# Quality issues, measures of interestingness and evaluation of data mining models

Philippe Lenca, Stéphane Lallich

# QIMIE 2009

Proceedings of
the first International Workshop

## Quality issues, measures of interestingness and evaluation of data mining models

In association with
The 13th Pacific-Asia Conference on
Knowledge Discovery and Data Mining (PAKDD'09)
April 27, 2009, Bangkok, Thailand

**Organized by**
**Philippe Lenca and Stéphane Lallich**

Philippe Lenca & Stéphane Lallich (Eds.)

QIMIE'09:
Quality issues, measures of interestingness
and evaluation of data mining models

First International Workshop
in association with PAKDD'09, Bangkok, Thailand, April 27, 2009
Proceedings
ISBN-13 978-2-908849-23-3 Telecom Bretagne

## Preface

The *Quality issues, measures of interestingness and evaluation of data mining models* Workshop (QIMIE'09) focuses on the theory, the techniques and the practices that can ensure that the discovered knowledge of a datamining process is of quality. This first edition of QIMIE'09 is organized in association with PAKDD'09 conference (Pacific-Asia Conference on Knowledge Discovery and Data Mining, Bangkok, Thailand, on April 27-30, 2009), a major international conference in the areas of data mining and knowledge discovery.

QIMIE'2009 would not have been possible without the work of many people and organizations. We wish to express our gratitude to: Telecom Bretagne, the University of Lyon, the Chairs and Co-Chairs of PAKDD'09 (Thanaruk Theeramunkong from SIIT/Thammasat University, Boonserm Kijsirikul from Chulalongkorn University, Nick Cercone from York University and Ho Tu Bao from Japan Advanced Institute of Science & Technology), the PAKDD'09 Workshop Chairs (Manabu Okumura from Tokyo Institute of Technology and Bernhard Pfahringer from University of Waikato), the QIMIE'09 Workshop facilitator (Juniar Ganis from from SIIT/Thammasat University), the QIMIE'09 Program Committee members & the external reviewers, the keynote speakers (Einoshin Suzuki from Kyushu University and Nitesh V. Chawla from University of Notre Dame), the QIMIE'09 Web master (Philippe Tanguy from Telecom Bretagne).

Last but not least we would like to thank all authors of the submitted papers. Each submission was reviewed by at least three members of the Program Committee. The papers presented in these proceedings were selected after a rigorous review process. At the end, the QIMIE'09 Program includes two keynote speakers and five regular papers.

Einoshin Suzuki who is a well known specialist of exception rule/group discovery gives an overview of Interestingness Measures, addressing Limits, Desiderata, and Recent Results in this domain. He points out the pitfalls to avoid in order to obtain a good quality of discovered patterns. In addition he proposes for a structured pattern an interestingness measure which has exhibited high discovery accuracy. This measure is parameter-free, exploits information from an initial hypothesis, and is based on the minimum description length principle.

It is important to understand the variability in data over time, since even the One True Model might perform poorly when training and evaluation samples diverge. Nitesh V. Chawla presents a very comprehensive framework to proactively detect breakpoints in classifiers predictions and shifts in data distributions through a series of statistical tests. He outlines and utilizes three scenarios under which data changes: sample selection bias, covariate shift, and shifting class priors.

To improve the support-confidence framework José L. Balcázar studies a complementary notion, to be employed jointly with the standard support and confidence bounds, which has the goal of measuring a relative form of objective novelty or surprisingness of each individual rule with respect to other rules that hold in the same dataset.

Prachya Pongaksorn, Thanawin Rakthanmanon, and Kitsana Waiyamai propose a new algorithm to discretize continuous attributes which uses both class information and order between attributes to determine the discretization scheme with minimum number of intervals. According to the experiments the new algorithm contains a smaller number of intervals than other supervised algorithms using less execution time, and the predictive accuracy is as high or higher.

Joan Garriga presents a new interestingness measure which is an alternative option to the support-confidence framework. The biases of this measure have not yet been thoroughly studied but the measure itself has proved to be quite effective as a heuristic when searching to optimize a sample in a simultaneous multi-interval discretization of continuous features. The empirical results show that the most relevant association or classification rules are revealed.

The integration of semantic relationship from the data domain into the knowledge evaluation is an important challenge in data mining. Jiye Li, Nick Cercone, Serene W. H. Wong, and Lisa Yan propose to enhance the rule importance measure issued from rough sets theory by incorporating a weight biased attribute concept hierarchy. By using a geriatric care data set, the authors show that this enhanced rule importance measure provides a knowledge oriented distinction of rules classified as important.

Sylvain Lespinats and Michaël Aupetit link different methods used to detect and avoid the main mapping defaults, i.e false neighbourhoods and tears. In addition, they suggest a new strategy to visualize tears and false neighbourhoods on a mapping by adapting well-tried tools.

Philippe Lenca & Stéphane Lallich

QIMIE'09 Chairs

# QIMIE'09 Committees

## QIMIE'09 Chairs

Philippe Lenca, Telecom Bretagne, France
Stéphane Lallich, Université Lyon 2, France

## QIMIE'09 Program committee

Hidenao Abe, Japan
Jérôme Azé, France
José L., Balcázar, Spain
Bruno Crémilleux, France
Sven Crone, England
Jean Diatta, La Réunion
Thanh-Nghi Do, Vietnam
Salvatore Gréco, Italy
Fabrice Guillet, France
Michael Hahsler, USA
Howard Hamilton, Canada
Martin Holena, Czech Republic
Stéphane Lallich, France
Ludovic Lebart, France
Philippe Lenca, France
Ming Li, China
Patrick Meyer, France
Annie Morin, France
David Olson, USA
Jan Rauch, Czech Republic
Gilbert Ritschard, Switzerland
Wanchai Rivepiboon, Thailand
Gilbert Saporta, France
Roman Slowinski, Poland
Robert Stahlbock, Germany
Athasit Surarerks, Thailand
Shusaku Tsumoto, Japan
Kitsana Waiyamai, Thailand
Dianhui Wang, Australia
Louis Wehenkel, Belgium
Gary Weiss, USA
Takahira Yamaguchi, Japan
Min-Ling Zhang, China
Djamel Zighed, France

## Additional reviewer

Izabela Szczech, Poland

# Contents

# Interestingness Measures - Limits, Desiderata, and Recent Results -

Einoshin Suzuki

Kyushu University, Fukuoka 819-0395, Japan,
suzuki@i.kyushu-u.ac.jp
http://www.i.kyushu-u.ac.jp/~suzuki

In the last two decades, interestingness measures, each of which estimates the degree of interestingness of a discovered pattern, have been actively studied e.g. [1–6, 8–16, 18–22]. A typical interestingness measure is more complex than a machine-learning measure which can be computed from statistics of the given data such as the accuracy, the recall, the precision, the F-value, and the area under the ROC curve. Interestingness measures can be classified as either objective or subjective depending on whether the measure uses only discovered patterns and the data from which the patterns are discovered, or the measure uses additional information such as domain knowledge.

Defining human's interestingness can be called as AI-hard as it is as difficult as all problems in artificial intelligence (AI). What is interesting depends on various factors including the task, the individual, and the context; and any problem and any combination of problems in AI can be used to invent a challenging situation for interestingness measures. We must beware of the hype of omnipotent interestingness measures and we must settle realistic objectives for research on interestingness measures.

Desiderata on interestingness measures can be classified into qualitative expressions such as generality, accuracy, simplicity, and comprehensibility; and quantitative relations such as [11, 21]. We have pointed out desirable properties in exception rule/group discovery such as interpretation of the evaluation measure, which supports the quality of discovered patterns [17]. Pitfalls to avoid are much less known than the desiderata and include four biases in evaluation [18] and a use of many parameters [7], the latter of which poses extra work on the users and results in a problem that is analogous to overfitting in classification.

Recently we have proposed, for a structured pattern, an interestingness measure which is parameter-free, exploits information from an initial hypothesis, and is based on the minimum description length principle [19]. The measure has exhibited high "discovery accuracy" i.e. the ratio that the measure discovers the true hypothesis from several data with up to 30 % of noise using incomplete initial hypotheses.

## Acknowledgment

A part of this research was supported by Strategic International Cooperative Program funded by Japan Science and Technology Agency (JST) and the grants-

P. Lenca and S. Lallich (Eds.): QIMIE/PAKDD 2009

in-aid for scientific research on fundamental research (B) from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

## References

1. J.-P. Barthélemy, A. Legrain, P. Lenca, and B. Vaillant. Aggregation of Valued Relations Applied to Association Rule Interestingness Measures. In *Modeling Decisions for Artificial Intelligence, LNCS 3885 (MDAI)*, pages 203–214. 2006.
2. R. J. Bayardo and R. Agrawal. Mining the Most Interesting Rules. In *Proc. Fifth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 145–154, 1999.
3. S. Brin, R. Motwani, and C. Silverstein. Beyond Market Baskets: Generalizing Association Rules to Correlations. In *SIGMOD 1997, Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pages 265–276, 1997.
4. D. R. Carvalho, A. A. Freitas, and N. F. F. Ebecken. Evaluating the Correlation Between Objective Rule Interestingness Measures and Real Human Interest. In *Knowledge Discovery in Databases (PKDD), LNCS 3721*, pages 453–461. 2005.
5. R. Gras. *L' Implication Statistique*. La Pensée Sauvage, 1996. (in French).
6. S. Jaroszewicz and D. A. Simovici. Interestingness of Frequent Itemsets Using Bayesian Networks as Background Knowledge. In *Proc. Tenth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 178–186, 2004.
7. E. J. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards Parameter-free Data Mining. In *Proc. Tenth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 206–215, 2004.
8. P. Lenca, P. Meyer, B. Vaillant, and S. Lallich. On Selecting Interestingness Measures for Association Rules: User Oriented Description and Multiple Criteria Decision Aid. *European Journal of Operational Research*, 184(2):610–626, 2008.
9. B. Liu, W. Hsu, and S. Chen. Using General Impressions to Analyze Discovered Classification Rules. In *Proc. Third Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 31–36, 1997.
10. B. Liu, W. Hsu, L.-F. Mun, and H.-Y. Lee. Finding Interesting Patterns Using User Expectations. *IEEE Trans. Knowledge and Data Eng.*, 11(6):817–832, 1999.
11. G. Piatetsky-Shapiro. Discovery, Analysis, and Presentation of Strong Rules. In *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, Menlo Park, Calif., 1991.
12. A. Silberschatz and A. Tuzhilin. On Subjective Measures of Interestingness in Knowledge Discovery. In *Proc. First Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pages 275–281, 1995.
13. A. Silberschatz and A. Tuzhilin. What Makes Patterns Interesting in Knowledge Discovery Systems. *IEEE Trans. Knowledge and Data Eng.*, 8(6):970–974, 1996.
14. P. Smyth and R. M. Goodman. An Information Theoretic Approach to Rule Induction from Databases. *IEEE Trans. Knowledge and Data Engineering*, 4(4):301–316, 1992.
15. E. Suzuki. Autonomous Discovery of Reliable Exception Rules. In *Proc. Third Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 259–262. 1997.
16. E. Suzuki. Undirected Discovery of Interesting Exception Rules. *Int'l Journal of Pattern Recognition and Artificial Intelligence*, 16(8):1065–1086, 2002.
17. E. Suzuki. Evaluation Scheme for Exception Rule/Group Discovery. In *Intelligent Technologies for Information Analysis*, pages 89–108. Springer-Verlag, 2004.

18. E. Suzuki. Pitfalls for Categorizations of Objective Interestingness Measures for Rule Discovery. In R. Gras, E. Suzuki, F. Guillet, and F. Spagnolo, editors, *Statistical Implicative Analysis: Theory and Applications*, pages 383–395. Springer-Verlag, 2008.
19. E. Suzuki. Negative Encoding Length as a Subjective Interestingness Measure for Groups of Rules. In *Proc. Thirteenth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. 2009. (accepted for publication).
20. E. Suzuki and M. Shimura. Exceptional Knowledge Discovery in Databases Based on Information Theory. In *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD)*, pages 275–278. 1996.
21. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the Right Interestingness Measure for Association Patterns. In *Proc. Eighth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pages 32–41, 2002.
22. B. Vaillant, S. Lallich, and P. Lenca. On the behavior of the generalizations of the intensity of implication: A data-driven comparative study. In R. Gras, E. Suzuki, F. Guillet, and F. Spagnolo, editors, *Statistical Implicative Analysis: Theory and Applications*, pages 421–447. Springer-Verlag, 2008.

# Confidence Width: An Objective Measure for Association Rule Novelty

José L. Balcázar

Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria, Santander, Spain
joseluis.balcazar@unican.es

**Abstract.** Most often, association rules are parameterized by lower bounds on their support and confidence, even though many other measures exist that evaluate the intensity of implication of a single association rule. We remain within the support-and-confidence framework in an attempt at studying a complementary notion, to be employed jointly with the standard bounds, which has the goal of measuring a relative form of objective novelty or surprisingness of each individual rule with respect to other rules that hold in the same dataset. To do this, we propose to measure the extent to which the confidence value is robust, taken relative to the confidences of logically stronger rules, as opposed to the absolute consideration of the single rule at hand.
Consider a statement such as: "any dataset in which this first rule holds must obey also that second rule"; if the confidences are the same, the second rule becomes redundant. It may still be nonredundant, though, if its confidence is considerably higher than the confidence of the first rule, whereas, if these quantities are very similar, the second rule is not really contributing novel knowledge. We formalize and characterize this intuition, show how it can be computed reasonably efficiently, and study some of the properties of this parameter on two public datasets.

*Keywords: association rules, redundancy, confidence, bases*

## 1 Introduction

Data mining is described sometimes as a collection of processes that explore existing data in order to find knowledge that is actionable, correct, and novel. Actionability is a pretty difficult notion to formalize, in that it is closely related to the human attitudes of the decision makers in front of the discovered knowledge. However, one simple sufficient condition for actionability is based on a plain syntactic condition: if knowledge is expressed in a form that is suggestive of cause-effect relationships, educated users, even nonexperts, can grasp the intuitions in an easy manner and proceed to take steps that, hopefully, redund in some sort of advantage. Therefore, the notion of association rule mining makes sense, even though the corresponding parameters measure co-occurrence rather than causality (see the interesting discussion in [6]). Association rules have been

P. Lenca and S. Lallich (Eds.): QIMIE/PAKDD 2009

long studied; their deterministic counterparts, implications, are actually a form of Horn logic and some communities have developed important progress about them [7]; the probabilistic approach was studied since [13], and fuzzy versions have attracted attention for long as well [8]. Their pervasiveness came from the support constraint, which allowed the design of algorithms able to handle very large and dense datasets; see the survey [5] and the references there. Even for confidence-bounded rules, the notions of frequent closed set (that cannot be enlarged while keeping the same support) and minimal generator (that cannot be reduced while keeping the same support), widely studied in the FCA field [7], remains crucial from an algorithmic perspective [4], [10], [17], [21], [22]. We assume the reader familiar with the Zaki basis and the exact min-max basis of [17], [18], [20], [21]. We will build on a notion of redundancy, related to the entailment in observation logics [8], proposed in terms of association rules in [1], [9], [19]. See [2] for further discussion and relationship with the representative basis employed below.

Correctness of association rules is often assessed, therefore, by a support parameter, that plays also an algorithmic role of pruning exploration of infrequent rules, and some parameter measuring intensity of implication or deviation from independence. The most frequently chosen such parameter is confidence, which is the empirical conditional probability of the consequent of the rule given the antecedent; but a large number of alternatives exist. We will denote $c(X \rightarrow Y)$ the value of the confidence of the rule $X \rightarrow Y$.

Even though we are aware of several objections raised against confidence, we prefer to develop our proposal in that context, for several reasons. First, conditional probability is a concept known to many educated users from a number of scientific and engineering disciplines, so that communication with the data mining expert is simplified. Second, as a very elementary concept, it is the best playground to study other proposals, such as our contribution here, which could be then lifted to other similar parameters. Third, we believe that, in fact, our relative measure will make up for many of the objections raised against confidence, if not immediately, upon further research. Additionally, it must be taken into account that the quantity of data is usually insufficient to test a large number of hypothesis, even if schemes more efficient than the Bonferroni guarantees are employed, and it has been observed and argued that the combination of support and confidence offers already very good results at pruning off statistical artifacts that do not really correspond to correlations in the phenomenon at the origin of the dataset [15].

We focus, then, on the third intuition, namely that of novelty. Again, like actionability, a wide spectrum of subjective considerations regarding the user's previous knowledge can be played, and, of course, novelty with respect to knowledge existing previously to the data mining process is hard to formalize. But one fact is clear: novelty cannot be evaluated in an absolute form; it refers to knowledge that is somehow unexpected, and therefore some expectation, lower than actually found, must exist, due to some alternative prediction mechanism. Ad-

ditionally, an intuitive "rule of thumb" is that the amount of novel facts must be low in order that novelty is actually useful.

We wish to offer here a complementary, objective proposal, to be combined either with other objective measures such as support and confidence or with other subjective measures of novelty, appropriate to the task at hand. We propose to measure the novelty of each rule with respect to the rest of the outcome of the same data mining process. To do this, we resort to recent advances in the construction of irredundant bases and in mathematical characterizations of the most natural notion of redundancy. As we shall see, a redundant rule is so because we can know beforehand, from the information in a basis, that its confidence will be above the threshold. Pushing this intuition further, an irredundant rule in the basis is so because its confidence is higher than what the rest of the basis would suggest: this opens the door to asking, "how much higher?". If the basis suggests, say, a confidence of 0.8 (or 80%) for a rule, and the rule has actually a confidence of 0.81, the rule is indeed irredundant and brings in additional information, but its novelty, with respect to the rest of the basis, is not high; whereas, in case its confidence is actually 0.95, much higher than the 0.8 expected, the fact can be considered novel, in that it states something different from the rest of the information mined.

### 1.1 Related Work, Notation, Redundancy, and Bases

The main notion to be defined below has some surface similarity with the notion of all-confidence [16] and the related concept of $m$-patterns [14]. However, a strong point of these notions, namely, their antimonotonicity, is lacking in our approach, so that we just employ a support bound and discuss our contribution in terms of a standard frequent closed set miner. On the other hand, these notions are very restrictive, and provide only strong "niches" where all the sets of attributes within an output pattern depend heavily pairwise among them. We wish to depart in lesser degree from the standard association rule setting: our proposal discusses a parameter defined for one single rule, but relative to other rules mined.

We denote itemsets by capital letters from the end of the alphabet, and use juxtaposition to denote union, as in $XY$. For a given dataset $\mathcal{D}$, consisting of transactions, each of which an itemset labeled with a unique transaction identifier, we can count the "support" $s(X)$ of an itemset $X$, which is the cardinality of the set of transactions that contain $X$. The confidence of a rule $X \to Y$ is $c(X \to Y) = s(XY)/s(X)$.

A set is closed if there is no proper superset with the same support. A set is a minimal generator if there is no proper subset with the same support. In the presence of a support threshold, frequent closed sets are closed sets whose support clears the threshold. Frequent closed sets are very crucial to the algorithmics of association rules and to the identification of irredundant bases [4], [10], [17], [21], [22]. Absolute optimality of certain versions of these bases is shown in [2].

We start our analysis from one of the notions of redundancy defined formally in [1], but employed also, generally with no formal definition, in several papers on association rules; thus, we have chosen to qualify this redundancy as "standard".

**Definition 1.** *[1]* $X_0 \rightarrow Y_0$ *has* standard redundancy *with respect to* $X_1 \rightarrow Y_1$ *if the confidence and support of* $X_0 \rightarrow Y_0$ *are always larger than or equal to those of* $X_1 \rightarrow Y_1$*, in* all *datasets. That is, for every dataset* $\mathcal{D}$*,* $c(X_0 \rightarrow Y_0) \geq c(X_1 \rightarrow Y_1)$ *and* $s(X_0Y_0) \geq s(X_1Y_1)$*.*

This definition is akin to the definition of entailment in purely logic-based studies, and we will use sometimes the phrase "logically stronger" to refer to a rule that makes another one redundant with respect to standard redundancy. Note that the rules $X \rightarrow Y$ and $X \rightarrow XY$ are mutually redundant, in fact fully equivalent because their confidence $s(XY)/s(X)$ and support $s(XY)$ always coincide. Therefore we consider all association rules where *the right-hand side always includes the left-hand side*, although for the purpose of showing them to the user the repeated items of the left-hand side will be removed from the right-hand side. This simple convention greatly simplifies the mathematical development.

There are several alternative notions of redundancy in the literature; many are closely related to standard redundancy, like the "cover" notion described briefly below; whereas others are somewhat different, such as those based on closed sets: see [2] for further comparisons. For this particular notion we have just given, the aim is clear: whatever the dataset under analysis, and the support and confidence parameters, if we find that rule $X_1 \rightarrow Y_1$ appears among the mined rules by passing the support and confidence thresholds, any other rule $X_0 \rightarrow Y_0$ showing standard redundancy with respect to it is known to be also in the set of mined rules without need to inspect them to check out. This is because the support and confidence must be at least the same as those of rule $X_1 \rightarrow Y_1$, whence it passes the thresholds as well. We will not use in this paper any other form of redundancy; therefore, we will omit often the adjective "standard".

As an example, it will follow from Theorem 1 below, taken from [2], that, for items $A$, $B$, $C$, and $D$, the rule $AB \rightarrow C$ is redundant with respect to rule $A \rightarrow BC$, and is also redundant with respect to $AB \rightarrow CD$. It is easy to check the inequalities in the definition.

This notion of redundancy suggests a notion of a basis, which turns out to be already proposed, independently and in different but equivalent ways, in [1], in [9], and in [19] (see again the discussion in [2]). We term the rules of this basis "representative rules" as per some of these references.

**Definition 2.** *Fix a dataset* $\mathcal{D}$ *and confidence and support thresholds. The corresponding basis of* representative rules *consists of all the rules that hold in* $\mathcal{D}$*, passing both thresholds, which are* not *redundant with respect to any other rule that holds in* $\mathcal{D}$ *for the same thresholds.*

Among several equivalent possibilities to define representative rules, we have chosen a definition so that the following claim becomes intuitively clear: every

rule that passes the thresholds for $\mathcal{D}$ is either a representative rule, or is redundant with respect to a representative rule. Indeed, any given rule that is not among the representative rules must be redundant with respect to some other rule, that again must be redundant with respect to a third, and so on, until finiteness enforces termination that can be only reached by finding a rule in the basis, making redundant all the others found along the way. The formalization of this argument can be found in [9] (Theorem 1 below must be taken into account).

Moreover, any basis, that is, any set of rules that makes redundant all the rules mined from $\mathcal{D}$ at the given thresholds, must include all the representative rules, since there is no other way of making them redundant. Thus, the representative rules form the smallest such basis.

The definitions given are not particularly operative in terms of providing ways to efficiently compute the representative rules. However, in a later section we will describe some properties of these definitions, taken from the references indicated above, which allow one to efficiently compute the representative rules and further quantities we will explain shortly.

## 2 Confidence Width

This section describes the foundations of our proposal. Our intuition is as follows: consider a rule $X \to Y$ of a given confidence, say $c(X \to Y) = \gamma \in [0, 1]$, in a given dataset $\mathcal{D}$. Assume a fixed support threshold is enforced throughout the discussion, and consider what happens as we vary the confidence threshold.

If we set it higher than $\gamma$, the rule at hand will not play any role at all, being of confidence too low for the threshold. As we lower the threshold and reach exactly $\gamma$, the rule becomes part of the output of any standard association mining process, but two different things may happen: the question is whether, at the same confidence, some other "logically stronger" rule appears. If not, $X \to Y$ will belong to the representative rules basis for that threshold; but it may be that, at the same threshold, some other logically stronger rule is found. For instance, it could be that both rules $A \to B$ and $A \to BC$ have confidence $\gamma$: then $A \to B$ is redundant and will not belong to the basis for that confidence.

Let's then assume that the rule at hand does appear among the representative rules at the confidence threshold given by its own confidence value $\gamma$; and let's keep decreasing the threshold. At some lower confidence, a logically stronger rule may appear. If a logically stronger rule shows up early, at a confidence very close to $\gamma$, then the rule $X \to Y$ is not very novel: it is too similar to the logically stronger one, and this shows in the fact that the interval of confidence thresholds where it is a representative rule is short.

To the contrary, a stronger rule may take long to appear: in that case, only rules of much lower confidence entail $X \to Y$, so the fact that it does reach confidence $\gamma$ is novel in this sense. The interval of confidence thresholds where $X \to Y$ is a representative rule is large. For instance, if the confidence of $A \to AB$ is 0.9, and all rules that make it redundant all have confidences below 0.75, the

rule is a much better candidate to novelty than it would be if some rule like $A \to ABC$ would have a confidence of 0.88.

This motivates the following definition:

**Definition 3.** *Fix a dataset $\mathcal{D}$ and a support threshold $\tau$. Consider a rule $X \to Y$ of confidence $c(X \to Y) = \gamma$ in $\mathcal{D}$, and assume that it has support at least $\tau$ and that it belongs to the set of representative rules of confidence $\gamma$. Consider all rules different from $X \to Y$ but such that $X \to Y$ is redundant with respect to them, and pick one with maximum confidence in $\mathcal{D}$ among them, say $X' \to Y'$ (thus $c(X' \to Y') < \gamma$). The* confidence width *of $X \to Y$ in $\mathcal{D}$ is*

$$w(X \to Y) = \frac{c(X \to Y)}{c(X' \to Y')}$$

Note that $\gamma$ is not a confidence threshold here but just the exact value of $c(X \to Y)$: in order to check for the existence of $X' \to Y'$, one should mine at lower confidence levels (but see Proposition 1 below). The confidence width can be defined equivalently as the ratio between the extremes of the interval of confidence thresholds that allow the rule to be representative. That is: the highest value where the rule can belong to the representative rule basis is the confidence of the rule; and the denominator is the highest value where there is a different representative rule that makes it redundant, thus forcing it out of the representative basis.

Observe that when $X \to Y$ is redundant with respect to $X' \to Y'$, its confidence must be at least the confidence of the latter, which implies that the confidence width is always greater than or equal to 1. For a rule $X'' \to Y''$, the confidence width is exactly 1 if and only if there is a rule making redundant $X'' \to Y''$ and having the same confidence: this is the same as saying that $X'' \to Y''$ is never among the representative rules.

Regarding upper bounds, in principle there is none, in that it may happen that a rule of as large confidence as desired is only redundant with respect to rules of as low confidence as desired. However, in many practical cases the width stays between 1 and somewhere between 1.5 and 10, although we will see some cases where width reaches the order of several hundreds.

## 3 Properties and Algorithms

We proceed to study some properties of the confidence width; by combining them with known properties of the standard redundancy and of the representative rules, we will obtain reasonably efficient ways to compute the width of the rules in the basis.

Specifically, we will need first the following known property of the standard redundancy. The formulation was proposed, in different but almost equivalent forms, in [1], in [9], and in [19], as a candidate to a weaker notion of redundancy; the equivalence with standard redundancy was proved only recently (see [2]).

**Theorem 1.** *Consider any two rules $X_0 \to Y_0$ and $X_1 \to Y_1$ where $Y_0 \not\subseteq X_0$. The following are equivalent:*

1. *$X_1 \subseteq X_0$ and $X_0Y_0 \subseteq X_1Y_1$;*
2. *rule $X_0 \to Y_0$ is redundant with respect to $X_1 \to Y_1$.*

This indicates that, in order to test for redundancy, it suffices to compare itemsets by inclusion as in the first of these two statements. Moreover, the following holds:

**Proposition 1.** *Consider a rule $X \to Y$ and a different rule $X' \to Y'$ that makes it redundant; assume $X' \to Y'$ has maximum confidence as in the definition of width, say $\delta$. Then $X' \to Y'$ can be chosen among the representative rules for confidence $\delta$.*

This proposition can be proved easily by resorting to the known fact [9] that every rule of confidence $\delta$ is redundant with respect to a representative rule of the same confidence (possibly itself). As indicated in the previous section, rules not in the representative basis have minimum width, namely 1. Thus, to know the confidence width of all the rules it suffices to find it for representative rules.

We do not need to scan all frequent sets: it is known that if $X \to Y$ is a representative rule, then $XY$ is a closed set (any strictly larger set has strictly smaller support) and $X$ is a minimal generator (any strictly smaller set has strictly larger support) [10]. There are several published algorithms that compute the frequent closed sets and the minimal generators (see the survey [5]); in one form or another, all of them employ the key and well-known fact of the antimonotonicity of the frequent itemsets. These closures and minimal generators can be used to find the representative rules whose width is to be computed, by using the algorithm in [10]. (Note that the algorithm in [11], in principle faster, may miss rules due to the incompleteness of the heuristic employed. This observation will be further elaborated in a later paper. However, the ideas in that paper will be crucial to our improvement in the next section.)

A naive algorithm follows immediately: combine these known algorithms to construct the representative rules and scan them repeatedly, applying Proposition 1 to find, for each rule $X \to Y$, the largest confidence $c$ of any representative rule that makes $X \to Y$ redundant; use Theorem 1 to test for standard redundancy. Once this largest confidence $c$ is known, the width of $X \to Y$ is clearly $w(X \to Y) = \frac{c(X \to Y)}{c}$ by definition. However, notice that this algorithm requires time quadratic in the number of representative rules.

### 3.1 An Improved Algorithm

We analyze now further properties of the confidence width to search for a faster computation. The key is to avoid much of the exploration in the naive algorithm by precomputing a small amount of side information in a single scan of the closures lattice. We explain now what side information would be sufficient; it

is the same as used as a heuristic in [11] to compute a large subset of the representative rules faster. The first step is to find out more about the rules $X' \to Y'$ that could be useful to compute the width of $X \to Y$. Recall our assumption that $X \subset Y$ and $X' \subset Y'$, that we discussed shortly after Definition 1.

**Theorem 2.** *Let $X \to Y$ be a representative rule for a fixed dataset $\mathcal{D}$ at some fixed values of support and confidence. Let $X' \to Y'$ be a different rule that makes it redundant, and assume $X' \to Y'$ has maximum confidence as in the definition of width. Then either $X = X'$ and $Y'$ is a closed set, immediate superset of $Y$ in the lattice of closed sets, and of maximum support among the closed supersets of $Y$; or else, $Y = Y'$, and $X'$ is a minimal generator properly included in $X$ and having minimum support among the proper subsets of $X$.*

The proof is as follows. First apply Theorem 1, but assume that we are in neither of the two cases, that is: $X' \subset X \subset Y \subset Y'$ where all the inclusions are proper. Consider the rules $X' \to Y$ and $X \to Y'$. Clearly, appealing again at Theorem 1, both make $X \to Y$ redundant as well. However, since $Y$ is closed, $s(Y') < s(Y)$, and this implies that $c(X' \to Y') < c(X' \to Y)$; similarly, since $X$ is a minimal generator, $s(X) < s(X')$, and again $c(X' \to Y') < c(X \to Y')$. Therefore, the confidence of $c(X' \to Y')$ is not maximum as required, and one of the two rules $X' \to Y$ and $X \to Y'$ will be the one having maximum confidence among those making $X \to Y$ redundant.

Now, the algorithmic alternative is as follows: along the antimonotonicity-based construction of the frequent closures lattice and the minimal generators (or along the reading from a file if constructed by a separate closed set miner), we keep track of the largest existing support of the frequent closed supersets of each frequent closed set $Y$, let us denote it $mxs(Y)$. Similarly, for each minimal generator $X$, we keep track of the smallest existing support among the minimal generators properly contained in $X$, let us denote it $mns(X)$. Note that we must compute $mns(X)$ for all minimal generators regardless of whether they are also closed, which is something that can happen (for instance, the empty set is often closed, and is always a minimal generator of the smallest closed set, possibly itself). Note that some closures $Y$ may not have frequent closed proper supersets, in the sense that all larger closures could fall below the support threshold; likewise, some minimal generators $X$, namely, the empty set, will lack minimal generators as proper subsets. For such cases, we leave $mxs(Y)$ and $mns(X)$ undefined. If we have all this information available, we can compute rather easily the confidence width:

**Proposition 2.** *Consider a rule $X \to Y$, and assume that both $mxs(Y)$ and $mns(X)$ are defined. Then the width of $X \to Y$ is the minimum of the two values: $\frac{mns(X)}{s(X)}$ and $\frac{s(Y)}{mxs(Y)}$. If only one of $mxs(Y)$ and $mns(X)$ is defined, then the corresponding quotient gives the width.*

This follows directly from Theorem 2 since each of the two cases corresponds to the two options for a rule of maximum confidence making $X \to Y$ redundant.

Algorithm: fast computation of confidence widths
Given: dataset, support threshold, confidence threshold
find all frequent closed sets $Y$ and all their minimal generators $X$
along the same pass, compute the values $mxs(Y)$ and $mns(X)$
for each representative rule $X \to Y$
resort to Proposition 2 to compute $w(X \to Y)$

**Table 1.** Algorithm to avoid quadratic exploration of the redundancy basis

Some rules may not have confidence width according to the definition. These are exactly the same where both $mxs(Y)$ and $mns(X)$ are undefined. They have not arisen in our empirical analysis, and further theoretical development regarding them is undergoing.

The properties just described lead us to the algorithm described in Table 1.

## 4 Empirical Validation

We have computed the widths of the association rules mined from two of the standard FIMI benchmarks, of very different characteristics: `chess`, which is a small but very dense dataset on which even high support constraints lead to many many rules, and the largish, much sparser dataset `retail` coming from a standard application domain (market basket analysis). We have computed the representative rules and their widths and we have plotted the number of rules passing each of a series of width thresholds. In all cases the computation has taken just a few seconds in a mid-range laptop.

If comparatively larger width values are expected to correlate in some sense with novelty, we wish the number of such rules above comparatively larger thresholds to decrease substantially. This is indeed the behavior we have found. With respect to the `chess` dataset, we have constructed rules of confidence 85% out of the closures lattice formed by frequent closed sets at support 80%. Even for such a large support, the number of closures is around 5083 and the representative rules amount to a number of 15067. It is known from the theoretical advances that all of them are fully irredundant, that is, omitting any of them loses information; however, it makes no sense to expect a human analyst to look at fifteen thousand rules.

We propose, instead, to look at the width values: for this dataset, they range in the quite limited interval between 1 and 1.22; and we see that if we impose a very mild bound of width above 1.005, only 2467 out of the 15067 rules reach it. This means that all the others, even if they are indeed irredundant, this is so due to a rather negligible confidence increase. Higher width bounds exhibit an interesting phenomenon of discontinuity, represented by each plateau of the graph in Figure 1 (left): the maximum confidence width of 1.22 is attained by two rules; a third comes close, and all three have high confidences (between 97% and 99%). Then seventeen more rules show up together near width 1.18, and nothing

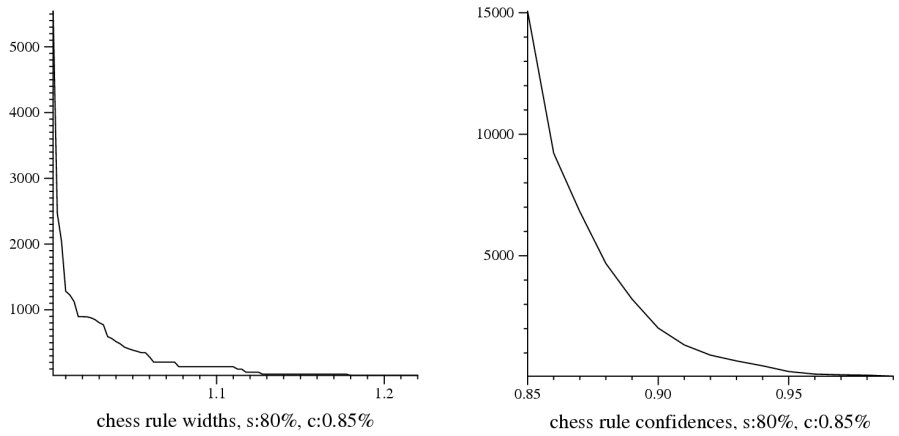



**Fig. 1.** Chess: Number of rules per width and confidence

happens until the width bound gets below 1.13 where a bunch of 31 rules show up together. Below 1.11 we are again at a stable figure of 134 rules, and seventy more appear together at the already quite low confidence width bound of 1.075. All the others, up to 15067, have extremely low width. But the same role cannot be filled directly by confidence: the plot in Figure 1 (right) indicates that there are no steep decreases, no plateau suggesting a good cutpoint shows up, no hint that really any novelty is at play, and, above all, the following fact: the 51 rules of width 1.13 or more all have confidence of 90% or higher, but there are around 1950 other rules, of lower width, attaining the same confidence. Just width is able to focus on the 51 more novel ones.

With respect to `retail`, the behavior of the notion of width is very different, and also very interesting. Huge widths are reached: there are 18 rules whose width is beyond 560 (up to 855.94), whereas the highest next width is just 29: no rule has width between 29 and 560. Another impressive plateau has 7 additional rules of width 21, and from there on the number of rules at each width threshold grows steadily. The different regimes below 29 (above or below width 2) are shown in Figure 2 (left and right, respectively).

## 5 Conclusions

We have proposed an objective approach to the analysis of the novelty of association rules. Although modest and conceptually limited, in the sense that there is no direct room to incorporate background knowledge, that must be the key to a subjective novelty analysis, it already provides some promising behavior on standard public benchmark datasets. We note an additional intuition that seems can be gleaned from the current early developments: it is known that, on the one hand, the standard support-and-confidence bound framework does a

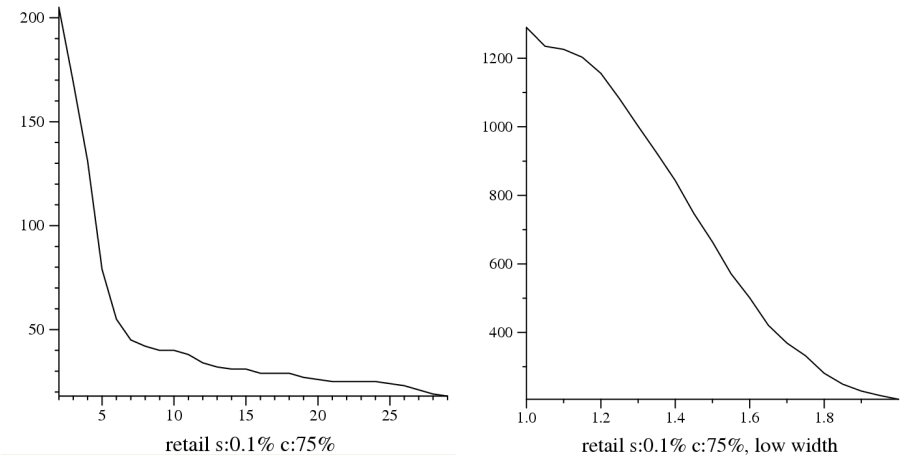



**Fig. 2.** Retail: Number of rules per width, mid and low ranges

good preliminary job for avoiding statistical noise, but, on the other hand, fail somewhat to focus on the really interesting facts; and this is the main reason that has led to a flourishing of variants of notions of intensity of implication to replace confidence, blaming into it the problem. However, we consider now that a viable alternative is to leave the standard support-and-confidence setting on, and complement it, in order to gain further focus, with a measure that does not check intensity of implication in an alternative way (thus, performing something intuitively analogous to confidence) but which checks a relative intensity of implication compared to the other rules mined in the same process.

Our experimental analysis is, admittedly, somewhat limited, and we plan to expand it and compare the two algorithms given here empirically, the naive one and the one in Table 1; but our work so far already suggests several interesting points. It shows that width has the ability to raise wide segments where a width threshold is very robust, and fixing it at a close but different value may select exactly the same rules. It tends to select rules of high confidence but is much more selective. A currently submitted paper of ours proposes a way of selecting in a consistent manner thresholds for confidence, support, confidence width, and a variant of improvement ([3], [12]) that we call "blocking factor". Also, confidence width opens a door to a more human-centered development where one can find ways of evaluating this formal notion of novelty with respect to user-conceived naive notions of novelty. One potential development could be to design an interactive knowledge elicitation tool that, on the basis of the theory described here, could tune in, up to focusing on the user's intuitions for novelty, by showing a handful of rules of high width, asking the user to label them as novel or not novel: we should develop further the theory to take into account facts such as rules of high width (or support or confidence) being labeled as

not novel, so that the labeling would have consequences on the values of these parameters for the rest of the rules.

## References


1. C C Aggarwal, P S Yu: A new approach to online generation of association rules. IEEE Transactions on Knowledge and Data Engineering 13 (2001), 527–540. See also ICDE'98.
2. J L Balcázar: Redundancy, Deduction Schemes, and Minimum-Size Bases for Association Rules. Pascal Report 4259 (2008). (Prelim. version in ECMLPKDD 2008.)
3. R Bayardo, R Agrawal, D Gunopulos: Constraint-Based Rule Mining in Large, Dense Databases. ICDE 1999, 188–197.
4. T Calders, C Rigotti, J-F Boulicaut: A Survey on Condensed Representations for Frequent Sets. Constraint-Based Mining and Inductive Databases 2004, 64–80.
5. A Ceglar, J F Roddick: Association mining. ACM Computing Surveys 38 (2006).
6. A Freitas: Understanding the crucial differences between classification and discovery of association rules. SIGKDD Explorations 2 (2000), 65–69.
7. B Ganter, R Wille: *Formal Concept Analysis*. Springer 1999.
8. P Hajek, M Holena: Formal Logics of Discovery and Hypothesis Formation by Machine. Theoretical Computer Science 292 (2003), 345–357.
9. M Kryszkiewicz: Representative Association Rules. Pacific-Asia KDD Conference, PAKDD'98, LNCS 1394, 198–209.
10. M Kryszkiewicz: Fast discovery of representative association rules. RSCTC, 1998, 214–221.
11. M Kryszkiewicz: Closed Set Based Discovery of Representative Association Rules. IDA 2001, 350–359.
12. B Liu, W Hsu, Y Ma: Pruning and summarizing the discovered associations. KDD 1999, 125–134.
13. M Luxenburger: Implications partielles dans un contexte. Mathématiques et Sciences Humaines 29 (1991), 35–55.
14. S Ma, J L Hellerstein: Mining Mutually Dependent Patterns for System Management. IEEE Journal on Selected Areas in Communications 20 (2002), 726–734.
15. N Megiddo, R Srikant: Discovering Predictive Association Rules. KDD 1998, 274–278
16. E R Omiecinski: Alternative Interest Measures for Mining Associations in Databases. IEEE Trans. on Knowledge and Data Engineering 15 (2003), 57–69.
17. N Pasquier, R Taouil, Y Bastide, G Stumme, L Lakhal: Generating a condensed representation for association rules. Journal of Intelligent Information Systems 24 (2005), 29–60.
18. J L Pfaltz, C M Taylor: Scientific Discovery through Iterative Transformations of Concept Lattices. Workshop on Discrete Mathematics and Data Mining at SDM 2002, 65–74.
19. V Phan-Luong: The Representative Basis for Association Rules. ICDM 2001, 639–640.
20. M Wild: A theory of finite closure spaces based on implications. Advances in Mathematics 108 (1994), 118–139.
21. M Zaki: Mining non-redundant association rules. Data Mining and Knowledge Discovery 9 (2004), 223–248.
22. M Zaki, M Ogihara: Theoretical foundations of association rules. Workshop on research issues in DMKD (1998).

# DCR: Discretization using Class Information to Reduce Number of Intervals

Prachya Pongaksorn, Thanawin Rakthanmanon, and Kitsana Waiyamai

Data Analysis and Knowledge Discovery Laboratory (DAKDL),
Computer Engineering Dept, Faculty of Engineering, Kasetsart University, Bangkok, Thailand.
{g4865394, fengtwr, fengknw}@ku.ac.th

**Abstract.** Discretization techniques for data set features have received increasing research attention. Results using discretized features are usually more compact, shorter, and accurate than using continuous values. In this paper, an algorithm called Discretization using Class information to Reduce number of intervals (DCR) is proposed. DCR uses both class information and order between attributes to determine the discretization scheme with minimum number of intervals. Attribute discretization order is determined based on information gain of each attribute with respect to the class attribute. The number of intervals is reduced by deleting training data at each step of attribute discretization. Experiments are performed to compare the predictive accuracy and execution time of this algorithm with several well-known algorithms. Results show that discretized features generated by the DCR algorithm contain a smaller number of intervals than other supervised algorithms using less execution time, and the predictive accuracy is as high or higher.

**Key words:** discretization, continuous feature, data mining, classification

## 1 Introduction

Data mining is a powerful approach to extracting meaningful information from large and unwieldy databases. However, for efficiency, appropriate pre-processing of the input databases is needed. The majority of these databases usually come in a mixed format called "mixed-mode data" containing both discrete and continuous features, as shown in Table 1. In the table, feature2 is discrete while feature1, 3, and 4 are continuous. Some learning algorithms [4, 10, 12, 18] can handle only discrete-valued attributes, while some others can handle continuous attributes but still perform better with discrete-valued attributes [6, 11]. This drawback can be overcome by using a discretization algorithm as a pre-processing step for data mining.

Discrete values offer several advantages over continuous ones, such as data reduction and simplification. Quality discretization of continuous attributes is an important problem that has effects on speed, accuracy, and understandability of the classification models [20].

Although much research has been done in the area of discretization, many algorithms still do not take advantage of class information to increase their

P. Lenca and S. Lallich (Eds.): QIMIE/PAKDD 2009

discretization performance. Thus, the resulting discretization schemes do not provide much efficiency when used in the classification process, e.g. they contain more intervals than necessary. Unsupervised discretization algorithms that do not use class information include the equal-width [1] and equal-frequency methods [17] that divide continuous ranges into sub-ranges. Supervised algorithms, such as statistics-based [11, 15], entropy-based [21], and class-attributes interdependency-based algorithms [13] use class information; however, these algorithms do not make use of relations between attributes in the database.

**Table 1.** Data set contanning both decrete and continuous attributes

| ID | feature1 | feature2 | feature3 | feature4 | Class |
|---|---|---|---|---|---|
| 1 | 17 | Yes | 49 | 33 | Z |
| 2 | 19 | No | 48 | 21 | Y |
| 3 | 21 | No | 50 | 50 | Y |
| 4 | 21 | No | 53 | 19 | X |
| 5 | 22 | Yes | 65 | 49 | Y |
| 6 | 35 | Yes | 70 | 55 | Y |
| 7 | 33 | Yes | 89 | 76 | Z |
| 8 | 42 | No | 48 | 80 | Z |
| 9 | 40 | Yes | 63 | 33 | Y |
| 10 | 22 | Yes | 72 | 21 | X |
| 11 | 23 | Yes | 80 | 10 | X |
| 12 | 20 | Yes | 73 | 9 | X |
| 13 | 19 | No | 65 | 43 | Y |
| 14 | 25 | Yes | 90 | 95 | Z |
| 15 | 29 | Yes | 73 | 21 | Y |

There are algorithms [2, 5] that use both class information and relations between attributes in their discretization process, however, they are not computationally efficient. Also, discretizations with the same accuracy but with fewer number of intervals are preferable to those with large number of intervals since they cause less fragmentation of the data in the sub-nodes of decision trees [16]. The Discretization Using Class Information to Reduce Number of Intervals (DCR) algorithm presented here uses both class information and order between attributes to determine an efficient discretization scheme. The order of attribute discretization is determined based on information gain of each attribute with respect to the class attribute. The number of intervals is successively reduced by deleting training data at each step of discretization. DCR is able to find the minimum number of discrete intervals while maintaining the accuracy of the classifier. Experimental results show that the discretized features generated using DCR nearly achieve the smallest number of intervals for a given level of accuracy. Further, DCR uses less execution time than well-known supervised discretization techniques such as CAIM and ChiMerge.

In the next section, we present the class-based discretization process. In Section 3 and 4 we present the DCR discretization concepts and algorithm, resp. we discuss the results of comparative experiments in Section 5. Finally, Section 6 gives the conclusion and further work.

## 2 Class-based Discretization

In this section, we present class-based discretization process. First we describe the discretization process for each attribute; then we describe the information gain used for sorting the attributes to be discretized by the process that we purpose; the order of attribute in discretization makes the result different. Lastly, we describe the quanta-matrix [9] that shows the relation between class and discretization scheme.

### 2.1 Univariate discretization process

Discretization can be univariate or multivariate. Univariate discretization quantifies one continuous feature at a time while multivariate discretization simultaneously considers multiple features. We mainly consider univariate (typical) discretization in this paper. A typical discretization process broadly consists of four steps:

1. Sort the values of the attribute to be discretized.
2. Determine a cut-point for splitting or adjacent intervals for merging.
3. Split or merge intervals of continuous values, according to some criterion.
4. Stop at some point.

### 2.2 Information Gain

Information gain is used in C4.5 [7] to choose the best attribute (maximum information gain) for splitting the data, but this method can handle only discrete values. For continuous valued attributes, there is a need for a discretization algorithm that transforms continuous attributes into discrete ones. Using the data in Table 1 as an example, calculate information gain of features 1, 3, and 4 using the equal width discretization method, where the range of values of a feature is evenly divided into equi-width intervals, to discretize these attributes before calculating the information gain. If we discretize the features in Table 1 into five intervals, then the information gains of feature1, 3, and 4 are 0.78, 1.023, and 1.53, resp.

### 2.3 Class-attribute interdependency discretization

The main objective of this paper is to find the discretization scheme for each continuous attribute that contains the minimum number of intervals while minimizing the loss of class-attribute interdependency. The quanta-matrix plays a major role in achieving this purpose. Table 2 shows a quanta-matrix of feature $f$ of a given discretization schema $D$. In this quanta-matrix, training dataset consists of $M$

examples, where each example belonging to only one of the $S$ classes. If we discretize attribute $f_i$ -- for which $d_0$ is the minimum value and $d_n$ is the maximum value -- into $n$ intervals, then the discretization scheme of attribute $f_i$ is

$$D_i = \{[d_0,d_1], (d_1,d_2], \ldots , (d_{n-1},d_n]\}$$

Using discretization scheme $D_i$, each value of attribute $f_i$ can be classified into one of the $n$ intervals.

**Table 2.** Quanta-matrix of feature $f$ of a given discretization scheme $D$

| Class | Interval | | | | | Class Total |
|---|---|---|---|---|---|---|
| | $[d_0,d_1]$ | .. | $(d_{r-1},d_r]$ | .. | $(d_{n-1},d_{dn}]$ | |
| $C_1$ | $q_{11}$ | .. | $q_{1r}$ | .. | $q_{1n}$ | $q_{1+}$ |
| : | : | .. | : | .. | : | : |
| $C_i$ | $q_{i1}$ | .. | $q_{ir}$ | .. | $q_{in}$ | $q_{i+}$ |
| : | : | .. | : | .. | : | : |
| $C_S$ | $q_{S1}$ | .. | $q_{Sr}$ | .. | $q_{Sn}$ | $q_{S+}$ |
| Interval Total | $q_{+1}$ | .. | $q_{+r}$ | .. | $q_{+n}$ | $M$ |

In Table 2, $q_{ir}$ is the total number of continuous values belonging to the $i^{th}$ class that are in interval $(d_{r-1},d_r]$ ; $q_{i+}$ is the total number of objects belonging to the $i^{th}$ class, and $q_{+r}$ is the total number of continuous values of attribute $f_i$ that are within the interval $(d_{r-1},d_r]$, for $i = 1, 2, ..., S$ and $r = 1, 2, ..., n$. Table 3 shows a quanta-matrix of feature4 (taken from Table 1) with discretization schema $D4 = \{[9,20], (20,65.5], (65.5,95]\}$.

**Table 3.** Quanta-matrix of feature4 from Table 1 with the discretization schema $D4 = \{[9,20], (20,65.5], (65.5,95]\}$

| Class | Interval | | | Class Total |
|---|---|---|---|---|
| | $[9,20]$ | $(20,65.5]$ | $(65.5,95]$ | |
| *X* | *3* | *1* | *0* | *4* |
| *Y* | *0* | *7* | *0* | *7* |
| *Z* | *0* | *1* | *3* | *4* |
| Interval Total | *3* | *9* | *3* | *15* |

# 3 Discretization using Class Information to Reduce Number of Intervals

In this section, we describe the DCR supervised discretization algorithm whose objective is to maximize predictive accuracy while generating a (possibly) minimal number of discrete intervals. To maximize the predictive accuracy, DCR uses class intervals, reduces training transactions at each step of attribute discretization.

## 3.1 Using class information to find the best cut points

To evaluate the relation between the discretization scheme and class for each attribute, a criterion called *DCR* is defined as (using the notation as in Table 2):

$$DCR = \frac{\sum_{r=1}^{n}\left(\sum_{i=1}^{S} q_{ir}^2 \Big/ q_{+r}\right)}{n} \qquad (1)$$

The DCR value is the average of the distribution of class ( $\sum_{r=1}^{n}$ ) in each interval ( $\sum_{i=1}^{S} q_{ir}^2 \Big/ q_{+r}$ ). It has the following properties:

- The algorithm is able to find the discretization scheme where each interval has one major class, consider each interval $r$ in the quanta-matrix, $q_{ir}$ value is in range [*0, $q_{+r}$*]. Thus, an interval has the maximal one major class when $q_{ir}$ equals $q_{+r}$. Since *DCR* depends on $\sum_{i=1}^{S} q_{ir}^2$ , *it* achieves its maximum when each interval has all of its values grouped within a single class label.

**Table 4.** Distribution of class for each interval

| Class | | Interval | | Class Total |
|---|---|---|---|---|
| | .. | $(d_{r-1},d_r]$ | .. | |
| $C_1$ | .. | *0* | .. | |
| $C_2$ | .. | *5* | .. | |
| $C_3$ | .. | *10* | .. | |
| Total | | *15* | | |

*(a)*

| Class | | Interval | | Class Total |
|---|---|---|---|---|
| | .. | $(d_{r-1},d_r]$ | .. | |
| $C_1$ | .. | *2* | .. | |
| $C_2$ | .. | *3* | .. | |
| $C_3$ | .. | *10* | .. | |
| Total | | *15* | | |

*(b)*

- The $q_{ir}^2$ values are used to compute the distribution of classes in each interval. To find the discretization scheme when each interval has one major class, the discretization scheme in Table 4*(a)* might be better than the scheme in Table 4*(b)*. DCR is able to distinguish which of the two scenarios is better. Because the interval in 4*(a)* has a smaller distribution of classes and it may possibly use only

one attribute to classify non-majority class. But the interval in Table 4*(b)* must use at least two attributes for classification.

- The $q_{ir}^2$ values are divided by $q_{+r}$ for normalization. Numerical overflow errors can be avoided by calculating $q_{ir}^2 / q_{+r}$ as $(q_{ir} / q_{+r})\, q_{ir}$, so the maximum value is $q_{ir}$.

### 3.2 Deleting training transactions to generate the minimum number of discrete intervals

We use the relation between attributes to reduce the size of training data by removing transactions for which the (continuous) values are in the interval having all of its values grouped within a single class label. Thus, the next continuous attribute has only unclassified transactions to be discretized. This process can also reduce the execution time for classifying other attributes. Further, the size of the training data is reduced at each attribute discretization. Fig. 1(a) shows the discretization scheme {[*9,20*], (*20,27*], (*27,38*], (*38,65.5*], (*65.5,95*]} of feature4.

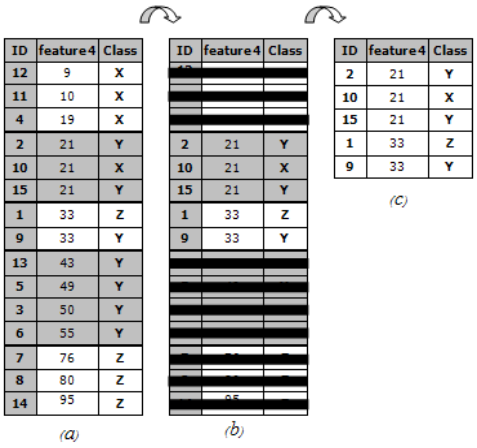



**Fig. 1.** At each attribute dicretization, trainning transactions are deleted to reduce the number of intervals

In Fig. 1, the class of continuous values that are in intervals [9,20], (38,65.5], and (65.5,95] are grouped within a single class label X, Y, and Z, resp, so these transactions are removed as they are classifiable transaction. Fig. 1(c) shows the remaining transactions that are used in next attribute to discretization.

In this method attributes are discretized one by one, and the order of attribute discretization may effect the final classifiers. Hence, the DCR algorithm uses the relation between attribute and class to compute the information gain value for each attribute with class and then discretize each attribute by its information gain value in descending order.

## 4 DCR Algorithm

The optimal discretization scheme can be found by searching over the space of all discretization schemes for the one with the highest *DCR* value; however, such a search is highly combinatorial and time consuming. Instead, the DCR algorithm uses a greedy approach, which searches for the approximate optimal value of the DCR criterion by finding locally maximum values of the criterion. Although this approach does not guarantee a global maximum, it is both computationally inexpensive and results in a near optimal discretization scheme, as shown in Section 5. The algorithm is composed of two principle steps:

1. Order the attribute to be discretized.
2. Discretize and reduce the size of current training data for each attribute

Pseudo code of the DCR algorithm is given in Fig. 2.

***Input:*** Training data set *DB* consisting of continuous attributes $F_i$, and class attribute *C* from a total of *s* classes

```
1.  for each Fi
2.    Ei = information_gain(Fi, C);
3.  arrange_order_desc(F, E);
4.  db = DB;
5.  for each fi          // feature Fi of current training data set db
6.    di0 = min(Fi);
7.    din = max(Fi);
8.    Di = {[di0, din]}
9.    if (db ≠ Ø) then
10.     k = 1;
11.     MaxDCR = 0;
12.     EB = essential_boundary_set(fi);
13.     Repeat
14.       DCR = compute_max_dcr_boundary(fi, Di, EB);
15.       If (DCR > MaxDCR) or (k < s) then
16.         Update Di with a boundary that has the highest DCR
17.         MaxDCR = DCR;
18.         k = k + 1;
19.     Until (DCR <= MaxDCR) and (k >= s)
20.   db = reducing_transaction(db, Di);
21.   D = D U {Di}
```

***Output:*** Set of all discretization scheme D

**Fig. 2.** Pseudocode of the DCR Algorithm

In steps 1-3 the algorithm orders attributes to be discretized based on the information gain value in descending order. Based on the information gain of attributes in section 2.2, the algorithm will discretize feature4, 3, and 1, resp., in step 5.

In the discretization process (steps 6-19), DCR starts with a single interval that covers all possible values of continuous attribute $F_i$ and divides it interactively. In step 12, form a set of all distinct values of $f_i$ in ascending order, and initialize all

possible interval boundaries B with all the midpoints of all the adjacent pairs in the set, denoted by B = $\{a_0, \ldots, a_m\}$. If the instances that fall into the intervals $(a_{i-1}, a_i]$ and $(a_i, a_{i+1}]$ belong to the same class, remove $a_i$ from set B until there are instances that fall into two adjacent intervals but do not belong to the same class. This results in an essential boundary set EB = $\{b_0, \ldots, b_n\}$, where $n < m$. For example, in Fig. 3 point 9.5 is the midpoint between transactions 11 and 12 both belonging to class X, so the point 9.5 is not included in set EB. The value 27 is the midpoint between feature value 21 (transactions 2, 10, 15) and feature value 33 (transactions 1 and 9) belonging to different class labels, hence 27 is added to set EB. Finally, the EB set for feature4 is {*20, 27, 38, 65.5*}.

| ID | feature4 | Class | |
|---|---|---|---|
| 12 | 9 | X | |
| 11 | 10 | X | 9.5 ✖ |
| 4 | 19 | X | 14.5 ✖ |
| 2 | 21 | Y | 20 |
| 10 | 21 | X | |
| 15 | 21 | Y | 27 |
| 1 | 33 | Z | |
| 9 | 33 | Y | 38 |
| 13 | 43 | Y | 46 ✖ |
| 5 | 49 | Y | 49.5 ✖ |
| 3 | 50 | Y | 52.5 ✖ |
| 6 | 55 | Y | 65.5 |
| 7 | 76 | Z | 78 ✖ |
| 8 | 80 | Z | 87.5 ✖ |
| 14 | 95 | Z | |

**Fig. 3.** Finding essential interval boundaries (EB) of feature4 in Table 1

From all possible division points that are tried (with replacement) in step 14, the algorithm chooses the division boundary that gives the highest value of the DCR criterion. For example, in finding the division points of feature4 the initial discretization scheme $D_4$ is {[*9, 95*]} and the set of essential interval boundaries EB is {*20, 27, 38, 65.5*}*;* as shown in Fig. 3, the algorithm adds an inner boundary value that is not already in $D_4$, from EB, and calculates the corresponding *DCR* value. The algorithm accepts the boundary value with the highest value of *DCR*, e.g., for the first element of EB, point 20, the new discretization scheme $D_4$ is {[*9,20*], (*20,95*]} and the data in the quanta-matrix are as in Fig. 4. Thus, the *DCR* value of this discretization scheme is 4.25.

| Class | Interval | | Class Total |
|---|---|---|---|
| | [9, 20] | (20, 95] | |
| X | 3 | 1 | 4 |
| Y | 0 | 7 | 7 |
| Z | 0 | 4 | 4 |
| Interval Total | 3 | 12 | 15 |

$$DCR = \frac{\lceil (3^2 + 0^2 + 0^2)/3 \rceil + \lceil (1^2 + 7^2 + 4^2)/12 \rceil}{2}$$

**Fig. 4.** The calculation of DCR value for feature4 in Table 1 where $D_4$ is {[*9,20*], (*20,95*]}

For boundary points 27, 38, and 65.5, the corresponding *DCR* values are 3.944, 3.41, and 4.25 resp. After all tentative additions have been tried, the point with the

highest *DCR* value (20 in this example) is added to $D_i$ in step 16. The algorithm assumes that every discretized attribute needs a number of intervals at least equal to the number of classes or that the *DCR* value shows improvement at each iteration, assuring that the discretized attribute can improve subsequent classification. Thus, the discretization scheme of feature4 is {[*9,20*], (*20,27*], (*27,38*], (*38,65.5*], (*65.5,95*]} as in Fig. 1(a). Step 20 creates a new training data set *db* by remove classifiable intervals as in Fig. 1(c).

# 5 Experimental Results

## 5.1 Experimental Set-up

The DCR algorithm is compared with five state of the art discretization algorithms including two unsupervised algorithms and three supervised algorithms. The unsupervised algorithms are equal width (EW) [1] and equal frequency (EF) [17]; supervised algorithms are the CAIM [13] splitting-based discretization, ChiMerge [11] merging-based discretization, and a discretization algorithm in the WEKA open-source data mining library.

Data for the experiments consist of six well-known continuous and mixed-mode data sets from the UCI repository of Machine Leaning Database [3]: Iris dataset (iris), Ionosphere dataset (ion), New-Thyroid dataset (thy), SatImage dataset (sat), Waveform dataset (wav), and Heart Disease dataset (hea). Properties of the data sets are listed in Table 5.

The unsupervised algorithms require the user to specify the number of discrete intervals. In the experiments, we set the number of intervals to be close to the number obtained with the DCR algorithm for purpose of comparison.

**Table 5.** Properties of data sets used in experiments.

| Properties | Datasets | | | | | |
|---|---|---|---|---|---|---|
| | iris | ion | thy | sat | wav | hea |
| Number of classes | 3 | 2 | 3 | 6 | 3 | 2 |
| Number of examples | 150 | 351 | 215 | 6435 | 3600 | 270 |
| Number of attributes | 4 | 34 | 5 | 36 | 21 | 13 |
| Number of continuous attributes | 4 | 32 | 5 | 36 | 21 | 6 |

## 5.2 Analysis of Results

In the experiments, we evaluated the results in terms of number of intervals, execution time, and accuracy of rules generated by the C5.0 algorithm.

### 5.2.1 Number of intervals

**Table 6.** Number of intervals for each discretization method.

| Discretization Method | Datasets | | | | | | Rank mean |
|---|---|---|---|---|---|---|---|
| | iris | ion | thy | sat | wav | hea | |
| EW | 12 | **64** | 15 | 180 | 63 | **12** | 2.0 |
| EF | 12 | **64** | 15 | 180 | 63 | **12** | 2.0 |
| CAIM | 12 | **64** | 15 | 216 | 63 | **12** | 2.3 |
| ChiMerge | 15 | 398 | 28 | 752 | 801 | 33 | 6.0 |
| WEKA | 10 | 117 | **14** | 475 | 81 | 13 | 3.8 |
| DCR | **9** | **64** | **14** | **154** | **62** | **12** | **1.0** |

Table 6 shows that the DCR algorithm generated discretization scheme with the smallest number of intervals for all data sets. A smaller number of discrete intervals reduces the size of the data and helps to better understand the meaning of discretized attributes. This is a major advantage of the DCR algorithm.

### 5.2.2 Discretization execution time

A comparison of the discretization times is given in Table 7. We implemented all discretization algorithms in the same programming language, except the WEKA algorithm and Built-in C5.0. Thus, they were not included in the comparison.

**Table 7.** Discretization execution time.

| Discretization Method | Datasets | | | | | | Rank mean |
|---|---|---|---|---|---|---|---|
| | iris | ion | thy | sat | wav | hea | |
| EW | 0.110 | **3.786** | 0.231 | 1233.254 | 381.999 | **0.300** | 1.8 |
| EF | **0.090** | 3.806 | **0.220** | 1198.744 | **337.575** | 0.320 | **1.5** |
| CAIM | 2.004 | 77.862 | 4.740 | 2140.000 | 1260.000 | 13.489 | 4.3 |
| ChiMerge | 8.362 | 2399.089 | 45.375 | **913.433** | 517.164 | 39.817 | 4.0 |
| DCR | 0.631 | 14.962 | 1.752 | 1477.004 | 864.213 | 3.305 | 3.3 |

The comparison of execution times shows that the unsupervised discretization algorithms exhibit the shortest execution times; this is to be expected since they do not process any class-related information. Among the supervised algorithms, DCR exhibited the smallest execution time for four out of six data sets, but the second highest execution time (after ChiMerge) for *sat* and *wav*. Still, based on average rank, DCR ranked fastest among the supervised algorithms.

### 5.2.3 Accuracy comparison

The discretized data sets generated in Section 5.2.1, were used as input to C5.0 algorithms to generate classification rules. The accuracy of the resulting classification rules were compared. Since C5.0 can generate data models from continuous attributes, we compared its performance using generated rules from raw data against the results achieved using discretized data produced by the six algorithms. A 10-fold cross-validation test was performed using all data sets: each data set was divided into 10 parts of which nine parts were used as training data and the remaining one part as test data. The experiments were performed for all 10 choices of the test data. The final predictive accuracy was taken as the average of the 10 predictive accuracy values.

**Table 8.** Comparison of the accuracies achieved by the C5.0 algorithm for six data sets using the seven discretization schemes.

| Discretization Method | Datasets | | | | | | Rank mean |
|---|---|---|---|---|---|---|---|
| | iris | ion | thy | sat | wav | hea | |
| EW | 97.3330 | 90.0285 | 86.0465 | 85.9518 | 74.6667 | 75.5556 | 5.0 |
| EF | 94.6667 | 81.7664 | 89.7674 | 85.2681 | 76.5000 | 80.0000 | 4.8 |
| CAIM | 94.0000 | 91.4530 | 94.8837 | 85.8430 | 77.0000 | 77.4070 | 4.2 |
| ChiMerge | **97.3333** | 92.0228 | 93.0233 | 83.3877 | 71.6111 | 76.2963 | 4.7 |
| Built-in C5.0 | 95.3020 | 90.8571 | 91.5888 | 85.9341 | 75.8544 | 78.8104 | 4.2 |
| WEKA | 93.2886 | 94.0000 | 94.3925 | **87.5971** | 77.6605 | **81.4126** | 2.7 |
| DCR | 94.6667 | **94.3020** | **96.2791** | 85.9518 | **78.7778** | 78.5185 | **2.2** |

The DCR algorithm exhibited the highest accuracy for three of the six data sets; WEKA was most accurate for two datasets and nearly as accurate as DCR for three other data sets.

## 6 Conclusions and Future work

Experimental results comparing several discretization algorithms using standard data sets indicate that the DCR algorithm performs discretization with fewer intervals and overall lower run time while still providing classifiers with high predictive accuracy. On average it had the fastest run-time of all supervised algorithms. The resulting discretized data and classifiers were not only more compact, but resulted in high predictive accuracy for all six experimental data sets. The WEKA algorithm also showed high predictive accuracy, but required more discretization intervals.

In the future work, we will focus on increasing the efficiency of discretization in the context of mixed-mode data. Another interesting research direction is to investigate other measures of interestingness [14, 19] as a way of optimizing the attribute discretization order.

**Acknowledgment.** Thanks to J. E. Brucker for reading and comments of this paper.

## References


1. A.K.C. Wong, D.K.Y. Chiu: Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data. IEEE Trans. Pattern Analysis and Machine Intelligence, pp. 796--805 (1987)
2. Chan, C.-C., Batur, C., Srinivasan, A.: Determination of Quantization Intervals in Rule Based Model for Dynamic. In Proceedings of the IEEE Conference on Systems, Man, and Cybernetics. Charlottesvile, Virginia, pp. 1719--1723 (1991)
3. C.L. Blake, C.J. Merz: UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.html
4. Cost, S., Salzberg, S.: A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. Machine Learning 10(1), pp. 57--78 (1993)
5. Ho, K.M., Scott, P.D.: Zeta: A Global Method for Discretization of Continuous Variables. In KDD97: 3rd International Conference of Knowledge Discovery and Data Mining. Newport Beach, CA, pp. 191--194 (1997)
6. J. Catlett: On Changing Continuous Attributes into Ordered Discrete Attributes. Proc. European Working Session on Learning, pp. 164--178 (1991)
7. J. R. Quinlan: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
8. J. R. Quinlan: Induction of Decision Trees. Machine learning, vol.1**,** pp. 81--106 (1986)
9. J.Y. Ching, A.K.C. Wong, K.C.C. Chan: Class-Dependent Discretization for Inductive Learning from Continuous and Mixed Mode Data. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 7, pp. 641--651 (1995)
10. K.A. Kaufman, R.S. Michalski: Learning from Inconsistent and Noisy Data: The AQ18 Approach. Proc. 11th Int'l Symp. Methodologies for Intelligent Systems (1999)
11. Kerber, R.: Chimerge: Discretization of Numeric Attributes. In Proc. AAAI-92, Ninth National Confrerence Articial Intelligence. AAAI Press/The MIT Press, pp. 123–128 (1992)
12. K.J. Cios, L. Kurgan: Hybrid Inductive Machine Learning: An Overview of CLIP Algorithms. New Learning Paradigms in Soft Computing, L.C. Jain and J. Kacprzyk, ed., pp. 276--322 (2001)
13. L. A. Kurgan, K. J. Cios: CAIM Discretization Algorithm. IEEE Trans. Knowledge And Data Engineering, February vol. 16, no.2, pp. 145--153 (2004)
14. Lenca P., Lallich S., Do T.-N., Pham N.-K.: A Comparison of Different Off-Centered Entropies to Deal with Class Imbalance for Decision Trees. Pacific-Asia Conference on Knowledge Discovery and Data Mining (eds. Washio T., Suzuki E., Ting K. M., Inokuchi A.), Lecture Notes in Computer Science, 5012, Springer, Osaka, Japan, pp. 634--643 (2008)
15. Liu, H., Setiono, R.: Chi2: Feature Selection and Discretization of Numeric Attributes. IEEE Computer Society, November 5–8, pp. 388--391 (1995)
16. M. Boulle: MODL: A Bayes Optimal Discretization Method for Continuous Attributes. Machine Learning, vol. 65, No. 1. pp. 131--165, (2006)
17. Nguyen, H.S., Nguyen, S.H.: Discretization Methods in Data Mining. In: Polkowski, L., Skowron, A. (Eds.), Rough Sets in Knowledge Discovery. Physica, pp. 451--482 (1998)
18. P. Clark, T. Niblett: The CN2 Algorithm. Machine Learning, vol. 3, pp. 261--283 (1989)
19. Pham N.-K., Do T.-N., Lenca P., Lallich S.: Using Local Node Information in Decision Trees: Coupling a Local Decision Rule with an Off-Centered Entropy. The International Conference on Data Mining (eds. Stahlbock R., Crone S. F., Lessmann, S.), CSREA Press, Las Vegas, Nevada, USA, pp. 117--123, July 14-17 (2008)
20. T. Qureshi, D. A. Zighed: Discretization of Continuous Features by Resampling. Proc. 8emes Journees fancophones Extraction et Gestion des Connaissances (2008)
21. U.M. Fayyad, K.B. Irani: On the Handling of Continuous-Valued Attributes in Decision Tree Generation. Machine Learning, vol. 8, pp. 87--102 (1982)

# A framework for monitoring classifiers performance: when and why failure occurs

Nitesh V. Chawla

University of Notre Dame, USA

**Abstract.** Classifier error is the product of model bias and data variance. While understanding the bias involved when selecting a given learning algorithm, it is similarly important to understand the variability in data over time, since even the One True Model might perform poorly when training and evaluation samples diverge. Thus, the ability to identify distributional divergence is critical towards pinpointing when fracture points in classifier performance will occur. Contemporary evaluation methods do not take the impact of distribution shifts on the quality of classifiers predictions. In this talk, I present a comprehensive framework to proactively detect breakpoints in classifiers predictions and shifts in data distributions through a series of statistical tests. I outline and utilize three scenarios under which data changes: sample selection bias, covariate shift, and shifting class priors.

P. Lenca and S. Lallich (Eds.): QIMIE/PAKDD 2009

# An Assertive Will for Seeing and Believing Introducing a Feature Cardinality Driven Distance Measure to Uninformative Distributions

Joan Garriga

Departament de Sistemes i Llenguatges Informàtics
Universitat Politècnica de Catalunya
jgarriga@lsi.upc.edu

**Abstract.** What regard should a learning algorithm hold for the different information traces found in a sample? Answering this question objectively is not easy. Moreover, given that a full range of traits can be found in a human learning analogy, from the most daring or ingenuous, to the most conservative or incredulous. But in AI domains it is a must to clearly state the right will for believing what is seen when mining data bases. A key concept in this matter is assertiveness. The aim of this work is to ponder an approach to assertive KDDB, based on a feature cardinality driven distance measure to uninformative distributions. From this perspective, we present an alternative option to the support-confidence framework. The biases of this measure have not yet been thoroughly studied but the measure itself has proved to be quite effective as a heuristic when searching to optimize a sample in a simultaneous multi-interval discretization of continuous features. The empirical results show that the most relevant association or classification rules are revealed. Also, optimal cardinalities and optimal subsets of parents are found for any feature, according to a natural bias toward the MDL principle. As a conclusion, it appears the measure assertively captures knowledge. This may be useful for other data mining issues.

## 1 Introduction

It is nothing new to point out that some kind of a disappointing shadow of confusion hovers over the data mining scene. The flurry of different measures as well as the comprehensive literature on selecting the right ones for each task at hand ([4],[6]) is no more than a symptom.

In my opinion, three basic objections are the culprit: (i) the stochastic essence of any sample is somewhat misunderstood, (ii) some subtleties about what knowledge is or, more precisely, what better knowledge is, are somewhat set aside, and (iii) the will for believing what is seen is not clearly stated.

These objections are further exposed in the next section as well as throughout this paper. They form the basis for introducing an alternative approach to

P. Lenca and S. Lallich (Eds.): QIMIE/PAKDD 2009

knowledge discovery wherein a new measure is suggested. The aim is to present this approach as an open door to further research while the expression given for the measure is yet to be considered an open question.

A thorough analysis on the properties [2] and biases [3] of this measure , as well as some examples should be presented, but unfortunately, space is limited.

In order to state a general framework addressing *association* and *classification rules*, as well as *feature subset selection*, *clustering* and *graphical modelling* issues we will use the following general terminology. Let's consider a domain or concept characterized by a set of $m$ multinomial features $X = \left\{X^1, X^2, \ldots, X^m\right\}$ and a set $\{D\}$ of $N$ examples over these features. Let's consider two any features of this domain and denote $X^p = \{x_1^p, x_2^p, \ldots, x_r^p\}$ and $X^q = \{x_1^q, x_2^q, \ldots, x_s^q\}$ as the set of possible outcomes of features $X^p$ and $X^q$ with cardinalities $crd\left(X^p\right) = r$ and $crd\left(X^q\right) = s$, respectively. Also, for any pair $\left(x_i^p, x_j^q\right)$ we denote $n_i^p, n_j^q$ and $n_{ij}^{pq}$ as the marginal and joint frequencies given in $\{D\}$ .

Additionally and for the purpose of clarity, we state three levels of relationship: (i) we refer to a *rule* whenever we are considering a relation like $x_i^p \rightarrow x_j^q$, (ii) we refer to a *subpattern* whenever we are considering the set of rules included in the relation $x_i^p \rightarrow X^q$, and (iii) we refer to a *pattern* whenever we are considering the whole set of rules included in the relation $X^p \rightarrow X^q$. These designations will hold, unless explicitly noted, independently of our intention when considering the relationships (association, classification or whatever).

## 2 Some Objections to Objective Measures

The most important group of objective measures is based on probability. Given a rule $x_i^p \rightarrow x_j^q$, *coverage* is given as the marginal probabilities of antecedent $P\left(x_i^p\right)$ and consequent $P\left(x_j^q\right)$ of the rule, *support* is given by the joint probability $P\left(x_i^p, x_j^q\right)$, and *confidence* is given by the conditional probability $P\left(x_i^p \left| x_j^q\right.\right)$. Down from here, all objective measures of interestingness combine in different ways these or directly related factors, taken from raw data.

Let's consider the simple example of a transaction data set given in Tab.1. [1]

| Milk | Bread | Eggs |
|---|---|---|
| 1 | 0 | 1 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 0 | 0 | 1 |

**Table 1.** Transaction Dataset

[1] This example is extracted from [2].

*Coverage* for Milk is 4/5 and hence *coverage* for NoMilk is only 1/5. Does it make any sense to consider a rule like Milk → Bread when there is no comparable evidence in the dataset for the rule NoMilk → Bread?

Some of the defined measures try to take this fact into account, introducing factors with the probabilities for counter facts in some way. But that is not the question. The real question is whether there is some evidence missing in the dataset in order to adeptly measure the significance of that possible rule. This topic is not new, and has two loose ends:

1. Due to its stochastic nature, any sample should be considered as being less than 100% reliable. Therefore, whenever we consider evidential support from raw data, the estimates we make are afected by the subjective consideration of the sample as being 100% reliable, even though they are estimates. In other words, would it be fair to always estimate a 0/100% of probability for a rule with a 0/100% of *support*?
2. On the other hand, a rule should always be considered, at least, within the framework of its subpattern [5]. If a dependence relationship between two features do exist, this dependence should be patent for the whole pattern. From this point of view, it is important to distinguish between *structural evidence* and *parametrical evidence*. The former relates to the pattern or subpattern levels and expresses whether a remarkable relationship may exist. The latter refers to each one of the rules in the pattern and expresses how this relationship acts whenever it exists.

Let's think again about the transaction example of bread, milk and eggs. For an association rule like Milk → Bread, we have a *support* of 3/5 and a *confidence* of 3/4 and for an association rule like (Milk, Bread) → Eggs we have a *support* of 2/5 and a *confidence* of 2/3. While the combination (Milk, Bread) has a total of four possible outcomes, the combination (Milk, Bread, Eggs) offers as much as eight possible outcomes, therefore with a much lower prior probability. Should we really believe that the former is better supported than the latter? Should we consider these levels of *confidence* from an absolute perspective? In other words, is the same kind, quantity/quality, of knowledge given by these two rules?

In this case, the argument is quite subtle and it has to do with the level of certainty/uncertainty associated with a feature as a function of its cardinality or what is also referred to as the quantity/quality of knowledge given by that feature. The larger the cardinality of the features involved in a rule, the more accurate and valuable is the information, but the lesser the prior probability of finding that rule in the dataset.

These topics have been somewhat overlooked, and this new approach tries to offer a way to address this omission.

## 3 Assertiveness by means of Objectivity

One really assertive measure should be defined by assuring an impartial comparison within any rule's evidence detected in the sample. Recalling the objections

raised above, three conditions should be met for this assumption to be true: (i) the sample should be 100% reliable and equilibrated or otherwise this should be taken into account in some way, (ii) the quantity/quality of knowledge expressed by the rule should be taken into account in some way, and (iii) the fairest balance between seeing and believing should be guaranteed.

In order to define such and impartial measure we state the following three concepts:

**Definition 1.** *A feature* $X^p \in X$*, with* $crd\left(X^p\right) = r$*, is in* perfect marginal distribution *(pmd) whenever all its possible outcomes are equally covered, that is,* $\forall\, x_i^p \in X^p$ *all marginal frequencies are* $n_i^p = N/r$

Ideally, if all features in a sample were in *pmd*, all rule's prior probability would be maximally equilibrated.

**Definition 2.** *Two features* $\left(X^p, X^q\right) \in X$*, with* $crd\left(X^q\right) = s$*, are in* absolutely incoherent conditional distribution *(aicd) whenever* $\forall \left(x_i^p, x_j^q\right) \in \left(X^p, X^q\right)$ *all joint frequencies are* $n_{ij}^{pq} = n_i^p/s$

Again, this is an ideal situation, possible only between features with equal cardinality, but clearly conveys a state of minimum information.

**Definition 3.** *The* knowledge factor $Q$*, which is only briefly introduced here, is defined as the degree of accuracy associated to a feature* $X^q$ *as a function of its cardinality,* $crd\left(X^q\right) = s$*, given by,*

$$Q = \left(s-1\right)/s \tag{1}$$

On one hand, independently from any sample or domain, *pmd* and *aicd* state two clearly defined uninformative distributions to take distances from:

1. for feature $X^q$, an expression of its marginal distribution distance to the *pmd* is given by,

$$\Delta\left(X^q\right) = \sum_j \left( \frac{n_j^q - \frac{N}{s}}{\frac{N}{s}} \right)^2 = \sum_j \left( s\,\frac{n_j^q}{N} - 1 \right)^2 \quad . \tag{2}$$

2. respect to feature $X^p$, an expression of $X^q$'s conditional distribution distance to the *aicd* is given by,

$$\Delta\left(X^q \,|X^p\right) = \sum_{i,j} \left( \frac{n_{ij}^{pq} - \frac{n_i^p}{s}}{\frac{n_i^p}{s}} \right)^2 = \sum_{i,j} \left( s\,\frac{n_{ij}^{pq}}{n_i^p} - 1 \right)^2 \quad . \tag{3}$$

What should be kept in mind, is that expressions (2) and (3) are measuring exactly the same concept.[2]

[2] Strictly speaking, this expressions don't hold the formal properties of a metric distance functional (particularly, the triangular inequality does not make sense). They should rather be regarded as deviations.I hope this is not going to be misleading.

On the other hand, raw distances given in (2) and (3) are clearly affected by a strong bias due to the cardinality of the features.

The philosophy behind this approach is that, taking the *knowledge factor* as a base expressing the quantity/quality of knowledge, a transformation can be applied in order to address this bias. The main contribution of this work is to present a general expression for this transformation, wherein alternative and significantly different measures to *coverage*, *support* and *confidence*, can be derived. These new measures intend to be as objective as possible and intend to state the most assertive will to believe what is seen. The final purpose is to allow an objective comparison between any trace of rule/pattern found, regardless of the actual reliability of the sample and regardless of the cardinality of the features involved, addressing the objections formerly exposed.

## 4 Defining the Measure

A useful transformation of such distances is given by the general function,

$$Z(x) = exp\left(\alpha \frac{ln(Q)}{Q^2}(s\,x-1)^2\right) \quad , \tag{4}$$

where $x$ can either refer to the marginal or conditional distribution, $n_j^q/N$ or $n_{ij}^{pq}/n_i^p$, whatever be the case.[3]

Aiming at simplicity, this expression can be rewritten as,

$$Z(x) = bxp\left(\frac{1}{Q^2}(s\,x-1)^2\right) \quad , \tag{5}$$

where *bxp (knowledge factor exponential base)* is a self allowed notation, derived from *exp (natural exponential base)* with analogous meaning, that is, $bxp(K) \equiv Q^K$.

The proximity of this function to a *normal* distribution, $N\left(\frac{1}{s}, \frac{Q}{s}\sqrt{\frac{-1}{2\,ln(Q)}}\right)$ is clear, with two obvious differences which are: (i) it is not a probability distribution, but a distance distribution, so not normalized as a *mass function*, and (ii) it makes sense only in the range $0 \leq x \leq 1$.

Therefore I call this function the *QNormal distance distribution*, $QN\left(\frac{1}{s}, \frac{Q}{s}\right)$, which is depicted in Fig.1 for different values of $s$.

At the mean, given by $1/s$, its value is 1, and at the boundaries the values are given by,

$$Z_z \equiv Z(0) = bxp\left(\frac{1}{Q^2}\right) \quad ; \quad Z_n \equiv Z(1) = bxp\left(s^2\right) \quad . \tag{6}$$

[3] The factor $\alpha$ has to do with the prior credibility we can give to the sample. Its thorough treatment lies beyond the scope of this work, so let's consider $\alpha = 1$.

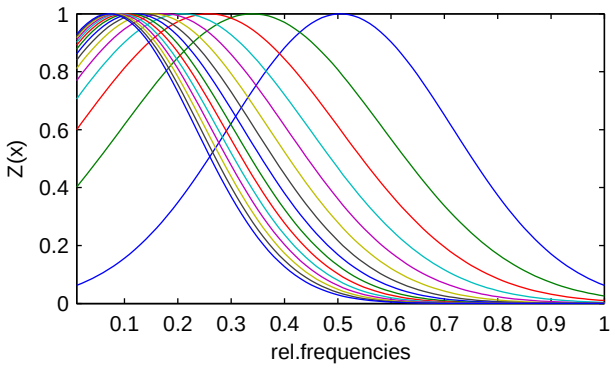



**Fig. 1.** $QN\left(\frac{1}{s},\frac{Q}{s}\right)$ for $2 \leq s \leq 15$ .

### 4.1 Presence

Applying the general expression given in (5) to the marginal distribution of feature $X^q$, we have,

$$\forall x_j^q \in X^q\,, \quad z_j^q \equiv Z\left(\frac{n_j^q}{N}\right) = bxp\left(\frac{1}{Q^2}\left(s\,\frac{n_j^q}{N}-1\right)^2\right) \quad , \tag{7}$$

Combining (7) with (6) in order to fit values into $(0,1)$, we can derive an alternative and significantly different measure of *coverage*, which I call *presence*, given by,

$$b_j^q = \frac{1}{s}\left(\frac{z_j^q - Z_z}{1-Z_z}\right) \;;\; 0 \leq \frac{n_j^q}{N} \leq \frac{1}{s} \quad , \tag{8}$$

$$b_j^q = \frac{1}{s}\left(\frac{z_j^q - Z_n}{1-Z_n}\right) \;;\; \frac{1}{s} \leq \frac{n_j^q}{N} \leq 1 \quad , \tag{9}$$

This function is depicted in Fig.2. The total *presence* of a feature is then given by $B^q = \sum_j \left(b_j^q\right)$ , with a maximum value of 1, given when all possible outcomes for the feature are equally covered. As long as *coverage* of that feature moves away from the *pmd* in any direction, the value of *presence* decreases, vanishing at the boundaries.

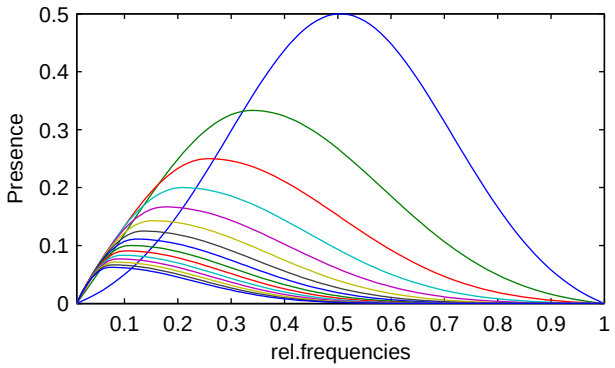



**Fig. 2.** Presence function for $2 \leq s \leq 15$.

### 4.2 Coherence

Applying the general expression given in (5) to the conditional distribution $(X^q \mid X^p)$, we have,

$$\forall \left(x_i^p, x_j^q\right) \in (X^p, X^q) \ , \quad z_{ij}^{pq} \equiv Z\left(\frac{n_{ij}^{pq}}{n_i^p}\right) = bxp\left(\frac{1}{Q^2}\left(s\,\frac{n_{ij}^{pq}}{n_i^p} - 1\right)^2\right) \ , \quad (10)$$

Combining (10) with (6) in order to fit values into $(0, 1)$, we can derive an alternative and significantly different measure of *confidence*, which I call *coherence* given by,

$$c_{ij}^{pq} = \frac{1}{r\,s}\left(1 - \frac{z_{ij}^{pq} - Z_z}{1 - Z_z}\right) \ ; \ 0 \leq \frac{n_{ij}^{pq}}{n_i^p} \leq \frac{1}{s} \ , \quad (11)$$

$$c_{ij}^{pq} = \frac{1}{r\,s}\left(1 - \frac{z_{ij}^{pq} - Z_n}{1 - Z_n}\right) \ ; \ \frac{1}{s} \leq \frac{n_{ij}^{pq}}{n_i^p} \leq 1 \ , \quad (12)$$

This function is depicted in Fig.3. The total *coherence* of pattern $X^p \rightarrow X^q$ is then given by $C^{pq} = \sum_{i,j}\left(c_{ij}^{pq}\right)$ , with a maximum value of 1, given when each subpattern is maximally coherent, as it is stated in the following definition.

**Definition 4.** *The conditional distribution* $(X^q \mid X^p)$ *is* maximally coherent *when* $\forall\, x_i^p \in X^p\,, \exists\, x_m^q \in X^q$ *, such that,* $n_{im}^{pq} = n_i^p$ *and* $\forall\, x_{j\neq m}^q \in X^q\,, n_{ij}^{pq} = 0$ .

And being both conditions necessary for the maximum *coherence*, they are both assigned the same value of *coherence* $1/\left(r\,s\right)$ .

Obviously, it is an asymmetric measure, so that most of the time it will be $c_{ij}^{pq} \neq c_{ji}^{qp}$ .

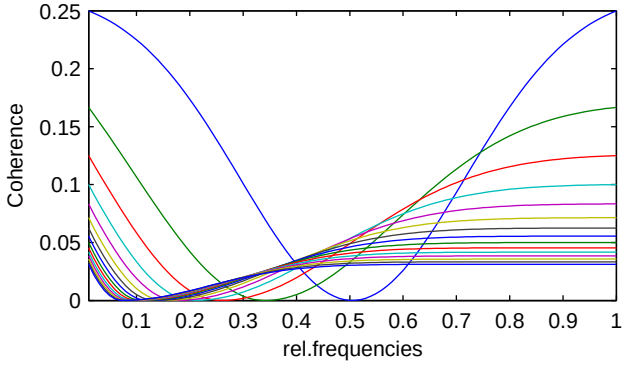



**Fig. 3.** Coherence function with $r = 2$ and for $2 \leq s \leq 15$.

### 4.3 Utility

Finally, combining the two former measures, we obtain the *utility* measure for the rule $x_i^p \rightarrow x_j^q$ , which is given by,

$$u_{ij}^{pq} = c_{ij}^{pq} \, \left(b_i^p \, r\right) \, \left(b_j^q \, s\right) \quad , \tag{13}$$

The total *utility* of pattern $X^p \rightarrow X^q$ is then given by $U^{pq} = \sum_{i,j} u_{ij}^{pq}$ , with a maximum value of 1, given when *coherence* is maximal and *presence* for both features is perfectly equilibrated.

A depiction example of the *utility* function for $x_i^p \rightarrow x_j^q$ with $(r = 2, s = 3)$ and being $X^p$ in *pmd* is given in Fig.4.

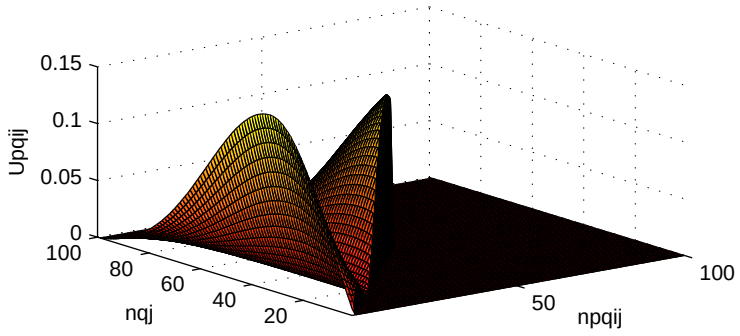



**Fig. 4.** Utility function for $r = 2$, $s = 3$, $N = 100$, $n_i^p = \frac{N}{r}$

By definition, *utility* is inversely related to the *total amount of uncertainty of the consequent given that the antecedent is known*, (see [1] for a related discussion). Even in the case of *independence*, $U^{p\perp q} \geq 0$, being zero only when $X^q$ is in *pmd*. This expresses the idea that even being independent it is still possible to get some certainty about the consequent, though coming from its own marginal distribution. In such a case, there exists a subspace in the set of all possible joint distributions, in the neighbourhood of *independence*, in which $U^{pq} \leq U^{p\perp q}$. This suggests the daring idea of expanding the concept of independence: it is not the single point where $P\left(X^p, X^q\right) = P\left(X^p\right)\, P\left(X^q\right)$ but the whole subset of joint distributions for which $U^{pq} \leq U^{p\perp q}$, that is, where the total uncertainty is even greater than that given in *independence*.

### 4.4 Parametrical Perspective

Finally, the *QNormal distance distribution* holds yet another possible derivation from the parametrical point of view, which clearly explains what it is conceptually being done.

From the inversion of the second half of the curve, we can derive the following expression,

$$\Theta\left(x\right) = \frac{Z\left(x\right)}{2}\,, \quad 0 \leq x \leq \frac{1}{s} \tag{14}$$

$$\Theta\left(x\right) = \left(1 - \frac{Z\left(x\right)}{2}\right)\,, \quad \frac{1}{s} \leq x \leq 1\,. \tag{15}$$

The depiction of this function is given in Fig.5.

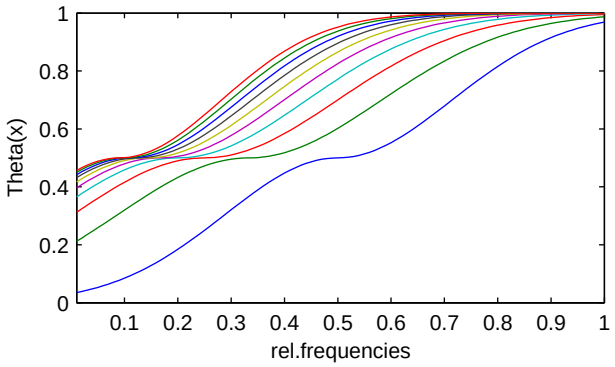



**Fig. 5.** Theta function for $2 \leq s \leq 15$.

In contrast to the raw interpretation resulting from measures like *coverage*, *support* and *confidence*, this function translates the parameters to a common

space where all of them can be seen relatively to the quantity/quality of knowledge they express.

There's a saddle point at the frequency given by $1/s$, which represents the equilibrium corresponding to the state of minimum information (*pmd* or *aicd*), and moving away from that point this equilibrium is consequently and gradually broken in one or other direction.

The breaking gradient is determined by the $\sigma$ parameter (depicted in Fig.6), given as,

$$\sigma = \frac{Q}{s} = \frac{(s-1)}{s}\,\frac{1}{s} \quad . \tag{16}$$

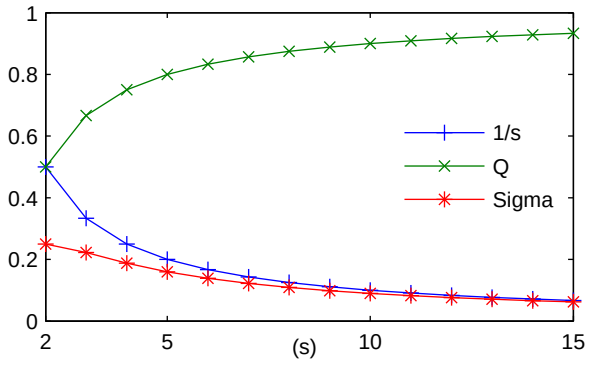



**Fig. 6.** Sigma function for $2 \leq s \leq 15$.

It combines two factors of $s$ expressing two clashing facts: (i) the $Q$ factor expresses the idea that the more the cardinality, the more accurate the information given by the feature, therefore $\sigma$ increases and the gradient decreases, so that more evidence must be seen in order to break the equilibrium, (ii) whereas the $1/s$ factor expresses the idea that the more the cardinality, the less the prior probability for the state of both minimum and maximum information (bigger entropy), therefore $\sigma$ decreases and the gradient increases, making it easier to reach it.

Still another notable difference is that this expression (as depicted in Fig.7) gives non-zero values at the zero frequency and non-one values at the frequency one, therefore providing a straight path to a full family of parameters, that is,

$$\Theta(0) = \frac{1}{2}\,bxp\left(\frac{1}{Q^2}\right) \quad ; \quad \Theta(1) = \left(1 - \frac{1}{2}\,bxp\left(s^2\right)\right) . \tag{17}$$

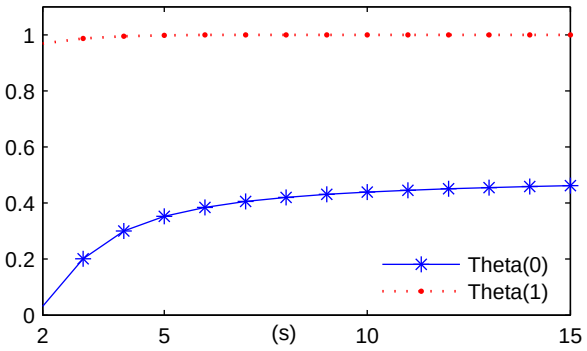



**Fig. 7.** Theta(0) and Theta(1) functions for $2 \leq s \leq 15$.

The non-zero values express the uncertainty associated to the fact of having no evidence of something. The more the cardinality, the more the prior probability of such a case, so the uncertainty increases. The non-one values express the uncertainty that should be regarded, in spite of having full evidence of something, given the stochastic nature of a sample. The more the cardinality, the less the prior probability of such a case, so the uncertainty decreases and the value tends to one.

Obviously, this expression is not normalized; it is not a *mass function*. It does not directly translate evidence into probabilities; rather, it translates traces of evidence into biases over the equilibrium. Anyway, normalization allows deriving a complete family of parameters from this expression. In classification issues, this conservative understanding of evidence usually turns to be enough and in most of the cases even better than a raw interpretation.

It's hardly worth mentioning, that an interesting option arises from the possibility of applying this parametrical model to any of the measures already existent.

## 5 Conclusions

This expression intends to give an equable, impartial and equilibrated measure of dependence relationship, taking into account its relative degree of *support* and its associated quantity/quality of knowledge.

*Coherence* is measured as a trace of dependence. It's to be assumed that whenever two features are dependent, this dependency should be patent for the whole pattern, moving away their conditional distribution from the *aicd*. On the other hand, high rates of *coherence* would be easily achieved with respect to a feature with a great bias in its marginal distribution toward or against one of its possible outcomes. That's the correction introduced into the expression of *utility* by the measure of *presence*. Good *coherence* but poorly or excessively supported

by the sample would be punished by the *presence* factor, giving poor rates of *utility*.

Equanimity is given by the fact that *presence* and *coherence* are measured exactly as the same concept, a distance to their respective uninformative distributions, guarantying this way the most possible assertive balance between seeing (*presence*, *coherence*) and believing (*utility*).

From a summarization point of view, being the measure defined at the least significant level, it can be summed up to whatever may be of interest, providing ranked classifications not only at pattern, subpattern or rule levels, but even at feature and sample levels. Therefore, relevance at each level can be objectively analyzed.

At pattern level, the *utility* measure relates to marginal dependence. However, this measure is directly extensive to relationships like $(X^p, X^q) \rightarrow X^c$. In this case, what is measured turns out to be the relation of conditional dependence $(X^p \perp X^q \,|X^c)$. Therefore, this extended measure of *utility* can be applied to any subset of parents of a feature, providing an ordered list of classification rules. Both matters have significant implications regarding to *clustering* and/or *graphical modelling*. At feature level, conclusions can be derived related to *feature subset selection* issues.

A striking practical application is to implement this measure as a heuristic in a search in order to optimize a simultaneous multi-interval discretization of a sample with some/all continuous features. This application has been tested both in real domain and synthetic data bases, and has shown that the measure leads to optimal cardinalities and optimal subsets of parents for each feature, according to a natural bias to the MDL principle.

## References

1. Julien Blanchard, Fabrice Guillet, Henri Briand, and Regis Gras. Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In Proc. of the 11th Int. Symposium on Applied Stochastic Models and Data Analysis (ASMDA'05), 191-200. ENST, 2005.
2. Liqiang Geng, Howard J. Hamilton. Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys, Vol.38, No.3, Article 9, September 2006.
3. Igor Kononenko. On biases in estimating multi-valued attributes. In Proc. of the Fourteenth Int. Joint Conference on Artificial Intelligence (IJCAI'95), 1034-1040, Montreal, Canada, 1995.
4. Philippe Lenca, Patrick Meyer, Benoît Vaillant, Stéphane Lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. European Journal of Operational Research. Vol.184., No.2, 610-626, January 2008.
5. Alex Tze Hiang Sim, Maria Indrawan, Bala Srinivasan. The importance of negative associations and the discovery of association rule pairs. International Journal of Business Intelligence and Data Mining, Vol.3 No.2, 158-176, September 2008.
6. Pang-Ning Tan, Vipin Kumar, Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In Proc. of the 8th Int. Conference on Knowledge Discovery and data Mining (KDD'02), 32-41, Edmonton, Canada, 2002.

# Enhancing Rule Importance Measure Using Concept Hierarchy

Jiye Li, Nick Cercone, Serene W. H. Wong, and Lisa Jing Yan

Faculty of Computer Science and Engineering, York University
4700 Keele Street, Toronto, Ontario, Canada M3J 1P3
{jiye, swong, jingyan}@cse.yorku.ca, ncercone@yorku.ca

**Abstract.** A rule importance measure is used to evaluate how important are the rules which characterize a data set. This measure was designed based on association rules and it has been proven to be effective to enumerate the most important rules of all rules generated. However, since rule importance is an objective measure, its usage as a rule interestingness measure relies on the interpretation of domain experts. We propose to enhance the rule importance measure previously used by incorporating a weight biased attribute concept hierarchy. The new measure better reflects the importance of a rule by integrating with the domain knowledge. A geriatric care data set is used as our experimental data set. We show that this enhanced rule importance measure provides a knowledge oriented distinction of rules classified as important.

**Key words:** Association Rules, Rule Interestingness, Rule Importance Measure, Concept Hierarchy, Rough Sets

## 1 Introduction

Association rule algorithms are well known for discovering item-item associations among the transaction data set, and have been widely used in fields such as business data analysis, transaction management, and medical research. One of the challenging problems for association rule algorithms is that, given the characteristics of the application data set, there are usually enormous number of rules generated by the algorithms. How can one interpret and identify interesting rules among all those generated? One solution to this problem includes using interestingness measures [12] to evaluate and rank the generated rules. For example, given a grocery transaction data set, rules such as "80% of male customers who bought beer also bought diaper" may have a higher interestingness measure than "80% customers bought bread and milk together".

To evaluate the interestingness of the association rules, both subjective measures and objective measures are commonly used [4]. Subjective measures rely on the human (usually the domain experts) effort to evaluate rules manually. This approach is more accurate, though it is also more expensive and time-consuming to involve the domain experts for evaluation. The objective measures

P. Lenca and S. Lallich (Eds.): QIMIE/PAKDD 2009

include measures from statistics, machine learning and information theory fields, and can automate the evaluation process without the involvement of domain experts. Objective measures alone are not sufficient to provide solid evaluations because the data domain knowledge is not taken into consideration for rule evaluation. Therefore the optimal solution would be to integrate both the subjective and the objective measures together into the rule evaluations. A Rule Importance Measure (RIM) [7] was designed as an objective rule measure similar to the interestingness measures to evaluate how important the rules are. This measure is designed based on rough sets theory and association rules, and is illustrated as follows. ROSETTA [9] rough sets software was first used to generate multiple reducts. Apriori [1] association rule algorithm was then applied to generate rule sets for each data set based on each reduct. Some rules were generated more frequently than the others among the total rule sets. Such rules were considered as more important. The rule importance was defined as the occurrence of an association rule across all the rule sets. Experimental results show the RIM reduces the number of rules generated and at the same time provides a diverse measure of how important a rule is.

In this paper, we propose an enhanced measure for the rule importance measure using concept hierarchy. The motivation of this research is to design a rule measure that integrates the domain experts' opinions into the objective evaluations. Given a data set, we first develop a concept hierarchy based on its domain, and then weights are assigned to the attributes according to their corresponding hierarchy. The Rule Importance Measure generates rules measures with its importance. Then from the RIM rules, rules with higher weighted attributes and higher occurrence are considered as more important. We name this enhanced rule evaluation approach ERIM (Enhanced Rule Importance Measure). This weight biased rule measure integrates the domain knowledge together into the rule evaluation, therefore a knowledge oriented distinction of rules are suggested. Note that the rules evaluated by ERIM are to be used for classification or prediction purpose. The rules we are interested to evaluate all contain the decision attributes on the right hand sides of the rules, and the condition attributes on the left hand sides of the rules.

The contributions of our work are summarized as follows. We propose a novel rule evaluation approach based on concept hierarchy which integrates both the subjective measures and the objective measures; the proposed ERIM provides a knowledge oriented distinction of rules demonstrated by our case study.

The rest of the paper is organized as follows. We review the related work in Section 2. The proposed new measure with the usage of concept hierarchy to combine domain knowledge into the rule evaluations is discussed in Section 3. The data set and the case studies are discussed in Section 4 and 5. Section 6 provides the conclusions and future work.

## 2 Related Work

### 2.1 Association Rules

The association rule algorithm was first introduced in [1], and is commonly referred to as the apriori association rule algorithm. This algorithm is used to discover rules from transaction datasets. The algorithm first generates frequent itemsets, which are sets of items that have transaction support greater than the minimum support; then based on these itemsets, the association rules are generated which satisfy the minimum confidence. Association rule algorithms can be used to find associations among items from transactions. For example, in *market basket analysis*, by analyzing transaction records from the market, we could use association rule algorithms to discover different shopping behaviours such as, when customers buy bread, they will probably buy milk. This type of behaviour can be used in the market analysis to increase the amount of milk sold in the market. The association rule $\alpha \rightarrow \beta$ holds in the transaction set $D$ with *confidence* $c$ if $c\%$ of transactions in $D$ that contain $\alpha$ also contain $\beta$. The rule $\alpha \rightarrow \beta$ has *support* $s$ in the transaction set $D$ if $s\%$ of transactions in $D$ contain $\alpha \cup \beta$.

### 2.2 Rule Importance Measure

The Rule Importance Measure applies rough sets theory to association rules generation in order to evaluate association rules and thus improve their utilities. Rough sets theory [10] was proposed to classify imprecise and incomplete information. Reduct and core are the two important concepts in rough sets theory. A reduct is a subset of attributes that are sufficient to describe the decision attributes. Core represents the most important information of the original data set. The intersection of all the possible reducts is called the core. The rule importance measure (RIM) is defined as the percentage of the number of times a rule is generated among all the rule sets (represented as $RuleSets$) over the number of available rule sets. The rule importance measure is obtained by $RIM_i = \frac{|\{ruleset_j \in RuleSets | rule_i \in ruleset_j\}|}{n}$. The Rule Importance Measure is simple, quick, easy to compute; it provides a direct and objective view of how important a rule is.

### 2.3 Concept Hierarchy

Much research effort has been found on using concept hierarchy towards databases management, text categorizations, natural language processing and so on. Algorithms on discover associations between different items from levels of taxonomy (which is represented in hierarchies) was introduced in 1995 as the mining approach for generalized association rules [11]. As an example of recent applications, a keyword suggestion approach based on concept hierarchy has been proposed [3] to facilitate user's web search. A data mining system has been proposed to induce the classification rules using concept hierarchy [2]. Concept

Hierarchy can reflect the concepts and relationships of a given knowledge domain. Such hierarchies are useful towards generalization and specialization.

## 3 Enhanced RIM using Concept Hierarchy

Our motivation is to enhance the RIM by integrating the subjective measure into the rule evaluation. We use a concept hierarchy to embed a semantic relationship from the data domain into the knowledge evaluation. In this section, we discuss given a problem domain, how to build a concept hierarchy and combine such hierarchy to enhance the rule measure.

Let $T$ be a data set. $T = (U, C, D)$, where $U$ is the set of data records in the table, and $U \neq \phi$, $C$ is the set of the condition attributes and $D$ is the set of the decision attributes.

Let $s$ be the total number of concepts for a given data set. $c(k)$ $(1 \leq k \leq s)$ is the $k$th concept categorized from the concepts. $Attr_{c(k)}$ denotes all the attributes that belong to the concept $c(k)$. The weight of the concept $w_{c(k)}$ denotes the importance of the concept $c(k)$ from the domain expert's opinion. For a set of rules, the new measure $ERIM_i$ for $rule_i$ can be obtained by Eq 1.

$$ERIM_i = \sum_{p=1}^{l_i} w_{c(k),p} \tag{1}$$

, where $l_i$ is the number of attributes contained by $rule_i$ and $w_{c(k),p}$ is the weight of the $p$th attribute in this rule.

Since the weights $w_{c(k),p}$ are assigned by the domain experts, the greater the value of $ERIM_i$, the more interesting a rule becomes from the domain expert's opinion. Therefore, the $ERIM_i$ measure integrates subjective measures based on concept hierarchy into the rule evaluations.

The concept hierarchy and the weight of the concepts are pre-determined by the domain expert. Concept hierarchies for a given data domain may contain more than multiple levels of hierarchies. For example, given a grocery data domain, concept hierarchies may contain "meat", "seafood", "vegetables", and "soft drinks" as the second level concepts; under each category, there exists more hierarchies. "Meat" may contain "pork", "beef", "lamb" and so on as the sub-hierarchies. In this paper we illustrate the utilities of ERIM by both a six-level and a eight-level hierarchy from a given domain. Domains with more or less hierarchies may use ERIM approach similarly.

The enhanced RIM approach thus consists of two steps. The first step is to obtain the RIM for the given data set; and the second step is for each of the rules from RIM set, derive the ERIM using the Equation 1. Therefore, for each rule generated from a given data set, we have an objective measure to evaluate how important the rule is, and at the same time, we obtain the subjective measure to evaluate which rule is indeed important from the domain expert's opinion.

The procedure of the ERIM measurement is shown as follows:

```
1. Derive concept hierarchies for the given data domain;
2. Assign attributes to concept categories;
3. Assign weights to attributes that belong to each concept category;
4. Calculate the RIM to obtain rule sets ranked by the importance measure;
5. Calculate ERIM for each rule;
6. Combining both RIM and ERIM into rule evaluation.
```

## 4 Data Set

The geriatric care medical dataset used is from Canadian Study of Health And Aging (CSHA). It has 8547 instances of a population of 65 years old and up, of whom 1865 died during the 72 months of follow-up. 3458 of them are male, and 5089 of them are female. 44 self-report attributes were used. The 44 attributes include factors such as disabilities, sicknesses and stress situations. Disabilities refer to attributes such as whether patients could prepare their own meal, or use the telephone, or take medication, or go grocery shopping. Sicknesses refer to attributes such as whether they have a chest problem, or a heart problem, or a kidney problem. Stress situations refer to attributes such as whether they have trouble in life. The class attribute is a binary value indicating whether an individual has died during the 72 months of follow-up. Detailed description of the 44 attributes are available [6].

The sample reduct set of this data is {edulevel, eyesight, hearing, shopping, housewk, health, trouble, livealone, cough, sneeze, hbp, heart, arthriti, eyetroub, eartroub, dental, chest, kidney, diabetes, feet, nerves, skin, studyage, sex}. The reducts are used for the calculation of RIM. There are 14 core attributes generated for this data set. They are *eartroub, livealone, heart, hbp, eyetroub, hearing, sex, health, edulevel, chest, housewk, diabetes, dental, studyage*. All of these reducts contain the core attributes. After removing 12 inconsistent data entries in the medical data set, we obtain the data containing $8,535$ records[1].

## 5 Case Study

We illustrate in more detail how to use the ERIM measure in this section. The geriatric care data is used as our experimental data set.

### 5.1 ERIM - 6 levels

We derive the concept hierarchy by classifying the 44 attributes into 6 categories: sickness, minor sickness, disability, attitude, symptom and others. Sickness refers to significant sickness such as heart problem or chest problem. Minor sickness refers to problems which are common among a lot of older adults but are not

[1] Notice from our previous experiments that the core generation algorithm cannot return correct core attributes when the data set contains inconsistent data entries.

significant such as ear trouble. Disability refers to how well they satisfy their daily activities such as walking, cooking and dressing. Attitude refers to how happy they are and how they feel about themselves. Symptom refers to having some signs medically, but not a sickness yet. Others include attributes that are not strongly related to the sickness, i.e., education level, gender, study age, and age group. For a detailed description on how the attributes are categorized into 6 categories, refer to Table 1. The six level concept hierarchy is shown in Figure 1.

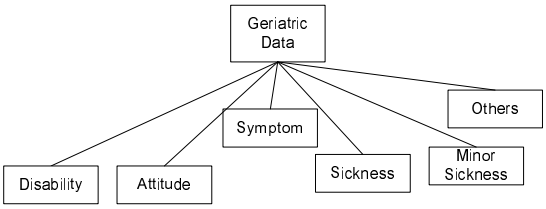



**Fig. 1.** 6-Level Concept Hierarchy

**Table 1.** Concept Hierarchy for the Geriatric Care Data

| Disability | Attitude | Symptom | Sickness | Minor Sickness | Others |
|---|---|---|---|---|---|
| dress, takecare | eyesight | eat | hbp | eyetroub | sex |
| walk, getbed | hearing | cough | heart | eartroub | studyage |
| shower, bathroom | health | tired | stroke | dental | age6 |
| phoneuse, walkout | trouble | sneeze | arthriti | stomach | edulevel |
| shopping, meal | livealon | | parkinson | bladder | |
| housewk, takemed | | | chest | bowels | |
| money | | | kidney | feet | |
| | | | diabetes | skin | |
| | | | nerves | fracture | |

**Table 2.** Weights for Concept Hierarchy of Table 1

| $c(i)$ | Disability | Attitude | Symptom | Sickness | Minor Sickness | Others |
|---|---|---|---|---|---|---|
| $w_{c(i)}$ | 6 | 2 | 1 | 30 | 1 | 1 |

We then assign weights for attributes of each concept category. Different weights are applied to different categories of attributes. The differences between weights are indications of different importance between attributes in terms of predicting the survival probability of an individual. For example, sickness is 30 times as important as symptoms, therefore the weight of sickness is assigned

as 30 and the weight of symptoms is assigned as 1. Thus, different weights are applied to the sickness, minor sickness, disability, attitude, and symptom. The weights are assigned as follows: the sickness category has a weight of 30, the minor sickness category has a weight of 1, the disability category has a weight of 6, the attitude category has a weight of 2, and the symptom category and other attributes category also have a weight of 1 [13]. These weights are set after consultation with the domain expert, and is shown on Table 2.

**Calculating RIM and ERIM** The Rule Importance is calculated on this data set. For each reduct set, association rules are generated with $support = 30\%, confidence = 80\%$ [2]. We are interested in rules with survival status on the consequent part of the rules. Rule templates [5] are defined to ensure the desired form of rules are generated [7]. ERIM is also calculated for each of the rules generated by the concept hierarchy from Table 1 and Table 2 using Eq. 1.

As an example of calculating ERIM, suppose we have a rule as follows: *If a person lives alone, has diabetes and nerve problems, then this person has a higher chance of not surviving at the end of the observation period.* This rule contains three attributes, "livealone", "diabetes" and "nerves". The ERIM is calculated as

$$ERIM_i = \sum_{p=1}^{3} w_{c(k),p} = w_{c(livealone)} + w_{c(diabetes)} + w_{c(nerves)} = 2 + 30 + 30 = 62$$

We list all the rules generated ranked by their RIM and ERIM in Table 3. In this table, the first column indicates the original rule number ranked by the RIM approach [7]. The lower the rule number, the more important this rule is. We keep this original number for comparison purpose. The second column contains the generated rules; the third column indicates the ERIM measure of this rule and the fourth column indicates the RIM measure of the same rule in this row. (Note that for comparison purpose, we use the percentage of ERIM divided by the largest ERIM value from all the generated rules. The percentage value of ERIM is also applied on Table 6.)

**Observations and Discussions** The rule importance is an indication of how significant a rule is in term of its classification ability for the decision attribute. The ERIM indicated in the third column is listed to specify the interestingness considered by the domain experts. We compare the two measures and show the differences between the ERIM and RIM. We have the following observations from the experimental results.

– Same important rules are not always considered as interesting by the domain expert. As noted, rule No.3 has the same ranking of the RIM as rule

[2] Note that the values of support and confidence can be adjusted to generate as many or as few rules as required.

**Table 3.** Sample Rules Generated from the Geriatric Care Data Set - 6-Level Hierarchy

| No. | Selected Rules | ERIM-6level | RIM |
|---|---|---|---|
| 159 | hbp, stroke, kidney, nerve problem → negative survival | 100% | 32.56% |
| 109 | hbp, dental problem, kidney problem, nerve problem, fractures → negative survival | 76.67% | 43.02% |
| 22 | oftensneeze, hbp, diabetes, nerve problem → negative survival | 75.83% | 81.40% |
| 44 | oftencough, hbp, kidney, nerve problem → negative survival | 75.83% | 66.28% |
| 53 | oftensneeze, hbp, kidney, nerve problem → negative survival | 75.83% | 61.63% |
| 66 | hbp, diabetes, nerve problem, anyfractures → negative survival | 75.83% | 58.14% |
| 158 | stroke, dental, kidney, nerve problem → negative survival | 75.83% | 32.56% |
| 89 | hbp, stroke, diabetes → negative survival | 75.00% | 48.84% |
| 93 | stroke, arthritis, diabetes → negative survival | 75.00% | 46.51% |
| 100 | stroke, diabetes, nerve problem → negative survival | 75.00% | 45.35% |
| 150 | stroke, arthritis, kidney problem → negative survival | 75.00% | 33.72% |
| 7 | livealone, diabetes, hbp → negative survival | 51.67% | 100% |
| 11 | livealone, diabetes, nerve problem → negative survival | 51.67% | 95.35% |
| 127 | hearing problem, phoneuse, nerve problem → negative survival | 31.67% | 39.53% |
| 216 | oftensneeze, dental, kidney, skin→ negative survival | 27.50% | 1.16% |
| 3 | hearing, diabetes → negative survival | 25.67% | 100% |
| 6 | heart → negative survival | 25% | 100% |
| 2 | chest → negative survival | 25% | 100% |
| 128 | hearing problem, phoneuse, dental problem → negative survival | 7.50% | 39.53% |
| 8 | housework problem → negative survival | 5.00% | 100% |
| 24 | troublewithlife → negative survival | 1.67% | 81.40% |
| 4 | ear trouble → negative survival | 0.83% | 100% |
| 5 | eye trouble → negative survival | 0.83% | 100% |
| 9 | feet → negative survival | 0.83% | 96.51% |
| | ... | | |

No.4, but the ERIM of rule No.3 is much higher than that of rule No.4. The attributes "hearing", "diabetes" and "ear" are all core attributes, therefore these two rules both have the RIM as 100%. However, from Table 2, $w_{diabetes}$ belongs to the sickness concept, and $w_{hearing}$ falls into the attitude concept. The sum of these two weights is greater than $w_{eartrouble}$, which is considered as minor sickness. The same observation goes to rule No.127 and No. 128. Rule No.127 and No.128 have the same RIM, but rule No.127 contains attributes with larger weights than those of rule No.128. Therefore rule No.127 is considered as more interesting by domain expert. *This demonstrates the domain knowledge is necessary to distinguish rules with the same classification ability.*

- Rules that are considered as interesting by the domain expert do not necessarily have the same RIM. Rule No.22 and rule No.44 have the same ERIM, which indicate they have the interestingness degree by the domain expert. However, the RIM for rule No.22 is greater than that of rule No. 44. Note that the only difference of these two rules is No.22 contain attribute "diabetes", and No.44 contains attribute "kidney". "Diabetes" is a core attribute, but not the "kidney". *This demonstrates that, what is considered less interesting by objective measures may be more interesting by the domain experts.*
- Rules having low RIM can be considered surprisingly interesting by the domain expert. Note that rule No. 159 has a low importance of 32.56%, however, it is the most interesting rule ranked by the ERIM measure from

Table 3. This is because attributes "hbp", "stroke", "kidney" and "nerve" all fall into the sickness concept with weight 30 and the ERIM is very high; however, among these attributes, only "hbp" is a core attribute from the RIM measure. Same observation applies to rule No. 216. This rule is considered less important because there are less core attribute contained in the rule, and it is generated less frequently across multiple reducts. However, the ERIM of this rule is 27.50%. *This demonstrate that the objective measures alone may ignore very interesting rules considered by the domain knowledge.*

– ERIM measure can be used together with the RIM for distinction of more knowledge oriented rules.

Although the Rule Importance is different from other objective measures and it provides a diverse ranking of how important the rules are [7], this measure can certainly be enhanced with ERIM for a more complete view of rules using the concept hierarchy. Concept Hierarchy is derived by the domain experts. According to the different purposes of the knowledge evaluation, there may exist more than one concept hierarchy for a data domain. For example, in our case study, a frailty index [8] may be considered for assigning the weighted concept categories for the geriatric care data, if the purpose of the study is to consider the proportion of the deficits instead of the nature of the deficits. Neither the objective measure (i.e., RIM) nor the subjective measure (i.e., ERIM) alone is sufficient for a thorough knowledge evaluation. Through the experiments, we observed an integration of both the objective and the subject measures is an optimal approach for knowledge evaluation.

### 5.2 ERIM - 8 levels

In this section, we study how the number of concept hierarchies affects the rule evaluations. We derive the concept hierarchy by classifying the 44 attributes from the geriatric care data into 8 categories: severe sickness, sickness, moderate sickness, minor sickness, disability, attitude, symptom and others. Severe sickness refers to severe sickness such as stroke and diabetes. Sickness refers to significant sickness such as heart problem or chest problem. Moderate sickness refers to moderate problems such as bladder or facture. Minor sickness refers to problems which are common among a lot of older adults but are not significant such as ear trouble. Disability refers to how well they satisfy their daily activities such as walking, cooking and dressing. Attitude refers to how happy they are and how they feel about themselves. Symptom refers to having some signs medically, but not a sickness yet. Other category includes age, gender and so on. For a detailed description on how the attributes are categorized into 8 categories, refer to Table 4. The concept hierarchy is shown in Figure 2.

We then assign weights for attributes of each concept category. Different weights are applied to different categories of attributes. The weights are as follows: the severe sickness category has a weight of 30, sickness category has a weight of 20, moderate sickness has a weight of 2, the minor sickness category has a weight of 1, the disability category has a weight of 6, the attitude category

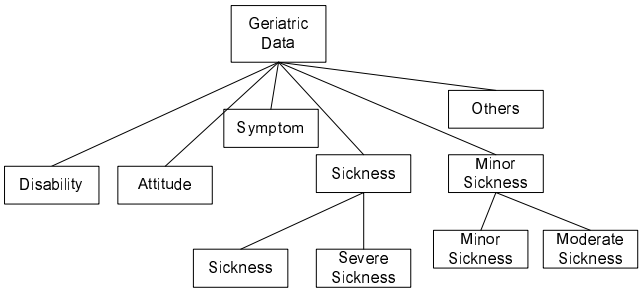



**Fig. 2.** 8-Level Concept Hierarchy

**Table 4.** Concept Hierarchy for the Geriatric Care Data

| Disability | Attitude | Symptom | Minor Sickness | Moderate Sickness | Sickness | Severe Sickness | Others |
|---|---|---|---|---|---|---|---|
| dress, takecare | eyesight | eat | eyetroub | bladder | hbp | stroke | Sex |
| walk, getbed | hearing | cough | eartroub | bowels | heart | parkinson | studyage |
| shower, bathroom | health | tired | dental | fracture | arthriti | diabetes | age6 |
| phoneuse, walkout | trouble | sneeze | stomach | | chest | | edulevel |
| shopping, meal | livealon | | feet | | kidney | | |
| housewk, takemed | | | skin | | nerves | | |
| money | | | | | | | |

**Table 5.** Weights for Concept Hierarchy of Table 4

| $c(i)$ | Disability | Attitude | Symptom | Minor Sickness | Moderate Sickness | Sickness | Severe Sickness | Others |
|---|---|---|---|---|---|---|---|---|
| $w_{c(i)}$ | 6 | 2 | 1 | 1 | 2 | 20 | 30 | 1 |

has a weight of 2, and the symptom category and others each has a weight of 1 [13]. These weights are set after consultation with the domain expert, and is shown on Table 5.

We list all the rules generated ranked by their ERIM and RIM in Table 6 according to the same approach as shown in Section 5.2.

From Table 6 we observe that rules are ranked by ERIM in the similar order as in Table 3. For example, rule No.159 are ranked as the highest ERIM in both approaches. By using more detailed concept hierarchy, rules may be differentiated in a deeper level. For example, rule No.22 and No.66 have the same ERIM by using 6-level hierarchy. However, with 8-level hierarchy, the "Sickness" and "Minor Sickness" in Table 1 are further differentiated by more hierarchies with more weights in Table 4. The two different attributes comparing No.22 with No.66 are "oftensneeze" and "anyfractures". In the 8-level hierarchy, "anyfrac-

**Table 6.** Sample Rules Generated from the Geriatric Care Data Set Ranked by 8-level Hierarchy

| No. | Selected Rules | ERIM-8 | RIM |
|---|---|---|---|
| 159 | hbp, stroke, kidney, nerve problem → negative survival | 100% | 32.56% |
| 89 | hbp, stroke, diabetes → negative survival | 88.89% | 48.84% |
| 93 | stroke, athritis, diabetes → negative survival | 88.89% | 46.51% |
| 100 | stroke, diabetes, nerve problem → negative survival | 88.89% | 45.35% |
| 66 | hbp, diabetes, nerve problem, fractures → negative survival | 80.00% | 58.14% |
| 22 | often sneeze, hbp, diabetes, nerve problem → negative survival | 78.89% | 81.40% |
| 7 | live alone, hbp, diabetes → negative survival | 57.78% | 100.00% |
| 11 | live alone, diabetes, nerve problem → negative survival | 57.78% | 95.35% |
| | ... | | |
| 3 | hearing, diabetes → negative survival | 35.56% | 100.00% |
| 4 | ear trouble → negative survival | 1.11% | 100.00% |
| 5 | eye trouble → negative survival | 1.11% | 100.00% |
| 9 | feet problem → negative survival | 1.11% | 96.51% |
| | ... | | |

tures" is assigned with a higher weight. Therefore, rule No.66 is ranked higher than No.22 using the 8-level hierarchy in Table 6. This results indicate that more concept hierarchies represent finer-grained domain knowledge, therefore the interestingness of the rules are differentiated in a greater detail comparing to using less hierarchies.

## 6 Conclusion

In this paper we have proposed a novel approach for rule evaluation based on concept hierarchy. An enhanced Rule importance measure ERIM is shown to be effective on evaluating interesting rules from the domain expert's opinion. We demonstrate through a real world data set that the integration of both the objective and the subjective measures can provide a knowledge oriented distinction of rules. The advantages of ERIM are as follows: it combines both the subjective and the objective measures for rule evaluation; in the situation where the two rules have the same RIM, ERIM can be used to provide a knowledge oriented distinction. The concept hierarchy based weights are indications of interestingness reflecting domain knowledge.

In the future we plan to continue developing rule evaluation measures that combine both the objective measures and the subjective measures. As discussed in Section 3, concept hierarchy is limited by the purpose of knowledge evaluation and it is not automatable at this stage. The constructing of concept hierarchy as well as the assigning of attribute weights depend on the particular problem domain. These two components of our approach are time consuming and sometimes difficult to obtain from the problem domain expert. Domain experts and statistics information should play an important role. We are also interested in

researching an automatic mechanism on developing the concept hierarchies to facilitate more efficient and more precise knowledge evaluations.

## Acknowledgements

We gratefully acknowledge the financial supports of the Natural Science and Engineering Research Council of Canada and Alpha Global-iT Inc.

## References

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.
2. M. E. M. D. Beneditto and L. N. de Barros. Using concept hierarchies in knowledge discovery. In A. L. C. Bazzan and S. Labidi, editors, *SBIA*, volume 3171 of *Lecture Notes in Computer Science*, pages 255–265. Springer, 2004.
3. Y. Chen, G.-R. Xue, and Y. Yu. Advertising keyword suggestion based on concept hierarchy. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 251–260, New York, NY, USA, 2008. ACM.
4. L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9, 2006.
5. M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. In N. R. Adam, B. K. Bhargava, and Y. Yesha, editors, *Third International Conference on Information and Knowledge Management*, pages 401–407. ACM Press, 1994.
6. J. Li. *Rough Set Based Rule Evaluations and Their Applications.* PhD thesis, University of Waterloo, Waterloo, Canada, 2007.
7. J. Li and N. Cercone. Introducing a rule importance measure. In J. F. Peters, A. Skowron, D. Dubois, J. W. Grzymala-Busse, M. Inuiguchi, and L. Polkowski, editors, *T. Rough Sets*, volume 4100 of *Lecture Notes in Computer Science*, pages 167–189. Springer, 2006.
8. A. Mitnitski, X. Song, and K. Rockwood. The estimation of relative fitness and frailty in community-dwelling older adults using self-report data. *J Gerontol A Biol Sci Med Sci*, 59:M627–M632, 2004.
9. A. Øhrn. *Discernibility and Rough Sets in Medicine: Tools and Applications.* PhD thesis, Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim Norway, 1999.
10. Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishers, Norwell, MA, USA, 1992.
11. R. Srikant and R. Agrawal. Mining generalized association rules. In *VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases*, pages 407–419, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
12. P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *KDD*, pages 32–41. ACM, 2002.
13. S. W. H. Wong and N. Cercone. Comparing classification methods for predicting survival probabilities in the elderly. *IEEE Conference on Bioinformatics and Biomedicine Workshop on Biomedical and Health Informatics*, 2008.

# False neighbourhoods and tears are the main mapping defaults. How to avoid it? How to exhibit remaining ones?

Sylvain Lespinats, Michaël Aupetit

CEA, LIST, Multisensor Intelligence and Machine Learning Laboratory. F-91191 Gif-sur-Yvette, France (sylvain.lespinats@cea.fr and michael.aupetit@cea.fr).

**Abstract.** Tears and false neighborhoods are the defaults that may occur when a mapping is set up. Three recent articles discuss about these risks and proposed various means to detect and avoid such penalizing situations. In the present paper we link these methods and suggest a new strategy to visualize tears and false neighborhoods on a mapping by adapting well-tried tools.

**Keywords:** Exploratory data analysis; MultiDimensional Scaling; High-dimensional data; Error visualization; False neighborhoods and tears.

## 1 Introduction

Since W.S. Torgerson and his famous "embedding theorem" [34] the distance preservation is the objective of most of mapping methods (indeed, Torgerson demonstrates that Principal Component Analysis (PCA) [30, 15] objective is equivalent to look for the data projection that preserves distances "as much as possible"). In that framework, many following methods proposed to especially account for small distances, which lead to non-linear mappings. There is a very high number of methods belonging to this category (known as Non-Linear MultiDimensional Scaling or NL-MDS) and we will only cite here the most known among them. For example, Sammon's mapping [26] and Curvilinear Component Analysis (CCA) [8] interactively minimize the weighted difference between distances in the input and output space. ISOMAP [33] computes the geodesic distance [10, 32] (which can be seen as a "non-linear distance" that follows the data manifold) before linearly embedding data according to the Torgerson's method [34]. Locally Linear Embedding (LLE), [31] and related methods [3, 12] account for nearest neighbour distances in a sparse matrix and find data position in the output space according to a spectral method. Generative Topographic Mapping (GTM) [4] approaches a lower dimensional manifold to data. Gaussian Process Latent Variable Model (GP-LVM) [20] results from a probabilistic interpretation of PCA. Many others methods start from various other paradigms such as Self-Organizing Map (SOM) [16, 17] that visualizes data on a discrete grid, Non-Metric MultiDimensional Scaling (NM-MDS)

P. Lenca and S. Lallich (Eds.): QIMIE/PAKDD 2009

[18, 19] and RankVisu [22] that preserve the ranking of distances, Kernel Principal Component Analysis (KPCA) [27] that searches for non-linear relationships between variables according to the "kernel trick" [28, 29].

Several recent papers [36, 2, 21], highlight two risks while a mapping is generated from distances. These risks are named here "false neighbourhood" and "tear" according to the terminology in [21] (that corresponds to "gluing" area and "tearing" area respectively for [2] and area with low "trustworthiness" and low "continuity" for [36]). A "false neighbourhood" occurs when a large distance in the original space is associated with a small distance in the output space (the corresponding data points seems neighbours whereas they are not). Respectively, a "tear" occurs when a small distance in the original space is associated with a large distance in the output space (true neighbours are mapped apart). Please report to section 6 for some intuitive examples of "false neighbourhoods" and "tears". Obviously, "false neighbourhoods" and "tears" are expected to be avoided in mapping framework. Subsequently, [36] and [21] proposed two mapping methods designed to avoid "as much as possible" false neighbourhoods and tears.

Moreover, when such penalizing situations occur anyway, exhibiting impacted areas should be a main concern as claimed in [2]. This article proposes then a sharp mean to visualize on the mapping the true neighbourhood of a given data point. However, a local index showing the risk level for each data point would be a critical improvement.

The present article connects dimensionality reductions methods that optimize item positions by minimizing a criterion based on distances preservation. Such approach highlights pros and cons of each method and leads to considerations on detection of mappings defaults. In particular, we set up here a couple of criteria that allow visualizing and characterizing local mapping defaults. Indeed, there are actually few tools available in order to locally analyze a mapping quality.

The present paper is organized as follows. Section 2 is dedicated to Sammon's mapping and Curvilinear Component Analysis, and how either false neighbourhoods or tears are penalized in such methods. Section 3 presents and compares two recent mapping methods accounting for both defaults: Data-Driven High Dimensional Scaling and Local MultiDimensional Scaling. Section 4 refers to several usual techniques to find out remaining defaults. Section 5 describes a method allowing characterising defaults as false neighbourhood or tear. An example on an intuitive dataset is subsequently presented in section 6.

## 2 Mapping within avoiding false neighbourhoods or tears

Two mapping methods are particularly interesting while considering mapping from the risk of false neighbourhood and tear point of view: the Sammon's mapping [26] and the Curvilinear Component Analysis (CCA) [8].

### 2.1 Sammon's mapping

Historically, Sammon's mapping [26] is one of the first Non-Linear MultiDimensional Scaling methods. Its purpose is to minimize the following function:

$$E_{Sammon} = C \times \sum_{i,j} \left( \left| d_{ij} - d_{ij}^{*} \right|^{k} \times F\left( d_{ij} \right) \right) \quad (1)$$

where $d_{ij}$ and $d_{ij}^{*}$ represent the distances between data points *i* and *j* in the input space and output space, respectively; *F* is the so called weighting function. *F* is designed to emphasize small distances. As a consequence, function *F*: $\Re_{+} \to \Re_{+}$ has to decrease. In case of Sammon's mapping, the traditional choice is $F(x) = 1/x$, $k = 2$ and $C = \sum_{i,j} d_{ij}$ (please note that *C* is a constant and has no impact on the resulting mapping).

In case where data points lie on a low-dimensional non-linear manifold, emphasizing small distance is expected to allow "unrolling" the dataset.

Although this idea is obviously powerful (after 40 years, the Sammon's mapping is still used and inspired many subsequent methods), it shows a major drawback while considering the risk of false neighbourhood and tears:

Sammon's mapping fairly penalizes tears, but it lightly penalizes false neighbourhoods. Indeed, let us suppose that there is a large distance between data points *i* and *j* in original space ($d_{ij}$), but, by misfortune, the corresponding distance in output space ($d_{ij}^{*}$) is small. $F\left( d_{ij} \right)$ is low and the difference between $d_{ij}$ and $d_{ij}^{*}$ does not much weight on $E_{Sammon}$. Such situation ($d_{ij}$ high and $d_{ij}^{*}$ low) corresponds to a false neighbourhood and could easily occur with Sammon's mapping.

### 2.2 Curvilinear Component Analysis (CCA)

In such framework, CCA [8] offers interesting behaviour. Indeed CCA is close to Sammon's mapping, but its weighting function relies on distance in the output space rather than in original space:

$$E_{CCA} = C \times \sum_{i,j} \left( \left| d_{ij} - d_{ij}^{*} \right|^{k} \times F\left( d_{ij}^{*} \right) \right) \quad (2)$$

where $k = 2$, $C = 1/2$ and various functions have been proposed for *F*.

The drawback highlighted in case of Sammon's mapping is fixed. However, even if false neighbourhoods are now fairly penalized, tears can easily occur: If $d_{ij}$ is low and $d_{ij}^{*}$ is high, $F\left( d_{ij}^{*} \right)$ is low. This case (which corresponds to a tear) is then lightly penalized.

## 3 Accounting for both false neighbourhoods and tears

Confronting to some datasets, false neighbourhood and tears cannot be avoided. For example, everyone knows that a perfect planisphere is unreachable: there is no mean to spread out a sphere onto a plan without tearing. However, lightly penalising one of false neighbourhood or tear can cause unnecessary defaults. Penalizing the both risks is then critical.

### 3.1 Data-Driven High dimensional Scaling (DD-HDS)

Related to both Sammon's mapping and CCA, DD-HDS [21] is designed to cumulate advantage of these methods while avoiding the previously presented drawbacks.

The minimized function is

$$E_{DD-HDS} = C \times \sum_{i,j} \left( \left| d_{ij} - d_{ij}^{*} \right|^{k} \times F\left( \min\left( d_{ij}, d_{ij}^{*} \right) \right) \right) \tag{3}$$

where $C = 1$ and $k = 1$ (note the choice of *k* differ for DD-HDS, mostly to be consistent with the used optimization process).

Thus, if a distance (in the original or in the output space) is low, the weighting function is high in order to account for a possible penalizing situation.

Moreover, DD-HDS is yet the only one mapping method that takes account for the "concentration of measure phenomenon" (one of the most important phenomena belonging to the famous "curse of dimensionality") [11, 1] through the weighting function: *F* is proposed to be a sigmoid function adjusted on the original distance distribution. DD-HDS is then an efficient method for mapping (especially in case of high dimensional data). The main drawback is (so far) the relatively high computational time.

### 3.2 Local MultiDimensional Scaling (Local MDS)

The Local MDS [36], is also closely related to Sammon's mapping and CCA. It offers to the user a trade-off tuning between penalization of false neighbourhoods and tears through a parameter (noted $\lambda$ in the following).

$$E_{LMDS} = C \times \sum_{i,j} \left( \left| d_{ij} - d_{ij}^{*} \right|^{k} \times \left( \lambda \times F\left( d_{ij} \right) + \left( 1 - \lambda \right) \times F\left( d_{ij}^{*} \right) \right) \right) \tag{4}$$

where $k = 2$, $C = 1/2$ and $F(x) = 1$ if *x* is lower than a chosen $\sigma$ and $F(x) = 0$ else. For $\lambda = 0$ the Local MDS corresponds to CCA and for $\lambda = 1$ the Local MDS corresponds to Sammon's mapping (except regarding the function *F*).

### 3.3 DD-HDS versus Local MDS

As shown above, solutions proposed by Local MDS and DD-HDS derived from a similar analysis. Although, they are somewhat close, advantage and drawback of each one should be highlighted.

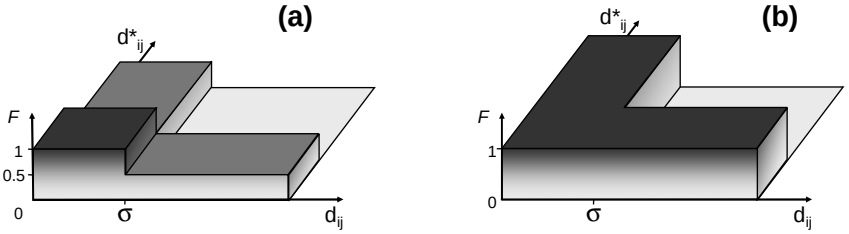
Local MDS proposes a trade-off between risks of false neighbourhood and tear. A user control ($\lambda$) allows balancing these risks. Moreover, placing mappings on a plot that quantify mapping defaults in terms of "trustworthiness" and "continuity" with varying the $\lambda$ value displays a curve which can be related to a ROC curve (Receiver Operating Characteristic). ROC curves are known to be very practical in order to select an optimal decision. To the opposite, DD-HDS does not allow such control.

Because DD-HDS equally penalizes false neighbourhoods and tears, it should be compared to Local MDS with $\lambda = 0.5$.

Note that, because *F* is a decreasing function,

$$F\left(\min\left(d_{ij}, d_{ij}^{*}\right)\right) = \max\left(F\left(d_{ij}\right), F\left(d_{ij}^{*}\right)\right) \quad (5)$$

Comparison between techniques accounting for false neighbourhoods and tears relies to comparison of weights: $\max\left(F\left(d_{ij}\right), F\left(d_{ij}^{*}\right)\right)$ in case of DD-HDS and $\left(F\left(d_{ij}\right) + F\left(d_{ij}^{*}\right)\right)/2$ in case of Local MDS. For sake of simplicity, function *F* is chosen as $F(x) = 1$ if $x < \sigma$ and $F(x) = 0$ else (as proposed in Local MDS).




(a) Local MDS (b) DD-HDS

**Fig. 1.** Weighting functions *F* (vertical axis), according to distance in original space (x-axis) and distance in output space (y-axis).

Let us consider these functions according to distances (Fig. 1).

We can observe that when one distance is small and the corresponding one is large (this corresponds to a false neighbourhood or a tear), the Local MDS weighting function is lower than when both distances are small together. To the opposite, the DD-HDS weighting function is maximal on the whole area where a distance is small. In other words, the use of maximum corresponds to the "OR" logical operation: the weighting function is at its maximum if a false neighbourhood OR a tear is found. As a consequence, a maximum effort is given in order to avoid simultaneously both false neighbourhoods AND tears.

## 4 Visualizing defaults in mappings.

Even if many solutions to visualize defaults in SOM have been developed, less solution exists in case of distances preserving methods.

Of course, each proposed stress corresponds to an index that quantifies the global mapping quality. The parting in two index called "trustworthiness" and "continuity" proposed in [35] and [36] allows for placing mappings on a plot according to its ability to avoid false neighbourhoods and tears.

The Shepard Diagram [18, 19, 7, 8] plots distances in original space versus distances in output space. It allows a totally model-free observation of the distance preservation according to distances. However, it is also a global analysis: the location of errors cannot be deduced from such diagram.

Few methods have been proposed in order to locally observing mapping defaults. In this framework, [6] proposes to analyse the stability of areas on mappings initialised randomly; and [37] compares surfaces of triangles resulting from a Delaunay triangulation in the output space to surfaces of corresponding triangles in original space.

For a consistent review, please report to [2].

### 4.1 Distortion visualization on Voronoï cells

A local comparison of a given data point neighbourhoods is made possible by [2]. In this method, the Delaunay triangulation is computed in the output space. Each Voronoï cell is then coloured according to the proximity of the related data point in the output space (light colours for cells related to close data points). If the chosen data point is fairly mapped, every light cell will lie together, apart from dark cells. Else, if some dark cells are embedded close to the chosen data point, it corresponds to a false neighbourhood; if some light cells are apart, it corresponds to a tear.

This method has the great advantage that the true neighbourhood of a given data point is immediately assessable in the mapping through an intuitive picture. However, the necessity of choosing a point of view makes it use somewhat irksome when we do not know a priori where to looking for a default.

### 4.2 Pressure defined in DD-HDS algorithm

The DD-HDS mapping is achieved by optimizing a stress (eq. 3) thanks to a Force Directed Placement (FDP) [13, 14, 9] algorithm. The FDP algorithm simulates a spring system: each data point corresponds to a mass and springs rely each couple of masses. Spring lengths at rest equal expected distances (i.e. original distances), in order to make original distances resemble to output ones after relaxation of the system. The spring stiffness allows for accounting for the weighting function $F$. The relaxation of the system is supposed to minimise the stress. Such algorithm is known to be robust to local minima and is somewhat few time consuming [24, 25]. The popularity of such algorithm increases in mapping community [5, 24, 23, 25, 21, 22].

An other advantage is that FDP allow defining the "pressure" that locally quantifies the stress level: the pressure on a data point is the sum of strength of forces applied on the corresponding mass [21]. Truly, this does not strictly correspond to the academic definition of a pressure, but this term allows an intuitive understanding of this index: the "pressure" is the quantity of forces applied on a data point.

Nevertheless, the pressure concept does not need an FDP optimisation, and could be defined from any mapping. Given a data point *i*, the pressure would be:

$$P_{DD-HDS}(i) = C \times \sum_j \left( \left| d_{ij} - d_{ij}^* \right|^k \times F(.) \right) \quad (6)$$

Moreover, because the weighting function *F* can easily be customized, Sammon's mapping, CCA, Local MDS or DD-HDS model (and so on) could be considered as well according to the pressure concept. The combination of the mapping and the pressure of data points through a greyscale allows for locally observing the stress level.

## 5 How to exhibit and characterize defaults?

Plotting DD-HDS pressure allows displaying area where the mapping shows defaults, but it is not informative about the nature of the default.

Furthermore, even if Sammon's and CCA weighting function have shown there limitation in order to drive the mapping [36, 21], they can be useful together to evaluate a mapping. Indeed, pressures related to Sammon's mapping and CCA can easily be defined:

$$P_{Sammon}(i) = C \times \sum_j \left( \left| d_{ij} - d_{ij}^* \right|^k \times F\left(d_{ij}\right) \right) \quad (7)$$

$$P_{CCA}(i) = C \times \sum_j \left( \left| d_{ij} - d_{ij}^* \right|^k \times F\left(d_{ij}^*\right) \right) \quad (8)$$

Due to the high similitude between formulas (7) and (8), values for $P_{Sammon}$ and $P_{CCA}$ can be compared if the choice of C, k and F are shared by two pressures.

For reason discussed in section 2.1 and 2.2, $P_{CCA}$ is an efficient mean to detect false neighbourhoods and $P_{Sammon}$ catches tears.

Note: DD-HDS pressure equals to

$$P_{DD-HDS}(i) = C \times \sum_j \left( \left| d_{ij} - d_{ij}^* \right|^k \times F\left(\min\left(d_{ij}, d_{ij}^*\right)\right) \right) \quad (9)$$

which corresponds to

$$P_{DD-HDS}(i) = \max\left(P_{Sammon}(i), P_{CCA}(i)\right) \quad (10)$$

## 6 Example

A classical test for mapping methods is the openbox dataset. Data points lie on sides of a 3-dimensional cube without upper side (Fig. 2, right insert with grey background) and are embedded in 2-dimensional spaces. These mappings are presented in Fig. 2 according to $E_{Sammon}$ and $E_{CCA}$ stresses (curves and points) and $P_{Sammon}$ and $P_{CCA}$ pressures (small inserts).

For a fair comparison, every methods used the same values for $C$ and $k$ ($C=1, k=2$) and function $F$ is chosen to be a rectangular function as proposed in Local MDS.

Situating mapping in a plot according to $E_{Sammon}$ and $E_{CCA}$ is in the spirit of Venna and Kaski "trustworthiness" and "continuity" visualisation [35, 36]. Note that, as expected, Sammon's mapping avoids tears to the cost of possible false neighbourhoods and CCA avoids false neighbourhoods to the cost of possible tears (see section 2). DD-HDS and Local MDS account for both defaults (section 3.1 and 3.2).

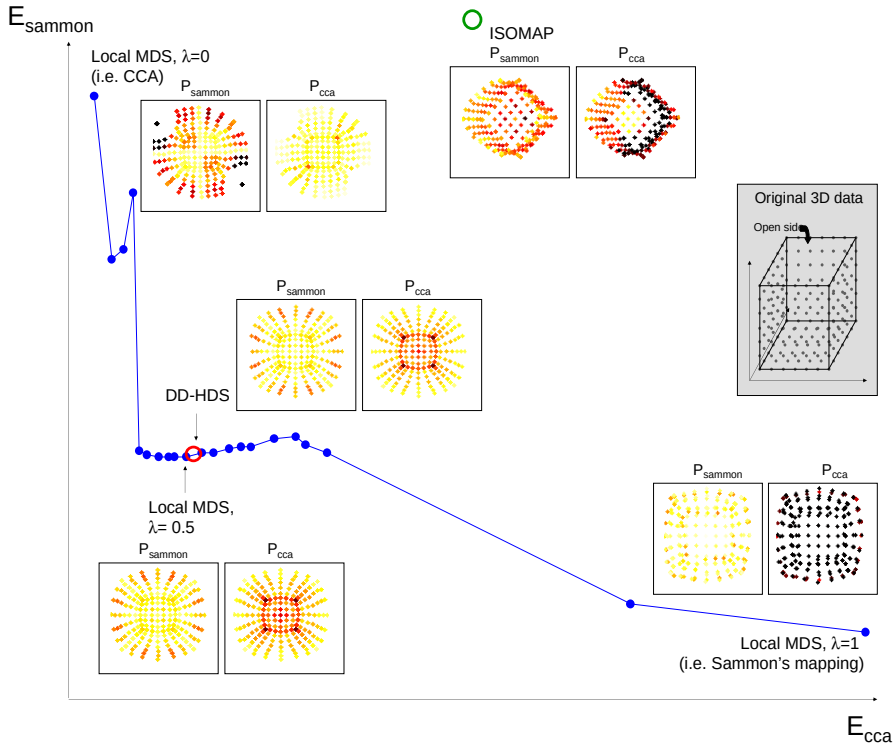



**Fig. 2.** 2D Mapping of a 3D open box (right insert with grey background) resulting from various methods as a function of Sammon's mapping and Curvilinear

Components Analysis stresses. ISOMAP and DD-HDS rely to empty circles, Local MDS for a $\lambda$ varying from 0 to 1 (with 0.05 as step) rely to full circles linked by lines. CCA corresponds to Local MDS with $\lambda = 0$ and Sammon's mapping corresponds to Local MDS with $\lambda = 1$. The most on the left the mapping, the most false neighbourhoods are avoided; the lowest the mapping, the most tears are avoided.

Mapping reached by five methods (ISOMAP, DD-HDS, CCA, Samon's mapping and Local MDS with $\lambda = 0.5$), are provided in five couples of inserts. Positions of items are similar for each couple, but colours are different: in left inserts, darker the data points, higher the pressure related to Sammon's stress (eq. 7); in right inserts, data darker the data points, higher the pressure related to CCA stress (eq. 8). Colorscales are similar for every mapping.

For five mappings (Sammon's mapping, CCA, Local MDS with $\lambda = 0.5$, DD-HDS and ISOMAP), pressures related to Sammon's and CCA stresses are presented in small inserts. It allows visualizing tears (left inserts, which corresponds to $P_{Sammon}$) and false neighbourhoods (right inserts, which corresponds to $P_{CCA}$) in the mappings (dark areas correspond to defaults). In case of CCA, two tears can easily be observed and correspond to black areas in left insert; no false neighbourhood appears (right insert). Sammon's mapping has smashed sides on the bottom face. It results many false neighbourhoods (highlighted by dark data points in left inserts) but few tears: $P_{Sammon}$, is low everywhere in the mapping (right insert). Two sides have been projected on the bottom by ISOMAP, creating false neighbourhoods. DD-HDS and Local MDS with $\lambda = 0.5$ reach close mappings. Light stretch can be observed at the top of the box (darker areas in left inserts) and some compressions in the bottom side (especially on the corners, right inserts).

## 7 Conclusion

The objective of the present paper is not to grade mapping methods. Indeed, it is obvious that: 1) the presented methods are closely related and will often reach close results (just as Local MDS and DD-HDS with the openbox dataset); 2) there respective originalities provide to each method its own advantage that should be exploited according to the situation.

On the one hand, a good point for DD-HDS on Local MDS when there is no reason for a priori favour false neighbourhood or tear is its capacity to accounting for both false neighbourhoods and tears within one single parameter. This advantage should not be neglected; indeed it permits to spare the balancing parameter. On the other hand, the Venna and Kaski's parameter give to Local MDS a unique chance to visually appreciate the trade-off between false neighbourhood and tears. In that framework, because the choice of $\lambda$ from resulting mapping falls under the data expert responsibility, tools presented here could be a useful supplementation.

The pressures proposed in section 5 are contenting themselves to finding out and characterizing mapping defaults. Such procedure should be combined with

visualization of neighbourhoods technique proposed in [2]. Indeed, $P_{Sammon}$ and $P_{CCA}$ can guide the user to the items for which such analysis is the most relevant.

## References


1. Aggarwal C.C., Hinneburg A. and Keim D.A., On the surprising behavior of distance metrics in high dimensional space, in J.V. Bussche and V. Vianu Eds. Lecture Notes In Computer Science, ser. 1973, (Berlin, Germany, Springer-Verlag, 2001), 420–434.
2. Aupetit M., Visualizing distortions and recovering topology in continuous projection techniques" Neurocomputing, vol. 10, no. 7-9 pp. 1304–1330, 2007.
3. Belkin M. and Niyogi P., "Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput., vol. 15, pp. 1373–1396, 2003.
4. Bishop C.M., Svensén M., and Williams C.K.I, GTM: The generative topographic mapping," Neural Comput., vol. 10, pp. 215–234, 1998.
5. Chalmers M., A linear iteration time layout algorithm for visualizing high-dimensional data, in Proc. 7th Conf. Visualization, R. Yagel and G. M. Nielson, Eds., San Francisco/Los Alamitos, CA, 1996, pp. 127–132.
6. Davidson G.S., Wylie B.N., Boyack K.W., Cluster stability and the use of noise in interpretation of clustering, in: Proceedings of IEEE Information Visualization (InfoVis'Ol), 2001, pp. 23–30.
7. Demartines P., Mesures d'organisation du réseau de Kohonen. presented at Congrès Satellite du Congrès Européen de Mathématiques: Aspects Théoriques des Réseaux de Neurones, Paris, France, 1992.
8. Demartines P. and Hérault J., Curvilinear component analysis: A selforganizing neural network for nonlinear mapping of data sets, IEEE Trans. Neural Netw., vol. 8, no. 1, pp. 148–154, Jan. 1997.
9. Di Battista G.D., Eades P., Tamassia R., and Tollis I.G., Graph Drawing: Algorithms for the Visualization of Graphs. Englewood Cliffs, NJ: Prentice-Hall, 1999.
10. Dijkstra E.W., A note on two problems in connection with graphs, Numerisch Mathematik, vol. 1, pp. 269–271, 1959.
11. Donoho D.L., High-dimensional data analysis: The curses and blessings of dimensionality, Los Angeles, CA, 2000, Amer. Math. Soc. Lecture: "Math challenges of the 21st century" [Online]. Available: http://www-stat.stanford.edu/~donoho/
12. Donoho D.L. and Grimes C., Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data, Proc. Nat. Acad. Sci., vol. 100, pp. 5591–5596, 2003.
13. Eades P., A heuristic for graph drawing, in Proc. 13th Manitoba Conf. Numer. Math. Comput. (Congressus Numerantium), D. S. Meek and G.H. J.V. Rees, Eds., Winnipeg, MB, Canada, 1984, vol. 42, pp. 149–160.
14. Fruchterman T. and Reingold E., Graph drawing by force-directed placement, Software-Practice Exp., vol. 21, pp. 1129–1164, 1991.
15. Jolliffe I., Principal Component Analysis, Springer-Verlag, New York, 2002.
16. Kohonen T., Self-organized formation of topologically correct feature maps, Biol. Cybern., vol. 43, pp. 59–69, 1982.
17. Kohonen T., Self-Organizing Maps., H.K.V. Lotsch, Ed. Heidelberg, Germany: Springer-Verlag, 1997.
18. Kruskal J.B., Non-metric multidimensional scaling: A numerical method, Psychometrika, vol. 29, pp. 115–129, 1964.
19. Kruskal J.B., Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika, vol. 29, pp. 1–27, 1964.

20. Lawrence N.D., Probabilistic non-linear principal component analysis with Gaussian process latent variable models, J. Mach. Learn.Res., vol. 6, pp. 1783–1816, 2005.
21. Lespinats S., Verleysen M., Giron A. and Fertil B., DD-HDS: a tool for visualization and exploration of highdimensional data, IEEE Trans. Neural Netw., vol. 18, no. 5, pp. 1265-1279, 2007.
22. Lespinats S., Fertil B., Villemain P. and Herault J., Rankvisu: Mapping from the neighbourhood network. Submitted to Neurocomputing.
23. Li J.X., Visualization of high-dimensional data with relational perspective map, Inf. Visualization, vol. 3, pp. 49–59, 2004.
24. Morrison A., Ross G., and Chalmers M., Fast Multidimensional Scaling through Sampling, Springs and Interpolation, Inf. Visualization, vol. 2, pp. 68–77, 2003.
25. Paulovich F.V., Oliveira M.C.F. and Minghim R., The projection explorer: A flexible tool for projection-based multidimensional visualization, In Proceedings of XX Brazilian Symposium on Computer Graphics and Image Processing – SIBIGRAPI 2007, Belo Horizonte, Brazil, 2007. IEEE Computer Society Press. pp. 27-36.
26. Sammon J.W., A nonlinear mapping for data structure analysis, IEEE Trans. Comput., vol. C-18, no. 5, pp. 401–409, May 1969.
27. Schölkopf B., Smola A.J., and Müller K.R., Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput., vol. 10, pp. 1299–1319, 1998.
28. Schölkopf B. and Smola A.J., Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA: MIT Press, 2002.
29. Shawe-Taylor J. and Cristianini N., Kernel Methods for Pattern Analysis. Cambridge, U.K.: Cambridge Univ. Press, 2004.
30. Pearson K., On lines and planes of closest fit to systems of points in space, Philosophical Magazine n°2, p. 559-572, 1901.
31. Roweis S.T. and Saul L.K., Nonlinear dimensionality reduction by locally linear embedding, Science, vol. 290, pp. 2323–2326, 2000.
32. Skiena S., Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica. Reading, MA: Addison-Wesley, 1990.
33. Tenenbaum J.B., de Silva V., and Langford J.C., A global geometric framework for nonlinear dimensionality reduction, Science, vol. 290, pp. 2319–2323, 2000.
34. Torgerson W.S., “Multidimensional scaling: 1. Theory and method,” Psychometrika, vol. 17, pp. 401–419, 1952.
35. Venna J. and Kaski S., Neighborhood preservation in nonlinear projection methods: An experimental study, In G. Dor_ner, H. Bischof, and K. Hornik, editors, Proceedings of ICANN 2001, International Conference on Artificial Neural Networks, pp. 485-491, Berlin, 2001. Springer.
36. Venna J. and Kaski S., Local multidimensional scaling, Neural Networks, vol. 19, no. 6-7, pp. 889-899, 2006.
37. Warnking J., Guerin-Duguet A., Chehikian A., Olympieff S., Dojat M. and Segebarth C., Retinotopical mapping of visual areas using fMRI and a fast cortical flattening algorithm, Neuroimage vol. 11, no. 5, S646, 2000.

# Authors index

# QIMIE 2009

There are a lot of data mining algorithms and methodologies for various fields and various problems. Each researcher is faced with assessing the performance of his own proposal in order to make comparisons with state of the art approaches. Which methodology, which benchmarks, which measures of performance, which tools, etc., should be used, and why?

The **Quality issues, measures of interestingness and evaluation of data mining models** Workshop (QIMIE'09) focuses on the theory, the techniques and the practices that can ensure the discovered knowledge is of quality. It thus covers the problems of quality and evaluation of data mining models.

These Proceedings contain the papers presented at the QIMIE'09 Workshop organized in association with the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'09), April 27, 2009, Bangkok, Thailand.

Philippe Lenca and Stéphane Lallich (Eds.)

QIMIE'09/PAKDD'09
ISBN-13 978-2-908849-23-3 Telecom Bretagne