What's Interesting About Cricket? – On Thresholds and Anticipation in Discovered Rules

John F. Roddick
School of Informatics and Engineering
Flinders University of South Australia
GPO Box 2100, Adelaide 5001
South Australia
roddick@cs.flinders.edu.au

Sally Rice
School of Computer and Information Science
University of South Australia
Mawson Lakes Boulevard,
Mawson Lakes 5095, South Australia
rice@cs.unisa.edu.au

ABSTRACT

Despite significant progress, determining the interestingness of a rule remains a difficult problem. This short paper investigates the lessons that may be learned from analysing the (largely manual) selection of interesting statistics for cricket (or any other data rich sport) by experts. In particular, the effect of thresholds on the interestingness of rules describing events in the sporting arena is discussed. The concept of anticipation is shown also to be critical in this selection and to vary the level of interest in events that may contribute to the achievement of a threshold value during a match, thus adding a temporal dimension to interestingness. This temporal aspect can be best modelled on the single-past-branching-future model of time. As a result of this investigation, a few new general ideas are discussed that add to the research in this area. Significantly, some of the new criteria are implicitly temporal in that they rely on a model of behaviour over time. The applicability of threshold values for detecting uncharacteristically poor performances are canvassed as areas of interest yet to be explored.

Keywords

Temporal Data Mining, Knowledge Discovery, Interestingness Criteria, Cricket Statistics.

1. INTRODUCTION

A number of researchers have investigated the important and difficult problem of determining interesting or relevant rules from spurious or unimportant information [2; 3; 4; 5; 6; 7; 8; 9; 10; 12; 13; 14; 15; 16]. This work has identified some general characteristics of interesting rules and some methods for isolating appropriate rules (in the context of the discovery process) have been proposed.

This paper looks at the practice of presenting statistics (in particular the methods of selection of those statistics) to viewers of sports such as cricket and the lessons that may be learned. Although cricket is used as the example in this paper, many data rich activities (including any sport in which the statistics collected are sufficiently large and multidimensional, such as baseball, basketball or (American or Australian Rules) football) may be substituted. The effect on interestingness of behaviour over time may also be relevant to some aspects of activities such as currency market analysis and could bear further investigation. Significantly, some presented statistics are simply the retrieval of information (For example, *How many overs have elapsed since a boundary has been scored?*) and others show trends/correlations through graphical representation. Although important, in this paper we are not concerned with such information retrieval aspects but rather the unsolicited presentation of interesting information.

This paper first summarises the types of data under discussion, showing that relative and absolute thresholds are important to identifying significant events within a match. The interestingness measures already in the literature are briefly reviewed. The cricket perspective is then given, showing how the concept of thresholds and behaviour can add to the meaning of interestingness, and that the concept of anticipation introduces a temporal aspect especially relevant to competitive real-time data mining of this type. A brief summary outlining possible future research pertaining to this work concludes the paper.

2. CRICKET STATISTICS

Watchers of cricket (basketball, baseball ...) would be used to the volume of statistics extracted and presented relating the current game to previous performances. The principal criterion used by the presenters is simple, *would the (average) viewer/listener be interested in this?* This criterion is somewhat recursive but does allow us to examine the presenter's interpretation of what is interesting to see if any general heuristics can be elicited. For those unfamiliar with cricket, Figure 1 gives some examples of the more common types of (non-information retrieval oriented) statistics reported and which require significant searching through (or prior knowledge of) historical cricket statistics.

In general two forms of mined rule are considered in this paper:

- *Static Facts* - commonly used after an event (or the match itself) has occurred.
- *Anticipated Events* - commonly used prior to or during the match.

Sports statistics are generally concerned with performances by players and by teams. These performances can be measured against previous performances by the same player (or team), or against performances by others. Best or worst

1. Gilchrist has just taken more wickets in a match than any other wicket keeper in history;
2. This is close to the longest winning streak ever;
3. Tendulkar has just passed his personal best;
4. Wilkinson needs just two more wickets to become Hampshire's highest wicket taker this season;
5. This is the highest third wicket partnership for England in Adelaide;
6. If Waugh gets his century here then he will have achieved this against every ICC nation.

Figure 1: Examples of Cricket Statistics

performances are both of interest. However, the notion of performance is multifaceted and can be qualified by location, weather, opponents, tournament, and many other variables in a variety of combinations.

One common thread running through many statistics which may be considered interesting is the presence of an event that achieves a threshold value of some type, either independent of the statistic (such as achieving 100 runs or 5 wickets in an innings) or dependent on previous performances (For example, the highest third wicket partnership in an Adelaide test).

Moreover, while these threshold events might be important, the statistic is often of interest before it occurs – there is a measure of anticipation which makes the likelihood of it occurring a factor in the assessment of its interestingness[1]. This aspect is discussed in more detail later.

3. CURRENT INTERESTINGNESS MEASURES

The initial work in data mining concentrated on the uncovering of rules of the form $A \rightarrow B$, where the relative frequency of occurrence of A together with B or the *accuracy* of the rule according to some measure was a gauge of the importance or 'interestingness' of the rule[2]. Early, measures associated with this aspect of interestingness are confidence ($|A \wedge B|/|A|$), coverage ($|A|$), completeness ($|A \wedge B|/|B|$) [1; 5], and support ($|A \wedge B|/N$, where N is the size of the data set) [1; 3].

Interestingness of a rule can be considered to be governed by the size of these measures without reference to the underlying meaning of the rule. There is an implication that the more support and confidence there is for a rule the more interesting it is.

Another aspect of interestingness is how unexpected the event is [5]. A rule that suggests that purchasers of disposable nappies are likely to buy beer as well is, perhaps, more surprising (and thus more interesting) than one suggesting that purchasers of potato chips are likely to buy beer.

Other aspects of interestingness that have been discussed include:

- *actionability* – whether knowing about the rule allows the environment to be changed in a beneficial way [16];
- *neighbourhood* – whether the context in which the rule appears affects its interestingness [3].

One common thread amongst many of these aspects, (such as support, confidence, coverage, etc.) is the static nature they give to the interestingness of rules - i.e. they are applied regardless of the semantics of the rule. In contrast, importance, unexpectedness, actionability and neighbourhood, take some account of the environment in which the data is being considered and the interests of the actor measuring the interest. For further research on the analysis of interestingness measure in data mining, please refer to [5; 7; 15].

4. THE CRICKET PERSPECTIVE

While the response to a direct question (or the graphical display of facts) is clearly information retrieval and the *ad hoc* searching for useful rules in large quantities of historical data is clearly data mining, cricket, in common with a number of domains, blurs the boundary between them. Commonly, the extraction of a piece of information starts with a hunch by an expert (perhaps the result of discussions by the commentators), which is followed by a check on a more or less complete database. It is, at least currently, rare that undirected mining occurs over the database.

Nevertheless, an analysis of the sorts of statistics found to be interesting in cricket shows that there may be some utility in mining and that some of the general characteristics of interestingness described above are applicable. For example, a rule of high confidence and support, such as *the third batsman will score higher than the tenth* can be of interest in determining thresholds, but a performance that follows the rule is not usually of interest – i.e. it is commonly the trait of 'unexpectedness' which is a factor in rules being considered interesting. However the introduction of the idea of a threshold value to measure performance leads to additional aspects to interestingness that need to be discussed.

Another aspect is that of context, both static and temporal. In a static sense statistics such as *most wickets by a Hampshire bowler in the national competition* is an example of the use of neighbourhoods to determine the interesting thresholds against which to measure performance. Temporally, the state of the match and the timing of a rule can be important. For example, the fact that a given batman has never scored a century at Lords has more significance when the batsman is has currently scored 98 than, say, 18. Behaviour and likelihood thus also become important. Note that in some cases rules that might become interesting can be mined in advance – however, the combinatorial effect of all possible outcomes could make this intractable.

4.1 Thresholds

4.1.1 Independent Thresholds

The attainment by a batsman of (multiples of) 50 runs is generally of more interest than other numbers. Considera-

[1]This aspect has been recognised by the sports networks and the use of anticipation is a real consideration in retaining viewers – few people turn off when a batsman is on 99.

[2]The semantics of the implication $\rightarrow$ might be association, cluster description, characterisation, and so on.

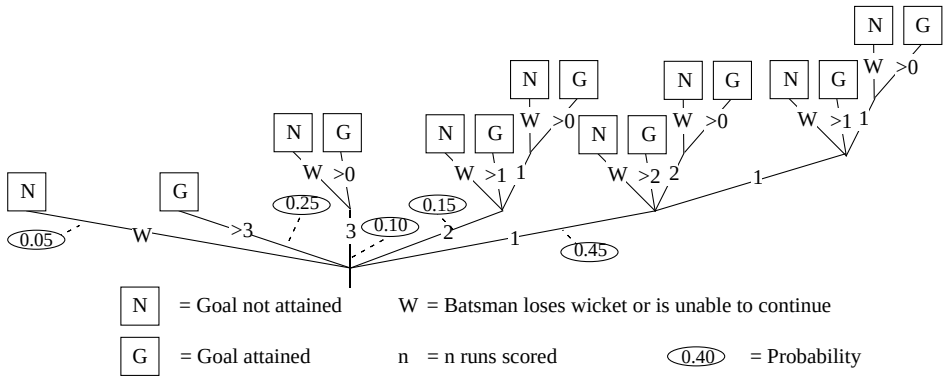

Figure 2: Simple Probability Graph for a Batsman on 96.

tion as to how interesting a given performance is with respect to this number leads to some surprising results. When using a value to measure performance, generally getting 'close enough' either above or below the value is not considered actionable [16], as only exceptionally poor or exceptionally good performances are actionable. In cricket, however, achieving 97 is of (relatively) little interest[3], and 100 is as interesting as 149 in the record books. Achieving the threshold overtakes the accurate measure of the performance in interestingness.

4.1.2 Context Dependent Thresholds

The use of context dependent thresholds are also of interest in proportion to the complexity of the qualification that needs to be applied. For example, *Gilchrist has taken more catches in an innings than any other Australian wicket-keeper* is of more interest than, say, *Bloggs has taken more catches in an innings than any other wicket-keeper in rain-shortened matches at Adelaide oval in the past three years.* One method by which the qualification can be quantified can be by examining the reduction in the number of applicable matches – the larger the reduction, the lower the interest. However, locality can sometimes also be important – New Zealand viewers are likely to be more interested in statistics relating to New Zealand players for example. However, whether or not rules are more or less interesting as their neighbourhood of applicability becomes smaller depends also on the situation and the existence of alternative rules; i.e. sometimes the most interesting rule *at the time* is needed.

4.1.3 Poor Performances – Low Thresholds

Another area of intense interest for sporting statistics relates to under-achievement. Here the lack of achievement of a threshold does become of interest, *a first-ball duck*[4] for example. This may be of more interest if the behaviour is frequent. (Note that while this seems to contradict an earlier statement regarding common behaviour not being of interest, what is frequent here is a rare event. That is, the *set* of performances being considered is rare in comparison with others.)

[3]Unless it impinges on other thresholds, (the highest score, a personal best, etc.) in which it may be of interest for other reasons.

[4]This involves not being able to face a single ball without getting out.

4.2 Time-Varying Interestingness – Accommodating Behaviour

4.2.1 Anticipation

Much of the interest or usefulness of a statistic is predicated on the behaviour of the actors in the game continuing their behaviour. Thus while statistics that report interesting facts are useful, so too are reasonable (i.e. likely) predictions (based on historical performance behaviour). The nature of interestingness thus has a distinctly temporal context to it [11] and an approximate interestingness factor can be calculated as the product of the significance of the event and the likelihood of it occurring.

Some examples of the manner in which anticipation affects the interest in an event that achieves a threshold value are given below.

1. *Independent thresholds.*

   The interestingness of an event that achieves a threshold value increases as the probability of that threshold being attained increases. For example, the interestingness of the attainment of a century in a match starts low, although past performance will indicate non-equal probabilities. As the batsman gets closer to the threshold the interest in the threshold and thus the rule increases. If the batsman loses his wicket before attaining the threshold, the performance generally loses interestingness, unless it satisfies some other rule as a result (perhaps he is making a habit of getting out close to the threshold). Once the threshold is reached, it moves into a different category of interestingness – that of an historical event.

   The interestingness can be modelled using a probability graph using weighted arcs to calculate the likelihood of the event as in Figure 2[5]. Determining the weightings for the graph can be difficult but one simple method would be an analysis of past performances of players in similar circumstances.

2. *Context Dependent Thresholds.*

   In a similar way records such as *the greatest number of catches by an South African wicket-keeper in a test match* can be equalled or broken at any time that

[5]This can be alternatively modelled using a NFSA network.

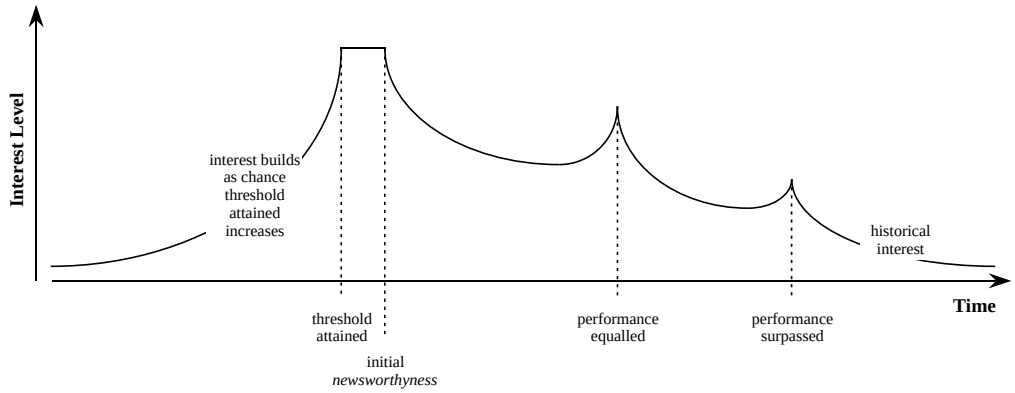

Figure 3: Rise and fall of interest in the attainment of a threshold over time.

South Africa plays a test. This threshold does not become interesting until the wicket-keeper gets close to it with enough of the match left to make it possible that it can be broken. Its interestingness is also affected by how outstanding it is, i.e. the number of times the record has been equalled, or how far is it from some further threshold such as the all-time record.

Factors affecting the interestingness during the match include:

- The relative significance of a player doing as well as he/she has done so far. For example, if the record for a wicket-keeper is 9 catches in a test and a large number of wicket-keepers have achieved, say, 7 catches in a test, the fact that this has been achieved again is less noteworthy.
- The opportunity for significance. For example, how much of the match is left? If a wicket-keeper achieves 8 of the required catches with 5 wickets left the performance is more interesting than if there are 2 wickets left, as the threshold can be surpassed by a larger amount.
- How close the record is to being broken. At the beginning of the match, breaking the record is possible, but unlikely. As the number of catches taken gets closer to the record the interestingness increases.

At first glance, performances which can be classified as winning (or losing) streaks appear to have a different pattern of interestingness. Unlike single match records, the possibility of a continuation of a run of good or bad performances is of interest from the start of a match. However, apart from the change in granularity of the time period from a single match to a sequence of matches, this is similar to the one-match record with regard to what makes it interesting, i.e. how likely the current state is, how much the record can be broken by, how close to the record it is, and whether achieving the record is still possible.

Achievements where a significant set of performances are completed (such as Steve Waugh becoming the only cricketer to achieve centuries in tests against all other test playing countries) appear at first glance to differ from both a single match performance and a winning streak as it does not involve a consecutive series of matches, and the threshold only has interest during a limited set of matches. However, in essence anticipation behaves in the same way as for the previous examples: it is not a single match achievement, and the granularity may be a complete career. The initial low interest builds as the threshold is neared, and is highest immediately before it is reached, and if the sequence of events considered is limited to the matches against countries not yet in the set, the pattern of interest is no different.

4.2.2 *Post-performance variation*

Once records such as these are achieved, their background interestingness as thresholds to consider is not static: interest is decreased a little each time it is equalled, and a lot when it is surpassed. However the background interestingness of an independent threshold does not change.
Figure 3 shows our hypothesis of how interest in a performance builds up as it approaches a threshold, attains it, passes it to become a new threshold, then drops post-performance as it is first equalled and then finally surpassed. Although it is difficult to see how to quantify this variation in the sporting arena (perhaps some analysis of mentions in the press could be used), an investigation of the amount of activity in money markets as values vary around significant thresholds could be illuminating[6] .

5. CONCLUSIONS AND FUTURE DIRECTIONS

Sport statistics are different in nature from other types of data in terms of interestingness measures. Threshold values control their interestingness, and very small differences above or below the threshold have a large effect.
The temporal nature of sport data adds a time dimension to its interestingness. During a performance anticipation increases its interestingness as it gets closer to a threshold. Its enduring post-performance interestingness depends on whether the threshold was reached, and its enduring value can decrease if other performances outclass it in the future.

[6]Certainly, some currency levels are seen as *confidence* levels below which behaviour of the market alters.

This paper represents some initial ideas only about the use of thresholds and anticipation in determining a rule's interestingness. At this stage, much of the area of using anticipation in rule evaluation is available for future research although some specific directions for this work might be as outlined below.

- An investigation of how thresholds operate in sporting and other fields, both with regard to the interestingness of the 'performance', and in how the interestingness of the threshold itself varies over time.
- Given the thresholds, how does this affect the manner in which data is mined? Most routines operate by finding rules first and testing their interestingness afterwards; would it be reasonable to guide mining routines such that they produce interesting rules as a result of a prior knowledge of thresholds?
- The current newsworthiness of various "match-fixing" enquiries suggests that investigating the use of low thresholds to uncover clusters of exceptionally poor performances could be of interest.
- Combining the idea of poor and exceptional performance leads to the idea of a normal performance range for players against which performances can be measured. Future work comparing mining using thresholds with ranges could be productive.

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD International Conference on Management of Data*, volume 22, pages 207–216, Washington DC, USA, 1993. ACM Press.

[2] G. Berger and A. Tuzhilin. Discovering unexpected patterns in temporal data using temporal logic. In O. Etzion, S. Jajodia, and S. Sripada, editors, *Temporal Databases - Research and Practice*, volume 1399 of *Lecture Notes in Computer Science*, pages 281–309. Springer-Verlag, Berlin, 1998.

[3] G. Dong and J. Li. Interestingness of discovered association rules in terms of neighbourhood-based unexpectedness. In X. Wu, R. Kotagiri, and K.B. Korb, editors, *Second Pacific-Asia Conference on Knowledge Discovery and data Mining: Research and Development*, pages 72–86, Melbourne, Australia, 1998. Springer-Verlag.

[4] A.A. Freitas. A multi-criteria approach for the evaluation of rule interestingness. In *International Conference on Data Mining*, pages 7–20, Rio de Janeiro, 1998. WIT Press.

[5] A.A. Freitas. On rule interestingness measures. *Knowledge Based Systems*, 12(5-6):309–315, 1999.

[6] R.J. Hilderman and H.J. Hamilton. Heuristic measures of interestingness. In J.M. Zytkow and J. Rauch, editors, *3rd European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'99)*, volume 1704 of *Lecture Notes in Artificial Intelligence*, pages 232–241, Prague, 1999. Springer.

[7] R.J. Hilderman and H.J. Hamilton. Evaluation of interestingness measures for ranking discovered knowledge. In D.W-L. Cheung, G.J. Williams, and Q. Li, editors, *5th Pacific-Asia Conference on Knowledge Discovery and Data Mining - PAKDD 2001*, volume 2035 of *Lecture Notes in Computer Science*, pages 247–259, Hong Kong, China, 2001. Springer.

[8] 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, Gaithersburg, Maryland, 1994. ACM Press.

[9] W. Klosgen. Efficient discovery of interesting statements in databases. *Journal of Intelligent Information Systems*, (4):53–69, 1995.

[10] B. Liu, W. Hsu, L-F. Mun, and H-Y. Lee. Discovering interesting missing patterns. In H. Lu, H. Motoda, and H. Liu, editors, *First Pacific-Asia Conference on Knowledge Discovery and Data Mining: Techniques and Applications*, pages 138–151, Singapore, 1997. World Scientific.

[11] J.F. Roddick and M. Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods. *IEEE Trans. Knowl. Data Eng.*, 2001.

[12] S. Sahar. Interestingness via what is not interesting. In S. Chaudhuri and D. Madigan, editors, *Fifth International Conference on Knowledge Discovery and Data Mining*, pages 332–336, San Diego, CA, USA, 1999. ACM Press.

[13] S. Sahar and Y. Mansour. Empirical evaluation of interest-level evaluation. In *SPIE - Data Mining and Knowledge Discovery: Theory, Tools and Technology*, volume 3695, pages 63–74, Orlando, FL, USA, 1999. SPIE.

[14] D. Shah, L.V.S. Lakshmanan, K. Ramamritham, and S. Sudarshan. Interestingness and pruning of mined patterns. In K. Shim and R. Srikant, editors, *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, Philadelphia, USA, 1999.

[15] A. Silberschatz and A. Tuzhilin. On subjective measures of interestingness in knowledge discovery. In U.M. Fayyad and R. Uthurusamy, editors, *First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pages 275–281, Montreal, Quebec, Canada, 1995. AAAI Press, Menlo Park, CA, USA.

[16] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems? *IEEE Trans. Knowl. Data Eng.*, 8(6):970–974, 1996.

7. ACKNOWLEDGMENT

This paper was conceived while watching the one-day international cricket match between India and Pakistan at the Adelaide Oval, Australia in January 2000 by one of the authors. The other author was too busy watching the match. We are grateful to the reviewers for their constructive comments on the paper.