Adv. & Appl. in Stat. **7**(3) (2007), 341-356

A META-ANALYSIS FRAMEWORK BASED ON HIERARCHICAL MIXTURE MODEL

A. DIMITRI and M. TALAMO

Nestor Lab.
University of Rome Tor Vergata
Rome, Italy
e-mail: dimitri@nestor.uniroma2.it

Abstract

Despite its widespread and growing use, meta-analysis continues to be a controversial set of not well-defined techniques to treat a set of well-known problems (publication bias, heterogeneous data and results, missing data, etc.). This study proposes a statistical framework that starts from the characterization of meta-analysis as the research of the maximum number of shared and large concepts. This research is made using hierarchical mixture models. In this general framework all the above well-known problems assume a new exact characterization.

1. Introduction

Meta-analysis is a quantitative method for synthesizing results from individual studies to estimate an overall effect. Since its introduction in the social sciences by Glass [2], there has been an enormous increase, both in epidemiology and in clinical trials, in the use of meta-analysis as a statistical technique for combining the results of individual analysis; and it has become an important tool for providing quantitative overviews of areas where many such individual studies have been carried out.

2000 Mathematics Subject Classification: 62H30, 62P10, 68Q32.

Keywords and phrases: meta-analysis, Bayesian meta-analysis, mixture models, statistical learning.

Received March 12, 2007

In the definition and more, in the analysis of meta-analysis studies, is unclear if meta-analysis is only a typical statistical application problem or is a specific field in statistical data analysis. Definition is vague. Despite its widespread and growing use, meta-analysis continues to be a controversial set of not well-defined techniques to face a set of well-known problems (publication bias, heterogeneous data and results, missing data, etc.). There is a use of different statistical techniques, without an effective explanation of the reasons behind the use of them. In many cases the methodological proposal and the applicative meta-analysis are confused, and then there are local solutions to general problems. This is the case of meta analysis studies that are at the same time methodological proposals.

In the analyzed meta-analysis studies, meta-analysis can be viewed as a set of inference techniques starting from a set of databases, one database for every study that composes the meta analysis. In this direction, common steps are:

- reconstruction of the database associated with every study;

- standardization of data;

- aggregated analysis of standardized data.

There are some questions associated with this point of view: first, the database is only a part of a study; a new meta-analysis approach has to consider the study as a whole, and has to build the model associated with it. Another related problem is the fact that, often, a study publics only a synthesis of the database. The fact that there is not a fully published database, but only a partial database and, more generally, a model, help us to understand meta-analysis. It cannot be the activity of making questions and finding answers in the available data. Meta-analysis has to be analysis of available data with the attempt to understand for which questions there are, in those data, the right answers. There are two phases in meta-analysis. In the first one, the phase of studies selection, there is a question, a problem object of study. In the second phase, after the studies selection, the previous question stays in background. The focus is in the set of studies or models. The objective is to learn, with them, the set of admissible questions and the associated answers. This is a learning problem.

The starting point of a statistical learning problem is a set X of statistical units. X can also be viewed as a input space or event space. For every unit $x \in X$ are defined n variables or attributes $A_1, ..., A_n$ and a conjunctive probability distribution $P(A_1, ..., A_n)$. Another element of this learning problem is a distribution D over X that represents the probability of an element $x \in X$ to be considered. Interaction between D and $P(A_1, ..., A_n)$ is fundamental in meta analysis and is a formalization of the publication bias problem. More precisely, given X, D and $A_1, ..., A_n$, there are many distributions, one for every subset S of X used to infer $P(A_1, ..., A_n)$.

In the previous defined space there are concepts and hypotheses about them. Let us consider a multidimensional random variable $A = A_1, ..., A_n$ with distribution $P(A)$ and two constant thresholds α and β. Let C be a probability distribution that can be inferred from $P(A)$ fully specified (for example $P(A_1/(A_2, A_3))$. Let MIS be the uniform measure on the support of C. Then, C is a concept if exists a region, subset of the support of C, with $\text{MIS}(r) < \alpha$ and $C(r) > \beta$. Intuitively speaking, a concept is a small area of high probability for a derived conditional variable. For example the probability of a defined symptom $(A_1 = 1)$ is high only for the subset of young $(A_3 = y)$ males $(A_2 = m)$, then $P(A_1 = 1/A_2 = m, A_3 = y) > \beta$. If C is the set of all true concepts, it is unknown and it has to be learned.

Given X, D, $A = A_1, ..., A_n$ and $P(A)$, then the set C is implicit in $P(A)$. Equivalently, a partial explanation of $P(A)$ can include a subset of C. For example, given X and a subset $B \subset A$ and $P'(B)$, then is defined a set of concepts that can be viewed as an approximation of C. A hypothesis is a model $M = \{X, B, P(B)\}$. In the proposed approach every study is a hypothesis.

At this point, it is fully defined the target learning problem: learn the set of concepts C, starting from a set of hypotheses $H = h_1, ..., h_m$, where m is the number of studies selected for the meta analysis.

In Bayesian Learning Theory the hypotheses space is a set of models h_i, $i = 1 \cdots M$, represented as probability distributions over elementary data X. The output of the learning is a new model or probability distribution P over X:

$$P'(\mathbf{y} \mid DB) = \sum_{i=1}^{M} P(\mathbf{y} \mid h_i) P(h_i \mid DB), \tag{1}$$

where DB represents the vector of sample instances, and h_i is an element of the set of models. The new distribution P' can be viewed as an averaged distribution function of all hypotheses. $P(\mathbf{y} \mid h_i)$ represents the probability of the instance $\mathbf{y}$ in the model h_i. The prior distribution $P(h_i \mid DB)$ defines the relative importance of every model in the explanation of the probability of the concept $\mathbf{y}$.

In the proposed approach there is a set of studies S_i; for every study there is a database D_i, that is the dataset declared in the study, and a model h_i, that is a probability distribution inferred from the study. The emphasis is on the gap between the dataset D_i and the hypothesis h_i because:

- for many reasons, every study does not report the original dataset D_i but only a collection of characteristics and relationships about it;

- every D_i has local errors and local typical characteristics.

There are three main properties behind this approach:

- in Bayesian Learning Theory a concept C is true if the average of $C \mid S_i$ is high; the aggregation phase of meta-analysis is done by summing probabilities; this is a response to the well-known problem of heterogeneity behind meta-analysis; there is a sum of probability over variables and not a sum of absolute levels of variables;

- for the publication bias problem, the learning approach permits us to define the exact set of questions a specific meta-analysis permits to answer.

Defined the new approach to meta-analysis and the model used to it (Bayesian learning via mixtures) the properties of the model have to be analyzed. In the analysis of meta-analysis studies there is the need of a tool to permit the emergency of a concept, where this one is not known a priori in its nature (for example a linear relation) and in the full specification of its parameters (the exact linear relation). The use of a specific tool, for example linear regression, does not say where is the information source that state linear correlation in some of the variables of meta-analysis. In other words, meta-analysis has to be a power tool of strongly unsupervised learning. In this direction the properties of the proposed model have been examined.

2. Model Definition

2.1. Mixture models

Let $x = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ be n independent vectors in R^d such that each $\mathbf{x}_i$ arises from a probability distribution with density

$$f(\mathbf{x}|\Theta) = \sum_{s=1}^{S} p(s)h(\mathbf{x} \mid \lambda(s)), \tag{2}$$

where the $p(s)$ are the mixing proportions $(0 < p(s) < 1$ for all s and $\sum_{s=1}^{S} p(s) = 1)$, $h(\cdot \mid \lambda(s))$ denotes a d-dimensional distribution parametrized by $\lambda(s)$. h is, for example, the density of a Gaussian distribution with mean vector μ and variance matrix Σ.

Another, more complex, definition is

$$f(\mathbf{x}|\Theta) = \sum_{s=1}^{S} a_s(w_s, \mathbf{x})h(\mathbf{x}|\lambda(s)). \tag{3}$$

The functions $h(\mathbf{x} \mid \lambda(s))$ are experts, each of which *operates* in regions of space for which the gating functions a_s are nonzero. Note that, assuming a_s to be independent of x leads to the standard mixture of (1).

The previous MoE (Mixture of experts) model is defined by a linear combination of S experts. An intuitive interpretation of the meaning of

this combination is the division of the feature space into subspaces, in each of which the experts are combined using the weights a_s. The Hierarchical MoE (HMoE) takes this procedure one step further by dividing the subspaces using a MoE classifier as the expert in each domain. Let $f(\mathbf{x})$ be the output of the HMoE, and let $g_s(\Theta_s, x)$ be the output of the s-th expert, $1 \le s \le S$. The parameter Θ_s is comprised of all the parameters of the s-th first level expert. This is described by

$$f(\mathbf{x}) = \sum_{s=1}^{S} a_s(w_s, \mathbf{x}) g_s(\Theta_s, x) \tag{4}$$

$$g_s(\Theta_s, x) = \sum_{s=1}^{Sj} a_{sj}(w_{sj}, \mathbf{x}) h_{sj}(\nu_{sj}, x), \tag{5}$$

where $a_{sj}(w_{sj}, \mathbf{x})$ is the weight of the $h_{sj}(\nu_{sj}, x)$, the j-th (sub-) expert in the second level. By defining $\Theta_{sj} = [w_{sj}, \nu_{sj}]$ we have that $\Theta_s = (\Theta_{s1} \ldots \Theta_{sSj})$. We also define $w = (w_1, \ldots, w_S)$ the parameter vector of the weights of the first level and $\Theta = [w, \Theta_1 \ldots \Theta_S]$ the parameter vector of the HMoE.

2.2. Application to Meta Analysis

Given m random variables $V_1 \cdots V_m$, a model H is the conjunctive probability distribution associated with the m variables.

For example if the m variables are SEX(S), AGE(A) and PANSS(P), given $P(S, A, P)$ we can infer $P(P = 3/S = \text{male and } A = 35)$. Then the conjunctive probability distribution $P(S, A, P)$ represents a general concept of the relationship model between variables.

Given s studies $S_1 \cdots S_s$, there are s associated datasets $D_1 \cdots D_s$ and s associated models $H_1 \cdots H_s$. In this context every model Hi represents an approximation of the target model H associated with the variables $V_1 \cdots V_m$, where every Di represents a set of instances of the conjunctive random variable over them. The target model is the model that shares all the regularities of the hypotheses Hi, because in the proposed approach every Hi is a partial explanation of H. We know $H_1 \cdots H_s$ but we do not know $D_1 \cdots D_s$ because studies do not publish their original datasets.

In Bayesian learning theory Hi are hypotheses regarding the distribution of $V = (V_1 \cdots V_m)$, the m-dimensional random variable. The output of the learning phase is not an element selected from $\{H_1 \cdots H_s\}$, but a new distribution $F(\mathbf{x})$, learned starting from a database $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ of n instances. The prior distribution over $\{H_1 \cdots H_s\}$ defines, for every $\mathbf{x}$ the relative contribution of the generic Hi to F. Then it can represent the expertise level of an associated context or study Si.

The behind hypothesis is that data are generated by S sources. In meta-analysis, S is the number of studies that can be viewed as data sources, models or hypotheses or experts about elementary data distribution.

Every study has associated Di, the dataset declared in the study, and Hi, a model inferred by the study. In the proposed approach Di is unknown; Si is known and Hi has to be inferred. There are two main explanations behind the gap between Di and the model Hi:

- for many reasons, every study does not report the original dataset Di but only a collection of characteristics and relationships about it. This fact hides a set of advantages and disadvantages, based on the sensibility of authors on highlight all and only true relationships in the data;

- every Di has local errors and local typical characteristics; Di are heterogeneous. The purpose behind the introduction of Hi is to deal with these elements.

Many approaches and statistical tools exist to infer Hi starting from Si and many of them were used in meta-analysis studies. The behind idea is to report all and only regularities and characteristics of Di reported in Si by its authors. We will see in the sample section an approach.

The second aspect of the model is that, for many reasons, every study does not report, generally speaking, the elementary data behind it. It reports only the belonging of a specific statistical unit to a class. Then, it is defined a set of classes $\{\mathbf{C}_1, ..., \mathbf{C}_s\}$, where $\mathbf{C}_i$ is the vector of classes

associated with the study Si. C_{ij} can be equal to C_{hl}: two classes associated with two studies can coincide, i.e., they correspond to the same sub partition in the space of features of the multidimensional variable V. Then, the unconditional probability of $\mathbf{x}$ can be represented by the following three-level hierarchical mixture model:

$$P(\mathbf{x} \mid \Theta) = \sum_{s=1}^{S} P_s \sum_{k=1}^{K_s} P(C_k \mid k, s) P(\mathbf{x} \mid C_k, s). \tag{6}$$

There are three set of parameters:

Σ: set of parameters for the priors over the sources, $\{P_s\}$;

Λ: set of parameters for the conditional probabilities of the classes, given a source, $\{P(C_k \mid k, s)\}$;

Ψ: set of parameters for the conditional probability of x, given a source and a class in the source, $\{P(\mathbf{x} \mid C_k, s)\}$.

2.2.1. Class definition in a data source

Let us start with an example, $X = AGE, SEX, V$, where AGE is the age of a statistical unit, SEX is its sex and V is a diagnosis result. Often, and generally speaking, a meta-analysis starts from a number of studies and, for every study, has not the data for the elementary statistical units, but only a classification of it. For example, for the study S1 data captured from it can be synthesized with the table:

SEX	AGE	V
male	34	12
female	28	11

Then, associated with the study S1 are defined two classes. The AGE is the mean age for the group of units of the same sex.

In the proposed methodology, meta-analysis has a class definition step. Depending on the studies and on the data that they report, class definition could not be trivial or unambiguous. For example, a study could

report two distinct variables V_i and V_j, and the others aggregated with respect to the modalities of V_i and with respect to the modalities of V_j. It is always possible to build a unique classification using the hypothesis of uniform distribution in case of absence of information.

Example. The starting point is a study with three variables $X1$, $X2$ and $X3$. To better understand the example, let us give a semantic to them: $X1$ is AGE and can have value (J (young), M (middle), O (old)), $X2$ is the index of presence of a symptom (1-30) and $X3$ is a presence/absence variable (telling if a drug has been used). The study reports two aggregated data tables:

AGE	PANSS	N	DRUG	PANSS	N
J	18	20	0	17	70
M	16	35	1	24	30
O	21	45			

There is a sample of 100 units. In a first analysis, common use of PANSS tell us that Prob(AGE = J and DRUG = 1) has to be low because units with AGE = J and DRUG = 1 have associated expected values of PANSS very different (18 vs 24).

From the first table above we can make the following assumptions:

$$AGE \propto MN(pj = 0.2,\ pm = 0.35,\ po = 0.45)$$

$$PANSS \mid AGE \propto N(\mu, \sigma),$$

where μ is the value reported in the above table for every class.

From the second table we can make the following assumptions:

$$DRUG \propto bN(p_0 = 0.7,\ p_1 = 0.3)$$

$$PANSS/DRUG \propto N(\grave{\imath}, \acute{o}),$$

where μ is the value reported in the above table for every class.

The object of the estimation phase is Prob(AGE, PANSS, DRUG), that plays the role of $\{P(C_k \mid k, s)\}$. From probability theory

$$P(AGE, PANSS, DRUG)$$

$$= P(AGE)*P(PANSS/AGE)*P(DRUG/AGE, PANSS). \quad (7)$$

The sample study does not report correlation data or correlation information regarding DRUG and AGE, then

$$P(DRUG/AGE, PANSS) = P(DRUG/PANSS). \quad (8)$$

From the Bayes theorem

$$P(DRUG/PANSS) = P(DRUG)*P(PANSS/DRUG)/P(PANSS), \quad (9)$$

where

$$P(PANSS) = P(PANSS/DRUG = 0)*P(DRUG = 0)$$

$$+ P(PANSS/DRUG = 1)*P(DRUG = 1). \quad (10)$$

Then we have all the elements to calculate Prob(AGE, PANSS, DRUG). This probability can also be used to estimate the elementary data matrix associated with the study. For example, we can use Monte Carlo methods to generate samples and $P(X)$ as a validation/ acceptation rule. Main expected result: the elementary data matrix reflects all and only the properties reported in the study.

2.3. Model properties

There are many studies about properties of the Bayes learning approach and of the mixture of expert models. The properties analyzed are risk bounds and convergence of the estimated concept to the correct unknown concept. In the introduction we said that the main problem in a meta-analysis is to highlight a concept, shared between all studies and latent in them. Here the goal is to highlight the ability of the Bayes learning approach to discover shared latent concepts. Concepts are borrowed from Kolmogorov complexity and MDL (minimum description length) theories.

Kolmogorov approach [4] to the analysis of the complexity of a specified pattern defines it the length of the simplest Turing machine that generates this pattern. A Turing machine, informally speaking, is an automaton that can be used to generate strings. The idea behind this approach, is that the generation of a regular pattern (for example a string that is a sequence of all and only '1') is a simple task and can be made by a simple Turing machine (for example constituted of few states). The string 1001000110111 is a more complex string and then a complex Turing machine is necessary to generate it. Inversely, if a simple Turing machine generates a pattern, this pattern has to be simple or regular.

This approach has been applied to the machine learning problem, viewed as the problem of model selection, using a set of data or database. There is a predefined class of models, for example a family of probability distributions, and there is a database to help selection of the generator model in this class. The minimum description length (MDL) approach [3] to model selection chooses a model that provides the shortest code length for the data that form the database. For example, if a model M is a linear relationship between two variables, given a couple $\langle x, y\rangle$, its length, given M, is $ln(x)+ln(y'-y)$, where y' is $M(x)$. If $M(x) = y$, then the length of $\langle x, y\rangle$ is $ln(x)$. The key is that a relatively simple quantitative measure, the length of $\langle x, y\rangle$ given M, captures a complex concept, the ability of M to describe $\langle x, y\rangle$.

Every model has associated a code that is a function to map strings to other strings. Every code and then every model have associated a code length function to map strings to integers that represent the length of their representation or code. Given a model, its associated code length function can be evaluated with respect to a predefined dataset. A model, giving a short representation of this dataset, captures its properties and regularities. And then this model highlights all concepts behind the dataset.

2.4. Inference using the model

In the inference step the obtained conjunctive distribution $P(X)$ has to be used to do meta-analysis and then to highlight meta-analysis concepts from it. There are three possible methodologies and associated sets of tools:

(1) separation of the group of variables in two subsets, dependent variables and explicate or meta-analytic variables and then incremental and exhaustive study of relationship between two subgroups, using marginal probabilities;

(2) starting from the distribution of $P(X)$, Monte Carlo methods can be used to find high probability regions and then local modes [5]. Last step is to attribute a semantic for every mode;

(3) starting from the distribution of $P(X)$, Monte Carlo methods can be used to generate samples [5] and to test specific theoretical model (i.e., a regression model).

2.5. An application

This sample is drawn from [6]. The starting point is the data matrix:

CASE/VAR.	DvS	SEX	PA–PO	PA–NE	AGE	N
CASE 1	–1	0.72	18.07	15.93	37.60	93
CASE 2	1	0.46	14.25	16.71	39.70	114
CASE 3	–1	0.89	28.5	18.9	23.50	8
CASE 4	1	0.89	28.6	26.6	23.50	8
CASE 5	–1	1.00	21.8	20.1	29.10	54
CASE 6	1	1.00	21.1	22.9	29.10	71
CASE 7	–1	0.64	18.13	17.03	42.87	70
CASE 8	1	0.58	16.75	17.08	44.48	52
CASE 9	–1	1.00	19	15.5	41.10	48
CASE 10	1	0.90	19.5	18.9	39.70	40
CASE 11	–1	0.67	17.1	22.2	32.20	18
CASE 12	1	0.62	15.9	23.3	33.70	32
CASE 13	–1	0.94	23.94	23.19	31.90	16
CASE 14	1	0.88	21.11	23.97	34.90	35
CASE 15	–1	0.88	15.3	14	34.30	33

DvS is a diagnosis type indicator (–1 dual, 1 single), SEX is the percentage of men in the sample, PA–PO are positive PANSS, PA–NE are negative PANSS, AGE is the averaged age of the sample and N is the number of statistical units in the sample. There are 8 studies, every one reports two classes described adjacently in the above table. Then the first two raw identify the first study and so on.

The main goal of the meta-analysis is the study of divergence effects related to the diagnosis (dual versus single).

The first step of the proposed methodology concerns the definition of $P(X)$, where $X = (DvS, SEX, PA - PO, PA - NE, AGE)$:

$$P(\mathbf{x} \mid \Theta) = \sum_{s=1}^{8} P_s \sum_{k=1}^{2} P(C_k \mid k, s) P(\mathbf{x} \mid C_k, s). \tag{11}$$

There are 8 studies. $P(s) = Ns/N$ for $s = 1, 2, ..., 8$ with $N = N_1 + \cdots + N_8$. $P(C_k \mid k, s) = N_s k / N_s$ for $k = 1, 2$. The values $N_s k$ are reported in the above table in the last column. Then

$$P(\mathbf{x} \mid C_k, s) = p(DvS \mid C_k, s) * p(SEX \mid C_k, s)$$
$$* p(PA - PO \mid C_k, s) * p(PA - NE \mid C_k, s) * p(AGE \mid C_k, s).$$

For the variables $X = (DvS, SEX, PA - PO, PA - NE, AGE)$ we made the following hypotheses:

$P(X/j, s) = P(X1/j, s) * \cdots * P(X5/j, s)$ local independence

$DvS \propto Bi(p = N_{sj}/N_s)$

$SEX \propto Bi(p = v_{sj})$ for example in the first study $p = 0.72$

$PA - PO \propto N(\mu = v_{sj}, \sigma = CV_j/\alpha)$

$PA - NE \propto N(\mu = v_{sj}, \sigma = CV_j/\alpha)$

$PA - AGE \propto N(\mu = v_{sj}, \sigma = CV_j/\alpha)$.

When the supposed distribution is the normal $N(\mu, \sigma)$, the μ parameter is the value collected for that study in that class. The σ parameter is the max value in the class minus the min value, all divided for α. The value of α chosen is 10. Greater values of α amplify differ between $DvS = -1$ and $Dvs = 1$ for a conditional variable in absolute order, but the relative ordering of conditional variables with refer to the DIFF value is the same. This fact gives absolute value to the tables below.

At this point we calculated the following groups of probabilities:

Probability	Values of parameters
$P(PA - PO/DvS)$	$PA - PO = (10, ..., 25)$, $DvS = \{-1, 1\}$
$P(PA - PO/DvS, SEX)$	$PA - PO = (10, ..., 25)$, $DvS = \{-1, 1\}$, $SEX = \{0, 1\}$
$P(PA - PO/DvS, AGE)$	$PA - PO = (10, ..., 25)$, $DvS = \{-1, 1\}$, $AGE = \{20, ..., 50\}$
$P(PA - PO/DvS, PA - NE)$	$PA - PO = (10, ..., 25)$, $DvS = -1, 1$, $PA - NE = \{10, ..., 25\}$
$P(PA - NE/DvS)$	$PA - NE = (10, ..., 25)$, $DvS = \{-1, 1\}$
$P(PA - NE/DvS, SEX)$	$PA - NE = (10, ..., 25)$, $DvS = \{-1, 1\}$, $SEX = \{0, 1\}$
$P(PA - NE/DvS, AGE)$	$PA - NE = (10, ..., 25)$, $DvS = \{-1, 1\}$, $AGE = \{20, ..., 50\}$
$P(PA - NE/DvS, PA - PO)$	$PA - NE = (10, ..., 25)$, $DvS = -1, 1$, $PA - PO = \{10, ..., 25\}$

Then, given for example

$$P(PA - PO = p/DvS = -1)$$

and

$$P(PA - PO = p/DvS = 1),$$

we calculate the absolute difference for the set of integer values reported in the above table. For example:

$$DIFF = \sum_{p=10}^{25} abs(P(PA - PO = p/DvS = 1) - P(PA - PO = p/DvS = -1)).$$

Below there is table with the highest values of DIFF. They represent the main concepts learned from data and then the goal of meta analysis.

Probability	DIFF
$P(PA - PO/DvS,\ SEX = \text{female})$	1.056
$P(PA - PO/DvS,\ AGE = 40 - 42)$	1.036
$P(PA - PO/DvS,\ PA - NE = 15 - 18)$	1.185
$P(PA - NE/DvS,\ PA - PO = 10 - 13)$	1.215

2.6. Conclusions

Main goals of this study are:

- the precise characterization of some well-known problems of meta-analysis like publication bias or comparison of heterogeneous variables. For the first problem, every meta-analysis must specify the exact set of problems it can answer to. Bias in the explicative variables with refer to complete populations, means a small set of concepts associated to a meta-analysis. For the second problem the comparisons between studies have to regard probability relationships and not absolute variables values.

- the precise characterization of the main goal of meta-analysis. Starting from a definition of concept, meta-analysis is defined the research of concepts that are shared and latent in a set of studies. With this definition, a general skeleton for meta-analysis has been proposed, general with respect to all possible inputs (studies) of meta-analysis. Specific statistical tools can be also used, but in the context of this skeleton.

References

[1] A. Azran and R. Meirn, Data Dependent Risk Bounds for Hierarchical Mixture of Experts Classifiers, The 17th Annual Conference on Learning Theory, COLT, 2004.

[2] G. V. Glass, Primary, secondary, and meta-analysis of research, Educational Researcher 5 (1976), 3-8.

[3] P. Grnwald, A tutorial introduction to the minimum description length principle, Advances in Minimum Description Length: Theory and Applications, P. Grnwald, I. J. Myung and M. Pitt, eds., MIT Press, 2005.

[4] M. Li and P. Vitanyi, An Introduction to Kolmogorov Complexity and its Applications, Springer-Verlag, 1997.

[5] R. M. Neal, Probabilistic Inference Using Markov Chain Monte Carlo Methods, Department of Computer Science, University of Toronto, 1993.

[6] A. Talamo, F. Centorrino, L. Tondo, A. Dimitri, J. Hennen and R. J. Baldessarini, Comorbid substance-use in schizophrenia: relation to positive and negative symptoms, Schizophrenia Research, 2006.

[7] M. K. Titsias and A. Likas, Mixture of experts classification using a hierarchical mixture model, Neural Computation 14 (2002), 2221-2244.

■