On the Arguments Against the Application of Data Mining to Medical Data Analysis

Anna Shillabeer, John F. Roddick and Denise de Vries
School of Informatics and Engineering
Flinders University,
PO Box 2100, Adelaide, South Australia 5001,
Email: {anna.shillabeer, roddick, denise.devries}@infoeng.flinders.edu.au

Abstract

There is a variety of criticisms of medical data mining which has led, in some cases, to the technology being overlooked as a tool. This paper presents a discussion of six of the strongest arguments against the application of data mining to the complex field of human medicine. The aim of the paper is to raise the predominant issues and suggest solutions whilst also opening the issues for further consideration by both medical and information technology communities.

The Arguments

1. Data mining outcomes are seen as generalisations and not verified for medical validity or accuracy. [Elwood and Burton, 2004; Milloy, 1995]

   Medicine is a highly complex domain for which data mining processes were not designed. In many cases they originated in response to changes in commerce or management practices and there was no real need to substantiate results on the basis of protocols or domain knowledge. Medicine has requirements which are outside of the original scope of the technology, and to be applicable to a science which is concerned with critical decision making there is a need to modify the technology to reflect this different environment. Whilst this first argument is a serious issue, it is often borne from misrepresentation of the results of data mining rather than from the process itself. There is a heightened need for careful consideration of the language used when reporting results [Raju, 2003; Mainonald, 1998]. It is possible for the results to be specific but for the language of reporting to generalise the message. For example, the MJA described a case where a mining outcome showed that smoking does not have a direct link with skin cancer [Elwood and Burton, 2004], however the resulting media story reported that smoking is not linked with cancer generally. While a scientific data mining process was applied the language of information presentation was misleading and the resultant reporting was inaccurate and medically invalid. Medicine is especially sensitive to this form of information distortion and the consequences have the potential to be life threatening, politically sensitive, costly and persistent which is rarely the case in other domains.

   There is little to sustain this argument in light of recent work in the field. By the application of suitable statistical methods, evaluation of all results and applying industry accepted standards there is no reason to believe that data mining cannot provide effective validation and accuracy checking processes [Shillabeer and Roddick, 2006; Gebski and Keech, 2003]. Three steps have been suggested to safeguard against this particular criticism [Smith and Ebrahim, 2002].

   1. Results should not be published on the basis of correlation alone.
   2. An explanation should be provided with the results to provide clarification e.g. A definition of the unique quality of the allergen that triggers the alleged immune response.
   3. Results should be replicated, confirmed and documented prior to publication.

   These steps are not part of standard data mining methodologies but are required to be undertaken if the mining of medical data is to overcome criticism, be viewed as ‘good science’ and gain trust in the medical community.

2. Associations are not representative of other similar attributes and do not consider other potential contributors. [Milloy, 1995; Raju, 2003; Smith and Ebrahim, 2002]

   In a medical context, relationships found between one allergen and symptoms must be substantiated through analysis of similar allergens or the same allergen in other temporal, special or demographic instances. If this cannot be shown it suggests that there is not a conclusive argument for cause and effect or that some other catalyst or cause has been missed [Raju, 2003; Smith and Ebrahim, 2002]. Again, data mining was not designed to do this however this should not be a preventive. Methods are available to achieve this where it is important to determine the semantic closeness of results [Shillabeer and Roddick, 2006]. Criticism often focuses on data dredgers who promote results as facts rather than being indicative of a possible scenario requiring further investigation [Raju, 2003]. Where an association is found it is important to compare this with other associations or to apply a clustering algorithm to group semantically and determine where there is similarity or otherwise to other attributes or rules.

3. P-values are set arbitrarily and therefore the results cannot be trusted. [Milloy, 1995; Smith and Ebrahim, 2002]

   The P-value is applied to the statistical testing of a null hypothesis to gauge the probability of the result happening by chance in a total population. Data mining provides a similar function through the use of support and confidence values although these apply only to the data set being mined, where support is the percentage of the data transactions under analysis that hold true for the association, and confidence (a.k.a. conditional probability) is the percentage of data transactions containing a specific attribute value that also contain another specific attribute value. Support and confidence values are thresholds set for reporting purposes.

IDAMAP 2006
and are not p-values, although they are liable to attract the same criticism. P-values, support and confidence may be applied in two ways: to evaluate and discriminate the acceptability of data analysis results for follow-up research and, as a guideline or tool for reducing the number of irrelevant outcomes. Data mining can also be applied in divergent modes; to show what the common patterns in data are, or to show where common patterns are refuted in the data. It is important to always set heuristic thresholds in context of the specific analysis being done and in fact a calculation applied should not be used alone [Shillaber and Roddick, 2006; Gebski and Keech, 2003]. In the medical domain attribute value relationships which occur frequently, and hence have high support and confidence, as well as a low p-value, are likely to be known already and would generally be of little if any interest. This is a major difference between traditional data mining applications, where generally the events which occur most frequently are of the greatest interest and hence have a similar support threshold, and applications in the medical domain where frequency is not a conclusive determinant in defining the usefulness, validity or applicability of results and hence may require varying threshold values.

4. Associations between attributes are dependent upon the data set being analysed and are not representative. [Raju, 2003; Smith and Ebrahim, 2002]

There is often a poor approach to the collection and description of data sources and samples which is not consistent with the process of data mining or other scientific methodologies [Milloy, 1995; Maimonald, 1998]. For results to be accepted the data source should be from an identifiable population with defined characteristics e.g. location, demographics, and proportions [Smith and Ebrahim, 2002]. In a clinical research setting this is overcome by the use of protocols and guidelines to ensure that results are representative and able to be replicated. One such protocol is CONSORT which is used globally by medical researchers and is endorsed by a number of prominent journals.

Data mining provides validation through the application of tools such as artificial intelligence and neural nets to the knowledge mining step to sample the data, provide outcomes then automatically test them on the whole data source to show that the outcome holds true for all available data not just one small subset [Smith and Ebrahim, 2002]. Data mining is a highly intensive machine process which utilises huge processing power, memory and time. Data sampling is often used as an initial step to reduce these constraints but correct utilisation may help to overcome this criticism also.

5. Data mining is simply a desperate search for something interesting without knowing what to look for. [Milloy, 1995; Smith and Ebrahim, 2002]

Exploratory mining, which is not constrained by user expectations, can uncover unexpected or unknown knowledge with wide reaching benefit and can be utilised to review and extend current medical knowledge. With the wealth of data being produced daily in the medical field the argument that it should not be used in an exploratory fashion to at least note important changes in data patterns demonstrates a misunderstanding of the potential value held therein. It is argued by some [Maimonald, 1998; Smith and Ebrahim, 2002; Shillaber and Roddick, 2006] that it can be beneficial to look simply for something interesting rather than make an assumption about what is present in the data as if we only ever look for what is known we will potentially never find anything new and progress cannot be made. Provided this is a result of a scientific process then further mining or clinical trials can be undertaken for evidence to substantiate the initial findings. This criticism is only valid where the search is for anything interesting even if only minimally and where there is little or no validation.

6. Data mining displaces research and testing and presents results as facts requiring no further justification. [Milloy, 1995]

Contrary to the criticism, data mining in medicine is generally viewed as an efficient tool for enhancing the work done in the field rather than as a replacement for it [Maimonald, 1998]. Its value is seen as a process of automated serendipity that stimulates and supports testing rather than replaces it. When considering the use of mining outcomes there are two questions often asked; is this result representative of what has been recorded over time?, and can the analysis outcome be verified through real world application? [Raju, 2003; Smith and Ebrahim, 2002]. Whilst the first can be answered with some conviction by data mining the second requires clinical input and hence the process of providing trusted knowledge from data requires a collaborative effort by automated and clinical processes. When we consider that time from hypothesis to application of new knowledge is often measured in decades we should feel compelled to find new knowledge as quickly as possible and data mining offers the ideal tool for this.

Conclusion
This paper has presented six common criticisms of medical data mining in an effort to demonstrate that as technologists we need to be aware of the social environment in which we work and to give a suggestion of the importance of continuing to work on making the technology applicable to this complex domain. We should not be disheartened by the criticism which surrounds the field in which we work but should take the criticisms on board, work with them and provide an outcome which is beyond reasonable reproach.

References