STAIRS 2014
U. Endriss and J. Leite (Eds.)

doi:10.3233/978-1-61499-421-3-71

On Evaluating Interestingness Measures for Closed Itemsets

Aleksey BUZMAKOV [a,b,1], Sergei KUZNETSOV [a] and Amedeo NAPOLI [b]
[a] *Higher School of Economics, Moscow, Russia*
[b] *LORIA (CNRS – Inria NGE – Université de Lorraine), France*

Abstract. There are a lot of measures for selecting interesting itemsets. But which one is better? In this paper we introduce a methodology for evaluating interestingness measures. This methodology relies on supervised classification. It allows us to avoid experts and artificial datasets in the evaluation process. We apply our methodology to evaluate promising measures for itemset selection, such as leverage and stability. We show that although there is no evident winner between them, stability has a slightly better performance.

Keywords. data mining, pattern selection, interestingness measures, stability, leverage, comparison

1. Introduction

One of the most important and frequent tasks in artificial intelligence is selection of the best option(s) among a huge set of possibilities. For example, in data mining one should often determine which patterns are of high interest. Usually patterns are evaluated w.r.t. a formal relevancy measure. Webb [1] says that measures cannot often reflect the true value of patterns because they “often depend on many factors that are difficult to formalize”. *Are we able to evaluate how well a measure approximates an expert interest?*

One way to do that is to evaluate patterns with experts [2]. In that case we evaluate how close the selected patterns approximate the expert knowledge. It is an expensive strategy requiring many experts for a domain. Thus, if one wants to compare measures on datasets from different domains the experiments become very expensive. Additionally such an experimentation requires a lot of time to be carried out.

Another way to evaluate patterns is to use artificial datasets [3], where the target patterns are known. The drawback of this approach is the relevancy of the artificial datasets w.r.t. real ones. Thus, *one goal* of this paper is to develop and evaluate a methodology for comparison of interestingness measures for itemsets without involving experts or artificial datasets (below we use indifferently “pattern” or “itemste”).

Our methodology *is based on* semi-supervised classification, where every data entry has a class label but labels are not directly involved into computation of a measure. A label is an additional information to entry description modeling domain knowledge or

[1] Corresponding Author: Aleksey Buzmakov, LORIA (CNRS – Inria NGE – Université de Lorraine), 615, Jardin Botanique street, 54600, Vandoeuvre-les-Nancy, France; E-mail: aleksey.buzmakov@inria.fr.

Table 1. A toy dataset

	a	b	c	d	e	f	Label
g_1	x				x	x	+
g_2		x			x	x	-
g_3			x		x	x	+
g_4				x	x	x	-
g_5		x	x			x	+
g_6		x	x		x	x	?(+)
g_7	x		x			x	?(+)
g_8	x		x	x		x	?(-)

expert intent. The basic idea is to rank patterns with an interestingness measure and, then, find among them the patterns that are relevant to classification. And if a measure $\mathscr{M}_1$ is better than another measure $\mathscr{M}_2$ w.r.t. expert interest, i.e. $\mathscr{M}_1$ attributes more systematically high ranks to more relevant patterns, thus increasing the performance of a classifier based on $\mathscr{M}_1$.

This methodology can be applied when a measure does not rely on class labels. Then it can find itemsets that are suitable for expert interest (and not biased towards the classification task).

The second goal of this paper is to evaluate some measures for itemset ranking. We evaluate leverage [4] and stability [5] measures that seem to be well adapted to itemset ranking. We also introduce difference measure that comes from an estimate of stability [6]. This measure is computed faster than stability. As it is widely used, the support of an itemset is used as a baseline measure. Finally, we also consider another leverage measure (rule leverage) which in contrast to the afore mentioned measures, relies on class labels. This rule leverage measure provides an idea of rule ranking measures w.r.t. itemset ranking measures.

Finally, we show that although there is no evident winner among stability and leverage measures, stability seems to be better on average. It is also shown that difference and stability have a similar behaviour. But according to previous studies, difference is faster to compute [6]. We can summarize *the novelty* of this paper as follows:

1. Methodology for comparison of measures for itemset ranking.
2. Comparison of leverage and stability measures for itemsets.

The rest of the paper is organised as follows. Section 2 introduces basic notions. Then related work is discussed. In Section 4 we define and discuss the evaluated measures. The next section describes our methodology. And finally before concluding the paper the experiments are discussed.

2. Preliminaries

In this section we discuss the basic definitions in terms of Formal Concept Analysis [7].

Definition 1. *A dataset or a context is a triple (G,M,I), where G is a set of objects, M is a set of attributes, and $I \subseteq G \times M$ is a relation between G and M.*

Every subset of attributes is called *a pattern* or *an itemset*. *The description* of a set of objects X is the set of attributes shared by all objects from X, $\phi(X) = \{m \in M \mid (\forall g \in X) gIm\}$. *The image* of itemset Y is the set of objects sharing Y, $\psi(Y) = \{g \in G \mid (\forall m \in Y) gIm\}$. The cardinality of the image of Y is called *support* of Y, $\texttt{Supp}(Y) = |\psi(Y)|$, while the value $\sigma(Y) = \frac{\texttt{Supp}(Y)}{|G|}$ is called *frequency* of Y.

Consider the toy dataset in Table 1. Let $G = \{g_1, g_2, g_3, g_4, g_5\}$, $M = \{a, b, c, d, e, f\}$ and I is shown in the table, then $\phi(\{g_1, g_2\}) = \{e, f\}$ is the set of attributes shared by g_1 and g_2 or the description of the set $\{g_1, g_2\}$. Similarly $\psi(\{e, f\}) = \{g_1, g_2, g_3, g_4\}$ is the image of $\{e, f\}$. We say that the support of the itemset $\{e, f\}$ is $\texttt{Supp}(\{e, f\}) = 4$ and its frequency is $\sigma(\{e, f\}) = \frac{4}{|G|} = 0.8$.

Definition 2. *An itemset X is closed if and only if there is no superset $Y \supset X$ such that* $\texttt{Supp}(Y) = \texttt{Supp}(X)$.

It means that an itemset X is closed if it is not possible to add any attribute to X preserving the set of objects that supports X. The operator $\phi \circ \psi$ is a closure operator and thus an itemset X is closed if and only if $\phi(\psi(X)) = X$.

Definition 3. *An association rule between an itemset X and an itemset Y is denoted by $X \to Y$, where X is called the premise and Y is called the conclusion of the rule.*

Rule $X \to Y$ means that if the description of some objects from G contains X, then it contains Y. There are two measures attached to an association rule: support (or frequency) and confidence.

Definition 4. *The support of a rule $X \to Y$ is* $\texttt{Supp}(X \cup Y)$ *and frequency of the rule $X \to Y$ is $\sigma(X \cup Y)$.*

Definition 5. *The confidence of a rule $X \to Y$ is* $\texttt{Conf}(X \to Y) = \frac{\texttt{Supp}(X \cup Y)}{\texttt{Supp}(X)}$.

The support and the frequency of a rule show how often one can find the premise in the dataset, while a rule $X \to Y$ with a high confidence means that in most of the cases if an object description includes X it is likely to include Y. For example, in the dataset in Table 1 with the set of objects $G = \{g_1, g_2, g_3, g_4, g_5\}$ the confidence of the rule $\{e\} \to \{f\}$ is 1 because in every case when an object description contains e it contains also f, while $\texttt{Conf}(\{c\} \to \{e, f\}) = \frac{1}{2}$.

A common objective in data mining is search for interesting patterns, i.e. for interesting itemsets or rules, that are usually related to a task. Among those different tasks, there are classification, clustering and expert analysis of the result. Here we focus on searching for patterns that are likely to be interesting to an expert. In the next section we describe the existing approaches for mining interesting itemsets.

3. Related Works

Probably, the most elaborated area of mining interesting patterns is association rule mining. Most of the measures created in this area optimize a formal criterion using statistical methods [8]. Although there is a huge number of interestingness measures, there are only few comparisons between them [2,9]. The main reasons for that are the diversity of formal criteria and the fact that no measure wins in all criteria. In [2] the authors evaluate different measures by means of expert interest. This is an important approach for pattern evaluation as it directly measures the relation between a formal measure and an expert interest. The drawback of this approach is the cost and diversity of datasets: for every dataset it is necessary to hire several experts which is costly. In [9], the authors evaluate measures by their performance in the classification task in a supervised setting, e.g. how confident is a rule concluding on a given class. In such case, the aim of an expert is expressed by labeling of a training set. This is in contrast with our approach which follows a semi-supervised setting, i.e. measures do not depend on labelling.

Another group of interestingness measures consists of measures created for itemset ranking. It is a less studied group. Several measures can be found in [3]. Some of them are related to the distribution of partitions induced by every attribute from the considered pattern. Others are related to measures of association rule mining. Another approach introduces the measure of leverage that corresponds to the difference between frequency of an itemset and the maximal expected frequency based on subsets of the itemset [4]. Finally, some measures can be found in the domain of formal concept analysis [10,11, 12]. Stability measure is one of the most interesting among them, because it is often used in domain specific areas where experts are often involved. Moreover, in contrast to all above mentioned measures for itemsets, stability is computed on object side making it possible to apply it for ranking any types of patterns, e.g. sequential patterns [13].

One comparison of interestingness measures of itemsets can be found in [3] where the authors introduce a Quest Generator, i.e. a tool generating a dataset from a given set of “goal” itemsets in the presence of possible noise. Then, the interestingness measures can be evaluated w.r.t. their ability for finding the “goal” itemsets. For all artificial tests there is always a question about the degree to which generated datasets reflect real data. Thus, in this paper we provide an alternative approach for evaluating interestingness measures of itemsets on real datasets without involving experts. In the next sections we consider and compare these measures in details.

4. Itemset Interestingness Measures

For comparison the stability and leverage measures are selected as the most recent and promising measures. Support is also included into the comparison as a baseline measure, since it is widely used in itemset mining.

Definition 6 ([11]). *Given a context (G,M,I), the stability of an itemset $Y \in 2^M$ is given by the following formula:*

$$\mathtt{Stab}(Y) = \frac{|\{X \subseteq \psi(Y) \mid \phi(X) = Y\}|}{2^{\mathtt{Supp}(Y)}}. \tag{1}$$

The intuition behind stability is the following [11]. Stability gives an idea of how much a closed itemset depends on an object in its image, i.e. if we remove an object does the closed itemset exists anymore? In other words, given a dataset of objects and an itemset, stability measures the probability that the same itemset can be found in a dataset built as a subset of the objects. Consider for example itemset $\{e, f\}$ for the dataset in Table 1. The image of this set is $\{g_1, g_2, g_3, g_4\}$ when $G = \{g_1, g_2, g_3, g_4, g_5\}$. There are

16 possible subsets of this image. The descriptions of the $\emptyset$, $\{g_1\}$,$\{g_2\}$,$\{g_3\}$,$\{g_4\}$, and $\{g_5\}$ are different from $\{e,f\}$. Then, stability of $\{e,f\}$ is $\mathtt{Stab}(\{e,f\}) = \frac{16-5}{16} = 0.69$. Similarly the stability of $\{d,e,f\}$ is 0.5. According to the definition, stability of a non-closed itemset is always 0.

Although, stability is a measure that is hard to compute [10], it can be efficiently estimated [6]: $\mathtt{Stab}(Y) \leq 1 - 2^{-(\mathtt{Supp}(Y)-\mathtt{Supp}(X))}$, for all $X \supset Y$. Thus, stability allows us to introduce the related measure of "minimal positive difference in support" between itemset Y and itemsets including Y:

$$\mathtt{Diff}(Y) = \min_{X \supset Y,\ \mathtt{Supp}(X) \neq \mathtt{Supp}(Y)} (\mathtt{Supp}(Y) - \mathtt{Supp}(X)). \tag{2}$$

Difference can be computed efficiently and the experiments show the interestingness of this measure. For example, difference of itemset $\{e,f\}$ is 3, because support of $\{e,f\}$ is 4 and any superset of $\{e,f\}$ has support at most 1. Similarly difference of $\{d,e,f\}$ is 1. For non-closed itemsets, difference is always zero.

The next measure that we evaluate is leverage for itemsets. For defining leverage we recall that a 2-partition of a set Y is a partition of Y in two subsets V and W and is denoted by $\mathtt{Part}_2(Y) = (V|W)$. For example, the pair $(\{a,b,c\},\{e,f\})$ is a 2-partition of the set $\{a,b,c,e,f\}$. Now we can define what the leverage of an itemset is.

Definition 7 ([4])**.** *The leverage of an itemset $Y \in 2^M$ is the difference between $\sigma(Y)$ and the maximal frequency that would be expected under assumption of independence of any subset of Y:*

$$\mathtt{Lev}(Y) = \sigma(Y) - \operatorname*{\mathtt{argmax}}_{(V|W)=\mathtt{Part}_2(Y)} \sigma(V) \cdot \sigma(W), \tag{3}$$

where $\mathtt{Part}_2(Y)$ is a 2-partition of Y.

According to the definition, leverage of an itemset can be applied to any itemset. If an itemset is non-closed then the leverage value is not zero and the next proposition holds.

Proposition 1. *The leverage of an itemset is not larger than the leverage of its closure,* $\mathtt{Lev}(Y) \leq \mathtt{Lev}(\phi(\psi(Y)))$.

Proof. Frequency can only decrease with addition of an attribute, i.e. $(\forall X \subseteq Y)\sigma(X) \geq \sigma(Y)$. Frequencies of an itemset and its closure are equal, $\sigma(Y) = \sigma(\phi(\psi(Y)))$. Given itemset X, a 2-partition of its closure $\mathtt{Part}_2(\phi(\psi(X))) = (V|W)$ induces the 2-partition of X, i.e. $(V \cap X|W \cap X)$ is a 2-partition of X. Then,

$$\begin{aligned}
\mathtt{Lev}(\phi(\psi(Y))) &= \sigma(\phi(\psi(Y))) - \operatorname*{\mathtt{argmax}}_{(V|W)=\mathtt{Part}_2(\phi(\psi(Y)))} \sigma(V) \cdot \sigma(W) = \\
&= \sigma(Y) - \operatorname*{\mathtt{argmax}}_{\substack{(V|W) = \mathtt{Part}_2(Y) \\ (P|Q) \qquad\qquad = \\ \mathtt{Part}_2(\phi(\psi(Y)))\setminus Y}} \sigma(V \cup P) \cdot \sigma(W \cup Q) \geq \\
&\geq \sigma(Y) - \operatorname*{\mathtt{argmax}}_{(V|W)=\mathtt{Part}_2(Y)} \sigma(V) \cdot \sigma(W) = \mathtt{Lev}(Y).
\end{aligned}$$

Thus, leverage maximizes its value on closed itemsets, and, consequently, we can compute it only for closed itemsets. □

Let us consider an example. In order to compute leverage of itemset $\{e,f\}$ we need to find all its 2-partitions. There is only one 2-partition $(\{e\}|\{f\})$. The frequencies are $\sigma(\{e,f\}) = 0.8$, $\sigma(\{e\}) = 0.8$, $\sigma(\{f\}) = 0.8$. Thus, $\texttt{Lev}(\{e,f\}) = 0.8 - 0.8^2 = 0.16$. For itemset $\{d,e,f\}$ we have three 2-partitions: $(\{e\}|\{d,f\})$, $(\{d\}|\{e,f\})$ and $(\{f\}|\{e,d\})$. The frequencies are $\sigma(\{d,e,f\}) = 0.2$, $\sigma(\{e\}) \cdot \sigma(\{d,f\}) = 0.8 \cdot 0.2 = 0.16$, $\sigma(\{d\}) \cdot \sigma(\{e,f\}) = 0.2 \cdot 0.8 = 0.16$, $\sigma(\{f\}) \cdot \sigma(\{d,f\}) = 0.8 \cdot 0.2 = 0.16$. Thus, $\texttt{Lev}(\{d,e,f\}) = 0.2 - 0.16 = 0.04$.

The leverage of an itemset is based on the notion of leverage of a rule. Hereafter, we use leverage of a rule in our comparison as a base line and, thus, we need to provide its definition.

Definition 8. *The leverage of a rule is defined as follows*

$$\texttt{Lev}(X \to Y) = \sigma(X \cup Y) - \sigma(X) \cdot \sigma(Y) \tag{4}$$

In this work rule leverage is applied to rules of the form $X \to \{\mathscr{C}\}$, where $\mathscr{C}$ is a class label in classification. Let us consider Table 1, where the target class is given by column "class". In order to define rule leverage of $\{e,f\} \to \{+\}$, first, we should find the frequencies: $\sigma(\{e,f,+\}) = 0.6$, $\sigma(\{e,f\}) = 0.8$, $\sigma(\{+\}) = 0.6$. Thus, $\texttt{Lev}(\{e,f\} \to \{+\}) = 0.6 - 0.8 \cdot 0.6 = 0.12$.

We are now ready to introduce our methodology.

5. Evaluation Methodology

In this work the classification task is used to estimate the interestingness of measures for itemset selection w.r.t. expert interest, by measuring the precision and recall of classifiers built with these measures.

The intuition behind the usage of classification for evaluating measures is the following. If an itemset is of high interest for an expert, then it should reflect basic dependencies in a domain. Thus, the performance of this itemset in classification should be better than an arbitrary itemset. Consequently, systematic good performances may mean that a measure is more suitable to find itemsets of high interest. Accordingly, the evaluation methodology consists of the following steps:

1. A dataset $\mathscr{D}$ is selected.
2. The dataset $\mathscr{D}$ is divided into training and test sets by random sampling 100 times. A training set contains 90% of the objects with class labels (but at most 1000 objects which is a limit of `Magnum Opus` demo [14] that is used for leverage computation). The test set contains the rest of the objects.
3. One target class label $\mathscr{C}$ is selected.
4. One target measure $\mathscr{M}$ is selected.
5. A training set built at step 2 is used to find itemsets and rank them w.r.t. the measure $\mathscr{M}$. However, during the search, class labels for objects are ignored.

6. Among the whole set of itemsets, the emerging patterns for class $\mathscr{C}$ are selected from the training set [15]. An emerging pattern is an itemset that is a characteristic of one class, i.e. it covers objects mostly labelled with the same class, w.r.t. a threshold θ. These emerging patterns are assumed to be good for classification purposes. The idea of emerging patterns is borrowed from [16], where emerging patterns are called hypotheses. Let say that there are N emerging patterns.
7. From these N emerging patterns we form N classifiers based on the first k patterns (with $k \leq N$). Each classifier works in the following way. Given a set of patterns $\{p_1, \cdots, p_k\}$, the classifier attaches the label $\mathscr{C}$ to any object whose description contains p_i for $i \in [1,k]$.
8. We compute precision and recall for these N classifiers in the test set. Then we interpolate 21 points of the form (p,r) where p stands for precision and r stands for recall, where $r \in \{0, 0.05, \cdots, 0.95\}$. These 21 points yield a curve.
9. Steps 6–8 are repeated for every pair of training and test sets. An average curve is computed for all the curves based on the pairs of training and test sets.
10. The area under this averaged curve is computed providing a numerical quality of the measure $\mathscr{M}$ in dataset $\mathscr{D}$ w.r.t. class $\mathscr{C}$.
11. We repeat steps 3–10 for all classes in $\mathscr{D}$ and all measures.
12. We repeat steps 1–11 for all available datasets.

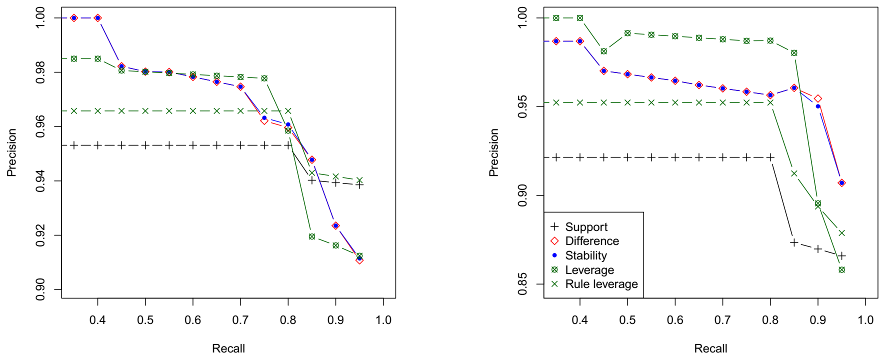
Thus, each measure is evaluated for every class label and for any division of a dataset. The precision and recall in step 8 are computed in a standard way, i.e. in terms of true/false positives/negatives where the precision is $\mathtt{Pr} = \frac{\mathtt{TP}}{\mathtt{TP}+\mathtt{FP}}$ and the recall is $\mathtt{R} = \frac{\mathtt{TP}}{\mathtt{TP}+\mathtt{FN}}$.

But how can one select the threshold θ? This is a tricky question. On the one hand, it is necessary to take the high θ in order to force a measure to select itemsets relevant for the classification. Thus, datasets where there are no patterns with high θ are not adapted for the methodology. On the other hand, it is necessary to have a sufficient number of emerging patterns to capture differences between measures. Here, we posed $\theta = 90\%$, i.e. at least 90% of objects in the image of a pattern are in the same class. However, the selection of an ideal θ is still an open question.

In [9] measures rely on class labeling and thus they are biased for classification task. In contrast in our work measures evaluate itemsets and after that a labeling is introduced. Thus, our approach appears to be closer to the expert interest.

Let us consider this methodology on the example in Table 1. We have a dataset containing 8 objects (step 1). This dataset is divided into training set, $Tr = \{g_1, g_2, g_3, g_4, g_5\}$, and test set $T = \{g_6, g_7, g_8\}$ (step 2). The target class label is $\mathscr{C} = +$ (step 3). The target measure is difference (step 4). In this example we consider an itemset to be an emerging pattern if 50% of objects in its image are labeled with the target class. Thus, we have five closed emerging patterns: $\{e,f\}$, $\{c,f\}$, $\{a,e,f\}$, $\{c,e,f\}$ and $\{b,c,f\}$ (step 6). The corresponding differences are 3, 1, 1, 1, 1. Thus, they are well sorted and we are ready to construct classifiers (step 7) and evaluate their performance (steps 8 and 9).

The first one is only based on $\{e,f\}$. This itemset is only included in the description of g_6, consequently only g_6 should be classified positively. The precision and recall of this classifier are 1 and 0.5. The next classifier is based on $\{e,f\}$ and $\{c,f\}$. The description of g_6 includes $\{e,f\}$ and, thus, it should be classified positively with the second classifier. The descriptions of g_7 and g_8 include $\{c,f\}$ and, thus, they should be also classified positively. The precision and recall of the second classifier is $\frac{2}{3}$ and 1. After



(a) Eatable mushrooms

(b) Poisonous mushrooms

Figure 1. Precision and Recall for mushroom dataset for classifiers built with different interestingness measures.

repeating this with all emerging patterns we can interpolate the value of precision for every recall of the form $0.05 \cdot K$, where $K \in \{1, 2, \cdots, 20\}$. Doing this several time for every division of the dataset we can obtain the averaged precisions corresponding to these recalls. Finally, we can compute the area under the average curves providing a numerical quality of the measure on this dataset.

Finally, any emerging pattern X for class $\mathscr{C}$ can be written as an association rule $X \rightarrow \{\mathscr{C}\}$. Thus, it is also possible to introduce the interestingness measures for rules in this framework as a baseline for evaluating interestingness measures of itemsets. We decided to add the leverage interestingness measure for rules, see Eq. (4). The results for rule leverage measure are provided only as a baseline because it uses the target class in the computation procedure and, thus, can be better adapted for classification purposes.

6. Experiment

The experiments are carried out with public available datasets from UCI [17]: `Mushroom`[2], `Congressional Voting Records`[3], `Nursery`[4] datasets. All datasets contain emerging patterns and thus we can apply our methodology. In the experiments we have compared 4 interestingness measures for itemset ranking, i.e. support, stability (1), difference (2) and leverage (3), as well as a measure for association rule ranking, i.e. rule leverage (4). Comparison of the computational efficiency is not the goal of this paper. Thus, we only mention that computations take less than a minute per experiment in every case.

Let us consider one dataset deeper. Figure 1 shows the results of two experiments on `Mushroom` dataset. Figure 1a shows precision and recall for predicting the class of edible mushrooms, while Figure 1b corresponds to poisonous mushrooms. Every line in

[2] `http://archive.ics.uci.edu/ml/datasets/Mushroom`

[3] `http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records`

[4] `http://archive.ics.uci.edu/ml/datasets/Nursery`

Table 2. The surface under the ROC-diagram for different datasets different target classes and different measures. The best measure in a row is bolded.

Dataset	Class	Support	Difference	Stability	Itemset Lev.	Rule Lev.
Mushroom	Poisonous	0.890658	0.945881	0.945665	**0.956895**	0.919898
Mushroom	Eatable	0.927239	0.953793	**0.953941**	0.946683	0.938007
Vote	Democrat	0.865279	0.862507	0.8645	0.904433	**0.953708**
Vote	Republican	0.883406	**0.921093**	0.921004	0.884818	0.883406
Nursery	Not Recommended	**0.975**	**0.975**	**0.975**	**0.975**	**0.975**
Nursery	Priority	**0.78503**	0.743039	0.725221	0.605405	0.525
Nursery	Special Priority	**0.875556**	0.850174	0.851127	0.699788	0.639793

this figure corresponds to a measure. Every point corresponds to a precision-recall pair at the end of step 9 of the proposed methodology.

In this figure we can see that stability and difference have nearly the same behaviour. It is the case for every tested dataset. The second point is that the support measure is not the best one for pattern selection, which is not surprising. The unexpected result here is that the rule leverage does not perform well. Logically it should be the best one because it is the only tested measure that can access the label information in the dataset. One explanation can be that the statistical significance (at least of the rule leverage type) is not directly related to the relevancy of an itemset to real patterns.

In Table 2 the numerical qualities for every dataset and every class label is given. Every column corresponds to a measure and every line corresponds to a combination of a dataset and a class label. First, the difference and stability measures have a similar behaviour, and the numerical quality has nearly the same value in every experiment. Second, although there is no evident winner between stability and itemset leverage and they often have a comparable result, but on `Nursery` dataset stability has a significantly better result.

7. Conclusion

In this paper we have proposed a methodology for evaluating interestingness measures for closed itemset selection. The proposed methodology has been applied to compare leverage, stability and difference measure. Although stability has a slightly better behaviour than leverage we cannot conclude that one is better than the other. It is also shown that stability and difference have very similar behaviours, but difference is computed faster.

It should be noticed that stability and difference have an important property that they can be applied to any kind of datasets as soon as support can be computed, e.g. datasets of sequences or graphs. This is not, for example, the case for leverage. Since difference is faster to compute, we should conclude that difference is the most convenient measure providing the same quality as stability and leverage.

There are several directions for future research. First, other measures and other datasets should be involved into comparison for more reliable results. Second, the correlation of ranking w.r.t. different measures should be studied. Finally, since stability and leverage have the best performances on different datasets, it can be an important task to develop a powerful measure based on both approaches.

Acknowledgements

This research received funding from the Basic Research Program at the National Research University Higher School of Economics (Moscow; Russia) and from the BioIntelligence project (France).

References

[1] Geoffrey I Webb and Songmao Zhang. K-Optimal Rule Discovery. *Data Min. Knowl. Discov.*, 10(1):39–79, 2005.

[2] Deborah R. Carvalho, Alex A. Freitas, and Nelson Ebecken. Evaluating the Correlation Between Objective Rule Interestingness Measures and Real Human Interest. In Alípio Mário Jorge, Luís Torgo, Pavel Brazdil, Rui Camacho, and João Gama, editors, *Knowl. Discov. Databases PKDD 2005*, volume 3721 of *Lecture Notes in Computer Science*, pages 453–461. Springer Berlin Heidelberg, 2005.

[3] Albrecht Zimmermann. Objectively evaluating interestingness measures for frequent itemset mining. In *Emerg. Trends Knowl. Discov. Data Mining-PAKDD 2013 Int. Work.*, 2013.

[4] Geoffrey I. Webb. Self-sufficient itemsets. *ACM Trans. Knowl. Discov. Data*, 4(1):1–20, January 2010.

[5] Sergei O. Kuznetsov. On stability of a formal concept. *Ann. Math. Artif. Intell.*, 49(1-4):101–115, 2007.

[6] Aleksey Buzmakov, Sergei O Kuznetsov, and Amedeo Napoli. Scalable Estimates of Concept Stability. In Christian Sacarea, Cynthia Vera Glodeanu, and Kaytoue Mehdi, editors, *Form. Concept Anal.*, volume 8478 of *Lecture Notes in Computer Science*, pages 161–176. Springer Berlin Heidelberg, 2014.

[7] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 1st edition, 1999.

[8] Adnan Masood and Stephen Soong. Measuring Interestingness – Perspectives on Anomaly Detection. *Comput. Eng. Intell. Syst.*, 4(1):29–40, 2013.

[9] Paulo J. Azevedo and Alípio M. Jorge. Comparing Rule Measures for Predictive Association Rules. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Mach. Learn. ECML 2007*, volume 4701 of *Lecture Notes in Computer Science*, pages 510–517. Springer Berlin Heidelberg, 2007.

[10] Sergei O. Kuznetsov. Stability as an Estimate of the Degree of Substantiation of Hypotheses on the Basis of Operational Similarity. *Autom. Doc. Math. Linguist. (Nauch. Tekh. Inf. Ser. 2)*, 24(6):62–75, 1990.

[11] Camille Roth, Sergei Obiedkov, and Derrick G Kourie. On succinct representation of knowledge community taxonomies with formal concept analysis A Formal Concept Analysis Approach in Applied Epistemology. *Int. J. Found. Comput. Sci.*, 19(02):383–404, April 2008.

[12] Radim Belohlavek and Martin Trnecka. Basic Level in Formal Concept Analysis: Interesting Concepts and Psychological Ramifications. In *Proc. Twenty-Third Int. Jt. Conf. Artif. Intell.*, IJCAI'13, pages 1233–1239. AAAI Press, August 2013.

[13] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In *Data Eng. 1995. Proc. Elev. Int. Conf.*, pages 3–14, March 1995.

[14] Geoffrey I. Webb. Discovering Significant Patterns. *Mach. Learn.*, 68(1):1–33, 2007.

[15] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proc. fifth ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, KDD '99, pages 43–52, New York, 1999. ACM.

[16] Sergei O. Kuznetsov. Mathematical aspects of concept analysis. *J. Math. Sci.*, 80(2):1654–1698, 1996.

[17] A. Frank and A. Asuncion. *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]*. University of California, Irvine, School of Information and Computer Sciences, 2010.