Analysis of Breast Feeding Data Using Data Mining Methods

Hongxing He[1] **Huidong Jin**[1,2] **Jie Chen**[1] **Damien McAullay**[1]
Jiuyong Li[3] **Tony Fallon**[3]

[1]CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra ACT 2601, Australia.
[2]National ICT Australia (NICTA), Canberra Lab, Canberra, Australia.
[3]University of Southern Queensland, Toowoomba QLD, 4350, Australia.
Email: hongxing.he@csiro.au, huidong.jin@nicta.com.au,jiechen@ieee.org,
Damien.McAullay@csiro.au,jiuyong@usq.edu.au, fallon@usq.edu.au

Abstract

The purpose of this study is to demonstrate the benefit of using common data mining techniques on survey data where statistical analysis is routinely applied. The statistical survey is commonly used to collect quantitative information about an item in a population. Statistical analysis is usually carried out on survey data to test hypothesis. We report in this paper an application of data mining methodologies to breast feeding survey data which have been conducted and analysed by statisticians. The purpose of the research is to study the factors leading to deciding whether or not to breast feed a new born baby. Various data mining methods are applied to the data. Feature or variable selection is conducted to select the most discriminative and least redundant features using an information theory based method and a statistical approach. Decision tree and regression approaches are tested on classification tasks using features selected. Risk pattern mining method is also applied to identify groups with high risk of not breast feeding. The success of data mining in this study suggests that using data mining approaches will be applicable to other similar survey data. The data mining methods, which enable a search for hypotheses, may be used as a complementary survey data analysis tool to traditional statistical analysis.

Keywords: Data Mining, Survey Data, Features Selection, Classification, Association Rule

1 Introduction

Breast feeding is acknowledged by the World Health Organisation (WHO 2001) to be the optimal method of infant feeding. It has been shown in past literature to provide physical and psychological benefits to both mother and child (Kramer & Kakuma 2003). Additionally, there is evidence to suggest that increasing initiation of breastfeeding and breastfeeding duration have environmental benefits, as well as economic benefits, both to health care systems and individual families (Riodan 1997, Smith 2001, Smith et al. 2002).

It is therefore important to study the factors leading to decisions on baby feeding method. A research team at the University of Southern Queensland has conducted research on this issue (Hegney et al. 2003). In the project, all mothers giving birth in the two Toowoomba hospitals between July 10 and November 30 in 2001 were approached to participate in the study prior to being discharged from the hospital. Mothers who decided to participate agreed to fill out a pre-discharge questionnaire prior to discharge or by telephone shortly after discharge. They were then contacted via telephone at three-months and six-months postpartum to complete follow-up surveys. Of the 940 mothers eligible to participate at discharge, 625 (67%) chose to participate. 554 (89%) mothers were able to be contacted at the three-month follow-up. Of the 372 mothers who were breastfeeding at three-month follow-up, 329 (88%) mothers could be contacted at six-month follow-up.

An extensive study has been carried out on the data collected in the project to study the factors which influence the decision on whether or not breast feeding is given by mothers. In the study, detailed uni-variate analyses were carried out to evaluate the role of individual factors on the output variable. Logistic regression has also been applied to the problem (Hegney et al. 2003). We take an alternative data mining approach in this paper. We apply a feature selection module to select the most discriminating and least redundant features from the original feature set. In selecting the feature subset we do not assume any prior domain knowledge; we let the data speak for themselves. The features selected are then used to train a decision tree model or logistical regression to classify the individuals with respect to an output variable. The output variable is used as the class variable of the individual subjects. In this approach, we do not consider features individually, rather we consider feature sets or subsets as a whole that will influence the decision of whether breast feeding or other feeding method will be used. We also apply a risk pattern mining approach to identify groups of mothers who are not likely to breast feed their babies. These rules should be able to help to conduct a targeted education program to promote breast feeding more effectively.

The remainder of this paper is organised as follows. Section 2 discusses the feature selection methods used as a data pre-processing step of data mining. Section 3 explains two types of classification tools used in the study. Section 4 describes the risk pattern mining method. Section 5 presents main results and discussions of the data mining application to the breast feeding data. Section 6 concludes the paper.

2 Feature Selection

Questionnaires often incorporate a large number of questions to capture as much information from respondents as possible. It is important to do so as there may be limited opportunities to gather this information from the respondents, so the study should be over-inclusive rather than exclusive. However, not all of this information is going to be useful when trying to answer a particular question. Some irrelevant or redundant features are likely to be included in the survey design. Therefore, feature selection becomes an important step before the data can be properly analysed . In the current study we consider two different approaches used in our feature selection procedure. One of them is a selection algorithm based on information theory and the other is a statistical approach (Chi-square).

2.1 Information Theory Based Feature Selection

Feature selection methodology based on information theory is a type of *filter* approach (Fleuret 2004, Wang et al. 2004). A filter approach is classifier independent, where relevance of features to the class variable and correlation between features are studied in order to select the most important features. The other type of general approach is *wrapper* (Kohavi & John 1997), which is classifier dependent and various subsets of the original features are compared to identify the best option in terms of the size of feature subset and classification accuracy using a classifier. For our studies, we use the filter approach. The feature selection method does not rely on any classifier. The feature subset selected can then be used to do classification with various classifiers. They can also be used for other data mining tasks, say, clustering Jin et al. (2005) and visualisation Jin et al. (2004).

The information theory based feature selection method uses concepts from *entropy* and *mutual information* (Shannon. 1948, Cover & Thomas. 1991) as a basis for selecting discriminating and non-redundant features. The entropy, measuring uncertainty of a variable, is defined in Equation 1.

$$H(x) = -\sum_{i=1}^{n} P_{x_i} \log P_{x_i} \qquad (1)$$

Where P_{x_i} is the probability of x taking the value x_i. Variable x takes n distinct mutually exclusive values. The Mutual Information (MI) or information gain is defined in Equation 2.

$$\begin{aligned} I(y;x) &\triangleq IG(x \mid y) = H(x) - H(x \mid y) \\ &= H(y) - H(y \mid x) = H(x) + H(y) - H(x,y) \qquad (2) \end{aligned}$$

Where $H(x, y)$ is defined by Equation 3.

$$H(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{x_i,y_j} \log P_{x_i,y_j} \qquad (3)$$

We apply a feature selection algorithm FIEBIT (Feature Inclusion and Exclusion Based on Information Theory) (He et al. 2005), developed recently to select the most discriminating and least redundant features. FIEBIT uses Conditional Mutual Information (CMI) while excluding irrelevant and redundant features according to the comparison among Individual Symmetrical Uncertainty (ISU) and Combined Symmetrical Uncertainty (CSU). The Conditional Mutual Information of y, x_n given x_m can be defined as:

$$\begin{aligned} I(y;x_n|x_m) &= H(y \mid x_n) - H(y \mid x_n, x_m) \\ &= H(y, x_m) - H(x_m) \\ &= -H(y, x_n, x_m) + H(x_n, x_m) \qquad (4) \end{aligned}$$

Furthermore, if we normalise mutual information, we may introduce some symmetric measures. For example, following Yu & Liu (2004), we may use Individual Symmetric Uncertainty (ISU) to describe the correlation between a feature x and class variable y. Basically, it is the mutual information (or information gain) between two variables normalised by the sum of their individual entropy.

$$ISU(x;y) = 2\frac{I(y;x)}{H(x)+H(y)}. \qquad (5)$$

The ISU compensates for mutual information bias toward features with more values and restricts its values to the range [0,1]. In addition, it still treats a pair of features symmetrically.

Similar to the ISU, we can treat feature $x_j \times x_i$ as the domain $x_{j,i}$ to define Combined Symmetric Uncertainty (CSU) with respect to class variable y.

$$CSU(x_j, x_i; y) = 2\frac{I(y; x_j, x_i)}{H(x_j, x_i) + H(y)}. \qquad (6)$$

The feature selection method uses Conditional Mutual Information Maximisation (CMIM) introduced by (Fleuret 2004, Wang et al. 2004) to select the $(k + 1)^{th}$ feature based on Equation 7 when k features have been selected.

$$f(k+1) = \arg\max_{n}(\min_{1\le l\le k} I(y; x_n \mid x_{f(l)})) \qquad (7)$$

The process continues until the desired number of features are selected.

Feature Inclusion and Exclusion Based on Information Theory (FIEBIT) chooses the features with the highest minimum conditional mutual information of the features not selected-so-far. If k features have already been selected, Equation 7 selects the $(k+1)^{th}$ feature.

After each new feature is selected, FIEBIT excludes the redundant features using ISU criteria Yu & Liu (2004). The candidate set then becomes smaller for each step. FIEBIT can therefore efficiently select a near-optimal feature subset without pre-defining the number of features to be selected. The detail of the algorithm implementing FIEBIT can be found in (He et al. 2005).

2.2 Statistical Approach in Feature Selection

We compare our data mining approach to a statistical approach which uses the Chi-square test for selecting the features (Yang & Pedersen 1997). The χ^2 defined by Equation 8 quantitatively measures the relevance of a condition to the outcome.

$$\chi^2 = \sum_{i=1}^{n} \frac{(E_i - O_i)^2}{E_i} \qquad (8)$$

where O_i is the ith actual value while E_i is its expected value. It takes value 0 if the feature has no effect whatsoever on the outcome, which is commonly called the null hypothesis in statistics. In other words, the output variable is independent of the input variable. A large χ^2 value implies a great importance in deciding the value of the output variable by the variable. Therefore, we select the variables which have high χ^2 values with the output variable. In order to overcome the bias of the population selection, we calculate the p-values which indicates the statistical significance of the χ^2 value.

Chi-square is a test of statistical significance for bi-variate tabular analysis. We use the following two criteria to select n_f features to form a selected feature subset. n_f is a user predefined number.

1. The P value is lower than 0.05.

2. Top n_f features in the list sorted by χ^2 in descending order.

The second criterion selects the features unlikely to be independent to the output variable. The first condition guarantees the result to be statistically significant at the level 0.05.

3 Classification

Classification is one of the most popular data mining tasks. It aims to classify subjects automatically by labelling each subject as a class index. All subjects are then divided into distinct classes. For example, in the breast feeding survey data we can use a feature indicating that the mother chooses to breast feed her baby or not as the class variable. The objective of classification is to predict the class variable using descriptive variables automatically. In order to classify automatically, a reliable model needs to be created. There is a learning process to train the model to perform the classification task.

There are generally two types of learning systems for training the model. Supervised learning uses training samples to optimise the parameters in the training model. Unsupervised learning automatically divides the subjects into various classes in such a way that the subjects belonging to the same class are similar to each other and subjects belonging to different classes are dissimilar. Supervised learning is the most popular approach in classification when study samples are available. It is therefore applied in the current study. Classification models may suffer from the over-fitting problem. The model may achieve very high accuracy on the training samples, however, it may perform poorly on generalisation. Therefore, we need some kind of validation method to test its generalisation accuracy. We use *leave-one-out* as a validation method. In the *leave-one-out* approach we use one data record in turn as the test data. All the other data records are used to train the classification model. The trained model is then applied to the single subject, which is not used in the training process. In general, there will be some correctly and wrongly classified subjects after N runs. The average error rates are then calculated on N runs (N is the total number of data records).

3.1 Decision Tree

Decision tree is a popular supervised learning method used in data mining. Decision tree describes a tree structure wherein leaves represent classifications and branches represent conjunctions of features that lead to those classifications. It is easy to visualise a decision tree. It has advantages over other so called black box classification tool, such as neural network, for having more explanatory power. It not only gives the decision on the classification but also presents the reasoning behind the decision. In our breast feeding data application, output variable y is a categorical variable. The decision is made by a classification tree rather than a regression tree where y takes continuous values. As mentioned above, we use the output variable as a class variable. The selected subset of features are used as input variables. We use the binary variable "M3FEEDAT" (Any breast feeding at 3 months postpartum") as a class variable. It takes two possible values; "Breast Feeding" and "Not Breast Feeding".

We use the commonly used C4.5 (Quinlan 1993) software to train and validate our model. C4.5 creates pruned and un-pruned decision trees based on the training data set. The decision trees are then used to predict the class of test data. The attractiveness of the decision model is judged by its prediction accuracy on the test data.

3.2 Generalised Linear Model

We compare decision tree model with the commonly used statistical approach logistic regression. In logistic regression, the dependent variable is a logit, which is the natural log of the odds.

$$logit(P) = \ln(\frac{P}{1-P}) = a + b\mathbf{X} \tag{9}$$

The log odds (logit) is assumed to be linearly related to $\mathbf{X}$, where $\mathbf{X}$ is short notation for all input variables used in the model. We use the freely available statistical package R to perform the generalised linear modeling. What makes our study different from traditional statistical approaches is that we do not use all data records in our training. Instead, we use data mining methodology to test the accuracy of the generalised linear model. As mentioned previously, we divide the data set into training and test data sets. The training data set is used to train all the parameters in Equation 9. The test data set tests the generalisability of the model.

4 Risk Pattern Mining (RPM)

Risk pattern mining (Li et al. (2005), Gu et al. (2003)) deals with data consisting of two unbalanced classes. The minor class (usually the high risk class) is the primary study group. Unlike the classification method, risk pattern mining is not used to build classifiers, but to generate an optimal risk pattern set. A pattern is excluded from the optimal risk pattern set when its relative risk is lower than a simpler pattern with fewer variables in it. Therefore, the optimal pattern set does not give highly accurate prediction, but indicates all interesting cohorts that are more likely to belong to the high risk class. In our risk pattern mining approach, the targeted class of the study is the mothers who do not breast feed their babies. RPM enables us to identify mothers with certain characteristics, leading to a high risk of not breast feeding their babies. Bi-variate analysis certainly helps to identify the single characteristic, which may lead to the baby feeding method decision. This has been done extensively by an earlier study (Hegney et al. 2003). The risk pattern mining method can find the factors associated with not only a single variable alone but also a combination of factors. The combination of these factors leads to the decision of not breast feeding their babies. Therefore, the risk pattern mining approach may complement the statistical approach. The risk pattern mining method also has the advantage of identifying the group automatically from available data alone. It does not assume any prior knowledge on what may pose a high risk. It allows the data to speak for themselves. The main interest of this study is in finding groups of higher occurrences of mothers not breast feeding their babies than the average. These mothers are classified as class C (target class). The other class is called class $\overline{C}$. We define the support of A, $supp(A)$, as the number of subjects satisfying the condition A. We can represent the population by contingency table 1.

The Risk Ratio (RR) values for Class C is defined as ratio of cross products of terms in Table 1 or expressed as follows.

$$RR(A \rightarrow C) = \frac{supp(A \rightarrow C)}{supp(A)} / \frac{supp(\overline{A} \rightarrow C)}{supp(\overline{A})} \tag{10}$$

Table 1: Contingency Table

	C	$\overline{C}$	**Total**
A	$supp(A \rightarrow C)$	$supp(A \rightarrow \overline{C})$	$supp(A)$
$\overline{A}$	$supp(\overline{A} \rightarrow C)$	$supp(\overline{A} \rightarrow \overline{C})$	$supp(\overline{A})$
Total	$supp(C)$	$supp(\overline{C})$	

RR specifies how many times more likely the subjects satisfying pattern A and belonging to the target class are than others. Its 95% Confidence Interval (CI) can be calculated by Equation 11 (Fleiss (1981)).

$$I(A \rightarrow C) = RR(A \rightarrow C) \pm \frac{supp(A{\rightarrow}C)supp(\overline{A}{\rightarrow}\overline{C}) - supp(A{\rightarrow}\overline{C})supp(\overline{A}{\rightarrow}C))}{\sqrt{supp(\overline{C})supp(C)supp(A)supp(\overline{A})}}. \quad (11)$$

5 Results and Discussion

We use the feature "M3FEEDAT" (Type of feeding at 3 months postpartum) as the output variable. There are 53 descriptive variables and 498 subjects. The purpose of the data mining is to decide the factors influencing the decision on feeding method.

5.1 Feature Selection and Classification

FIEBIT and Chi-square methods are applied in the feature selection step. C4.5 and logistic regression are used as classification models for data with features selected by a feature selection module. In logistic regression, the functions provided in R package is used to establish and apply the model in deciding the classification. We use *leave-one-out* as the testing method to decide the accuracy of the models. The classification accuracies on training and test data are listed in Table 2.

The results of logistic regression using all features are not available due to the restrictions of the software. From the decision tree analysis, the following branch covers 123 subjects, of which 106 are breast feeders. Only 7 are not breast feeders.

```
Q3.9.11 (Breast Feeding is convenient) =Yes
Q3.9.1  (My mother breastfed)          =Yes
Q3.9.14 (I do not want to have to mix
        formula/sterilize bottles)     =Yes
```

It is likely that these factors are important in deciding the feeding method.

The following observations can be made out of the results of various feature selection and classification methods.

- Classification accuracy is improved by using a kind of feature selection as a data preprocessing step. The use of the whole feature set leads to high classification accuracy on training data, but low accuracy on test data. This implies the over-fitting problem associated with irrelevant or redundant features included.
- Feature subsets selected by the information theory based method lead to a bit higher classification accuracy on test data than that selected by the Chi-square method. More experiments, say using an n-fold cross-validation, may help draw a sound conclusion, which are left as future work.
- The highest classification accuracy on test data is achieved by using FIEBIT as the feature selection method followed by C4.5 as the classifier. The generalisation accuracy is 77.91%. The accuracy on training data is only slightly higher (80.94%) than that of the test data. This indicates that the over training problem is largely overcome by selecting a good feature subset.
- A decision tree can be used as a feature selection method since it can be applied to the original data set. However, features selected by decision trees are not as good as FIEBIT and ChiSQ on this data set since the accuracy on the test data set is lower. More experiments may help draw a sound conclusion. For example, we may impose some sort of regularisation on the classification models (ridge regression or something similar in the logistic regression, or more aggressive pruning of the decision tree) which would be likely to lead to nice results too. This is left as a future work direction.

5.2 Risk Pattern Mining

The rules identified by the algorithm can find the cohort with high risk ratio of not breast feeding. We use 8 features selected by FIEBIT in the following example.

Results of simple rules with one or two variables from Risk Pattern Mining (RPM) can be also found by statistical analysis, and both findings agree. However, in a statistical approach it is difficult to explore interactions of three or more variables systematically whereas RPM method can. The following rules are some cohorts that are overlooked by the previous statistical analysis.

```
Rule 1
 Q3.9.13  (I enjoy breast feeding)  = No
 Q3.9.14  (I do not want to have to
           mix formula/sterilize
           bottles)                 = No
 FEEDDECI (At what time did the
           mother make the decision
           about feeding method)    = Late
                   in pregnancy/after baby
                       born/still deciding
     Cohort size = 11
 Contingency table
            not breast  breast
            feeding     feeding
 pattern        10        1
 non-pattern   119       368
```

Length = 3, $RR = 3.72 \pm 0.22$

There are a total of 11 subjects in the cohort, 10 of them are not breast feeding. The subjects in the cohort are 3.72 times more likely not to be breast feeding than the subjects not satisfying the rule.

```
 Rule 2
Q3.9.11  (Breastfeeding is more
         convenient)             = Yes
Q3.9.1   (My mother breastfed)   = No
 MOAGEREC (Mother's age)          = Under 25
 Q3.9.13  (I enjoy breast
           feeding)               = No
 Q3.9.14  (I do not want to have
          to mix formula/sterilize
          bottles)                = No
 Q3.8.1  (general anaesthetic)    = No
     Cohort size = 8
 Contingency table
           not breast   breast
           feeding      feeding
 pattern         7         1
 non-pattern   122       368
```

Length = 6, $RR = 3.51 \pm 0.18$

There are total 8 subjects in the cohort, 7 of them are not breast feeding. The subjects in the cohort are

Table 2: Prediction accuracies of various feature selection and classification methods

Feature Selection	**Number of Features**	**Classification Method**	**Accuracy(%) Training**	**Accuracy(%) Testing**
FIEBIT	8	C4.5	80.94	77.91
FIEBIT	8	LogReg	77.70	76.49
None	53	C4.5	87.70	71.66
None	53	LogReg	NA	NA
ChiSQ	11	C4.5	81.97	75.05
ChiSQ	11	LogReg	77.58	76.20

3.51 times more likely not to be breast feeding than the the subjects not satisfying the rule.

```
Rule 3
Q3.9.11  (Breastfeeding is more
          convenient)           = no
MOAGEREC (Mother's age)         = 25-30
Q3.9.14  (I do not want to have
          to mix formula/sterilize
           bottles)             = no
FEEDDECI (At what time did the
          mother make the
          decision about feeding
          method)               = Before
                                 pregnancy
   Cohort size = 18
Contingency table
                 not breast  breast
                 feeding     feeding
pattern            14          4
non-pattern       115        365
```

Length = 4, $RR = 3.25 \pm 0.23$

There are total 18 subjects in the cohort, 14 of them are not breast feeding. The subjects in the cohort are 3.25 times more likely not to be breast feeding than the subjects not satisfying the rule.

These results show that the risk pattern mining method enables us to identify a cohort of mothers who are more likely not to breast feed their babies. This will enable us to conduct a focused education program on a targeted group of mothers to increase the rate of breast feeding. For example, based on rule 2, the cohort identified are 3.51 times more likely not to breast feed their babies. We should therefore target the young mothers (under 25) whose mother did not breast feed their babies. Specific information can be provided to address the concerns of each cohort.

6 Conclusions

Data mining aims at extracting novel, valuable and actionable knowledge from a database. In this study, we attempt to use data mining techniques on a survey data set, rather than traditional statistical analyses. We claim that the application of data mining methods may extract knowledge that statistical analysis may find difficult to identify, say, logistic regression for all the variables. As a complementary approach to statistical analysis, data mining methods can be used as a viable tool in survey data analysis.

1. Feature selection is an important step in data preprocessing as it enables irrelevant and redundant features to be eliminated. It substantially reduced the data dimensions for modelling. It not only makes the modelling procedure more efficient but also improves the accuracy of the model developed by data mining or statistical methods.

2. The feature selection process followed by a classification module is able to build a classifier which classifies the data on whether or not the breast feeding method is selected automatically with reasonable accuracy. The classification accuracy using properly selected feature subsets reached over 75% on test data. This implies that when we apply the model to a new subject (a mother), we can predict her likelihood of breast feeding or not with reasonable confidence. We can therefore take proper measures to provide education surrounding this prediction.

3. Risk pattern mining is able to discover a number of rules which identify groups of patients with a high risk of not beast feeding. The knowledge discovered by risk pattern mining may be used by doctors or nurses to assess the risk associated with a new mother based on her characteristics. A real world application is expectable by using the information discovered by risk pattern mining. For example, we may use a graphic interface McAullay et al. (2005), Chen et al. (2005), to enable medical practitioners to gain knowledge effectively based on the risk patterns mined, which is left as future work. A proper targeted education or other measures may then be taken.

4. The application of various data mining methods to breast feeding survey data helps to discover knowledge and the understanding of the decisions made by mothers. It complements and enhances the statistical analysis. Statistical analysis is powerful in hypothesis testing. Data mining methods, on the other hand, help in hypothesis generating Jin et al. (2006) It generates decision trees, rule sets etc. automatically without assuming any prior knowledge. The knowledge discovered becomes more comprehensive when both statistical analysis and data mining methods are applied to the same data set.

References

Chen, J., He, H., Li, J., Jin, H., McAullay, D., Williams, G., Sparks, R. & Kelman, C. (2005), Representing association classification rules mined from health data, in *Proceedings of 9th International Conference on Knowledge-Based & Intelligent Information & Engineering Systems (KES2005)*, Melbourne, Australia, pp. 1225–1231.

Cover, T. M. & Thomas., J. A. (1991), *Elements of Information Theory*, Wiley-Interscience.

Fleiss, J. L. (1981), *Statistical Methods for Rates and Proportions*, Wiley.

Fleuret, F. (2004), 'Fast binary feature selection with conditional mutual information', *Journal of Machine Learning Research* **5**, 1531–1555.

Gu, L., Li, J., He, H., Williams, G., Hawkins, S. & Kelman, C. (2003), Association rule discovery with unbalanced class, in *Proceedings of the 16th Australian Joint Conference on Artificial Intelligence (AI03), Lecture Notes in Artificial Intelligence*, Perth, Western Australia, pp. 221–232.

He, H., Jin, H. & Chen, J. (2005), Automatic feature selection for classification of health data, in *Proceedings of The 18th Australian Joint Conference on Artificial Intelligence (AI2005)*, Sydney, Australia, pp. 910–913.

Hegney, D., Fallon, T., O'Brien, M., Plank, A., Doolan, J., Brodribb, W., Hennessy, J., Laurent, K. & Baker, S. (2003), *The Toowoomba Infant Feeding Support Service Project: Report on Phase 1 A Longitudinal Needs Analysis of Breastfeeding Behaviours and Supports in the Toowoomba Region.*

Jin, H., Chen, J., Kelman, C., He, H., McAullay, D. & O'Keefe, C. M. (2006), Mining unexpected associations for signalling potential adverse drug reactions from administrative health databases, in *PAKDD'06*, pp. 867–876.

Jin, H.-D., Shum, W., Leung, K.-S. & Wong, M.-L. (2004), 'Expanding self-organizing map for data visualization and cluster analysis', *Information Sciences* **163**, 157–173.

Jin, H., Wong, M.-L. & Leung, K.-S. (2005), 'Scalable model-based clustering for large databases based on data summarization', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(11), 1710–1719.

Kohavi, R. & John, G. (1997), 'Wrappers for feature selection', *Artificial Intelligence* pp. 273–324.

Kramer, M. S. & Kakuma, R. (2003), *Optimal duration of exclusive breastfeeding*, The Cochrane Library.

Li, J., Fu, A. W.-C., He, H., Chen, J., Jin, H., McAullay, D., Williams, G., Sparks, R. & Kelman, C. (2005), Mining risk patterns in medical data, in *Proceedings of KDD'05*, pp. 770–775.

McAullay, D., Williams, G., Chen, J., Jin, H., He, H., Sparks, R. & Kelman, C. (2005), A delivery framework for health data mining and analytics, *in* V. Estivill-Castro, ed., *Twenty-Eighth Australasian Computer Science Conference (ACSC2005)*, Vol. 38 of *CRPIT*, ACS, Newcastle, Australia, pp. 381–390.

Quinlan, J. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Riodan, J. M. (1997), 'Commentary. the cost of not breastfeeding: a commentary.', *Journal of Human Lactation* **13**(2), 93–97.

Shannon., C. E. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**, 379–423,623–656.

Smith, J. (2001), *Mothers milk, money and markets*, Ann Congress Perinatal Society Australia and New Zealand.

Smith, J. P., Thompson, J. F. & Ellwood, D. A. (2002), 'Hospital system costs of artificial infant feeding: Estimates for the australian capital territory', *Australian and New Zealand Journal of Public Health* **26**(6), 543–551.

Wang, G., Lochovsky, F. H. & Yang, Q. (2004), Feature selection with conditional mutual information maxmin in text categorization, in *Proceedings of CIKM'04*, Washington, US, pp. 8–13.

WHO (2001), *The optimal duration of exclusive breastfeeding*, World Health Organization.

Yang, Y. & Pedersen, J. O. (1997), A comparative study on feature selection in text categorization, in *Proceedings of International Conference on Machine Learning*, Nashville, TN, USA.

Yu, L. & Liu, H. (2004), Redundancy based feature selection for microarray data, in *Proceedings of KDD'04*, ACM Press, New York, NY, USA, pp. 737–742.