Methods for Interpretation of Data in Medical Informatics

Boris Mirkin

Department of Data Analysis and Machine Intelligence, National Research University Higher School of Economics, 11 Pokrowski Boulevard, 109028, Moscow RF
Department of Computer Science, Birkbeck University of London, Malet Street, WC1E 7HX, London UK
mirkin@dcs.bbk.ac.uk

Abstract. An outline of a few methods in an emerging field of data analysis, "data interpretation", is given as pertaining to medical informatics and being parts of a general interpretation issue. Specifically, the following subjects are covered: measuring correlation between categories, conceptual clustering, and generalization and interpretation of empirically derived concepts in taxonomies. It will be shown that all of these can be put as parts of the same inquiry.

Keywords: data analysis, association between categories, clustering, hierarchical ontology, taxonomy, computational interpretation.

1 Introduction

In spite of the fact that medical informatics is one of the fastest growing areas both in research and in practice, as of this moment, there is no well developed system for the medical informatics domain. However, a number of focus areas are of interest to medical informatics:

- patient-centered systems: medical records and images;
- patient safety: error prevention and handling;
- clinical research informatics including new drugs and treatment methods;
- healthcare organization and administration;
- knowledge organization, updating and use.

So far most efforts and results have been related to the personal health support systems. However, each of the subjects is important in the health related efforts and can benefit significantly of informatics tools. Moreover, one cannot help but see the medical informatics as a pathfinder, a leader, in such computer-intensive areas of current interest as knowledge organization, updating, and use (see SNOMED CT ontologies development [12] and related efforts).

Currently, the issues of organization and maintenance of e-records are of urgent priority in medical informatics. Possibly, even more urgent are matters of reorganization of health services such as developing classifications of diseases and disorders

R. Kountchev & B. Iantovics (Eds.): *Adv. in Intell. Anal. of Med. Data & Decis.*, SCI 473, pp. 15–22.
DOI: 10.1007/978-3-319-00029-9_2

matching the common treatment practices. Yet there is a permanent need in automating of all aspects of data interpretation, which will become much apparent after the organizational issues have been addressed.

These are the subject of this presentation. Data of a set of patients may comprise tables, texts, and images. This paper refers mostly to the tabular data format corresponding to results of various tests over a set of patients, and, further down, to the author's attempts at developing methods for data interpretation. The current level of digitalization leads to growing popularity of exploratory data analysis and data mining approaches oriented towards finding patterns in data rather than testing hypotheses; the latter are prevailing in classical statistics frameworks still dominating many areas of the medical informatics discourse. Yet finding patterns is just an intermediate goal, the real challenge lies in developing data analysis methods in such a way that the result can be formulated in a way that a medical practitioner may find acceptable, understandable and reasonable. This is the niche I like to focus at.

I am going to present here a few data analysis methods oriented towards data interpretation issues:

(a) measuring correlation between categories,
(b) conceptual clustering and
(c) generalization and interpretation of empirically derived concepts in taxonomies.

It will be apparent in the end that these three are not as diverse as they seem to be. In fact, they are parts of the same inquiry.

2 Exploring Correlation between Categories: Interpretation Versus Statistics

This subject is of finding those sets of categories that are most correlated with each other. A simplest would be finding just pairs of correlated categories. First of all, I'd like to bring in an example showing the difference between the mathematical statistics and data interpretation approaches. A case for mathematical statistics: a lung cancer sufferer making a claim against an industrial company plant that they are responsible for the condition. To support their claim, the lung cancer sufferer's team refer to a statistical table (in the left part of Table 1).This table brings forward statistical testing of the plant company claim that the proportions of the condition near the plant and faraway from it, 0.05 and 0.03 on the sample, differ only because of the sampling bias and are, in fact, equal in the population. A two-sided z-test, like that in [4], would show that, under the assumption that the sampling has been random and independent, the hypothesis that the proportions are equal should be rejected, at 95% confidence level. Data analysis relates to a very different data and problem setting. The data comes from a database which has been collected from various sources, not necessarily independent or similar. There are many features in the database of which those mentioned in Table 1 could be just a couple. Moreover, the data may be much less balanced than in a goal-oriented sample. This is the case of data on the right in Table 1: only 60 cases from near the plant are in the dataset while the number of far-away-from-plant cases remains a thousand. Because of such a bias in the sample, the very

same z-test now decidedly supports the idea that the hypothesis that the proportions of the condition on the sample are the same cannot be rejected anymore, even in spite of the fact that they remain very much the same: 0.05 and 0.03. This is because the near-plant dwellers sample size is greatly reduced here so that the uncertainty of the situation increases.

Table 1. An illustrative example of contingency data in health statistics: (a) testing proportions, on the left side; (b) as happens in a data base, on the right side

	Classical statistics case			Data interpretation case		
Residence	No LC	LC	Total	No LC	LC	Total
Near plant	950	50	1000	57	3	60
Far from plant	970	30	1000	970	30	1000
Total	1920	80	2000	1027	33	1060

In contrast, the data interpretation view pays no attention to the classical mathematical statistics cause (except sometimes for the lip service only). The goal here is to capture the extent of correlation on the sample, rather than to see how the sample differs from the population – the latter is of no concern at all. The conditional probabilities, like those mentioned, 0.05 and 0.03 could be a good choice sometimes. Yet they can be used only in the case at which one subsample is compared to the other. A more universal measure has been proposed by the founding father of statistics A. Quetelet almost 200 years ago. Quetelet index compares the conditional probability of the event l at a given category k, with that on the entire set, not at a different subsample ([8,9]:

$$q(l/k) = \frac{P(l/k) - P(l)}{P(l)} \tag{1}$$

That is, Quetelet index expresses correlation between categories k and l as the relative change in the probability of l when k is taken into account. In our case, q(LC/Near_Plant) = 3*1060/ (33*60) -1 = 0.606. That means that living near the plant increases the chances of acquiring LC by 60.6% - this should be taken into account in the court whatever considerations of the statistical significance are! (As one can notice, both parts of Table 1 are subject to this interpretation device, not just the part on the left!)

It appears, the average Quetelet index, that is, the sum of $q(l/k)$ weighted by their probabilities, $P(k,l)$, coincides with the value of the well-known Pearson's chi-square coefficient which is widely used for assessing statistical independence, but not association, between categorical features (e.g., Daniel 1998 [1]). This sheds a different light over the Pearson's coefficient – that is an association measure, after all – and this is exactly the criterion for deriving a decision classification tree in some packages such as SPSS. A similar meaning can be assigned to other popular association measures such as Gini index.

3 Hierarchical Grouping: Conceptual Clustering

Hierarchical grouping with conceptual clustering was developed in 80es and recently has enjoyed some revival due to the emergence of ontologies and other conceptual structures (see, for example, Fanizzi et al. 2009 [2]). Our experience is based on the original developments in Russia in 80es for the analysis of data of large-scale sociological and health related surveys [7].

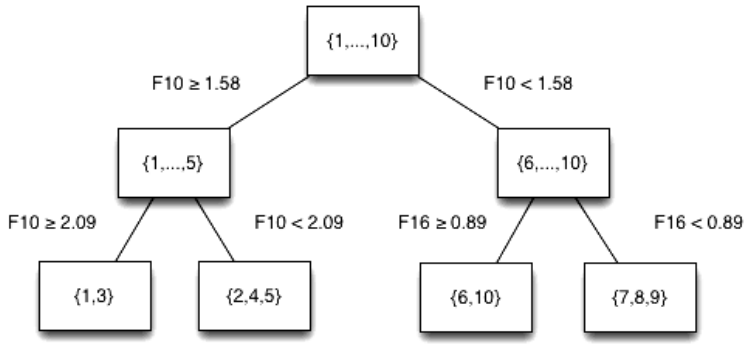

Fig. 1. An illustrative example of a conceptual grouping hierarchy

The result of application of a hierarchical grouping algorithm can be represented by a hierarchy resembling that of a decision tree [6]. Yet it is built automatically by sequential divisions of clusters, starting from the entire dataset, over features from a specified subset according to a criterion that is much similar to those used in clustering. Yet, in contrast to the classical cluster analysis, the clusters are split not over a multidimensional distance between entities but rather over just one of the features. If the feature x is quantitative, then the two split parts correspond to predicates "x>a" and "x≤a" for some feature value a. For a categorical feature, the split parts correspond to predicates "x=a" and "x≠a" for a category a. The algorithm tests all the candidate clusters and all the candidate features and chooses the split maximizing the summary association of that with all the features or, equivalently, the Ward's distance between the split parts' centroids [Mirkin 2011]. The obtained conceptual tree is much intuitive and, also, serves as an informative features selector (those actually used in the splits). The association of the hierarchic partition with the features is measured with the so-called correlation ratio, for quantitative features, or the Pearson chi square association coefficient, for categorical features [9]. The latter is to be modified to Gini coefficient depending on the data normalization, to keep the mathematical equivalence of the criterion to the so-called quadratic error criterion of k-means and similar clustering approaches. This is based on representation of the categories by the corresponding dummy variables with a follow-up standardization of them [8, 9].

In a large-scale survey conducted at Novosibirsk area (Russia) in early 80-es with regard to pneumonia, tuberculosis and other respiratory diseases, more than a dozen altogether, P. Rostovtsev and I built a hierarchical classification of the sample of more than 50000 individuals over the respiratory diseases and related features to find a final

conceptual clustering respiratory disease partition of about 20 clusters/disease types [11]. This partition was further used to find those features of the individuals' conditions that have been most correlated with it. The medical researchers were thinking of alcohol consumption and smoking as the two most important risk factors. In fact, the found partition of the individuals over respiratory conditions had no correlation with these whatsoever, which was very unfortunate because our findings could not be published at that time as being at odds with the dominating paradigm. Instead, we found two other features: "bad housing" and "the same disease in the family", as the real risk factors. The Quetelet coefficient for the former was about 600%.

4 Interpretation of Clusters over a Hierarchical Ontology of the Domain

Hierarchical ontologies, or taxonomies, are currently becoming a major format for computationally handling, maintaining and updating knowledge. A very recent international effort is being resulted in a set of hierarchical ontologies for the medicine SNOMED CT [12]. In fact, this is the very first example of the concept of ontology being developed as a device for practical purposes.

A hierarchical ontology is a set of concepts related by a tree-like hierarchical relation such as "A is a B" or "A is part of B". Of course, ontology of a domain may contain a rather small number of the domain concepts while many others, especially those new ones, remain out of the tree. The concept of ontology is much relevant to the medicine domains because it can encompass the mechanism of a disease and related disorders. The medical diagnostics process can frequently be put in terms of a decision tree related to a hierarchical ontology. The Manual [6] is an example of such an approach applied in the mental health domain.

Therefore, a problem of interpretation of concepts -"outsiders" in terms of the "insider" concepts emerges. Take, for instance, the International Association for Computing Machinery (ACM) classification of computing subjects – a hierarchical four-layer taxonomy of the computing world ACM-CCS. I realize that this may be considered as somewhat far from the medicine, but at this moment I have no application of the approach to be presented in the medical domain.

A recently emerged concept P, say "intuitionist programming", does not belong to the current ACM_CCS. To interpret that in terms of ACM_CCS take a look through a search engine like Yahoo! (because Yahoo was so much research-friendly) to find a profile of P.

Fuzzy profile of P (illustrative):

F.1 Computation by abstract devices - 0.60
F.3 Logics and meaning of programs - 0.60
F.4 Mathematical logic and formal languages - 0.50
D.1 Programming languages - 0.17.

(A fuzzy set, unconventionally normed in Euclidean metric so that the squares of the membership values sum to unity, because of another development by the author and S. Nascimento.)

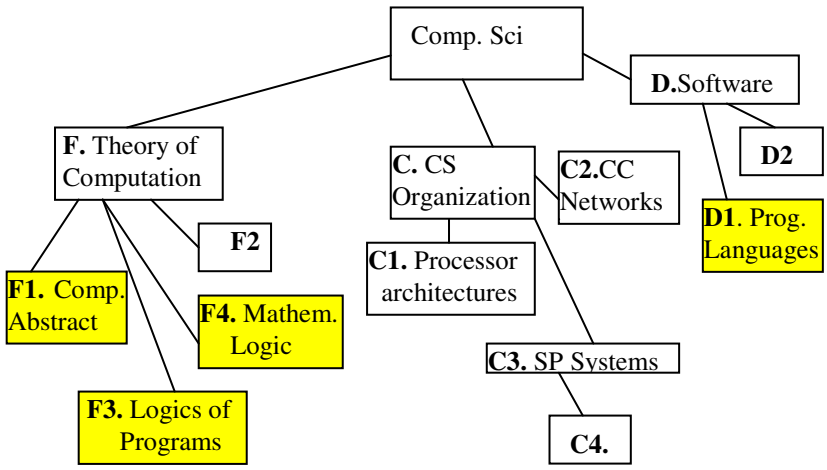

Fig. 2. A fragment of ACM-CCS taxonomy, along with the contents of a fuzzytopic set mapped to it (yellowish)

Mapped to ACM-CCS taxonomy as is (see Fig. 2), the contents of the fuzzy profile can be looked at through the taxonomy structure. Yet, when the contents counts a dozen or more topics well dispersed through the taxonomy tree, the mapping has some obvious drawbacks as being: (a) fragmentary, (b) not scalable, and (c) not quite cognition-friendly.

This is why we propose to interpret such a profile (fuzzy topic set) by lifting it to higher ranks of the hierarchy to minimize the number of subjects it embraces (Fig. 3). However, the lifting may make apparent some discrepancies, namely, gaps and offshoots. Therefore, the lifting penalty function should involve three types of elements: the "head subjects", the "gaps" and the "offshoots" so that their total, appropriately weighted, should be minimized at the interpretable result.

An algorithm, PARL, has been developed for optimally lifting a fuzzy topic set over a hierarchical ontology by recursively moving from the leaves to the root [10]. At each tree node, the algorithm specifies parsimonious events according to each of the two different scenarios: (a) the head subject has been inherited from the node's parent; (b) the head subject has not been inherited from the node's parent. The parsimony criterion is but an operational expression of the celebrated Occam's Razor principle of simplicity. To make the choice of the weights of different elements of the optimal scenario meaningful in a substantive domain, as many as possible concepts should be interpreted via the lifting process so that the probabilities of "gain" and "loss" of head subjects could be derived for the nodes. Then the "maximum parsimony" criterion can be changed for a "maximum likelihood" criterion at which the weights are defined by the maximum likelihood principle.

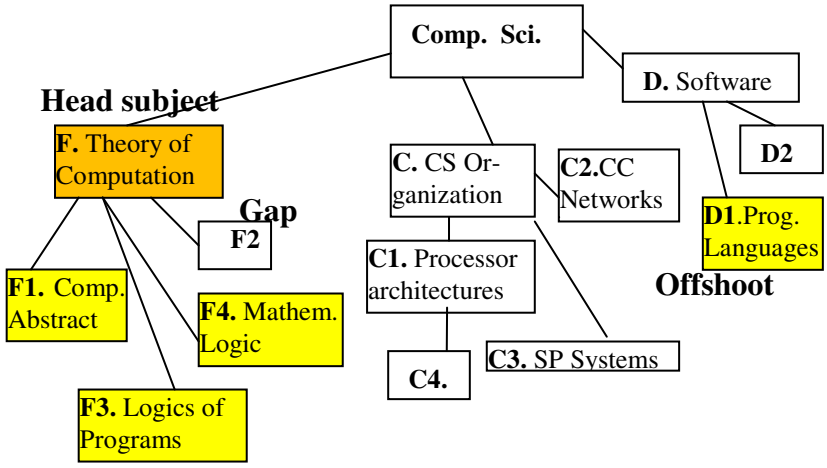

Fig. 3. Interpretation of the topic set by lifting it to "Head subject" F. Theory of computation (highlighted by a darker filling), with the price of having a "gap", F2, and an "offshoot", D1.

5 Conclusion

The discipline of computationally handling both data and knowledge is emerging as driven, to a large extent, by the medical informatics needs. The models and methods for interpretation of various patterns and facts will be an integral part to it. A few topics I just outlined are related quite closely: a topic set to be interpreted by lifting in a hierarchical ontology (section 4) can be derived with a conceptual clustering approach (section 3) which itself heavily relies on the ways for scoring category-to-category correlation (section 2).

References

1. Daniel, L.G.: Statistical Significance Testing: A Historical Overview of Misuse and Misinterpretation with Implications for the Editorial Policies of Educational Journals. Research in the Schools 5(2), 23–32 (1998)
2. Fanizzi, N., d'Amato, C., Esposito, F.: Metric-based stochastic conceptual clustering for ontologies. Information Systems 34(8), 792–806 (2009)
3. García, M.M., Allones, J.L.I., Hernández, D.M., Taboada Iglesias, M.J.: Semantic similarity-based alignment between clinical archetypes and SNOMED CT: An application to observations. International Journal of Medical Informatics (2012) (Available online March 13, 2012)
4. Joosse, S.A.: Two-proportion Z-test calculator (2011), `http://in-silico.net/statistics/ztest`
5. Ludwick, D.A., Doucette, J.: Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries. International Journal of Medical Informatics 78(1), 22–31 (2009)
6. Manual of Mental Disorders (DSM-IV-TR), American Psychiatric Association (2000)

7. Mirkin, B.: Grouping in Socio-Economic Research. Finansy I Statistika Publishers, Moscow (1985) (in Russian)
8. Mirkin, B.: Eleven ways to look at the chi-squared coefficient for contingency tables. The American Statistician 55(2), 111–120 (2001)
9. Mirkin, B.: Core Concepts in Data Analysis: Summarization, Correlation, Visualization. Springer, London (2011)
10. Mirkin, B., Nascimento, S., Fenner, T., Felizardo, R.: How to Visualize a Crisp or Fuzzy Topic Set over a Taxonomy. In: Kuznetsov, S.O., Mandal, D.P., Kundu, M.K., Pal, S.K. (eds.) PReMI 2011. LNCS, vol. 6744, pp. 3–12. Springer, Heidelberg (2011)
11. Rostovtsev, P.S., Mirkin, B.G.: Hierarchical grouping in socio-economic research. In: Mirkin (ed.) [7], Section 5.4, pp. 126–133 (1985)
12. SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) (2012), `http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html` (visited May 27, 2012)