Data mining scenarios for the discovery of subtypes and the comparison of algorithms

Colas, F.P.R.

Chapter 2

A Scenario for Subtype Discovery by Cluster Analysis

In this chapter, we present our subtyping scenario*. First, we discuss data processing issues when preparing the data before analysis. Next, we motivate our choice for a particular clustering method. Then, to select for a number of subtypes or a model, we describe a computational approach that repeats data modeling. Finally, we report on methods to characterize, compare and evaluate the most likely subtypes.*

2.1 Introduction

To identify homogeneous subtypes of complex diseases like Osteoarthritis (OA) and Parkinson's disease (PD) and to subtype chemical databases, we developed a scenario mimicking a cluster analysis process: from data preparation to cluster evaluation, see Figure 2.1 for an illustration of our scenario. It implements various data preparation techniques to facilitate the analysis given different data processing. It also features a computational approach that repeats data modeling in order to select for a number of subtypes or a type of model. Additionally, it defines a selection of methods to characterize, compare and evaluate the top ranking models.

The outline of the rest of the chapter is as follows. First, we describe data preparation issues with methods to answer them, as well as the clustering method. Second, we report methods to characterize, compare and evaluate cluster results. Illustrations of our scenario throughout this chapter are from medical research on OA and PD.

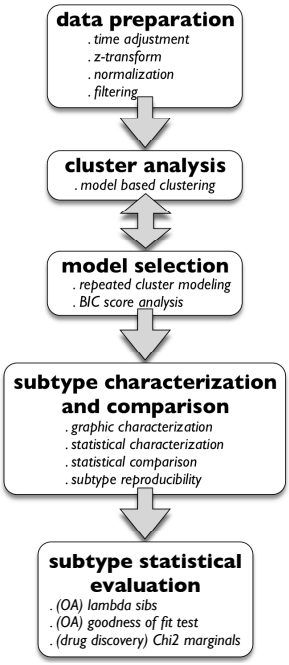

Figure 2.1: *Workflow of a subtype discovery analysis.*

2.2 Data preparation and clustering

We aim to identify homogeneous and *reliable* subtypes. Hence, cluster results should be reproducible and the clusters should characterize true underlying patterns, not the incidental ones. We discuss in this section the removal of the *time* dimension in the OA and PD datasets, the *reliability* and *validity* of cluster results and give a brief overview of model based clustering.

2.2.1 Data preparation

As data preparation can influence largely the result of data analysis, our scenario implements various methods to transform and process data, e.g. computing the z-scores of variables to obtain scale-invariant quantities, normalizing according to the Euclidean norm (L_2), the Manhattan distance (L_1), the maximum and centering with respect to the empirical mean, the median or the minimum.

As in the overal severity of OA and PD, respectively age or disease duration (thereafter, the *time*) are known to play a major role, we may want to remove their dimension in the data because we do not want to model clusters only characterized

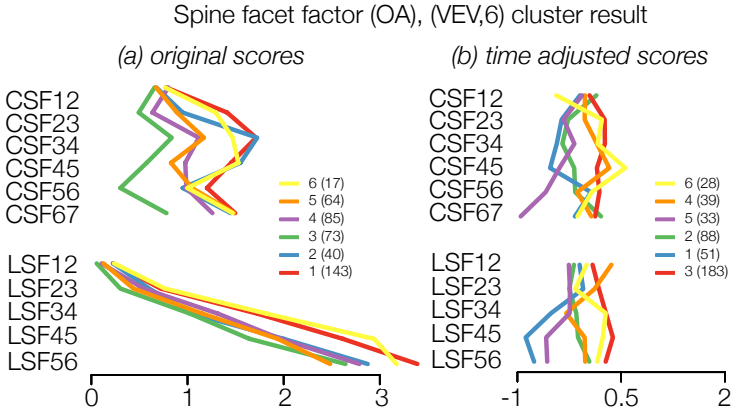

Figure 2.2: *For OA, we show results of two cluster analyses on the spine facets factor with a VEV model having six mixtures (the VEV model will be explained in section 2.2.3). In (a), the modeling is on the original ROA scores, i.e. between* $[0, 4]$ *and in (b), on the time adjusted scores, i.e. z-scores. This illustrates how the time influences the cluster results. The arrangement of the variables mimicks the disposition of the cervical and lumbar vertebrae, from top to bottom.*

by them. In Figure 2.2, we report a visualization of two cluster analyses on OA data: we conducted the clustering on the original scores and on the time adjusted scores; it shows how much the time influences the modeling. So, to remove the *time* dimension for the data, we first perform regression on the *time* for each variable and next, we conduct cluster analyses on the residual variance.

If we denote by α and β the estimated intercept and coefficient vectors of the regression, by the matrix $\mathbf{X}$ the data where x_{ij} refers to measurement j of observation i, then the regression is given by

$$x_{ij}(t_i) = \alpha_j + \beta_j g(t_i) + \varepsilon_{ij}, \tag{2.1}$$

$$\varepsilon_{ij} = x_{ij}(t_i) - \alpha_j - \beta_j g(t_i). \tag{2.2}$$

The ε_{ij} refer to the residual variation and $g(t) \in \left\{log(t), \sqrt{t}, t, t^2, exp(t)\right\}$ (the time effect is not necessarily linear). Additionally, residuals ε_{ij} should distribute normally around zero for each variable j, as illustrated in Figure 2.3.

2.2.2 The reliability and validity of a cluster result

In our data mining scenario for subtype discovery by cluster analysis, hierarchical clustering [Sne73] or k-means [Has01] did not match our expectations in terms of *reliability* and *validity* (see discussion below). Instead, we selected model

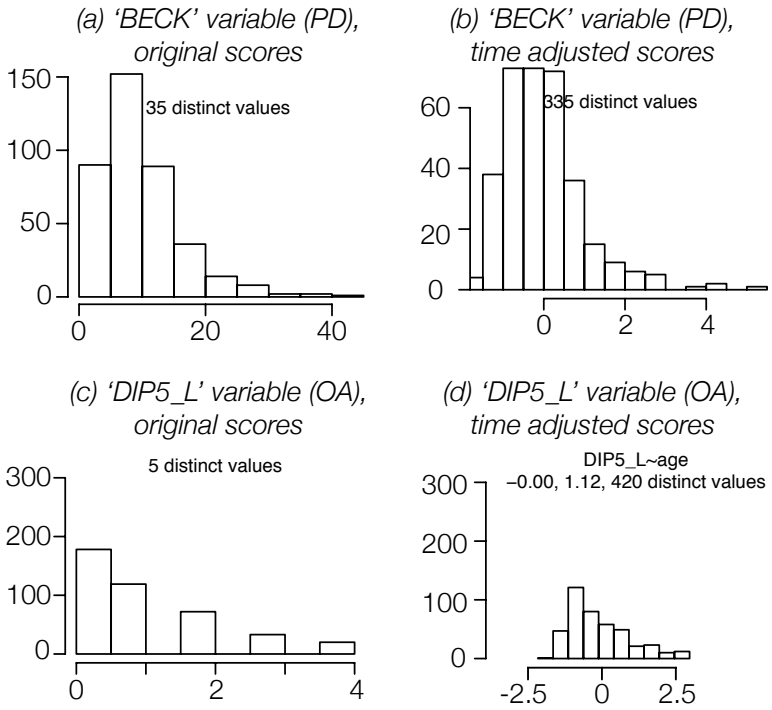

Figure 2.3: *These four figures illustrate the original and the time-adjusted data distributions of variables DIP5_L and beck, which respectively pertain to OA and PD analyses. Such histograms are obtained when plotting a dataset class (*`cdata`*) of the* R Subtype-Discovery *package. To be valid, the residuals ε_{ij} of the regression on the time should distribute normally around zero for each variable j.*

based clustering that relies on the EM-algorithm (Expectation Maximization) for parameter estimation [Fra99; Fra02b; Fra03; Fra06]. In the following two paragraphs, we discuss the *reliability* and *validity* of the k-means, the hierarchical and the model based clustering.

k-means and hierarchical clustering First, in terms of *reliability*, the cluster results should be consistent when we repeat the analysis. For example, when we repeat the k-means, solutions may differ because of the different starting values. Second, both the hierarchical clustering and the k-means clustering depend on distance measures which do not necessarily mimic the data distribution of the clusters; however, to be *valid*, the clusters should be understandable which is not evident when they are defined in terms of distances, especially for non-euclidean ones.

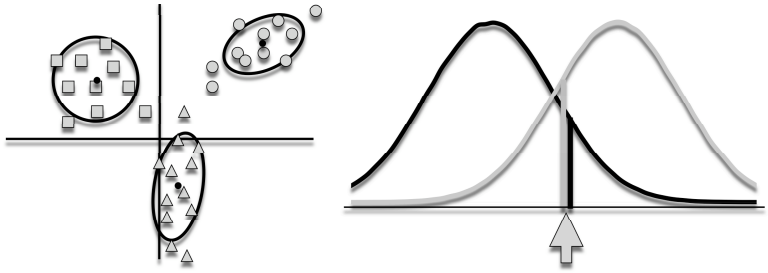

Figure 2.4: *On the left, we show a simple modeling with three mixtures in two dimensions which are defined by their center μ_k and their geometry Σ_k with $k = 1, 2, 3$. On the right, we illustrate two mixtures on a single dimension. Membership of the gray is most likely. Membership of the black is less likely.*

In fact, the clusters should also be distinguishable, which becomes an issue as the modeling takes place in higher dimensions because the distance-based algorithms are sensitive to the curse of dimensionality [Bey99]. And finally, another aspect that hampers especially the hierarchical clustering, concerns the numerous parameters that can only be set subjectively. The book [Sne73] gives a detailed description of all the possible parameters.

Model based clustering To be fair, *reliability* issues also exist for clustering by mixture of Gaussians because it relies on the EM-algorithm. To estimate model parameters, EM optimizes iteratively the model likelihood and as a matter of fact, different starting values for EM may lead to different cluster results. Therefore, an important issue concerns the sensibility to different starting values of the mixture modeling. In this regard, Fraley and Raftery decided to initiate systematically their EM-algorithm by a model based hierarchical clustering [Fra99]. This choice ensures the reproducibility of the cluster results because two repeats of the mixture modeling will initiate EM equally.

Concerning the *validity* issue, mixture modeling not only reports the estimated center of each mixture but also it estimates the covariance structure. Therefore, it also yields estimates of the cluster membership certainty. Further, as shown in [Ban93] and as illustrated with an example in Figure 2.4, the framework relies on the concept of reparameterization of the covariance matrix which enables to select and adapt the level of complexity of the covariance by controlling its geometry. Hence, the analysis offers a range of models that involve varying number of parameters to estimate. For instance, a particular model may set an equal data distribution for all mixtures, while another may discard the estimations of the covariates in the model.

2.2.3 Clustering by a mixture of Gaussians

In this subsection, as in [Fra99; Fra02b; Fra03; Fra06], we describe clustering by mixture modeling.

First, the likelihood function of a mixture of Gaussians is defined by

$$\mathcal{L}_{MIX}(\theta, \tau|\mathbf{x}) = \prod_{i=1}^{N} \sum_{k=1}^{G} \tau_k \phi_k(\mathbf{x}_i|\mu_k, \Sigma_k), \tag{2.3}$$

where $\mathbf{x}_i$ is the i^{th} of N observations, G is the number of components and τ_k the probability that an observation belongs to the k^{th} component (hence $\tau_k \geq 0$ and $\Sigma_{k=1}^{G} \tau_k = 1$). Then, the likelihood of an observation $\mathbf{x}_i$ to belong to the k^{th} component is given by

$$\phi_k(\mathbf{x}_i|\mu_k, \Sigma_k) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\}}{\sqrt{det(2\pi\Sigma_k)}}. \tag{2.4}$$

The reparameterization proceeds by eigenvalue decomposition of the covariance matrix Σ_k

$$\Sigma_k = D_k \Lambda_k D_k^T. \tag{2.5}$$

This decomposition depends on the diagonal matrix Λ_k of the eigenvalues and on the eigenvector matrix D_k which determines the orientation of the principal components. The matrix Λ_k can be decomposed further into

$$\Lambda_k = \lambda_k A_k \tag{2.6}$$

with A_k the geometrical shape and λ_k the largest eigenvalue.

In their framework, Fraley and Raftery control the structure of Σ_k using constraints on the three parameters λ_k, A_k and D_k. The constraints are expressed in letters $\{I, E, V\}$ which stand for identical, equal and variable respectively.

- λ_k refers to the relative size or the *scale* of the k^{th} mixture which may be equal for all mixtures (E) or vary (V).
- A_k specifies the *geometrical shape* which may limit the mixtures to spherical shapes (I), to equally elongated shapes for all mixtures (E), or to varying ones (V).
- D_k characterizes the principal orientations of the covariance which may simply coordinate along the axes (I) and therefore neglect estimation of the covariates; but when considering covariates, we may select an equal orientation for all mixtures (E) or a different one (V).

Hence, a constraint is expressed by three letters, one for each parameter. For example, the constraint VVI refers to a model where a diagonal covariance matrix will be estimated for each cluster; therefore, no covariate is estimated.

EM-algorithm For a given number of mixtures and a covariance model, the EM-algorithm is used to estimate the model parameters [Dem77]. It alternates iteratively between a step of *Expectation* to estimate for each observation its cluster membership likelihood, and a step of *Maximization* to identify the parameters that maximize the model likelihood. The iterative process stops as likelihood improvements become small.

An important concern for the EM algorithm is the dependency on the starting point. As mentioned before, Fraley and Raftery propose to systematically initialize EM with a model based hierarchical clustering. Though, a common strategy is to start EM from several random points and then to study the sensibility of the cluster results to these changes. We selected this second strategy for our data mining scenario.

2.3 Model selection

The larger the number of parameters, the more likely it is that our model may overfit the data which restricts its generality and comprehensiveness. In this section, we discuss a score that we use as a guidance to compare models involving different numbers of parameters and an approach to conduct a *valid* model selection.

2.3.1 A score to compare models

For model selection, Kass and Raftery [Kas95] prefer the Bayesian Information Criterion (BIC) to the Akaike Information Criterion (AIC) because it approximates the Bayes Factor. Therefore, our analyses also rely on the guidance provided by the BIC. It is defined by

$$BIC = -2 \log \mathcal{L}_{MIX} + \log \left(N \times \#params \right), \tag{2.7}$$

with $\mathcal{L}_{MIX}$ the Gaussian-mixture model likelihood, N the number of observations and $\#params$ the number of parameters of the model.

2.3.2 Valid model selection

In our data mining scenario, we found it inappropriate to conduct model selection on the basis of a single BIC value because it left several questions unanswered. We give some of them:

1. What is the statistical significance of BIC scores differences that are less than 5%?
2. If EM was initialized from different starting values, how reliable would the cluster results be?

3. Did the EM-algorithm end in a local or a global likelihood maximum?

For this reason, we decided to further validate our choice for a particular model by repeating the data modeling process for different starting values; our approach proceeds as follows:

1. Set an integer that fixes the starting point of the random number generator.
2. Draw from a uniform distribution a matrix of cluster membership probabilities.
3. Proceed to a *maximization* step (M-step) to identify the parameters of the most likely model.
4. Start EM-algorithm from its *expectation* step (E-step).

This way, by repeating EM initialization from many different starting points, we can select the most likely model and consider it as the *optimal* one.

Then, to conduct a valid model selection, we aggregate the BIC scores in a number of ways. In first place, we report the average rank of the model (respectively the average rank of the number of clusters) when a particular number of clusters (resp. a model-type) is chosen. These rankings may enable to select for a particular type of model and a number of clusters. We also report tables that characterize statistically the BIC scores in terms of the empirical *mean*, the standard deviation and different quantile statistics. Finally, two more tables present the starting values and the BIC scores of the most likely models for each combination.

2.4 Characterizing, comparing and evaluating cluster results

Because cluster models may take different spatial-shapes, we need methods to report their characteristics and to compare them. Further, when analysing data from the medical domain, we consider as important to evaluate the clinical relevance of the subtypes by some additional characteristics. Therefore, in this section, we present our techniques to address these different aspects.

2.4.1 Visualizing subtypes

To check the effect of changing the settings (the type of cluster model and the number of clusters), we need visualization tools to see the characteristics of the cluster results. Being influenced by Tukey [Tuk77] and Tufte [Tuf83; Tuf90] for scientific data visualization and by Brewer's suggestions for color selection in geography [Bre94], we selected three visual-aids to address this issue: *heatmaps* [Eis98], *parallel coordinates* plots [Ins85] and *dendrograms* [Sne73].

Heatmaps In the analysis of micro array data, heatmaps are often used to display and cluster data. However, as heatmaps depend on hierarchical clustering, there are many parameters that need to be set rather subjectively. Besides, as we do calculations with distance measures, the variables should be scale-free and comparable; this may be awkward when variables are not scale-homogeneous. On top of that, as variables are correlated, the distances will mostly reveal patterns in the principal component dimensions of the data.

For the OA data, we can illustrate this by considering a large joint factor that consists of hips and knees and another one that consists of the spine joints. Simply because there are only four variables in the first factor and about 20 in the second, the spine has a larger "contribution" than the large joints in the distance. So, simple distances lack sensitivity to manifest changes in the small principal component dimensions. We limit the use of heatmaps to report statistical patterns of the clusters, e.g. the mean, the median or quantiles.

Next, as hip left and right pertain to the hips in OA or as both urinary and cardiovascular problems reflect autonomic symptoms in PD, we can often group variables into main factors. Indeed, we may expect the variables to correlate in each factor; yet, standard heatmaps do not exploit the grouping of the variables, this makes the comprehesion of the cluster results more difficult.

Parallel coordinate plots In parallel coordinates plots, we can make use of this grouping information in factors to order the variables appropriately. For each cluster, we use a different color and, as Figure 2.2 illustrates OA data, we characterize each center (μ_k) by lines connecting the different variables (the parallel axis). In this Figure, we notice the particular ordering for the cervical and the lumbar spinal joints that reflects the natural ordering of these joints from top to bottom. An interesting additional property of this type of plots is that besides each cluster center (the mean pattern), we can also report quantile-statistics using connected lines of a different shape (e.g. the 2.5% and 97.5% patterns of a cluster).

Dendrograms Finally, in spite of the many disadvantages of hierarchical clustering, we find it a useful addition to the heatmaps and parallel coordinates because dendrograms can illustrate the similarity between the center patterns or between the variables. In fact, a dendrogram on the cluster centers can help to order the clusters by similarity, whereas a dendrogram on the variables can provide a rudimentary factor analysis. Therefore, both kinds of dendrograms are included and provide additional understanding.

2.4.2 Statistical characterization and comparison of subtypes

First, using the *log of the odds*, we report the main statistical characteristics of the clusters. Second, to cross-compare the cluster results, we rely on regular

association tables from which we estimate the usual χ^2 statistics. Next, we use further the χ^2 statistics to calculate a single measure in terms of the *Cramer's V* coefficient of nominal association. Finally, as a way to assess the *reproducibility* of cluster results, we estimate the generalization of the cluster result by training common machine learning algorithms on the clustered data.

Statistical characterization For each application domain, we group variables by main factor such as the main joint sites in OA (the spine facets, the spine lumbars, the hips, the knees, the distal and the proximal interphalengeal joints), the impairment domain in PD (the cognitive, the motricity and the autonomic disorders) and the class of molecular descriptors in drug discovery. Then, to characterize statistically the cluster results, we compute the *odd* of the cluster data distribution as compared to the one of the dataset; the data distribution is the sum of the scores in each group of variables (the factors).

In practice, one might refer to the *log of the odds* as the cross-product because we calculate it from tables similar to Table 2.1. We express the log of the odds of a cluster k on a factor l as

$$logodds_{kl} = \log \frac{A \times D}{B \times C}. \tag{2.8}$$

Table 2.1: *For each sum score l, we consider a middle value δ_l such as the dataset mean or median. For cells A and B, we use it to count how many observations i in the cluster S_k have a sum score above and below its value. For cells C and D, we proceed to a similar count but on the rest of the observations $i \in \{S - S_k\}$.*

	$x_i < \delta_l$	$x_i \geq \delta_l$
$i \in S_k$	A	B
$i \in \{S - S_k\}$	C	D

Statistical comparison of cluster results In order to compare cluster results, we report association tables that describe the joint distribution between the two cluster affectations of the observations (nominal variables). If the table has many empty cells, then the two cluster results are highly related. However, if the joint distribution over all cells is even, then the two cluster results are unrelated (independent).

Further, to summarize the association tables, we calculate the Cramer's V. Similarly to Pearson's correlation coefficient, the Cramer's V takes values in $[0, 1]$; one stands for completely correlated variables and zero for stochastically independent ones. The measure is symmetric and it is based on the χ^2 statistics of

nominal association. Therefore, the more unequal the marginals, the more V will be less than one. Alternatively, the measure can be regarded as a percentage of the maximum possible variation between two variables. It is defined by

$$V = \sqrt{\frac{\chi^2}{n \times m}}, \tag{2.9}$$

where n is the sample size and $m = min(rows, columns) - 1$.

In our table-charts, we will embed in the top left the joint distribution and in the lowest row the *Cramer's V* coefficient.

Estimating the cluster result reproducibility When performing unsupervised cluster analysis, it is important to know whether the cluster result generalizes, for instance to the total patient population in the case of medical research. Therefore, we chose to assess the cluster result *learnability* by training machine learning algorithms like the naive Bayes, the linear Support Vector Machines or, as a baseline, the one nearest neighbor classifier.

To evaluate these algorithms, we use the average classifier accuracy estimated by training ten times the classifiers on datasets splitted randomly into training (70%) and test set (30%). To split the data, we chose to preserve in every training and test set the cluster proportions from the original sample.

Stratifying the samples enables to reduce the variability of the accuracy estimates which is coherent with the practice in machine learning because we primarily aim to compare algorithms. However, in medical research, we might prefer to include the variability inherent to the cluster proportions in the estimation of the accuracy.

2.4.3 Statistical evaluation of subtypes

When conducting a subtype discovery analysis, a key concern is the evaluation of the clusters. For that purpose, we implemented a simple mechanism to add study-specific evaluation procedures of the clusters.

In OA for instance, as the study involves sibling pairs, we defined two statistical tests that assess the level of familial aggregation in each subtype and its significance. Our first test relies on a risk ratio which we refer to as the λ_{sibs}, whereas the second test makes use of a χ^2-test of goodness of fit.

In drug discovery, χ^2 cell-statistics between the human-defined classification and the one identified by the subtyping are reported; we search for χ^2 cell-statistics showing a large marginal.

The λ_{sibs} risk ratio in OA research First of all, we characterize each individual as *proband* or *sibling* depending on whether this individual was the first sibling involved in the study or not.

Then, this test quantifies the *risk* increases of the second sibling given the characteristics of the proband. For instance, a $\lambda_{sibs} = 1$ means that the risk does not increase and that the cluster membership of the proband does not influence the one of his sibling. On the other hand, if $\lambda_{sibs} = 2$, then the risk increase is two-fold. Finally, a λ_{sibs} is significant when the lower bound of the 95% confidence interval is above 1. In the following, we describe formally the λ_{sibs} and we derive its confidence interval analytically by the delta method.

Take two siblings s_1 and s_2 with s_1 being the proband. A proband is the first affected family member who calls for medical attention. We consider the probability of a sibling to belong to a group S_k as $P(s_i \in S_k)$ with $i \in \{1, 2\}$, or for short $P(s_i)$. Then, the conditional probability that the second sibling is in S_k given that the first sibling is also in S_k is referred to as $P(s_2|s_1)$. Therefore, the λ_{sibs} is expressed by

$$\lambda_{sibs}(S_k) = \frac{P(s_2|s_1)}{P(s_2)} = \frac{P(s_1, s_2)}{P(s_1)P(s_2)} = \frac{P(s_1, s_2)}{P(s)^2}. \tag{2.10}$$

where $P(s_1) = P(s_2) = P(s)$ if the population is considered to be infinite. Next, we derive a confidence interval by the delta method using

$$\lambda_{sibs} = \frac{\hat{\alpha}}{\hat{\beta}} \tag{2.11}$$

where $\hat{\alpha} = P(s_1, s_2)$, $\hat{\beta} = P(s)$ (the hat denotes quantities estimated from the data). Then, the variances and covariance of $\hat{\alpha}, \hat{\beta}$ have the form

$$\sigma_\alpha^2 = \frac{\hat{\alpha}(1 - \hat{\alpha})}{n_i}, \tag{2.12}$$

$$\sigma_\beta^2 = \frac{\hat{\beta}(1 - \hat{\beta})}{N}, \tag{2.13}$$

$$cov(\hat{\alpha}, \hat{\beta}) = \frac{\hat{\alpha}(1 - \hat{\beta})}{N}, \tag{2.14}$$

with n_i the sibship size and N the number of observations. The first order Taylor approximation of $f(\alpha, \beta)$ in $(\hat{\alpha}, \hat{\beta})$ is expressed by

$$f(\alpha, \beta) = f(\hat{\alpha}, \hat{\beta}) + \sum_{\delta=\alpha,\beta} (\delta - \hat{\delta}) \frac{\partial f(\hat{\alpha}, \hat{\beta})}{\partial \delta} + R_1. \tag{2.15}$$

If we move the zeroth derivative to the left and we raise everything to the square, then we obtain

$$\Big(f(\alpha, \beta) - f(\hat{\alpha}, \hat{\beta})\Big)^2 = \left((\alpha - \hat{\alpha}) \frac{\partial f(\hat{\alpha}, \hat{\beta})}{\partial \alpha} + (\beta - \hat{\beta}) \frac{\partial f(\hat{\alpha}, \hat{\beta})}{\partial \beta}\right)^2. \tag{2.16}$$

Provided that $\partial f(\hat{\alpha}, \hat{\beta})/\partial \alpha = 1/\hat{\beta}^2$ and $\partial f(\hat{\alpha}, \hat{\beta})/\beta = -2\hat{\alpha}/\hat{\beta}^3$, we obtain

$$\begin{aligned}\left(f(\alpha,\beta)-f(\hat{\alpha},\hat{\beta})\right)^2 = \quad & (\alpha-\hat{\alpha})^2\left(\tfrac{1}{\hat{\beta}^2}\right)^2 \\ & +(\beta-\hat{\beta})^2\left(\tfrac{-2\hat{\alpha}}{\hat{\beta}^3}\right)^2 \\ & +2(\alpha-\hat{\alpha})\left(\tfrac{1}{\hat{\beta}^2}\right)\left(\tfrac{-2\hat{\alpha}}{\hat{\beta}^3}\right).\end{aligned} \tag{2.17}$$

Finally, taking the expectation, the variance is expressed by

$$\sigma_\lambda^2 = \frac{1}{\hat{\beta}^4}\left(\sigma_\alpha^2 - 4cov(\hat{\alpha},\hat{\beta})\frac{\hat{\alpha}}{\hat{\beta}} + 4\sigma_\beta^2\frac{\hat{\alpha}^2}{\hat{\beta}}\right), \tag{2.18}$$

or equivalently

$$\sigma_\lambda^2 = \frac{1}{\hat{\beta}^4}\left(\sigma_\alpha^2 - 4cov(\hat{\alpha},\hat{\beta})\hat{\beta}\lambda + 4\sigma_\beta^2\lambda\right). \tag{2.19}$$

A χ^2-test of goodness of fit for OA research We also implemented a simple χ^2 test of goodness of fit to assess the level of familial aggregation in each cluster k.

This test counts the pairs of siblings in each group and compares them to the ones expected when cluster membership would be random. If we first define N as the number of individuals and if we let S be a random draw of size $|S|$, then the probability that an individual i belongs to S is

$$P(i \in S) = \frac{|S|}{N}. \tag{2.20}$$

Next, if we consider a second individual j which is independent of i, then the probability that both i and j belong to S is expressed by

$$P(i,j \in S) = P(i \in S)P(j \in S) = \left(\frac{|S|}{N}\right)^2. \tag{2.21}$$

Further, if we denote by $E(i,j \in S)$ the expected number of sibling pairs under random cluster membership which relies on the total number of pairs $(N/2)$, then

$$E(i,j \in S) = P(i,j \in S)^2\frac{N}{2}. \tag{2.22}$$

Finally, the Grand Total of the χ^2 test is

$$GrandTotal = \sum_{k=1}^{G}\frac{(O(i,j \in S_k) - E(i,j \in S_k))^2}{E(i,j \in S_k)} = \sum_{k=1}^{G}\chi_k^2, \tag{2.23}$$

where k indices over the different clusters and χ_k^2 refers to the separate χ^2 statistics of each cluster. The number of degrees of freedom of our test is

$$df = G - 1 \tag{2.24}$$

with G the number of clusters.

Association tables in drug discovery In order to better understand the relationship between the bioactivity classes, we decided to study the joint distribution between the subtypes and the bioactivity classes. For this purpose, the joint distribution between the cluster affectation and the bioactivity class is reported both in terms of cell-counts and χ^2 cell-statistics; we are interested in the cells with high χ^2-statistics.

2.5 Concluding remarks

We presented a data mining scenario that facilitates and enhances the search for subtypes with application to medical research and drug discovery. This scenario involves techniques to prepare data, a computational approach repeating data modeling to select for a number of clusters and a particular model, as well as other methods to characterize, compare and evaluate the most likely models. Therefore, our scenario does not solely cluster data but it also produces a set of results to conduct a subtype discovery analysis: from data preparation to subtype evaluation.