set of patterns. For each application domain, a specialist assigns weights to each property of measures. Each weight shows the importance of that property in the given domain. Then using all properties and also the weights assigned to each of them, measures are ranked by applying a multi-criteria decision process. The measure with the highest rank can be selected to be used in that specific domain.

In another work, Lenca et al. [22] compare interestingness measures based on formal definitions and experimental results. 5 interestingness measure properties were described. Three of these properties along with three other properties were used to group 20 different interestingness measures. 5 different groups have been obtained with a hierarchal ascendant clustering using the average linkage and Manhattan distance. Another clustering was done based on experimental results from 10 different datasets. They also found 5 clusters that match the previous clustering.

There are also two excellent surveys [12, 26] with useful information about interestingness measures.

3.2 IM for associative classifiers

In an effort to present better alternatives to the confidence measure in classification based on associations, Lan et al. proposed two novel interestingness measures, intensity of implication and dilated chi-square [20]. These measures, which are used to sort generated rules, statistically reveal the interdependence between the antecedent of a rule and its class. They showed that on 16 UCI datasets the resulting rule set is 90% more compact that the one generated by a confidence-based classifier and the accuracy improved between 1% and 4%.

After showing that even confident rules can have negative correlations, Arunasalam and Chawla propose a new measure called Complement Class Support (CCS) which guarantees rules to be positively correlated [5]. Their experiments on 8 UCI datsets show that their measure is better suited than simple confidence for imbalanced datasets.

Verhein and Chawla utilize the Fisher Exact Test's (FET) ρ-value to extract only statistically significant rules and introduce a new measure, Class Correlation Ratio(CCR), to select only the rules that are more positively correlated to the class they predict rather than the other classes[34]. For classification, they use a strength score to rank the rules, a combination of ρ-value, confidence and CCR. They show on 6 UCI datasets that they can outperform other classifiers including a confidence-based associative classifier.

Azevedo and Jorge compared 10 different interestingness measures in the selection phase of an associative classifier[6]. Using 17 different datasets from the UCI repository they showed that the strategy selecting the applicable rule with the highest conviction measure yielded the best result. Overall, conviction, confidence and Laplace were the only measures that could produce competitive classifiers in their experiments.

Azevedo and Jorge also attempted to find which measure, conviction, confidence or Laplace, is better suited in the selection phase given some features about the dataset at hand [6]. While the study was not conclusive, it was noted that conviction is the best measure for unbalanced datasets.

While there is a wide range of interestingness measures, most are only studies in the context of association rule mining. Few are proposed exclusively for associative classifiers but all studies on interestingness measures for associative classifiers consider only the rule selection phase. We investigate all these measures on all 3 phases of an associative classifier.

4. INTERESTINGNESS ON ALL PHASES OF AN ASSOCIATIVE CLASSIFIER

As mentioned above, an associative classifier is composed of three major phases: rule generating, rule pruning and rule selection. It is possible that the proper choice of an interesting measure could have an impact on each of these phases. At a higher level though, one can generalize the associative classifier into two stages: the learning, which encompasses the rule generation and pruning; and the classification, which is the selection of appropriate rules to fire.

In the first step of an associative classifier, using an association rule mining technique, rules with class labels as consequent are generated. For generating rules, an anti-monotonic measure, am, and a threshold, t_a, is required to efficiently prune the search space. Only measures with an anti-monotonic property can be used in this phase. This property is not obviously known for all 53 measures listed in Table 1. In this study, only global support (i.e. support in the whole dataset) and local support (i.e. support in a class) are used for this phase. However, the interestingness measures are used as filters after generating all possible rules using a low support. In other words, using a support generates a learned model (i.e. a set of classification rules). Applying the different interestingness measure filters generates other learned models. Conflicting and redundant rules are later eliminated in the rule pruning phase. Redundancy removal pruning depends upon the strategy used in the rule selection phase and we refer the reader to [16] for more details. In brief, one strategy in rule selection during classification to order rules that apply based on some ordering and select the highest ranked rule. Pruning rules would not have an effect on the accuracy as rules lower in the rank are weeded out. However, when using a strategy where the interestingness measure of all rules predicting the same class is averaged to select the class label based on highest obtain average, a pruning would impact the calculated average and thus may lead to a different prediction.

The last phase of an associative classifier is to select a rule or a set of rules for predicting a class label of an object. For classifying a new object, an ordering should be defined. In this work, we use an ordering that is based on the used interestingness measure. Let sm be an interestingness measure. Based on this ordering and considering the highest ranked rule, $r_j <_{or} r_i$ (r_i gets a higher rank than r_j), if:

- $sm(r_j) < sm(r_i)$
- or $sm(r_j) = sm(r_i) \wedge support(r_j) < support(r_i)$
- or $sm(r_j) = sm(r_i) \wedge support(r_j) = support(r_i) \wedge length(r_i) < length(r_j)$

Where sm is the selecting measure, $support$ is the support of the rule, and $length$ denotes the length of the rule which is equal to the number of attribute-value pairs in the antecedant.

For taking into account the average of measures for all matchable rules, first all rules that apply to the unknown object should be grouped based on their class labels. If $\mathcal{R}_c$

			Highest ranked		Average	
Datasets	# of rules	Max acc %	FM%	Acc%	FM%	Acc%
Anneal	309,828	98.11	64.32	88.98	66.64	89.76
Breast	6,936	100.00	94.55	95.12	96.08	96.42
Census	63,226	98.35	65.11	81.69	73.50	83.89
Colic	188,278	98.63	62.48	72.57	80.36	82.57
Credit	299,311	99.42	74.55	76.68	87.64	87.97
Diabetes	923	97.79	66.87	73.83	68.54	74.09
German	223,508	99.00	48.80	71.50	43.62	70.10
Glass	1,599	88.51	55.35	69.31	54.85	66.97
Heart	41,096	100.00	66.05	70.27	80.37	80.85
Hepatitis	1,150,690	100.00	44.26	79.42	67.17	83.65
Iris	108	99.33	95.19	95.33	91.06	91.33
Labor	44,203	100.00	50.10	68.67	82.02	82.33
Led7	473	86.98	73.26	74.00	70.97	72.00
Pima	988	97.53	66.96	74.35	68.64	74.48
Tictactoe	7,398	100.00	70.30	78.72	88.06	90.19
Vote	955,659	99.77	85.17	87.12	95.69	95.87
Vowel	18,501	87.88	61.00	62.73	56.24	58.38
Waveform	35,626	100.00	79.98	80.32	75.57	76.54
Wine	185,942	100.00	76.76	79.87	95.60	95.48
Zoo	971,581	100.00	66.00	81.16	91.26	94.99
Average	225,294	97.56	68.35	78.08	76.69	82.39

Table 2: Results on 20 datasets using global support with threshold of 1%, with selecting based on the highest ranked rule and the rules' average of measures. "Max acc" denotes the maximum possible accuracy, "FM" denotes the macro average f-measure and "Acc" denotes the accuracy of the classifier.

denotes a rule set which all rules have the class label c, based on this ordering, $\mathcal{R}_c <_{orAvg} \mathcal{R}'_{c'}$, if:

- $Avg_{r_j \in \mathcal{R}_c}\{sm(r_j)\} < Avg_{r_i \in \mathcal{R}'_{c'}}\{sm(r_i)\}$
- or $Avg_{r_j \in \mathcal{R}_c}\{sm(r_j)\} = Avg_{r_i \in \mathcal{R}'_{c'}}\{sm(r_i)\} \wedge Avg_{r_j \in \mathcal{R}_c}\{support(r_j)\} < Avg_{r_i \in \mathcal{R}'_{c'}}\{support(r_i)\}$
- or $Avg_{r_j \in \mathcal{R}_c}\{sm(r_j)\} = Avg_{r_i \in \mathcal{R}'_{c'}}\{sm(r_i)\} \wedge Avg_{r_j \in \mathcal{R}_c}\{support(r_j)\} = Avg_{r_i \in \mathcal{R}'_{c'}}\{support(r_i)\} \wedge Avg_{r_i \in \mathcal{R}'_{c'}}\{length(r_i)\} < Avg_{r_j \in \mathcal{R}_c}\{length(r_j)\}$

Both these approaches are used in this study. If no rule can match the object, the dominant class is assigned to it.

5. EXPERIMENTAL RESULTS

The impact of using different interestingness measures is explored individually for each phase. Then, the combination of the best measures in each phase is studied.

20 datasets having different characteristics, were chosen from the UCI repository. To convert the relational datasets into transactional datasets, all numeric attributes are discretized using the entropy-based discretization method [10] used in CBA [24] to categorize the continues attributes.

For evaluation, each classifier is assessed based on the number of rules its model contains, the macro average f_1-measure, accuracy and maximum possible accuracy. Henceforth, f-measure refers to macro average f_1-measure.

The maximum possible accuracy shows the maximum accuracy that is achievable if for each test object, the right rule is selected from the set of available rules. Hence, if for a test object there exists at least one rule that applies to that object with the same class label, that object is considered as a correct classification, otherwise, it is a misclassification. This evaluation measure is useful to evaluate the pruning and see whether the essential rules are pruned or preserved.

All the results are based on 10-fold cross validation and the folds used for all classifiers are the same for each dataset.

5.1 Global vs. Local Support

Local and global supports with a threshold of 1% are used as anti-monotonic measures.To remove the conflicting rules, a minimum confidence threshold of 51% is used. No pruning method is used here and the measure used in the selection phase is confidence with two different approaches, selecting based on the "highest ranked rule" and based on the "average of rules". Rule sets generated only using local/global support and confidence are called "original rule sets". All other results are compared with the results of these rule sets.

Datasets	RR %	FC%	AC%
Anneal	99.86	-38.98	-9.78
Breast	95.23	-1.18	-1.04
Census	92.49	-22.09	-5.43
Colic	96.69	-8.23	-4.24
Credit	97.62	-1.68	-1.64
Diabetes	83.42	-2.51	-0.88
German	95.36	-5.61	-0.14
Glass	82.47	-12.11	-10.86
Heart	96.72	-0.03	+0.33
Hepatitis	99.74	-34.10	-5.06
Iris	69.65	+0.90	+0.73
Labor	99.23	-8.19	-6.88
Led7	31.85	-0.30	-0.78
Pima	83.47	-2.20	-0.88
Tictactoe	69.79	-49.92	-25.81
Vote	99.53	-6.21	-5.29
Vowel	88.52	-6.23	-6.75
Waveform	71.58	-3.57	-2.85
Wine	99.14	-13.73	-11.69
Zoo	99.77	-32.45	-21.31

Table 3: Percentage of rule reduction while using redundancy removal pruning on rule sets generated with global support as well as the change of f-measure and accuracy while the rules' average of measures are used for prediction. RR, FC and AC are short forms for rule reduction, f-measure change and accuracy change respectively.

The results of the original rule sets show that using local support yields a very large number of generated rules, especially when the class labels are imbalanced (i.e., when the standard deviation of class distributions is high), but it also creates more accurate models for this kind of datasets as it also finds frequent patterns in small classes. The results of the original rule sets using global support is shown in Table 2 in terms of number of rules, maximum possible accuracy, f-measure and accuracy. The results for local support can be found in [16].

5.2 Using Redundancy Removal Pruning

While using the highest ranked rule in selection phase, this pruning can be used to remove the rules that are never used in prediction. Hence, the f-measure and accuracy does not change. Table 3 shows a huge percentage of rule reduction while using the redundancy removal pruning on original rule sets generated with global support. On the other hand, using this pruning method while predicting based on the rules' average of measures, changes the number of rules as well as f-measure and accuracy. The percentage of change in f-measure and accuracy are also shown in Table 3. The results show large reduction of f-measure in some datasets. Hence, although redundancy removal pruning can reduce a

Datasets	RR %	FC%	AC%	MPAC%	Measure
Anneal	-16.51	21.64	7.27	-1.81	KM (0.1)
Breast	-9.67	1.84	1.52	-1.43	Lev (0.8)
Census	-55.70	18.64	3.03	-17.99	Zhang (0.8)
Colic	-99.99	34.81	18.32	-72.98	2WaySup (0.3)
Credit	-69.82	17.63	14.54	-5.38	Lap (0.9)
Diabetes	-40.43	8.44	1.40	-19.04	AddVal (0.2)
German	-93.62	35.34	-2.38	0.00	Acc (0.5)
Glass	-44.74	11.37	2.71	0.00	FM (0.2)
Heart	-99.87	28.01	20.68	-0.67	CollStr (9)
Hepatitis	-99.86	62.66	1.68	-0.59	CF (0.05)
Iris	-58.67	0.80	0.70	-0.67	Klos (0.2)
Labor	-1.11	85.92	35.92	-1.67	Lev (0.9)
Led7	-16.86	0.70	0.76	-0.07	HConf (0.9)
Pima	-40.13	9.00	1.05	-18.97	AddVal (0.2)
Tictactoe	-98.91	40.91	25.97	-0.84	Corr (0.2)
Vote	-99.91	12.35	10.02	-0.92	CF (0.4)
Vowel	-8.90	0.15	0.16	-2.76	CCD (0.1)
Waveform	-33.55	0.15	0.15	-0.02	Zhang (0.7)
Wine	-10.57	19.42	14.71	-0.53	Lev (0.95)
Zoo	-94.30	19.44	10.50	0.00	LC (0.7)

Table 4: Impact of measure-based filtering when maximizing the f-measure and the winning measure (with its minimum threshold used). Global support is used for rule generation and the selection phase is based on the highest of rules. RR, FC, AC and MPAC are short forms for rule reduction, f-measure change, accuracy change and maximum possible accuracy change respectively.

large number of rules, it is not safe to be used while prediction is based on the rules' average of measures.

5.3 Measure-based Filtering

Three different experiments are conducted to find the impact of using 53 different measures in measure-based pruning. Two of these experiments are based on rule reduction. In the first experiment, the goal is to find the minimum number of rules without jeopardizing the f-measure (keeping it above 95% of its original). In the second experiment, the aim is to find the minimum number of rules without changing the maximum possible accuracy. The goal of the last experiment is to eliminate the misleading rules in order to improve the f-measure. For these experiments the selection measure is fixed on confidence.

In these experiments the number of rules shrunk by up to 99.99% while the f-measure in some cases even improved. For each experiment, we ranked the measures for each dataset based on rule reduction achieved (or f-measure improvement in the third experiment). To find the measures that can have the highest impact in rule reduction, the number of times a measure ranked between 1 and 3 are counted. Based on these counts, IWD, Kappa, GK, Corr, Klos, 2WaySup, CF, Gini, and Spec are measures that have the highest impact on rule reduction without jeopardizing the f-measure, while FM, Cos, Jacc, CollStr, Acc, and Spec have the highest impact on rule reduction whilst not changing the maximum accuracy possible. Lev, Kappa, Zhnag, Acc, GK, 1WaySup, CF, Cos, FM, LC, and Spec were the high achievers in the experiment maximizing the f-measure. Interestingly, many measures never achieved a high enough ranking on any dataset. Table 4 shows the maximum percentage of f-measure improvement and the measure used for this achievement. The results show that there are some significant improvements in f-measure and the rule reduction is still substantial. The complete results are detailed in [16].

	Highest ranked			Average		
Datasets	FC%	AC%	Measure	FC%	AC%	Measure
Anneal	20.25	6.02	Klos,	11.45	2.85	ConfC
Breast	0.72	0.47	Lev,	0.68	0.59	ConfC
Census	19.99	3.74	CCS,	5.21	-0.08	ConfC
Colic	30.74	14.21	IntImp,	0.89	0.35	ConfC, Lev
Credit	15.77	12.84	Lap,	0.00	0.00	CCD, Conf, Gan
Diabetes	8.64	0.86	DChi2,	6.45	0.86	CCS
German	35.97	-0.56	Klos,	32.87	3.71	Conv, Loe
Glass	6.84	-1.45	Lev,	8.25	4.01	ConfC
Heart	25.21	17.91	IntImp,	1.57	1.23	ImpInd
Hepatitis	63.74	-0.89	DChi2,	6.95	-1.45	ConfC
Iris	0.80	0.70	1WaySup, Loe, CCC, CCD, Conv, ConfC, InfoGain, Lift, OddMul, Gan,	4.54	4.38	Conv, Lev, Loe, OddMul, SS
Labor	77.68	30.58	IntImp,	1.48	5.26	Lap
Led7	0.31	0.33	CnfrmC,	1.73	1.60	SS
Pima	8.37	-0.36	OddMul,	7.09	1.22	Lev
Tictactoe	40.91	25.97	IntImp,	8.93	6.83	ConfC
Vote	12.65	10.30	CnfrmC,	0.00	0.00	ConfC, CCC, CCD, Lev, Conf, Gan
Vowel	0.90	0.64	CCS,	6.98	5.36	CCS
Waveform	0.02	0.02	Lev,	4.36	3.55	CCS
Wine	14.35	10.50	Lap,	0.04	0.07	Ex&Cex
Zoo	19.49	11.53	CnfrmC,	0.14	0.00	Ex&Cex

Table 5: Percentage of f-measure change, accuracy change and the measure used in selection phase to get the maximum f-measure. Global support is used for rule generation and the selection phase is based on both the highest ranked rule and rules' average of measures. FC and AC are the short forms for f-measure change and accuracy change respectively.

5.4 Measures in Rule Selection Phase

The effect of using different selection measures, in the third phase of the associative classifier, is only on the improvement of the f-measure. There is no change in the number of rules per se as the learning model is already built. Table 5 shows the best measures for f-measure improvement for each dataset. From the results, it can be inferred that there are some significant improvements in f-measure, specially when predicting is based on the highest ranked rule.

The measures are ranked based on the f-measure improvements in each dataset. OddMul, CCS, CnfrmC, Conv, Lap, Loe, and Zhang are the measures with the most top ranks when the highest ranked rule is used for selection and ConfC, CCC, Lev, Conv, CCS, Loe and Ex&Cex are the measures with the most top ranks when the rules' average of measures strategy is used. Many measures never achieve a top rank with any dataset.

5.5 Using IM in Both Pruning and Selection

The impact of using different interestingness measures on each individual phase of the associative classifier was highlighted above. Here the goal is to study the impact of using different interestingness measures both in the pruning and the selection phases together. For this reason, the best measures found in measure-based pruning are combined with the best measures found in selection phase for each dataset. For cases where the strategy using highest ranked rule is adopted for prediction, redundancy removal is also used after the measure-based pruning. The results for f-measure changes are shown in Table 6. In this table, the percentage of f-measure changes using measure-based pruning and using different measures in the selection phase are compared with that of the combination of these two phases. The results show that not only combining the best interestingness measure of each phase does not improve the f-measure, but, there are some cases with significant decrease in the f-measure.

Datasets	Pruning measure	Selecting measure	FC % prune	FC % select	FC % combine
Anneal	KM	Klos	21.64	20.25	17.62
Breast	Lev	Lev	1.84	0.72	0.72
Census	Zhang	ccs	18.64	19.99	19.77
Colic	2waySup	IntImp	34.81	30.74	24.82
Credit	Lap	Lap	17.63	15.77	15.58
Diabetes	AddVal	DChi2	8.44	8.64	9.23
German	Acc	Klos	35.34	35.97	34.78
Glass	FM	Lev	11.37	6.84	4.36
Heart	CollStr	IntImp	28.01	25.21	22.64
Hepatitis	CF	DChi2	62.65	63.74	39.37
Iris	Klos	1WaySup	0.80	0.80	0.80
Labor	Lev	IntImp	85.92	77.68	91.54
Led7	HConf	CnfrmC	0.70	0.31	0.31
Pima	AddVal	OddMul	9.00	8.37	10.16
Tictactoe	Corr	IntImp	40.91	40.91	40.91
Vote	CF	CnfrmC	12.35	12.65	12.08
Vowel	CCD	ccs	0.15	0.90	1.09
Waveform	Zhang	Lev	0.15	0.02	0.09
Wine	Lev	Lap	19.41	14.35	18.69
Zoo	LC	CnfrmC	19.44	19.49	19.49

Table 6: Comparing the changes of f-measure with the best measure used in measure-based pruning for f-measure improvement, the best measure used in selection phase, and the combination of these two measures. Global support is used for rule generation and the selection phase is based on the highest ranked rules. FC is the short form for f-measure change.

Hence, a suitable selecting measure based on an original rule set is not necessarily a suitable selecting measure for a pruned version of that rule set. The tables for rule reduction are not shown for lack of space. What is noteworthy is that the redundancy removal could even prune more rules from rule sets already pruned by measure-based pruning.

To summarize the results, there are interestingness measures that can be used as filtering measures and be able to reduce the number of rules significantly in all datasets without jeopardizing the accuracy of the model. In other words, the filters are capable of identifying unnecessary rules from the model. However, this drastic improvement in the number of rules is not necessarily observed in terms of accuracy. The change in accuracy remains stable, but some positive improvements in the accuracy (f-measure) were noted. Another observation is that, no single measure can be declared as a winner for all types of datasets. There are some measures that have more impact than others.

6. CONCLUSION AND FUTURE WORK

Associative classification is a relatively new paradigm for classification relying on association rule mining and naturally inherits the most commonly used interestingness measures, support and confidence. These are not necessarily the best choice and no systematic study was undertaken to identify the most appropriate measures from the myriad measures already used as filters or rankers for relevant rules in different fields.

This study is to answer the question whether other measures are more suited for the different phases of the associative classifier, and an attempt to identify the best measure for each phase. The results clearly indicate that many interestingness measures can indeed provide a better set of classification rules (i.e. a drastic reduction in the number of rules) and a more accurate classifier. However, there was no single measure that was consistently impacting the rule set for all datasets tested, even though for each dataset, some interestingness measure was successful in reducing the rule set or improving the effectiveness of the classifier. These measures are introduced for each individual phase. The results show that the measures that are the best in one phase are not necessarily the best measures for the other phase.

Another observation is that using the combination of the best measures in pruning and selection phases does not improve the accuracy of the classifier which means that the best selecting measure for an original rule set is not the best for the pruned version of that rule set. This observation shows that there might exist some rule set characteristics that have effect on selecting the best measure. Hence, for each pruned rule set, the appropriate selecting measure should be probed.

All the measures were clustered in different experiments. Some of the measures behave similarly in all the cases. Hence, in future work, selecting only one measure from each group as a representative, is sufficient.

An interesting future study would be to identify the relevant features of a dataset or a rule set that would help indicate the appropriate interestingness measure to use, and in this way exploit these features to build a predictor for best measure to use in the associative classifier given a specific training set.

7. REFERENCES

[1] J.M. Adamo. *Data mining for association rules and sequential patterns: sequential and parallel algorithms*. Springer-Verlag, 2001.

[2] C.C. Aggarwal and PS. Yu. A new framework for itemset generation. In *PODS: Proceedings of the 17th symposium on Principles of Database Systems*, pages 18–24. ACM, 1998.

[3] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *The International Conference on Very Large Databases*, pages 487–499, 1994.

[4] M.L. Antonie and O.R. Zaïane. Text document categorization by term association. In *Proc. of the IEEE 2002 International Conference on Data Mining*, pages 19–26, Maebashi City, Japan, 2002.

[5] B. Arunasalam and S. Chawla. Cccs: A top-down associative classifier for imbalanced class distribution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 517–522. ACM, 2006.

[6] P. J. Azevedo and A. M. Jorge. Comparing rule measures for predictive association rules. In *ECML '07: Proceedings of the 18th European conference on Machine Learning*, pages 510–517, Berlin, Heidelberg, 2007. Springer-Verlag.

[7] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 265–276. ACM, 1997.

[8] W. Buntine. Graphical models for discovering knowledge. *Advances in knowledge discovery and data mining*, pages 59–82, 1996.

[9] S. Chiusano and P. Garza. Selection of high quality rules in associative classification. In C. Zhang Y. Zhao and L. Cao, editors, *Post-Mining of Association RUles: Techniques for Effective Knowledge Extraction*. Information Science Reference, Hershey, NY, USA, 2009.

[10] U.M. Fayyad and K.B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on AI*, pages 1022–1027, 1993.

[11] M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani. Mining the stock market: Which measure is best? In *proceedings of the 6 th ACM Int'l Conference on Knowledge Discovery and Data Mining*, pages 487–496, 2000.

[12] L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9, 2006.

[13] M. Hahsler and K. Hornik. New probabilistic interest measures for association rules. *Intell. Data Anal.*, 11(5):437–455, 2007.

[14] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 1–12. ACM, 2000.

[15] R.J. Hilderman, H.J. Hamilton, and B. Barber. Ranking the interestingness of summaries from data mining systems. In *In Proceedings of the 12th Annual Florida Artificial Intelligence Research Symposium (FLAIRS'99*, pages 100–106, 1999.

[16] Mojdeh Jalali-Heravi. A study on interestingness measures for associative classifiers. Master's thesis, University of Alberta, 2009.

[17] Y. Kodratoff. Comparing machine learning and knowledge discovery in databases: An application to knowledge discovery in texts. In *In: ECCAI summer*, pages 1–21. Springer, 2000.

[18] I. Kononenko. On biases in estimating multi-valued attributes. In *in Proc. 14th Int. Joint Conf Artificial Intelligence*, pages 1034–1040. Morgan Kaufmann, 1995.

[19] S. Lallich, O. Teytaud, and E. Prudhomme. Association rule interestingness: Measure and statistical validation. In *Quality Measures in Data Mining*, pages 251–275. Springer, 2007.

[20] Y. Lan, D. Janssens, G. Chen, and G. Wets. Improving associative classification by incorporating novel interestingness measures. In *ICEBE '05: Proceedings of the IEEE International Conference on e-Business Engineering*, pages 282–288, Washington, DC, USA, 2005. IEEE Computer Society.

[21] P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. A multicriteria decision aid for interestingness measure selection. Technical Report LUSSI-TR-2004-01-EN, LUSSI Department, GET/ENST, France, 2004.

[22] P. Lenca, B. Vaillant, P. Meyer, and S. Lallich. Association rule interestingness measures: Experimental and theoretical studies. In *Quality Measures in Data Mining*, pages 51–76. Springer, 2007.

[23] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *IEEE International Conference on Data Mining (ICDM'01)*, San Jose, California, November 29-December 2 2001.

[24] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *KDD*, pages 80–86, 1998.

[25] J. A. Major and J. J. Mangano. Selecting among rules induced from a hurricane database. *Journal of Intelligent Information systems*, 4:39–52, 1995.

[26] K. McGarry. A survey of interestingness measures for knowledge discovery. *Knowl. Eng. Rev.*, 20(1):39–61, 2005.

[27] K. McGarry and J. Malone. Analysis of rules discovered by the data mining process. In *Applications and Science in Soft Computing Series: Advances in Soft Computing.*, pages 219–224. Springer, 2004.

[28] M. Ohsaki, S. Kitaguchi, H. Yokoi, and T. Yamaguchi. Investigation of rule interestingness in medical data mining. In *Active Mining*, pages 174–189, 2003.

[29] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In G. Piatetsky-Shapiro and W.J. Frawley, editors, *Knowledge Discovery in Databases*. AAAI/MIT Press, Cambridge, MA, 1991.

[30] W. Romão, A. Freitas, and I. Gimenes. Discovering interesting knowledge from a science and technology database with a genetic algorithm. *Appl. Soft Comput.*, 4(2):121–137, 2004.

[31] P. Tan and V. Kumar. Interestingness measures for association patterns: A perspective. Technical Report 00-036, Department of Computer Sciences, University of Minnesota, 2000.

[32] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 32–41. ACM, 2002.

[33] P .Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Inf. Syst.*, 29(4):293–313, 2004.

[34] F. Verhein and S. Chawla. Using significant positively associated and relatively class correlated rules for associative classification of imbalanced datasets. In *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM '07)*, pages 679Ű–684, Los Alamitos, 2007. IEEE Computer Society Press.

[35] K.Y. Yeung and W.L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[36] M. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In *Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining*, pages 283–Ű296.

[37] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. Technical report, Department of Computer Science, University of Minnesota, 2002.