eJHI
electronic Journal of Health Informatics
http://www.ejhi.net
2006; Vol 1 (1): e6

Towards Role Based Hypothesis Evaluation for Health Data Mining

Anna Shillabeer, John F. Roddick
School of Informatics and Engineering
Flinders University of South Australia,
Adelaide, SA, Australia

Abstract
Data mining researchers have long been concerned with the application of tools to facilitate and improve data analysis on large, complex data sets. The current challenge is to make data mining and knowledge discovery systems applicable to a wider range of domains, among them health. Early work was performed over transactional, retail based data sets, but the attraction of finding previously unknown knowledge from the ever increasing amounts of data collected from the health domain is an emerging area of interest and specialisation. The problem is finding a solution that is suitably flexible to allow for generalised application whilst being specific enough to provide functionality that caters for the nuances of each role within the domain. The need for a more granular approach to problem solving in other areas of information technology has resulted in the use of role based solutions. This paper discusses the progress to date in developing a role oriented solution to the problem of providing for the diverse requirements of health domain data miners and defining the foundation for determining what constitutes an interesting discovery in an area as complex as health.

Keywords: Data mining, data analysis, health records, role based, medicine

1 Introduction

Health is considered by many working in the data mining field as the most complex and problematic domain yet to be conquered (Cios & Moore 2001; Roddick, Fule, & Graco, 2003). It is a field in which current data mining technologies are often used minimally and are most in need of adaptation if they are to be fully applicable (Imberman & Domanski, 2002; Hagland 2004). The task for data mining researchers is to manipulate data mining technologies to make them applicable to the requirements of the health profession and those who work within it. One of the predominant issues is the difficulty in developing a sufficiently detailed understanding of both medicine and data mining to produce a system using current conventions which addresses this task. Health data mining is thus a collaborative enterprise built upon the common foundation of needing to turn data into information and information into knowledge (Kuonen, 2003; Bresnahan, 1997; Hendrix, 2004).

The major challenge presented by health and medicine is to develop a technology that can provide trusted hypotheses based on measures which can be relied upon in medical health research and applied in a clinical environment (Ashby & Smith, 2002). Due to the broad nature of the health profession and the diversity of professional roles present, it is necessary to provide a system that can be applicable for many roles whilst incorporating their individual requirements. To date, systems have been developed specifically for a particular role or field of investigation and are not transferable to other user roles within the domain. The requirement is thus to provide a health data mining system which can facilitate the production of statistically valid hypotheses which are appropriately targeted to the role of an individual user and which can provide a sound foundation for further research or clinical trials.

2 Terminology

This paper utilises new terminology and definitions for some traditional data mining terms. The literature describes many criticisms of medical data mining (Lord, Gebski & Keech, 2004; Milloy, 1995). Some of this is as a result of the inappropriate use of terminology and we believe that some of these criticisms can be overcome

The *electronic Journal of Health Informatics* is an international journal committed to scholarly excellence and dedicated to the advancement of Health Informatics and information technology in healthcare. ISSN: 1446-4381

through the use of new terms.

To discriminate between the outputs traditionally produced by mining algorithms and those resulting from our proposed hypothesis evaluation tool, we have adopted the term hypothesis rather than rule. In a health context this is of importance as the traditional term 'rule' is misleading. The use of 'hypothesis' ensures that the results of data mining are seen merely as prompts for further investigation, rather than infallible statements about the world. This has been a criticism of data mining in health to date.

This paper also promotes the term 'information strength' to replace 'interestingness' when referring to the heuristics used in quantifying the acceptability of mining results to the user. Our reasoning is twofold. First, we are not simply evaluating and discriminating by the level of interest a user may have in the hypotheses, but by a combination of many factors. Second, while a strong hypothesis is more likely to be interesting, interest is subjective and thus no combination of heuristics, however powerful, can guarantee an interesting result.

3 Role based hypothesis evaluation

The concept of role

The issue of developing a flexible data mining system with the intention of enabling the production of hypotheses with an acceptable information strength based on the requirements of each role within the health domain is the focus of current work by the authors. As discussed elsewhere (Roddick, Fule, & Graco, 2003; Ordonez et.al. 2001) the widely varying requirements and approaches of the many roles in the domain teamed with the complexity of the data, mean that generic (albeit powerful) data mining algorithms generally produce too many and/or inappropriate sets of hypotheses. It is thus essential to firstly understand what determines the value of a hypothesis to different user roles. This knowledge can then be applied to measure the strength of any hypothesis produced as a result of data mining and facilitate an increase in specificity of hypotheses presented to each role and, as a side effect, a reduction in numbers of hypotheses will be achieved. The emphasis in the authors work is therefore not primarily focussed on reducing the numbers of hypotheses but on increasing the value of hypotheses presented as determined by the needs of the role.

It is necessary here to define the term and application of role. Each of us plays many roles in our daily lives as a student, doctor, nurse, teacher, parent, guitar player, amateur photographer etc. etc. Each of these roles will define a level of interest in the world around us, which will be determined by how relevant to each role the world is at any specific time. As each of us has a unique set of roles then each individual will have a corresponding unique set of interests or interest triggers, and different information will appeal to us at different times. Our role is therefore defined by how we measure the value of our interest in information that is presented to us. For example an Oncologist would most likely have been interested in an article on new cancer treatments in a recent Weekend Australian newspaper entitled "Hype or Hope?" (Cornwell, 2005), as it relates to their professional role and, based upon the simplistic evaluation of keywords contained in the article, would have been a member of their set of interests. In contrast an Electrical Engineer would most likely not have had the same level of interest unless they also held the role of cancer sufferer or carer of a cancer sufferer for example. Therefore we can define a role as being a collection of quantifiable interests. Thus the focus of current work is to develop a system that will allow the identification of these interest sets and develop a method for determining how to measure and evaluate how strongly the information is able to trigger interest. Hence it is necessary to provide a quantitative evaluation of interest by evaluating the requirements of the interest set and measuring the applicability of the information or hypothesis to that unique set.

The application of role

The application of a role in determining which hypotheses are of relevance is more complex than simply looking for the presence of keywords. In the field of epidemiology for example the simple presence of the word 'flu' is not sufficient to trigger interest, there needs to be statistical support for the information. In particular it needs to be shown that the incidence of the condition is higher than expected for that population at that time. The difficulty is in determining which heuristics will give an acceptable measure by which we can include or exclude hypotheses for each role. A role based hypothesis evaluation engine has been designed in preference to other options including new heuristics and new heuristic combinations as these fixed solutions can not provide a generic answer to the issue of evaluating the level of interest or information strength for the health domain as a whole given the complexities noted above. It was necessary for an evolution in current thinking in the area and for single-user solutions to be discarded. Foundation work done by the authors was based upon the concept that there could be a generalised system of measuring information strength for the health domain. This has since been discarded. It was realised that there can be no generalised solution to the issue of hypothesis reduction for this domain due to the broad range of roles and requirements to be addressed. Our work has since evolved into developing a system that can incorporate the range of roles without the need for a separate system for each. As it is the role that determines how the strength of the hypothesis is measured, it was a natural step to discriminate the analysis by the role of the current user. Users should be able to analyse and focus their data mining outputs using a single system regardless of their speciality or analytical requirements. Whilst the needs of each role are unique there is also considerable overlap and the heuristics required to determine information strength varies from role to role and also within each user role depending upon the nature of the analysis being undertaken (Kuonen, 2003; Bresnahan, 1997; Tan, Kumar & Srivastava, 2002). A single flexible system provides the best use of resources to accommodate this.

Role based access models have been successfully implemented in a wide range of domains and have demonstrated an ability to overcome issues such as those seen in health including a need for careful management of sensitive information and the need to provide enterprise level security policies which discriminate on a local level based on the role of the user (Cabri et. al., 2004; Ferrailo & Kuhn, 1995). Role based access models have provided a fixed framework from which to apply highly flexible system definition and this concept was the major attraction in the creation of a role based evaluation of hypotheses produced from data mining applications in the health domain.

The following features of role based systems have been adapted and incorporated into the hypothesis evaluation engine proposed in this paper.

- The accommodation of roles that allow for overlapping requirements and measures.
- The ability for a user to be a member of more than one role at a time.
- The ability to enforce constraints on data access where required to accommodate ethical sensitivities.
- The ease of modifying the role of a group of users to accommodate new technologies or methodologies.
- The ability to constrain at a global level and provide flexibility at a local level.

Whilst the choice of information strength heuristics will often fluctuate little across mining runs, some vary greatly, become redundant or require supplementation by new or existing measures. This high level of flexibility is not currently available in documented data mining systems. Algorithms and selected heuristics are applied singly or in a fixed combination with others within a specific system designed for a specific use. The ability to utilise the users role as a means of selecting and applying a significant number of the range of measures in a unique combination as required has not been documented and is believed to be a novel approach to the issues presented here. Support for such an approach has been provided by health domain professionals and domain based publications including the Medical Journal of Australia (MJA) which stated that appropriate statistical methods for analysing trial data are critical and suggested that the statistical methods used in each trial should be specifically tailored to each analysis (Gebski & Keech, 2003). Each specialist field and role has its own requirements and hence the level of flexibility in data mining software packages must be equally flexible, open to adaptation and tailored for the user role at run time.

4 Determining hypothesis strength for each role

Researchers in health and medicine apply many measures to scrutinise and determine the strength of a derived hypothesis (Imberman & Domanski, 2002; Gebski & Keech, 2003). There are presently over 80 information strength measures documented which can be grouped according to the manner in which they dictate, constrain and inform the user (e.g. Bayardo & Agrawal, 1999; Dong & Li, 1998; Freitas, 1998 & 1999; Hilderman & Hamilton, 1999 & 2001; Klemettinen et. al., 1994; Ordonez, Santana & de Braal, 2000; Piatetsky-Shapiro & Matheus, 1994; Sahar, 1999; Shah et. al., 1999; Silberschatz & Tuzhilin,1995 & 1996). An initial aim was to group these measures based on the role that uses them, this was rejected due to the overlaps discussed earlier. Current work groups the measures into classes based upon their type and the characteristic of strength they are able to quantify thus allowing each role to select and apply them as required (for sample see Table 1. Many other heuristics have been documented but those shown clearly demonstrate an ability to address all classes and characteristics). The measures were classified and grouped based upon the description and application of each as provided by the originator or subsequent users. This evaluation resulted in the creation of a tabular description of each measure, a sample from which is shown in Table 2. Only the first paper in which the measure was mentioned is listed unless a different use was discussed elsewhere. Mention of a characteristic or use in a paper is denoted by a cross in the appropriate box in the table, absence of a cross denotes no evidence of use. The authors chose to rely on the documented utilisation of each measure as it represents a verified example of how the measure could be applied and supports our aim to reflect actual needs and proven methods.

While there is no fixed notion of what defines information strength for a particular role in each instance, there is agreement on the characteristics that indicate strength and these can be grouped and measured to quantify their level of expression in hypotheses presented. These classes were verified in discussion with a range of medical professionals during a work in progress workshop in population health (May 19th 2005, Flinders Medical Centre) presented to by one of the authors. In attendance were medical specialists from several fields including cardiology, epidemiology, biostatistics, nursing, clinical research and government and all agreed that the proposed classes defined the qualities they looked for during hypothesis development and testing. It was noted that whilst most of those present could not adequately describe what determined a strong hypothesis in the domain generally, it was felt that the classes presented would provide an acceptable quantification for any role within the domain if applied uniquely for each role or field. It was also proffered that each hypothesis was considered individually depending upon their needs at the time.The heuristics employed were often not selected until the time of evaluation thus strengthening the argument away from a generalised approach. This, in fact, emphasised the need to utilise traditional measures but in a flexible combination for each hypothesis evaluation.

This paper proposes a method of determining hypothesis strength through the application of classes of traditional information strength heuristics. Each class will measure a characteristic or criterion that is currently used to determine the value of information in a hypothesis. These classes are measured uniquely and flexibly for each role rather than utilising a unique but fixed sub set of heuristics for each system. Based upon the values achieved by each heuristic, unqualified mining outputs can be eliminated from presentation thus providing only those hypotheses developed from data mining rules which meet the requirements of the role and adequately represent the desired characteristics.

Strength Test Classes	Information Strength Classes					
	Novelty	Applicability	Relativity	Provability	Understandability	Validity
Statistical						
Support			x			x
Confidence			x			x
Significance						x
Standard Deviation	x					x
P test	x	x	x			
Chi Square						x
Gray & Orlowsky	x					x
Odds ratio		x	x			x
Domain Testing						
Thresholds		x				x
Zoom in	x					
Zoom out	x					
Trends	x	x	x	x		
Sensitivity	x					
Templates						
History Log	x	x				
Inclusive rule template	x	x			x	
Restrictive rule template	x	x			x	
General impressions		x				
Comparative						
Anomaly detection	x					
Difference detection	x					
Longitudinal analysis			x			
Spatial	x	x				
I measure	x	x				x
Distance	x					
Reliable exception	x	x	x			
Peculiarity	x	x				
I variance	x	x				
I Simpson	x	x				
I Total	x	x				
I McIntosh	x	x				
Unexpected						
Shannon measure	x		x			
Freitas surprisingness	x	x				
Dong & Li	x	x				
Evidence Theory						
Probability assignment	x					x
Belief	x					x
Doubt	x					x
Plausibility	x			x		x
Credibility			x			x
Measure of effect						
Projected savings				x		
Tuzhilin interestingness		x		x		
Kamber & Shinghal						
Necessity	x					
Sufficiency	x	x				
Protocols/Standards						
No ambush			x	x		
Consort			x	x		
PH CDM		x		x	x	
Data Formatting						
Canonical forms		x			x	
IDC9		x			x	
SNOMED		x			x	

Table 1. Sample of measure classification.

Source	Measure	Type	Application	Domain
Poulin et. Al, 1998	P	Statistical	Determine degree of difference in results	Medicine
Poulin et. al., 1998	chi2	Statistical	Determine degree of difference in results	Medicine
Brosette et al. 1998	chi2	Statistical	Result comparison	Medicine
Biedl et. al., 2001	Pearson's Correlation	Comparative	Measure of difference	Bio-inf
Biedl et. al., 2001	Euclidian Distance	Comparative	Measure of difference	Bio-inf
Karypis et. al., 2004	Cosine Similarity	Comparative	Measure of similarity of text	Linguistics
Savasere et. al.,1998	Support	Statistical	Probability, Frequency	Retail
Han & Fu, 1995	Confidence	Statistical	Probability, Frequency	Retail
Cios & Moore, 2002	Accuracy	Domain	Determination of class membership	Medicine
Cios & Moore, 2002	Sensitivity	Domain	Measure of ability to find true positives	Medicine
Cios & Moore, 2002	Specificity	Domain	Measure of ability to reject true negatives	Medicine
...	...	...	...	...

Table 2. Sample measure evaluation table.

Six classes or criteria are provided for information strength measurement and each of these may require a number of statistical, comparative or other tests to determine the overall information strength for each criterion. The individual information strength values are then combined to provide a comprehensive measure of information strength for that hypothesis based on the total requirements for the user role. A greater strength suggests a hypothesis that is more likely to be of value to the role that defined the heuristics and their scopes. The criteria for measuring information and hence hypothesis strength are discussed following.

Novelty

Is it unknown in the body of domain knowledge? This is more complex than simply not duplicating existing knowledge or presenting expected patterns. A hypothesis based upon existing knowledge may still be of interest if the strength or pattern of the hypothesis differs sufficiently from that which is expected. For example, medical professionals would reject a hypothesis that proposes that 3.6% of pregnant women develop gestational diabetes mellitus (GDM) as this is known and expected knowledge even though it would have sufficient strength by some traditional measures to warrant further investigation (Stone et. al., 1996). However if a hypothesis were provided which showed that the incidence rate of GDM in a data set was 3.6% but from a data set primarily for a North Asian population then the interest in this may be greater as the rate would be expected to be higher. Hence it is the pattern novelty as a whole which is being evaluated and which thus determines the strength. There are a number of measures that can be applied to quantify the expectedness or similarity of hypotheses to existing knowledge and it may be necessary to test this criterion using several classes of tests to adequately assess the novelty of a hypothesis.

Applicability

Is it relevant to the current user? This infers that either some contextual information is required, or that previous hypotheses are tagged as interesting (or not) so that the system can learn and reference. The definition of applicability (or relevance) is context based and should be maintained on an individual level. An outlier may not be relevant to an epidemiologist as it is not representative of the population but still may potentially be of interest to a clinical specialist or medical researcher and should be tagged for reference by that role. The implication is that any derived hypothesis of sufficient strength produced from a medical data store is potentially valuable to some role in the medical domain. If accepted, this suggests the importance of strength determination at a role based level to ensure that each role sees only hypotheses they are most likely to have an interest in and be able to act upon but that no strong hypothesis is omitted completely from consideration.

Relativity

Is it valid relative to the data from which it originated or a class of object that it describes? Once again the applicability of this criterion is determined by the context within which it is measured. Within epidemiology it is important for hypotheses to be shown to be applicable for a generalised population. Results therefore need be demonstrated to apply across the human race or a definable sub section of it. A recent study published in the MJA discussed a potential link between passive smoking and breast cancer (Elwood & Burton, 2004). Whilst the link was biologically plausible in 1999 it was not deemed to be representative of the female population in an epidemiological sense and hence was not deemed interesting. Further work was done which focussed on the effect of environmental tobacco smoke across the age variable specifically. It is now accepted that there is enough evidence to suggest that passive smoking specifically in the early years of a females' life has a measurable impact upon the incidence of breast cancer later in life. Investigation at a finer granularity resulted in a hypothesis that is accepted as representative of a defined sub section of the population. This suggests that strength should be measured for all applicable classes, not only the most obvious or highest ranking. Hypotheses also need to be shown to be representative of the data set from which it came and there are standardised checklists such as that provided by CONSORT (Consolidated Standards of Reporting Trials) (Lord, Gebski & Keech, 2003 & 2004; Altman et. al., 2001) which are widely used within medical research and should be incorporated into the planning of data mining systems.

Provability

Can it be proven through clinical testing? This reflects the actionability of the outcomes of data mining and incorporates the need to adhere to guidelines such as CONSORT discussed earlier. Whilst there are perceived difficulties in automatically determining what could be tested clinically, there are several requirements which define what the foundations of a clinical hypothesis should be and these should be present in hypotheses derived through data mining also (Lord, Gebski & Keech, 2004). For organisations that adhere to research guidelines, it is important that the prerequisites are met for further work so that the potential for follow up clinical testing is not prevented. This criterion aims to ensure that hypotheses are not rendered inactionable due to the methodology employed for their derivation rather than trying to determine what will be actionable.

Understandability

Can it be understood through appropriate presentation? A hypothesis that cannot be described easily or accurately is of less use. The inability for the human brain to assimilate and perform functions upon large amounts of complex data is the very foundation upon which the field of data mining was based. When presenting hypotheses, this must be given due consideration. An overly complex or lengthy hypothesis may be overlooked in favour of those that can be read and understood quickly. Determining the maximum length of a hypothesis will be an important aspect of our work in this area. Consideration must also be given to domain specific terminology and semantic hierarchies (Ashby & Smith, 2002). This will ensure that hypotheses are presented using uniform, accurate and appropriate terminology. Hypotheses should also be presented via a medium that is accepted as standard by each role or domain. Whilst this is outside of the scope of current work future consideration by those working in the area of visualisation would be warranted.

Validity

Is it statistically valid according to trusted domain measures? As shown in Table 1 there are a wide range of statistical measures available to test these classes and each user role should be able to apply measures to each analysis based on the nature of the analysis and personal experience. This paper argues for the use of role based metrics which are manipulated and utilised according to the individual needs of each user and suggests that pre-defining specific heuristics for statistical validity is redundant and archaic.

5 The hypothesis engine

Human intuition often plays a strong role in determining what is interesting in a health context and many breakthroughs are born out of a serendipitous discovery that is subsequently validated through further research not by statistical validity (Bresnahan, 1997). A final decision on what is interesting may be based on little more than gut feeling and hence the need for flexibility in determining the metrics for inclusion and exclusion of a hypothesis is required to mirror the natural process. The foundation for applying the theories presented here is the development of a hypothesis engine. The engine will provide a flexible means to discriminate hypotheses based upon individual role based requirements. The engine aims to provide a system that will allow the following functionality:

- The flexible application of a wide range of heuristics
- The run time selection and scoping of heuristics
- A role based default heuristic selection
- A means of discriminating hypotheses based on critical and non-critical requirements
- The measurement of information strength based on trusted classes of heuristics.

There are numerous statistical measures and methods which are available to test the six proposed criteria, and many are used in medicine including support, confidence, c^2, PERFEX expert knowledge matching, boolean analyser, probabilistic interestingness, stepwise discrimination analysis, CART decision trees and others (Imberman & Domanski, 2002; Beals, Gross & Harrell, 1999). However not all are critical measures of hypothesis strength. Hence there is a need to define both needs and wants and be able to discriminate on that basis. For example, a confidence of between 80 and 95% may be wanted, however hypotheses with a lower confidence would still be of interest if other heuristics achieve acceptable levels. Conversely in an epidemiological context the incidence of a condition would need to be greater than background levels to be of interest, and anything less than or equal to background levels will not be interesting regardless of other factors. These heuristics are deemed to be critical measures as the scope for acceptance is inflexible.

Worthy of note here is work done by Han, Lakshmanan & Ng (1999) in which they describe the use of constraints with antimonotonicity properties which are a specific type of critical measure. These constraints have inflexible scopes and can be used to cull rules as they are evaluated. For example if the incidence of early onset menopause were being analysed then any record with gender = 'male' could be discarded as this record will never be part of the set in which this condition would occur. Similarly a record

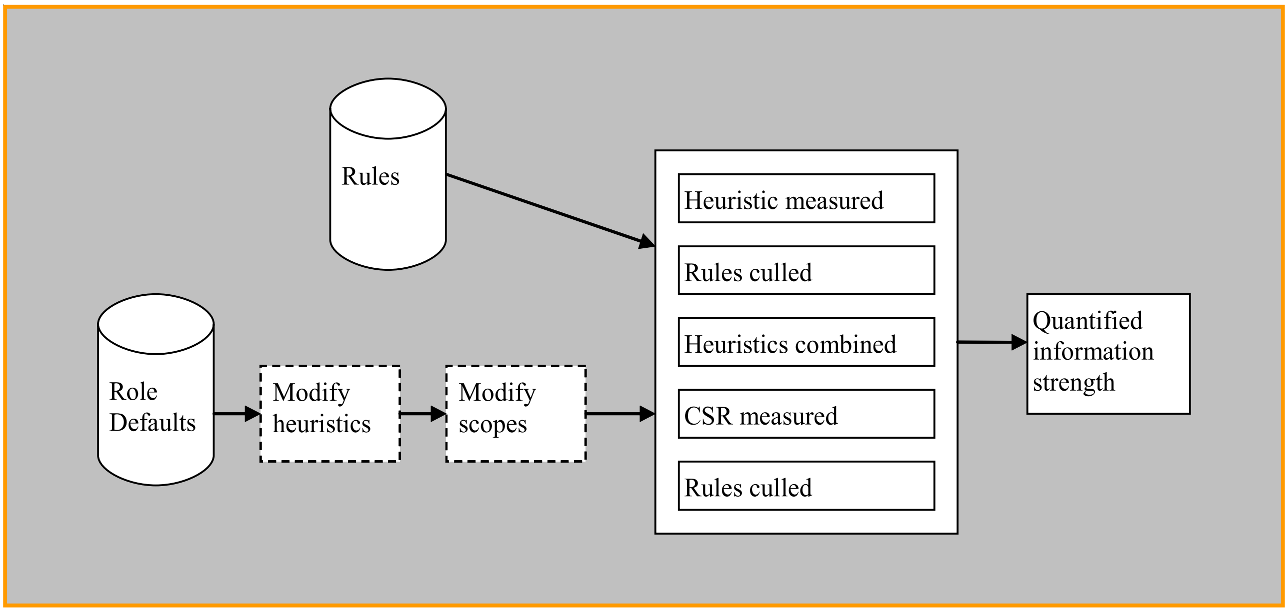

Figure 1: Hypothesis engine overview.

for a person over 50 could be discarded early as again they would be outside of the fixed scope of interest. The epidemiological example given earlier is not antimonotone as adding more incidents or records could allow the value to exceed background levels at some point. The use of this form of constraint would allow each role to set their own rapid culling rules in a subjective manner and thus reduce the resource overhead and rapidly eliminate hypotheses that are not applicable to the user role.

A three phase hypothesis culling process has been defined as shown in Figure 1. The hypotheses are systematically culled as they attempt to propagate upwards through the three levels of measurement from individual heuristics to a class strength rating (CSR) to the quantified information strength value. Any individual heuristic achieving within the role defined scopes would automatically propagate to the next level. A metric that does not achieve a level within scope would be filtered through a switch. If the metric is critical (needed) to determine strength then the switch would cull that hypothesis. If the metric were not critical (only wanted) then the switch would not be activated resulting in that heuristic value and the hypothesis being included in the next level if no other critical heuristics fail for that hypothesis. Whilst it is believed that the switches will have the dual purpose of reducing the numbers and increasing the validity of hypotheses presented we are yet to test how stringently we can do this before the risk of eliminating hypotheses of value becomes untenable and this forms the focus of future work.

6 Conclusion

Data mining has to date been constrained and governed by an algorithm and a fixed set of specific statistical methods and measures. This is not practical in the health domain, and its eradication as the predicate is believed to be a catalyst in increasing the effectiveness of automated data analysis and knowledge discovery to medicine.

As there is little data mining documented for medicine it is difficult to determine the exact number of heuristics that have been tested in a data mining capacity but research suggests that there are a variety of measures which could be successfully applied to measure the strength of hypotheses produced by medical data mining. However, as they are not available together in a single entity it is currently difficult to determine overall hypothesis strength according to all required criteria as discussed in this paper and mining results are often criticised as data dredging for this reason. Hypotheses produced are a result of the specific methods and metrics applied but do not necessarily comply with or measure all required criteria. Some hypotheses potentially fail criteria that are unable to be tested as part of the run-time set, but as they complied with all criteria that could be tested they may be considered strong on that basis. This is deemed to be one of the most important reasons why purely exploratory medical data mining is not yet accepted practice and the development of the hypothesis engine presented in this paper is believed to be a key in overcoming this issue.

References

Altman, D.G., Schultz, K.F., Moher, D., Egger, M., Davido_, F., Elbourne, D., Gotzsche, P., Lang, T.: Consort explanation. Annals of Internal Medicine (2001) 663–694.

Ashby, D., Smith, A.: The Best Medicine? Plus Magazine - Living Mathematics. Issue 22 November 2002. Retrieved March 12 2005 from http://plus.maths.org/issue22/features/medical/index.html

Bayardo Jr, R., Agrawal, R.: Mining the most interesting rules. In Chaudhuri, S., Madigan, D., eds.: Fifth Int. Conf. on KDDM, San Diego, CA, USA, ACM Press (1999) 145–154.

Beals, M., Gross, L., Harrell, S.: Diversity indices: Simpsons (1999). Retrieved September 25 2004 from http://www.tiem.utk.edu/

~gross/bioed/bealsmodules/simpsonDI.html

Biedl, T., Brejova, B., Demaine, E. D., Hamel, A. M., Vinar, T. Optimal arrangement of leaves in the tree representing hierarchical clustering of gene expression data. Technical Report 2001-14. University of Waterloo, ON, Canada.

Bresnahan, J.: Data mining - a delicate operation. CIO Magazine (1997). Issue June 15 1997

Brossette, S. E., Sprague, A. P., Hardin, J. M. et al. Association rules and data mining in hospital infection control and public health surveillance. JAMIA 1998 July; 5 (4): 373-381.

Cabri, G., Ferrari, L., Leonardi, L., Zambonelli, F. Role based approaches for agent development. Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3 (AAMAS'04, July 19-23 2004. New York USA) 1504-1505.

Cios, K.J., Moore, G.W.: Medical data mining and knowledge discovery: Overview of key issues. In: (Cios, 2001). (2001) 1–20.

Cios, K.J., ed.: Medical data mining and knowledge discovery. Volume 60 of Studies in Fuzziness and Soft Computing. Physica-Verlag, New York (2001).

Cios, K.J., Moore, G.W. Uniqueness of medical data mining. Artificial Intelligence in Medicine Journal. Vol26. No. 1-2, September-October 2002 1-24.

Cornwell, J. Hype or hope? The Weekend Australian Magazine. Aug 27-28 2005. 36-41.

Dong, G., Li, J.: Interestingness of discovered association rules in terms of neighbourhood-based unexpectedness. In Wu, X., Kotagiri, R., Korb, K., eds.: Second Pacific-Asia Conf. on KDDM: Research and Development, Melbourne, Australia, Springer-Verlag (1998) 72–86.

Elwood, J., R.C.Burton: Passive smoking and breast cancer: is the evidence for cause now convincing? MJA 181 (2004) 236–237

Ferrailo, D., Kuhn, R. An introduction to role-based access control. NIST/ITL Bulletin December 1995.

Freitas, A.: A multi-criteria approach for the evaluation of rule interestingness. In: Int. Conf. on Data Mining, Rio de Janeiro, WIT Press (1998) 7–20.

Freitas, A.: On rule interestingness measures. *Knowledge Based Systems* Journal: Special Issue 14-16 December 1998 Volume 12, Numbers 5-6 (1999) 309-315.

Gebski, V., Keech, A.: Statistical methods in clinical trials. MJA 178 (2003) 182–184.

Hagland, M.: Data mining. Healthcare Informatics Online (2004). http://www.healthcare-informatics.com/issues/2004/04_04/hagland.htm. Accessed March 28th 2005.

Han, J., Lakshmanan, L.V.S., Ng, R.T. Constraint based multidimensional data mining. Computer 08/1999. 2-6.

Han, J., Fu, Y. Discovery of multiple-level association rules from large databases. In: Umeshwar Dayal, Peter M. D. Gray, Shojiro Nishio (Eds.): VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland. Morgan Kaufmann 1995, 420-431.

Hendrix, M.: "Without research, there is no hope". MJA 181 (2004) 594–595.

Hilderman, R.J., Hamilton, H.J.: Heuristic measures of interestingness. In 3rd European Conf. on Principles and Practice of KDD (PKDD'99). Volume 1704 of Lecture notes in Artificial Intelligence (LNAI), Prague, Springer (1999) 232–241.

Hilderman, R.J., Hamilton, H.J.: Evaluation of interestingness measures for ranking discovered knowledge. In 5th PAKDD 2001. Volume 2035 of Lecture Notes in Computer Science (LNCS), Hong Kong, China, Springer (2001) 247–259.

Imberman, S.P., Domanski, B.: Using dependancy/association rules to find indications for computerised tomography in a head trauma dataset. Artificial Intelligence in Medicine 26 (2002), 55-68.

Karypis, G., Han, E., Kumar, V. Multilevel refinement for heirarchical clustering. Retrieved July 26. 2004 from http://glaros.dtc.umn.edu/gkhome/fetch/papers/clrefineTR99.pdf

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., Verkamo, A.: Finding interesting rules from large sets of discovered association rules. In Adam, N., Bhargava, B., Yesha, Y., Eds. Proceedings of the Third International Conference on Information and Knowledge Management, Gaithersburg, Maryland, ACM Press (1994) 401–407.

Kuonen, D.: Challenges in bioinformatics for statistical data miners. Bulletin of the Swiss Statistical Society (2003) 10–17.

Lord, S., Gebski, V., Keech, A.: Multiple analyses in clinical trials: sound science or data dredging? MJA 181 (2004) 452–454.

Milloy, S.: Science without sense. The risky business of public health. Cato Institute,Washington DC (1995).

Ordonez, C., Omiecinski, E., deBraal, L., Santana, C., Ezquerra, N., Taboada, J.,Cooke, C., Krawczynska, E., Garcia, E.: Mining constrained association rules to predict heart disease. In: IEEE International Conference on Data Mining (ICDM'01), San Jose, California. (2001) 433–440.

Ordonez, C., Santana, C.A., de Braal, L.: Discovering interesting association rules in medical data. In Gunopulos, D., Rastogi, R., eds.: ACM SIGMOD Workshop on Research Issues in Special Interest Group on Management of Data Workshop on research issues in Data Mining and Knowledge Discovery 2000 (DMKD 2000), Dallas, Texas (2000) 78–85.

Piatetsky-Shapiro, G., Matheus, C.: The interestingness of deviations. In Fayyad, U.M., Uthurusamy, R., eds: AAAI-94 Workshop on KDD, Seattle,Washington, USA, IEEE Press (1994) 25–36.

Poulin, C., Stein, J., Butt, J. Surveillance of drug overdose deaths using medical examiner data. Chronic Diseases in Canada Vol 19 No. 4 (1998). 177-182.

Roddick, J.F., Rice, S.P.: What's interesting about cricket? - on thresholds and anticipation in discovered rules. SigKDD Explorations 3 (2001) 1–5.

Roddick, J.F., Fule, P., Graco, W.J.: Exploratory medical knowledge discovery: Experiences and issues. SigKDD Explorations 5 (2003) 94–99.

Sahar, S.: Interestingness via what is not interesting. In Chaudhuri, S., Madigan, D., eds.: Fifth Int. Conf. on KDDM, San Diego, CA, USA, ACM Press (1999) 332–336.

Sahar, S., Mansour, Y.: Empirical evaluation of interest-level evaluation. In: The International Society for Optical Engineering Conference on Data Mining And Knowledge Discovery: Theory, Tools and Technology. Volume 3695., Orlando, FL, USA, SPIE (1999) 63–74.

Savasere, A., Omiecinski, E., Navathe, S. Mining for strong negative associations in a large database of customer transactions. Proceedings of the Fourteenth International Conference on Data Engineering February 23-27, 1998, Orlando, Florida, USA. IEEE Computer Society 1998. 494 - 502.

Shah, D., Lakshmanan, L.V.S., Ramamritham, K., Sudarshan, S.: Interestingness and pruning of mined patterns. In: ACM SIGMOD Workshop on Research Issues in DMKD, Philadelphia, USA (1999).

Silberschatz, A., Tuzhilin, A.: On subjective measures of interestingness in knowledge discovery. In Fayyad, U.M., Uthurusamy, R., eds.: First Int. Conf. on KDDM (KDD-95), Montreal, Quebec, Canada, AAAI Press, Menlo Park, CA, USA (1995) 275–281.

Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems? IEEE TKDE 8 (1996) 970–974.

Stone, C., McLachlan, K., Halliday, J., Wein, P., Tippett, C.: Gestational diabetes in Victoria in 1996: incidence, risk factors and outcomes. MJA 177 (2002) 486–491.

Tan, P., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of the Eighth Association of Computer Machinery Special Interest Group on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining July 23 - 26, 2002 Edmonton, Alberta, Canada. 32-41.

Correspondence

Anna Shillabeer, John F. Roddick
School of Informatics and Engineering
Flinders University of South Australia
PO Box 2100
Adelaide SA 5001
Australia

anna.shillabeer@infoeng.flinders.edu.au
roddick@infoeng.flinders.edu.au