

# Analysis of Primary Care Data

Ying-Lie O<sup>1</sup>

**Abstract.** Intelligent data analysis methods have been used to determine the knowledge models for medical decision support. In primary care all related data are entered in the computer-based patient record (CPR) as events in the journal. The overall characterisation of the care provision is based on patient groups with specific healthcare related conditions and needs.

The development consists of the following steps: problem formulation, database configuration, and data analysis. The features are chosen using a heuristic strategy: initially based on domain knowledge, and then the contribution of the remaining attributes is tested.

The data-set for analysis is count-based. The patient groups are obtained using a modified nearest neighbour cluster analysis method. The proposed approach is mainly data-driven. Only a very limited domain knowledge has been used for the initial selection of features, correction of outliers, and interpretation of the results.

## 1 INTRODUCTION

Intelligent data analysis methods [3, 7] have been used to determine the knowledge models for medical decision support. Most of the use concern specific health conditions [4, 8, 1], such as chronic diseases, or critical care. The required information is generally extracted from data-sets that contain cases. Each case consists of attributes with values that represents a possible condition associated to the disease.

A computer-based patient record (CPR) supports the care provision as a journal of events. The lack of standards in vocabulary, incompleteness, and inaccuracies limits the overall analysis of the data. The composition of a data-set for analysis requires well designed processing.

The analysis methods are bound by the inexact nature of the data. This means that “soft” methods prevail above logic methods. Appropriate methods would be association rules from data mining and cluster analysis from pattern recognition, and for a limited number of variables also regression, interpolation, and neural networks.

## 2 PROBLEM STATEMENT

The provision of primary care is generally provided in local health care centres or out-patient clinics of regional hospitals. These health care centres are typically staffed by GPs (general practitioners) and practice nurses.

The typical setting of routine primary care is the current visit and possible follow-up visits. All related data are entered in the CPR as events in the *journal*. The journal is a layered composition of care activities, short notes, diagnostic tests, and other data. The main entry

consists of activity types and further reference to detailed information. Primary patient characteristics are stored in the patient identification record. To allow analysis, the layered event-based structure must be converted to a data-set in a single layered structure or *universal relation* as opposed to the highly normalised database.

In order to provide proper health care, an overall characterisation of the required care provision is needed. The characterisation is based on patient groups with specific health care related conditions and needs. Each group is distinguished by health care related features, some of which specifically determine outcomes. All features are specified by variables that are attributes of the *item-set*.

In a top-down approach, first of all the most general patient groups that can be directly determined from the main entries of the journal are determined. Analysis of more detailed information not necessarily yield a hierarchical structure, because the structure of the data may be different from the functional association. For instance, a disease-based item-set consists of all attributes from different levels of detail related to the disease.

## 3 OVERALL CHARACTERISATION OF PATIENT GROUPS

The overall characterisation is determined from the main entries of the journal in conjunction with primary patient characteristics in the patient identification record.

In this study, the posed question is: “What is the amount of provided care activities according to general patient characteristics?”

The development consists of the following steps: problem formulation, database configuration, and data analysis.

### 3.1 Problem Formulation

The above posed question is a common data mining problem [2]. With respect to the covered data, the *problem formulation* encompasses

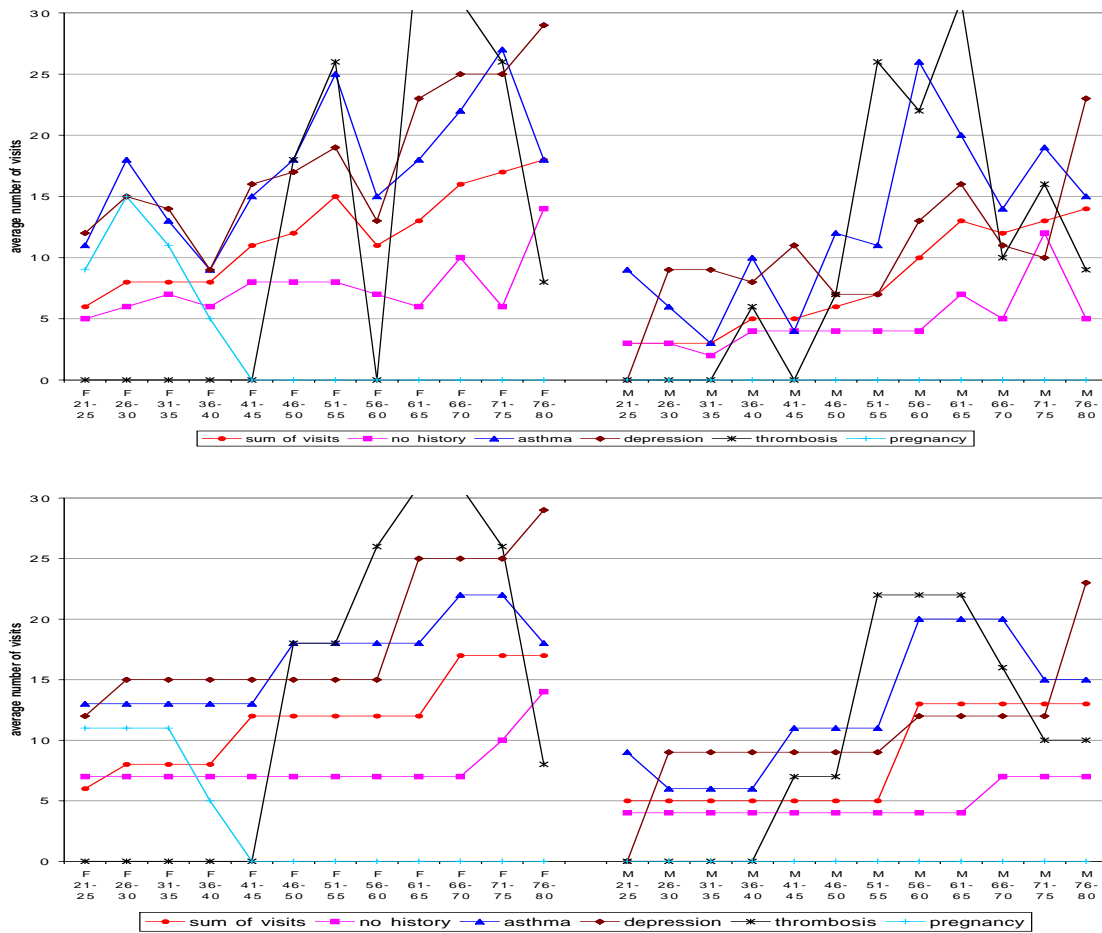
- *Description* divides most of the data into a limited number of large partitions that gives the dominant concepts for feature selection.
- *Grouping* divides most of the data into a reasonable number of partitions with clear concepts that specifies the features.

The selected features is a subset of available attributes that directly contribute to health care activities:  $\{age, gender, history, activitytype\}$ . In this item-set, *activitytype* is the *outcome*, and all other variables are considered to contribute to this outcome.

The problem is formulated as follows: find the patient groups based on the association  $\{age, gender, history, activitytype\} \rightarrow \#activities$ , where # stands for the number of counted instances.

---

<sup>1</sup> Julius Centre for General Practice and Patient-Oriented Research, University Medical Centre Utrecht, the Netherlands, email: y.o@jc.azu.nl



**Figure 1.** Upper: Data-set containing average numbers of visits for different history conditions in age groups of 5 years for females (left) and males (right). Lower: Analysed data-set depicting the clusters of patient groups.

Instead of an attribute set-based approach, a *count-based* solution is applied by replacement of the above association by the number of counted instances in the item-set  $\{age, gender, history, activitytype\}$ .

The description problem regards a summary of occurrences in the item-set and the grouping problem concerns the division of into a reasonable number of patient groups.

### 3.2 Database Configuration

Database configuration is the process of making a data-set that is suitable for analysis according to the problem formulation. This includes the definition of the data structure according to the item-set, retrieval, pre-analysis for formatting, correction and preparation of the data, and pre-analyses regarding the choice of attributes. These tasks are mainly database operations using queries and built-in functions.

The selection of the attributes can be performed in two ways: by analysis of the contribution or association of all attributes to the outcome, or a heuristic strategy.

In the *heuristic* strategy, first a hypothesis is posed, then the alternatives are tested. Failing the test implies adaptation of the initial hypothesis. In case of the item-set, first the predefined attributes are

considered, then the contribution of the remaining attributes are analysed. If the contribution is significant, then the attribute is included.

The first action is the creation of a *count-based* table from the original normalised database tables. Preliminary analyses have shown that several attributes do not affect the number of activities. Also, a limited number of activity types is given to a majority of patients characterised by age and gender. The most important activity type are the regular visits. Hence, the item-set is reduced to the desired attributes  $\{age, gender, history, activitytype = visit\}$ .

The genders have significantly different behaviour. There is a functional tendency in the age dependence, but it is disturbed by fluctuations. Therefore, the ages are stratified into bins of 5 years, and ages below 20 and above 80 are excluded because of different dependencies. Age is treated as the running variable, separately for each gender.

It is generally not possible to correct for outliers, unless it is an obvious error such as a value in the midst of zeroes, or the condition “pregnancy” for the gender “male”. Missing values can be corrected by statistical imputation, or nonlinear filtering such as median or opening (closing) of mathematical morphology. Nonlinear filtering removes outliers and smoothes fluctuations. Closing fills ditches related to neighbouring values, while opening cuts peaks.

The description problem regards the distinction between different history conditions. History conditions are generally only registered in case of a chronic disease, or if the patient has been treated. For the grouping problem several different types of history conditions with high numbers of instances are selected:  $history = \{nohistory, asthma, depression, thrombosis, pregnancy\}$ .

Thus, the final data-set in a count-based table containing the number of visits for different history conditions according to the item-set  $\{age = \{bins\ of\ 5\ year\}, gender = \{F, M\}, \{nohistory = \#visits, \dots, pregnancy = \#visits\}\}$ .

The reduction of the amount of original data to the final data-set is shown in Table 1.

**Table 1.** The number of occurrences in each data-set

	original in journal normalised	associated to visits count-based	selected age bins count-based	final data-set count-based
activities	390945	81086	62143	31775
patients	11218	8921	6506	4452

### 3.3 Data Analysis

The pre-analyses in the database configuration part, the selection of features and attributes, and the resulting item-set and data-set can be considered as an answer to the description problem.

Grouping divides the data into partitions that are not necessarily non-overlapping. In this case, the number of clusters are not known beforehand (unsupervised).

Clustering can be performed by association rules [6] or cluster analysis [5]. Analysis reveal a large variety of different history of health conditions, with low fractions of occurrence that are not discriminative to distinguish different clusters.

Therefore, the grouping problem will be based on the transformed data-set  $\#data = f(\#visits, \#patients)$ , where the function  $f$  is the rounded *specific average* of the number of visits  $\#visits$  on the number of patients  $\#patients$  for each bin. The resulting data-set is shown in Figure 1.

Cluster algorithms are based minimising the dissimilarity between items “within” the clusters, and if non-overlapping maximising the dissimilarity “between” clusters. It is common to use a metric as dissimilarity measure. For this purpose, the  $d_1$  metric  $d(x, y) = \sum |x - y|$  will be used. The clustering algorithm is a modified *nearest-neighbour* algorithm that takes into account that the items are values instead of spatial points.

1. Initialisation: Specify the nearest-neighbour threshold  $t$ , typically 1 or 2 times the standard deviation. Select a number of initial clusters points, for instance by taking local maxima.
2. Nearest neighbour:  $\forall$  item  $x_i$  in a cluster, find its nearest neighbour  $x_j$ .
3. Assignment: If  $d(x_i, x_j) < t$ , assign  $x_j$  to the cluster, otherwise assign  $x_j$  to a new cluster.
4. Stopping criteria: If every item has been assigned to a cluster, stop. Else repeat the process and go to step 2.
5. Completion: Remove small or outlier clusters, and assign the items to the clusters of its nearest neighbours if  $d(x_i, x_j) < 3t$ , else assign to a new cluster. If every item has been reassigned, stop, and if desirable assign the median value of the clusters to the items in the cluster.

The resulting clusters are shown in Figure 1. In comparison with the rather disturbed data, the different patient groups are now clearly distinguished including the tendencies.

The overall number of visits increases with age for both genders, women have a higher contribution. If no history is registered, then it is constant with a slight increase for the elderly.

Asthma and depression occur at the whole range of ages, and is increasing with age. Thrombosis occurs at a certain age, has a peak, and then decreases. As can be expected, pregnancy only occurs for females in the reproductive age.

## 4 CONCLUSION

A strategy for the selection of features, and a modified nearest neighbour clustering method for count-based data-sets have been proposed. The approach seems to be appropriate for the analysis CPR data.

The heuristic strategy of first choosing the features based on domain knowledge, and then testing the contribution of the remaining attributes is satisfactory. Only a very limited domain knowledge has been used for the initial selection of features, correction of outliers, and interpretation of the results.

The clustering algorithm gives promising results, its choice is motivated by the failure of the much used association rules method. This is caused by the main property of this data set: large number of variables with relatively small instances. Using a specific average compensates for fluctuations in the data. The nearest-neighbour clustering is robust and able to handle “bad” data and favors the grouping into homogeneous age groups.

This study is a pilot to gain experience in the application of analysis methods on primary care data in particular and CPR data in general. Future work should include the automatic selection of useful features for the item set, inclusion of domain knowledge, and correction for outliers. Also, for each type of data, an adaptive cluster algorithm should be designed.

## ACKNOWLEDGEMENTS

The author is indebted to Dr. M.C. de Bruijne who posed the question, and kindly provided the data and additional domain information.

## REFERENCES

- [1] R. Belazzi, B. Zupan, S. Andreassen, E. Keravnou, W. Horn, C. Larizza, N. Lavrac, X.H. Liu, S. Miksch, and C. Popow, editors. *Intelligent Data Analysis In Medicine and Pharmacology*, 1998.
- [2] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI Press / MIT Press, 1996.
- [3] D.J. Hand, J.N. Kok, and M.R. Berthold, editors. *Advances in Intelligent Data Analysis*, LNCS 1642. Springer, 1999.
- [4] W. Horn, Y. Sharar, S. Lindberg, G. Andreassen, and J. Wyatt, editors. *Artificial Intelligence in Medicine*, LNAI 1620. Springer, 1999.
- [5] A.K. Jain, R.P.W. Duin, and Mao J.-C. Statistical pattern recognition: a review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.
- [6] W.A. Kosters, E. Marchiori, and A.A.J. Oerlemans. Mining clusgters with association rules. In Hand et al. [3], pages 39–50.
- [7] X.H. Liu, P.R. Cohen, and M.R. Berthold, editors. *Advances in Intelligent Data Analysis: Reasoning about data*, LNCS 1280. Springer, 1997.
- [8] Y. Sharar, S. Anand, S. Andreassen, L. Asker, R. Belazzi, W. Horn, E. Keravnou, C. Larizza, N. Lavrac, X.H. Liu, S. Miksch, C. Popow, and B. Zupan, editors. *Intelligent Data Analysis In Medicine and Pharmacology*, 1999.