

# Contextual Comparison of Discovered Knowledge Patterns

A.G. Büchner\*, M. Baumgarten\*, J.G. Hughes\*, W.D. Patterson\*

\* Northern Ireland Knowledge Engineering Laboratory  
Faculty of Engineering, University of Ulster  
{ag.buchner, m.baumgarten, jg.hughes, wd.patterson}@ulster.ac.uk

**Abstract:** *Contextual comparison of discovered patterns deals with the interpretation of outputs from data mining algorithms. The vehicle provided to perform such operations is that of contextual interestingness, which allows the allocation of importance and direction to each attribute in a result set. Applying these mechanisms it is not only possible to detect trends in results across time, but also to compare individual result elements.*

**Keywords:** data mining, knowledge comparison, interestingness, context

## 1 Introduction

Recently, dissemination, application and deployment of results, which have been generated by knowledge discovery, have been of interest to the research community. The main objective is that once useful and novel information has been discovered it can be utilized in a given domain. A problem which occurs frequently in commercial and research scenarios, is that results from knowledge discovery exercises carried out in separate contexts have to be compared. The objective of this paper is to address this issue.

For example, a classification is carried out on the customer base of a retail outlet to discover distinguishing behavior between customers with and without loyalty cards. After the introduction of special offers targeted at loyal customers, another analysis is carried out. The question one would ask subsequently is “How much better are the results based on the introduced campaign?” However, no mechanism exists at present to carry out such a comparison operation. Similarly, when evaluating different algorithms for the discovery of the same type of patterns, the results cannot be compared without mundanely stepping through the details of each result set.

Section 2 defines the problem and outlines the scope of the paper. In Section 3, a generic interestingness framework is presented and notational issues are addressed. Section 4 applies the proposed framework before an application in the web mining arena is presented in Section 5. Section 6 concludes the paper.

## 2 Problem Definition

Results from knowledge discovery stem from different contexts. The types of context which are relevant are *algorithm contexts* (same data is applied to algorithms, e.g. ID3 and C5), *data contexts* (different data / same algorithm and threshold settings, e.g. from different time spans or samples), *parameter contexts* (parameters such as

thresholds are modified using the same algorithm and data), or any permutation thereof. Additionally, analysts interpret results from different viewpoints (*user context*).

In order to allow the interpretation of results generated in such disparate situations, it is necessary to have access to a flexible, yet powerful, mechanism which allows the comparison of knowledge. Issues which arise are

- What types of knowledge can be compared with each other?
- How can contextual information be incorporated?
- What is the most appropriate equivalence mechanism to be applied in order to perform comparisons?

This work will resolve these issues and provide a novel contextual interestingness measure which can be used for comparison of results from knowledge discovery.

The high-level calculation of the contextual interestingness calculation  $\theta$  of any knowledge component of *type* in a certain *context* is formulated as follows.

$$\theta_{Type}^{Context} \rightarrow [0..1] \quad (1)$$

The greater the result the more interesting it is. The objective is to specify this calculation with the greatest degree of flexibility and support of contextuality.

Silberschatz & Tuzhilin [1] have tackled the central problem of ‘good’ measures to identify the interestingness of a pattern by introducing two different kinds of interestingness. Objective measurements relate to the structure of a pattern object and the underlying data used to discover them, while subjective measurements depend on the user’s needs, the domain the data is analyzed in, and the scenario to which they are applied. While their approach allows the comparison of interestingness values, it neither provides a vehicle to allow a user to define the concept of comparative or contextual interestingness. Further work related to the proposed approach covered by the three areas of *knowledge fusion* which deals with the combination of knowledge, *knowledge sharing* which refers to the process of locating and extracting knowledge from multiple sources and transforming it so that the union can be applied in problem-solving and *sequence alignment methods* which calculate the distance between sequences, which is reflected by the numbers necessary to convert a source sequence into its target counterpart.

## 3 Interestingness Framework

This section introduces a framework which provides structures and operations for the comparison of multiple

results from knowledge discovery. The principle idea is not to compare knowledge per se, but to compare the results which are derived from discovered knowledge.

### 3.1 Result Comparison Structure

The outcome of a knowledge discovery exercise is a predictive model. Each model is of a certain type, for instance a neural network or a set of sequences. Models of some type contain also the information about the data it has been derived from (associations, sequences, episodes), while most types only provide information about the model itself (rules, clusters, neural nets, regression, etc).

While it is in principle possible to compare results of different type e.g., a neural net with a decision tree, the scope of this work is restricted to the comparison of compatible results, i.e. results of the same type. All results of the same type  $t$  are organized in a result space  $\mathcal{R}$ .

**Definition 1.** Result Space  $\mathcal{R}$

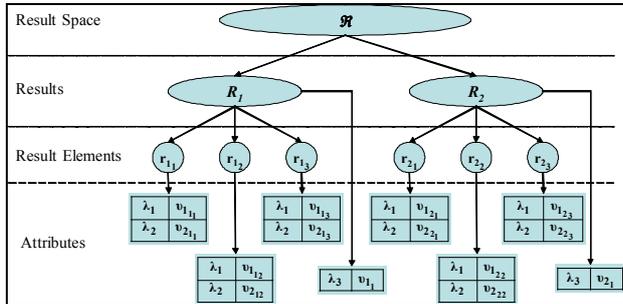
$$\mathcal{R} = \{R_1, R_2, R_3, \dots\}, \text{ such that } \left| \bigcup_{i=1}^{|\mathcal{R}|} t(R_i) \right| = 1. \quad \blacklozenge$$

Each  $R \in \mathcal{R}$  contains a set of result elements which describe the result with quantitative and / or qualitative values. Quantitative measures represent information about the result element per se e.g. support and confidence. Qualitative measures provide information about the content of a result element e.g. quantification of values.

**Definition 2.** Result  $R$  and Result Elements  $r$

$R = \langle \{r_1, r_2, r_3, \dots\}, a \rangle$ ; each result element  $r = \langle I, a \rangle$ , where  $I$  is an optional set of items and  $a$  a set of attribute tuples such that  $a = \{ \langle \lambda_1, v_1 \rangle, \langle \lambda_2, v_2 \rangle, \dots \}$ , where  $\lambda$  represents a label and  $v$  its normalized value ( $0 \leq v \leq 1$ ).  $\blacklozenge$

Quantitative and qualitative values are treated holistically and are referred to as attributes. A set of attributes can be attached to each element in  $\mathcal{R}$  and  $R$ . The range of available attributes depends on the type of knowledge that has been generated. A topology containing results and allotted attributes is depicted in Figure 1.



**Figure 1.** Result Space Topology

In order to illustrate the outlined concepts an example is provided dealing with the results of a decision tree. The objective of the exercise is to apply a model on two data sets (1998 and 2003) of the same patients, which predicts the likelihood of developing Alzheimer's disease. The

input data contains information on patients' gender, age, height, weight, smoking habits, alcohol consumption, etc.

A rule induction example has been applied with three classification labels (low, medium and high) for the predictability of the disease, which are represented as items. The output sets contain two attributes, namely support and confidence.

$R_i$	$I_{1998}$	$a_{support}$	$a_{confidence}$	$R_j$	$I_{2003}$	$a_{support}$	$a_{confidence}$
$r_{11}$	Medium	0.04	0.4	$r_{21}$	High	0.05	0.5
$r_{12}$	High	0.05	0.2	$r_{22}$	High	0.02	0.8
$r_{13}$	Low	0.07	0.8	$r_{23}$	Medium	0.07	0.6
$r_{14}$	Medium	0.01	0.4	$r_{24}$	High	0.02	0.2

**Table 1.** Rule Induction Results

The sample result elements are shown in the table above. Questions that are feasible to ask are: "Has the health of patients improved or deteriorated?" and "Which patient's condition has improved / deteriorated over the last 5 years?"

### 3.2 Contextual Interestingness

In order to compare attributes and results, the notion of contextual interestingness is introduced on attribute, element and result level. In order to allow the user to specify contextuality for a given problem, the concept of contexts is introduced. A context describes a given phenomenon, scenario or problem, using knowledge discovery specific attributes. Contexts are organized in a context space  $\Gamma$ .

**Definition 3.** Contexts

$\Gamma = \{\gamma_1, \gamma_2, \gamma_3, \dots\}$ , where each  $\gamma = \{\alpha_1, \alpha_2, \alpha_3, \dots\}$ . Each attribute  $\alpha = (\lambda, \delta, \iota)$ , where  $\delta \in \{0, 1\}$  and  $0 \leq \iota \leq 1$ .  $\blacklozenge$

The label  $\lambda$  is the name of a phenomenon in a given problem space, aka context identifier. Examples in data mining are thresholds (support, confidence) or quantitative information (weight, size, and length). The label is the logical link between result attributes and context attributes. The direction  $\delta$  is a binary value that states whether an increase of the value is positive (0) or negative (1) in the context the data mining exercise is carried out. The importance factor  $\iota$  states the relevance of the direction  $\delta$ .

For example, when discovering sequences in a web mining application, long sequences are attractive when the host's remuneration is based on the number of page impressions. Contrarily, short sequences are appealing if the objective is that customers solve their problem with as few clicks as possible. Both importance and direction are adjustable within the given limits.

The interestingness of an attribute comprises the degree of interest associated with it in a given context. That is when putting a certain result in a certain context, its interestingness can be calculated.

**Definition 4.** Attribute interestingness  $\theta_a$

$$\theta_a(v, \delta, \iota) = |(v - \delta) * \iota| \rightarrow [0 \dots 1] \quad \blacklozenge$$

Interestingness embraces both objective and subjective measurements. Former relates to the structure of a pattern and the underlying data used to discover it, while latter

depends on the user's needs, the domain the data analyzed is in, and the scenario to which it is applied. Thus, the interestingness of a pattern is by no means an objective value that remains constant across comparisons; interestingness is a subjective representation of the user's priorities in conjunction with the raw pattern values.

In order to compute the interestingness  $\theta_o$  of the attributes of either a result element  $r$  or a result  $R$  all attributes are taken into account.

**Definition 5.** Object interestingness  $\theta_o$

$$\theta_o(a) = \left( \sum_{i=1}^{|a|} \theta_a(v_i, \delta_i, t_i) \right) / |a| \rightarrow [0..1] \quad \blacklozenge$$

$|a|$  represents the amount of attributes in  $a$ . This operation calculates the arithmetic mean of all attribute values in a given context. This measure is used as basis for the calculation of the result element interestingness  $\theta_e$  and the result interestingness  $\theta_r$ .

**Definition 6.** Element interestingness  $\theta_e$

$$\theta_e(r) = \theta(a(r)) \rightarrow [0..1] \quad \blacklozenge$$

**Definition 7.** Result interestingness  $\theta_r$

$$\theta_r(R) = \theta(a(R)) \rightarrow [0..1] \quad \blacklozenge$$

Using the earlier introduced Alzheimer's prediction example, the classification labels have been quantified to low=0.3, medium=0.6 and high=1. The following can be calculated, given that for  $\lambda_{\text{risk}} \delta=0$  (in this context lower is more interesting),  $\iota=100\%$  (highest importance), for  $\lambda_{\text{support}} \delta=1$ ,  $\iota=50\%$  and for  $\lambda_{\text{confidence}} \delta=1$ ,  $\iota=80\%$ .

$$\theta_e(r_{11}) = \frac{|(0.6-1)*1.0| + |0.04*0.5| + |0.4*0.8|}{3} = 0.247 \quad (2)$$

$$\theta_e(r_{21}) = \frac{0 + |0.05*0.5| + |0.5*0.8|}{3} = 0.142$$

Due to the fact that interestingness of the result in 1998 is greater than the one of 2003, this means that patient  $r_1$ 's condition has worsened. Calculating the four result element interestingness measures for 1998 and 2003, and building the arithmetic mean results to 0.252 and 0.18, respectively. Those two values are accepted as new attributes  $\lambda_{\text{risk}}$  for  $R_1$  and  $R_2$  ( $\delta=0$  and  $\iota=100\%$ ). Given that Alzheimer's disease is an age related illness the data set of 2003 is 'punished' via an age attribute ( $\delta=1$ ,  $\iota=20\%$ ).

$R$	$\left( \sum_{i=1}^4 \theta_e(r_i) \right) / 4$	$\delta_1$	$\iota_1$	age	$\delta_2$	$\iota_2$
1998	0.252	0	100%	0.5	1	20%
2003	0.18	0	100%	0.8	1	20%

**Table 2.** Result Attributes

Calculating  $\theta_r(R_1)=0.176$  and  $\theta_r(R_2)=0.038$  it can be shown that the overall population has also deteriorated; in fact the situation has worsened substantially.

### 3.3 Attribute Generation

As outlined previously, the set of attributes  $a$  associated

with each  $r$  and  $R$  of the result space form the basis of calculating contextual interestingnesses. Attributes on each level can be provided through the result sets themselves, they can be specifically set by the user or they can be derived from other attributes, e.g. average or coverage values. Average values represent the arithmetic mean of sub values, that is the average interestingness of all  $r \in R$  calculated as follows.

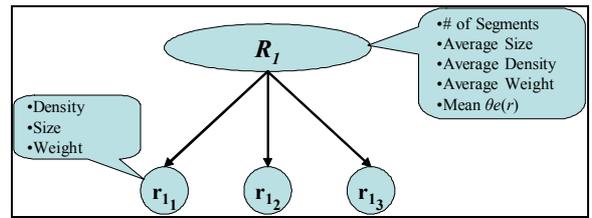
$$\left( \sum_{i=1}^{|R|} \theta_e(r_i(R)) \right) / |R| \quad (3)$$

Coverage values are derived using the scope of the overall result space. For instance, given all distinct result elements of  $r$ , a coverage attribute for each  $R$  is calculated indicating the exposure of  $r$  in each  $R$ .

## 4 Specific Knowledge Types

The proposed framework has been applied to results from a rule induction exercise. In order to show that the introduced mechanisms can be adapted to every common type of knowledge discovered by a predictive modeling exercise, the specific knowledge type of sequences is presented in detail, before other types are briefly covered.

In addition to the generic structural and operational artifacts presented in Section 3, a collection of standard segment-specific attributes is introduced. A pictorial summary of the constructs is depicted in Figure 2 below. Typical standard attributes on result element level are segment density, size and weight. More domain (or context) specific attributes can be added to this set, for instance average price per segment. Attributes on result level include the number of segments in all  $r$ 's or average segment size, density and weight. The inclusion of the mean element interestingness of all  $r$ 's has proven useful (see example above).



**Figure 2.** Segment-specific Structure

It is possible to introduce attributes at result space level, for example, total number of segments, total number of unique segments or average segment size, density and weight. These attributes have proven useful when calculating the interestingness of  $\mathcal{R}$ , for instance, to determine the quality of the entire result space. However, due to the fact that they cannot be used for comparison purposes (only one knowledge space exists for each type), they were not included in this work.

Due to space restrictions it is not possible to cover a wide range of knowledge types in detail. The table below lists a range of attributes for a list of important knowledge types.

Type	Element	Result
Associations	# of items	# associations
	Support	Avg. / unique # items
	Confidence	Support/ confidence
Sequences	# transactions / sets	# sets
	Max. time	Occurrence
	Max. support	Support / Confidence
Decision Trees	# nodes	Score
	# scores	Record count
	Avg. predicate number	# predicates
Neural Networks	# neurons	Avg. bias / threshold
	Avg. bias	Avg. connection weight
Regression	Avg. # predictors	# predictors
	Max. intercept	Intercept
	Avg. coefficient	Avg. coefficient
Naïve Bayes	Threshold	# pair counts
	# Bayed input	Avg. value / count
	Avg. value / counts	Max. value

**Table 3.** Knowledge Type Attributes

It must be stressed that the set of covered attributes and methods is by no means complete and can be extended at any time. For instance, in the application described in Section 5, an additional ‘coverage’ measure has been introduced, which conveys the number of items in a rule related to the number of items in the respective result.

## 5 Application

The presented domain-agnostic interestingness framework has been integrated in a knowledge comparison architecture, where a web mining application has been conducted. The goal was to show the changes of visitor clusters in two time frames (contexts). After the analysis of the first period, changes were carried out to the site and the objective was to demonstrate the impact of those changes. A representative investor relations segment from each time frame is used to show the application of the framework.

During the first period the cluster contained eight pages, which are listed in Table 4a in conjunction with the rounded average number of seconds spent on each page.

Page	Seconds spent	Page	Seconds spent
Home	5	Home	5
Staff/CEO	22	Seniorstaff	33
Staff/CIO	20	Investment	48
Staff/CFO	20	Endorsements	11
Investment	45	Board	9
Board	16		
Awards	8		
Endorsements	8		

**Table 4.** Cluster from (a)Context<sub>1</sub> and (b)Context<sub>2</sub>

The three staff pages were amalgamated to a senior staff page and the content of the awards page was included in an endorsement page (Table 4b). The following table shows the attributes and its values for both contexts and their respective importance factors and directions.

Attribute	r <sub>1</sub>	r <sub>2</sub>	ι	δ
λ <sub>#Pages</sub>	8 (1.0)	5 (0.625)	100%	↓
λ <sub>TimeSpent</sub>	101 (0.73)	139 (1.00)	100%	↑
λ <sub>Density</sub>	0.42	0.37	75%	↓
λ <sub>Size</sub>	84 (0.74)	113 (1.00)	10%	↑
λ <sub>Weight</sub>	0.17	0.22	75%	↑

**Table 5.** Attributes for r<sub>1</sub> and r<sub>2</sub>

Following the element interestingness  $\theta_e$  calculation in Definition 6, the comparison of the two segments representing investor relations, was calculated as follows.

$$\theta_e(r_1) = \frac{0 + 0.73 + 0.435 + 0.074 + 0.623}{5} = 0.372 \quad (5)$$

$$\theta_e(r_2) = \frac{0.375 + 1 + 0.473 + 0.1 + 0.585}{5} = 0.5065$$

The cluster from the second time frame (r<sub>2</sub>) is significantly greater (that is, better within the scope of the analysis) than the first ( $\theta_e(r_1) < \theta_e(r_2)$ ). Thus, it was possible to measure the impact to the investor segment of the changes which were made to the web site structure.

A further comparison was carried out which compared the two clusters in their entirety (all results of both contexts). The result interestingness  $\theta_r$  was performed according to Definition 7, based on the derived avg. element interestingness of all r’s, the number of clusters, the avg. time spent, the avg segment size and weight.

Attribute	R <sub>1</sub>	R <sub>2</sub>	ι	δ
λ <sub>Avgθ<sub>e</sub></sub>	0.44	0.58	100%	↑
λ <sub>#Cluster</sub>	6 (0.75)	8 (1.00)	0%	↑
λ <sub>AvgTimeSpent</sub>	34 (0.89)	38 (1.00)	100%	↑
λ <sub>AvgSize</sub>	417 (1.00)	364 (0.87)	50%	↑
λ <sub>AvgWeight</sub>	0.25	0.22	75%	↑

**Table 6.** Attributes for R<sub>1</sub> and R<sub>2</sub>

Calculating the result interestingness for R<sub>1</sub> and R<sub>2</sub> results in  $\theta_e(R_1)=0.404$  and  $\theta_e(R_2)=0.436$ , which shows that the new site is more interesting. However, the change is not as significant as the one for the investor relationship segment, which was the intention of the restructuring.

If the target of the analysis would be the improvement of the click-to-close ratio, the direction of  $\lambda_{AvgTimeSpent}$  would be reversed. Keeping all other values static results in  $\theta_e(R_1)=0.248$  and  $\theta_e(R_2)=0.236$ . This hypothetically example demonstrates the simplicity and flexibility of the framework and shows how the same results of knowledge discovery can be analyzed in multiple contexts.

## 6 Conclusions

A generic framework has been presented that allows the comparison of results, which are discovered by knowledge discovery. The framework is algorithm agnostic, that is it covers all common types of knowledge (*generality*). Due to its flexible structure, full *extensibility* and *re-usability* are guaranteed. The vanilla approach of the model and its calculations assure *simplicity* and *integratability*. All contextual values can be adjusted interactively (*flexibility*) providing a solid domain-agnostic basis (*applicability*).

## References

- [1] A. Silberschatz, A. Tuzhilin, What Makes Patterns Interesting in Knowledge Discovery Systems, in *IEEE TKDE*, 8:970–974, 1996.