Data Fusion for Decision Making

Stéphane Chauvin
Doctor Mathematic Applied
stephane.chauvin@r2c-system.com
January 2000

Keywords: *Fusion-Optimisation, Knowledge Discovery, Multi-Classifiers, Bayesian Network, Data mining.*

Abstract

This paper presents a fusion model called System Optimisation by Fusion of Information (SOFI) which is derived from Bayesian networks. It is designed to improve user decision-making through the help of the more active and powerful Decision Support System (DSS) that is used to reduce the "Information Gap" for data mining applications. We show here the interesting approach for data mining to integrate the SOFI fusion module into a data mining platform taking the example of o a new platform called Track Toolkit gathering decision tree, cluster tool and neuronal network. This paper specifies the SOFI software which is a combining classifier tool and which can be integrated in every kind of data mining platform that gathers several classifications coming from data mining tools.

Technically speaking, we introduce here an original manner to both learn and build automatically a relational graph to represent the information. We focus our presentation over the logical process and review the components of data mining technology to provide the material for fusion that is used to discover knowledge and for which the a priori information is integrated step by step. Then the purpose of the SOFI model is to combine decisions to deliver one classification optimising the whole of both knowledge and decision discovered. The SOFI generates automatically a General Bayesian Network (GBNs) classifier optimised by a stochastic process in order to discover decision rules into data.

We reserve a large section for the SOFI application which is dedicated here to produce a model classification for forecasting problems. This part shows that both SOFI is especially designed for the Knowledge Discovery in Databases (KDD) and has the advantage to evaluate each separated decision and to propose a quick access to a synthetic result for the end-user being able to rebuild a complete analysis. We compare the result between some other data mining technique like cluster and decision tree and illustrated by SPSS modules. We conclude with the model performance of the SOFI model.

1 INTRODUCTION

This paper presents a fusion model called System Optimisation by Fusion of Information (SOFI) which is derived from Bayesian networks [1]. It is designed to improve user decision-making through the help of the more active and powerful Decision Support System (DSS) for data mining applications. It is taking as an example the data mining platform called the Track Toolkit ([1]), which provides the fundamental component processes in data mining used to manage temporal data for the knowledge discovery. The fusion concept is used on such workstation in order to optimise the knowledge gathered through classifications and/or segmentations on a given set of data (see [2]) in order to gauge expert knowledge by gathering events step by step. Fusion gain is the increase in the quality of available both information and knowledge for a general Knowledge Discovery problem (see the report summarises state of the art in [3]).

[1] TRACK is an ESPRIT project (European Strategic Program for Research in Information Technology) with seven partners involved in its development: three industrial software developers companies from Spain (Ibermatica), UK (AniteSystem) and France (Isoft), three representative financial companies form each country (Banco Santander, NationWide, Caisse d'Épargne) and the ITC R&D Institute.

An example of the use of these data mining platforms, taken from the financial sector, is the effective "top-down strategic systems planning" which is complemented by fast-track technology: return of investment measures, predictive style, pro-active stance, etc. Analyses can be made by the step by step integration of automatic data processing and natural feedback in order to locate the knowledge appearing in the data. The useful of the data mining platform can be seen into the well-know churn problem initially investigated for telecom applications to detect and predict the a priori reason of changing customer's behaviour (see amongst others [4]). Method to answer is processed in two steps: first process given customer valorisation and second given customer activities, both are measured and classified. First step consists to both discover the customer profiling with their social states and score the reliable set of customer for generating a significant and macro-economic return of investment measures. This analyse need clearly a calculation module which integrates every kind of both numerical and logical functions implemented within the data mining platform to perform, for example, the Life Time Value of each customer. After that, segmentation and/or classification process are required to specify the expectation group onto the whole of customer. Second step, cumulating the historical marketing data of the customer, data mining tools as the decision tree are used to detect the active and no-active population calculating the predictive the following factors of the attrition: as linear lowliness consummation, artefact into the behaviour changing the initial product by a cheaper product, number of calling received to the customer attention service, do not pay at time, no answer at the cumulated advantage offered from marketing department. End-user result is, using the SOFI module, the optimisation between the numerical data measuring both the customer value and risk activities and, the symbolic data showing the groups of both customer reliable and the level of activity for the company. Classification combination allows the end-user to separate their reasoning in sub-task and takes the optimal decision finding the optimal set of customer marketing target. This method for stopping the customer "churner" is implemented in a European CRM company.

The interest of the fusion process appears for the Online analytical processing method [5]. Effectively, OLAP delivers a multidimensional cube which describes the dependency between variables in order to discover knowledge. Fusion is done here to reduce the dimensionally problem of the OLAP approach (also called OLAP data mining) and reduces the data solution space by aggregating a priori pre-analysis and focusing the solution in an appropriate way (the reduced data space and decision space is here called System Information Layer – SIL). Track toolkit appears here as an intelligent user friendly data mining OLAP. The final process is concluded by the combining classifier allowing to user to have access at an optimised view integrating its own a priori knowledge.

Then, many data mining tasks can be viewed as classification each described by a set of features. Learning accurate classifiers from pre-classified data is the purpose of the SOFI model showed as a fusion of analyses and based on ICA technology (see the Independent Component Analysis in [6] and [7]). We detail in this paper the whole fusion process which includes everything from the data both measurement (data preparation) and expert decision scheme to the optimisation scheme. In 4 parts, we specify the general data mining platform and the material for fusion, the type of measure used to evaluate a single classification, the Bayesian Network for fusion. We present an original application of a time series analysis problem.

More specifically, section 2 deals with the fusion design within the toolkit. First, it is reviewed the concept of expert which refers to everything from individual data mining technology to end-users offering an analysis based on specific knowledge and/or experience. Secondly, SOFI model is introduced as a mixed fusion model (or a multi-dimensional classifier, see [8]) which establishes a link between the informational content of a database (numerical data source) and the knowledge provided by a set of experts (symbolic a priori data) constructing the SIL space and avoiding the well-know learning phase problem. Thirdly, we detail the architecture which is centralised and applied for the accumulating partial analyses in parallel. Section 3 describes the mathematical formalism of the accuracy measurement of the independence between the set of classes which is the degree of explanation of the classes by the numerical variables. We take here the most useful function to measure the dependency between class coming from [9]. Section 4 deals with the process of fusion by introducing the General Bayesian Networks (GBNs) model [10]. A GBN supports the information (SIL space) and is the material to consult knowledge and to merge information (see, amongst others [11] and for a general view [12]). We then deduce from the aforementioned model a mechanism to consult and evaluate each individual analysis. Finally, we introduce the commonly process used for the energy minimisation relaxing independence assumptions for a combinatorial problem. It is specified the calculus of the probability of transition between two states which is guided by the symbolical knowledge coming from the experts. Fusion benefits are obtained by the reduction of the solution space thanks to the building up of intermediate steps of knowledge.

In order to illustrate the process of fusion, we take the example of a time series analysis problem in section 5. It deals with the medium-term forecasting ability of various volatility models in the foreign exchange market. Advances in time series modelling such as, amongst others, ARCH/GARCH and/or stochastic volatility models, have made it possible to integrate the time-varying nature of volatility and correlation and thus to relax such embarrassing assumptions as constant volatility and constant correlation. The predictive power of several time series models of currency volatility (homoskedastic, ARMA, GARCH and stochastic volatility) are considered here, using daily data from January 1991 through March 1999 for 6 major exchange rates: DEM/JPY, GBP/DEM, GBP/USD, USD/CHF, USD/DEM and USD/JPY ([2]) [62]. Using the Track platform, fusion involves a readable classification of the data series forecasting models according to criteria of performance. Results are compared between results coming from both supervised clustering tool and decision tree tool illustrated by SPSS graphic module.

In conclusion, we summarise the characteristics of the SOFI model which uses a non-supervised learning method and non-parametric formalism on both numerical (variables to be explained) and symbolic data (the analyses involved by R experts). We conclude by the time CPU performance which is good for particular application.

2 DATA MINING PLATFORM FOR FUSION

The Track toolkit mechanism is designed to aggregate analyses made in parallel. Two basics steps are considered in this section: 1) data collection and data preparation delivered by R analyses, 2) data knowledge discovered and analysed through fusion, making the optimal link between data sources and the R a priori knowledge sets. The fusion functionality is that of a search engine designed to integrate all kinds of symbolic and numeric information [14]. We briefly review the overall quality of this kind of platform through the definition of Expert, the fusion techniques and the architecture of the fusion process.

2.1 Expert Specification

The general objective of data mining tools is the search for the desired state of a vector in a set of profile identifiers. Therefore, an individual data mining technology is a sensor, which filters both symbolic and numerical vectors of information in order to insert calculus into the data, adding measures of information for the understanding of the data set. Data mining tools involve either a scalar number solution (for example scoring identifiers by a value into 0 to 1) or a decision label that enhances the meaning of each class in a more formal language (i.e. label "AGE > 32" gathering people who are more than 32 years old). To have a complete data mining platform, some data mining tools are need which should be able to manage, filter and project temporal data in one scalar measure as has been done in the Track Data Manipulation Module (TDMM).

Each individual tool uses both its number of parameters, decreasing or increasing in system performance, and its ergonomic interface to increase the users participation. New software production is designed to have the best compromise between time processing and data visualisation allowing the users to have access to the systems solutions with the degree of accuracy that they require. Most of these tools use a specific technology, e.g. neural networks, genetic algorithms, induction, statistics to resolve the problem of the optimum characteristic choice of a data sample. The best way to resolve a pool process would be a statistical one (Systat, Sas, Spss) and to resolve a KDD problem, it is preferable to use the hierarchical modelling of a figure tree that involves a particular point of view of the database (Decision Tree, OLAP, Bayesian Network). Artificial Intelligence and Expert Systems manage database knowledge by using syntax rules for rapid decisions integrating the symbolic information (inaccurate and uncertain information). Supervised Learning Neural Nets have been amongst the most universally applied and successful neural network models, with applications in many industries. They are used for modelling complex processes, forecasting and decision making where historical data is available (see [15] for a comparison between Neural Network and a Nearest-Neighbor for classification). This kind of tool has a bad reputation for being a black-box but could be made more user-friendly through the use of adaptive methods along with prior knowledge that takes fuzzy rule forms. Artificial Intelligence systems compete with statistical and stochastic processes in the race to find solutions for encoding, computing and architectural

[2] We use the notation of the International Organisation for Standardisation (IOS), respectively DEM/JPY for the Deutsche Mark against the Japanese Yen, GBP/DEM and GBP/USD for the Deutsche Mark and the US Dollar against the Pound Sterling, USD/CHF, USD/DEM and USD/JPY for the Swiss Franc, Deutsche Mark and Japanese Yen against the US Dollar.

flexibility. Besides, statistical systems display a number of difficulties in interpreting results in a natural language and systems using reasoning have weak control over the true measurements. There is a significant number of different approaches taken by commercial tools with the aim of identifying clusters of data behaviour. Their goal is to allow for population segmentation and decision analysis. Hence, for the same application, the individual software calculations rarely give a consensus solution (see [16]).

An expert is a mechanism sensor-user which increases Shannon's entropy (an increase of knowledge) and, for data mining purposes, provides a X-out-to-Y matching classification with |X|>|Y| such that X is the data set and Y is a classification result in a labelling space $\Re$. We propose in section 3 a measure deriving from [17], which takes into account the intrinsic "depth process" used to lead to the evaluation of each classification Y.

2.2 Data Fusion Techniques

Each model that synthesises knowledge refers to the fusion concept: see, for instance, mobile robots (Stanford Mobile Robot, Crowley's mobile robot [18], [19]), dense co-operative environment systems used by the military investigation for targeting information through diagnosis and control [20], [21], [22], medical applications (Mycin system) and multi-temporal applications [23]. Above and beyond the great heterogeneity of the applications, some research teams try to give a scientific definition of the fusion concept and/or a set of concepts defined in the following manner [24]:

1) Data Fusion consists of data streams of row measurements coming from different sensors. For financial data, marketing departments building profile rows (or Identifier) according to some characteristic fields would do this.
2) Feature Fusion concerns the combination of features extracted from a set of rows. Here, we speak of topology features for data mining purposes. The complexity of the features is intuitively measured according to the profile discrimination characteristics delivered by an expert.
3) Decision Fusion is directly related to the main purpose of the SOFI model. It is the process of combining partial, soft or hard decisions that involve the relevant features introduced by the different experts.

The data source and the number of analyses (noted by R) provide the material for the fusion for SOFI. Hence, the data fusion problem could be expressed as a decision problem concerning a proposition truth or a probability coming from different experts (see, for a general view, [25]). Therefore, the challenge is to take into account the complicated character of the information: redundant, complementary analyses with inaccurate, incomplete and uncertain information. In the literature on the subject, the fundamental choice between models is still analysed and the Fuzzy set approach [26], [27], Evidence Theory [28], [29], [30] or the probabilistic view [31], [32] are in competition. In [33], method combines several Neural Networks and shows that the performance of individual neural networks could be improved significantly. The main difficulty with classical data fusion techniques like [34] is that these methods require the estimation of a lot of parameters. They often require statistical NP-Complete tests as the Neyman-Pearson Test [35] or Bayesian rule [36]. Then, model must be either non-parametric or parametric. Industrial requirements justify the use of non-parametric methods, which open the way to fast algorithms in order to directly estimate a classifier from the data [9].

A specific fusion model should be agreed with the "Objective synergy" terms due to [37]: "fusion is a formal framework in which are expressed means and tools for the alliance of data originating from different sources. It aims at obtaining information of greater quality; the exact definition of "greater quality" will depend upon the application. For our purposes, the quality results are in decrease of error measurements and perform a non-disturbed information system. For data mining, "greater quality" would be defined as the best decision rule. For our application, a rule appears when it optimises the "depth process" viewed in section 3.

The initial X-out-to-Y problem resolved by each expert is the same for fusion but with a different decision space called System Information Layer (SIL). SIL is obtained by projecting the data into the same decision space (see a similar view in [38]). It will be presented in section 4. The SOFI process becomes the following matching problem: SIL-out-to-Y* with Y* being a particular combination of the R classifications (cumulative learning) used to produce a set of independent rules. This involves a specific model of fusion of distributed experts, which can be viewed as a multi-dimensional classifier [39] using the GBN technology [40] and [41].

2.3 Architecture Work Flow for Fusion

In order to initialise the SIL-out-to-Y problem, data mining platforms involve efficient systems if user objectives are taken into account. Users wish to broaden their general knowledge of their own data in order to include multiple strategies and, also, to target knowledge.

- Population Segmentation: the tool analyses information in the database from which individuals share similar characteristics.
- Decision Data Analysis: a decision support toolkit may be more appropriately used to design a decision procedure that forecasts the behaviour of a particular or new individual, or to explain the variation of certain variables against other variables.

Then, in order to reach a solution, the SOFI module offers the following scheme which accumulates R parallel analyses (figure 2.1).

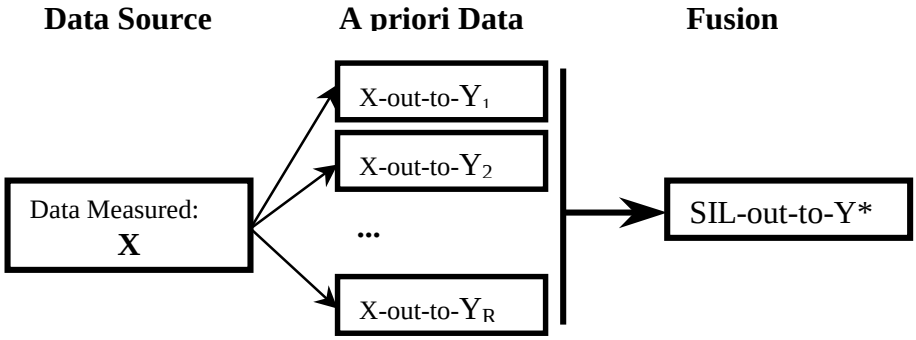


Figure 2.1: *SOFI manages both Data Base information (Data Source X) and R parallel analyses (A Priori Data); SOFI module evaluates each individual analysis and optimises a global solution according to the objectives established by the user (multi-classifier).*

To avoid the complexity of the fusion architecture for a serial aggregation flow, a parallel scheme is required to obtain feedback in order to eliminate the experts or to add new ones and experiment the "when" the fusion is appropriate. In this case, fusion consists of calculating a final score by mixing a partial one in an associative way. Then, the user explores the data according to the questions used to direct the search for the dependence/independence between identifiers and variables in order to gauge decision rules.

The notion of dependency is well supported by the Bayesian model which integrates the complexity of the information system by the statistical relational graph [42]. A graph provides an efficient method for reasoning ([43], [44], [45], [31]) and allows data mining platforms to integrate most types of information [46]: trees, rules. The four Lukasiewiez's laws allow both logical relations and probabilistic interpretation generating a GBN: negation, implication, conjunction and disjunction. As with the connectionist approaches [47], a mathematical module of fusion is given in order to include R analyses into a symbolic-numerical relational graph that contains the topological link between valued objects [48]. The system of integration of information is done by the re-normalisation of nodes in a relational graph. This dual graph is deduced and allows to formulate the probability link onto the framework, which will be described in section 4, as a Bayesian network.

3 MEASURING EXPERT MODELS

Let $\mathbf{X}$ be the original data set rearranged according to N row measurements described by M random variables such that $\mathbf{X} = \{ \mathbf{X}_{ij} \}_{i = 1,...N \; ; \; j = 1,..., M}$. $\mathbf{X}$ is associated with a distance measuring the dissimilarity between two records, according to the M characteristics. There exists a function $\boldsymbol{\phi}$ which associates the population $\mathbf{X}$ in a set $\mathbf{Y}$ and gathers the records according to selected fields and constraints. For our purposes, $\mathbf{Y} \subseteq \Re$ is a classification of a finite class number K associated with a label vector: $\mathbf{Y = \{ Y_1, Y_2, \ldots, Y_K \}}$. $\Re$ is a part of the discrete finite space 3^K ($K \in N^*$). The marginal partition such as $Y_k = \{Y_{k,j}\}_{j = 1,..., M}$ is defined by the upper and lower boundary values of the M variables:

(1) $\quad \forall \mathbf{k} = \{1, 2, \ldots, K\}, \qquad \mathbf{Y_k} = (\{ Y_{kj} = (Y^-_{kj} \; ; Y^+_{kj}) \}_{j=1,...,M} ; c(k))$

where $Y^-_{kj} = \min_{i=1,...,N} [X_{ij} \in Y_{kj}]$ and $Y^+_{kj} = \max_{i=1,...,N} [X_{ij} \in Y_{kj}]$ are the boundaries of the variable *j* associated with a label c(k) of the class k. The lattice generated by the K classes described by (1), is called System Information Layer (SIL). The SIL format could also bear temporal variables. The truncature Min-Max object is also called Hoxel (High Order of pixels [49]).

Each class $\mathbf{Y_k}$ cuts the hyper-cube **X** over the M directions. The "greater quality" aforementioned in section 2, is calculated according to the discrimination power of the class k. This would be the "depth process" generated by an expert. Then, the underlying mathematical hypothesis of each individual technology produces the criteria that the function ϕ optimises. To enhance the evaluation of experts, the probability measure is introduced in order to deduce a general formula, which will be the final decision rule ϕ.

3.1 Measure of Reference

Let ϕ be a positive function. A configuration **Y** over **X** is the application ϕ such that:

$$3^M \times 3^K \rightarrow 3^{*+}; (x,y) \mapsto \phi(x,y).$$

Let the probability model be $P(\phi(x,y)) = \exp(-\max(\phi(x,y),\varepsilon))$ ($\varepsilon>0$) (see [50]). The function ϕ can be estimated as a Chi-square distribution considering the reference probability equal to the frequency density:

$$\mu(\mathbf{Y_k}) = \mathbf{N_k}/\mathbf{N} \quad \text{with} \quad \sum_{k=1,...,K} \mu(\mathbf{Y_k}) = \sum_{k=1,...,K} \mathbf{N_k}/\mathbf{N} = 1$$

with $\mathbf{N_k}$ the number of identifiers contained in the class number k. $\mu(\mathbf{Y_k})$ will hereafter be referred to as μ_k. We draw up a number of N rows, which cover K levels of energy. The probability of the N-Training configuration is equal to the multinomial law:

$$(2) \qquad P(\phi(x,y)) = \Pi_{k=1,...,K} \mu_l^{NP_k} = \exp(-N.(D(P\|\mu) + H(P)))$$

where P is the empirical distribution, H(P) is the information rate of the distribution P (Shannon Entropy) and $D(P \| \mu)$ is the Kullback Leibler Entropy measuring the "distance" between both the probability distributions P and μ (see [51] for a complete description). In the case that the distribution P has a great probability to be repeated, the state of P is called "type" [52] and the distance between μ and P is close to zero implying the probability model $P(\phi(x, y)) = \exp(-N.H(P))$. Focusing on the set of the "type" series distributions noted T(P), the probability for having only the distributions concentrated around the μ distribution is such that:

$$P(\phi(x, y) / T(P)) = n(P) . \exp(-N. D(P \|\mu))$$

where $n(P) \in [1/(N+1)^K, 1]$ ([3]). This can be re-written using the law of large numbers when μ is close to P, $\mu/P = 1+\varepsilon/P$. Using the development of the log in the Kullback Leibler formula, we derive the Chi-Square distribution. (see, amongst others, [53]):

$$(3) \qquad \phi(x, y) \approx \log[P(\phi(x, y) / T(\mathbf{P}))] = N.\Sigma_{k=1,...,K} (\mu_k - P_k)^2 / \mu_k^2 \rightarrow \chi^2_{K-1}$$

Then, this formula is the $\phi(x, y)$ positive function that evaluates the classification Y and corresponds to a Chi-Square distribution with K-1 degrees of freedom. This is used to formalise a Chi-Square test for which a classification Y has a distribution P close to μ if $\phi(x, y) < V_\alpha$. V_α is read in the χ^2_{K-1} table or calculated from the Wilson-Hilferty's Chi-Square approximation formulae. α is the probability threshold fixed to 0.05.

3.2 The Decision Rule

Taking the hypothesis that the best model is that which concentrates all elements $x \in X$ in K classes without ambiguity, the research probability is the result of the independence between the classes: $\forall k \neq k'$, $P(Y_k, Y_{k'}) = P(Y_k).P(Y_{k'})$. Defining the complementary class $\neg Y_k$ such that $\neg Y_k = \cup_{k' \neq k} Y_{k'}$, the conditional probability $P_k = P(Y_k / \neg Y_k)$ is calculated by a counter function such that: $P_k = \tilde{N}_k/N$ with $\tilde{N}_k = \sum_{x \in X} \{ \mathbf{x} / P(\phi(x, y_k)) > 0 \}$ and

[3] Using the thermodynamic multiplicity $N!/(N_1 ! N_2 ! ... N_K !)$ and the Stirling approximation when one only considers the typical series P.

the reference distribution calculated by $\boldsymbol{N}_k = \sum_{x \in Y_k} \{ \mathbf{x} \}$. Using the (3) aforementioned formulae, the evaluation function of a classification **Y** is equal to:

(4) $$\phi(x, y) = \Sigma_{k=1,...,K} (\boldsymbol{N}_k / N - \tilde{N}_k / N)^2 / (\boldsymbol{N}_k / N)^2 = \Sigma_{k=1,...,K} (\tilde{N}_k / \boldsymbol{N}_k - 1)^2 < K$$

with $\tilde{N}_k / \boldsymbol{N}_k \leq 1$.Therefore, a good solution is when $P=\mu$ and $\tilde{N}_k / \boldsymbol{N}_k$ is equal to 1 for each class and involves the independence between two classes: $P_k = \mu_k = \boldsymbol{N}_k / N$. If $\phi(x, y) = K-1$, there is functional dependency between the K classes: $P(Y_k / \overline{\ }Y_k) \neq P(Y_k)$. Practically, the SOFI process looks for the minimum expectation using the weight N_k / N of the class (see similar decision rule in [13], [17]).

3.3 The Depth Process ϕ

The ϕ function is a Chi-Square for an arbitrary number M of variables. That is also true for every marginal direction of the hyper-cube Y_k. The model becomes:

(5) $$\phi (x, y) = \Sigma_{k=1,...,K} . \Sigma_{j = 1,...,M} (\tilde{N}_{kj} / \boldsymbol{N}_{kj} - 1)^2 \leq M.K$$

The $\phi(x, y)$ follows a $\chi^2_{d=inf(M-1,K-1)}$ distribution. Then, there exists an intrinsic value $\boldsymbol{m} \leq M$ which is the number of variables having the independence hypothesis accepted by a $\chi^2_{d=inf(M-1,K-1)}$ test. $\boldsymbol{m}$ is the measure of the depth criterion of the ϕ function (measure of the discrimination power of the ϕ function), which has a value lower than $(M-\boldsymbol{m}).K+\boldsymbol{m}.V_\alpha$. An expert discriminates $\boldsymbol{m}$ variables in K classes and can be either a "simple Expert" ($\boldsymbol{m} = 1$) or a "complete Expert" ($\boldsymbol{m} = M$), or somewhere in between. Then, optimisation goal is the reducing correlation between classes for all numerical variables.

A simple expert would separate groups according to one variable (poor criterion). For example, an expert would ask about people's age: "people that are between 3 to 30 years old are put in a group (young people), another group of people that are between 31 to 60 (older people)". An expert with a depth equal to $\boldsymbol{m}=2$ would separate the people' classification in new subgroups. Practically, decision trees allow to move from simple experts to complete experts. In fact, trees methods characterise a population's subgroups by highlighting their most distinctive criteria. It automatically breaks down the whole population into groups, and groups into sub-groups, integrating variables step by step. Decision trees are often used for the learning phase for Bayesian Network.

4 BAYESIAN NETWORK FOR FUSION

This part focuses on the aggregation of R classifiers. First, we analyse the expert flow aggregation, the structure of the graph underlying the network {SIL, Λ} where each node $Y_k \in$ SIL represents a domain variable and attribute, and each arc into Λ between nodes represents a binary link. This graph is deduced from the dual graph of the expert matching. The Bayesian Network is completed by the training phase of a the conditional probability of nodes measuring the dependency between them. Finally, we introduce the merging node process for synthesising R classifications and minimising the **ϕ** function.

4.1 Graph Representation

Let Γ = {¢, Ξ} be a graph matching all the experts on the labels' nodes: ¢ = **{** $k_r^{h(r)}$ **}**$_{r=1,...,R\ ;\ h(r)\ =\ 1,...,Kr}$ with the matrix of edge Ξ ($\prod_{1\le r\le R}$ |Kr| x $\prod_{1\le r\le R}$ |Kr|). Let $c_r(k) = \{k_r^1, k_r^2, k_r^3, \ldots, k_r^{|Kr|}\}$ be the label vector coming from the classification Y_r of the expert r.

The dual graph of Γ is G = {SIL, Λ} defined by the SIL space which gathers K nodes such that K is the number of separated vector label: ∀ Y_k , $Y_{k'} \in$ SIL, ∃ r $k_r \neq k'_r$ .This is the operation of the intersection (or disagreement) between the experts. The number K is inferior or equal to the product $K \le \prod_{1\le r\le R}$ |Kr|. Each node of the G graph coming from the definition (1) is defined as follow:

(6) $$Y_k = (\ \{\ Y_{kj} = (Y^-_{kj}\ ;\ Y^+_{kj})\ \}_{j=1,...,M}\ ;\ c(k) = \{k_1^{h(1)}, k_2^{h(2)}, \ldots, k_R^{h(R)}\}\ )$$

The space SIL stores the K classes and is then graphed where Λ is the matrix KxK: $\Lambda_{k,k'} = 1$ if ∃ r such that $k_r = k'_r$. Figure 4.1 shows the underlying network displayed by the graph Γ (the sub-matrix of the expert r being Γ_r) and the dual graph G.

Graph G involves the K chains of length of R in the graph Γ. The space SIL regroups all connections coming from the fine knowledge derived from the R analyses. Graph G increases in size according to the increase in the number R of experts. It is as if new data were added to a message of length R. Let G_R be a graph associated with the probability distribution P^R. In information theory, the Shannon entropy is equal to $H(P^R) = \sum_{k=1,...,KR} P^R_k . \log(P^R_k)$. The marginal information rate of a new expert is the difference $H(P^R)-H(P^{R+1})$. If this value is close to 0, the graph resulting from the aggregation of a new expert does not grow much (redundancy of information and high dependency between expert solutions). If the application introduces an expert that has a different analysis, the above difference between graphs is large and the amount of knowledge grows significantly. Therefore, the space is finite and for large R, $H(P^R) = \log(N)$.

The choice between Γ and G graphs made for the computing phase can be done according to the application. For the first representation (graph Γ), the size of the matrix Ξ always increases with a new expert. We commented before that it is not the same for graph G in the case of redundant experts. Therefore, the data mining applications involve analyses with a small number of classes. Then, the representation graph G is preferred to graph Γ.

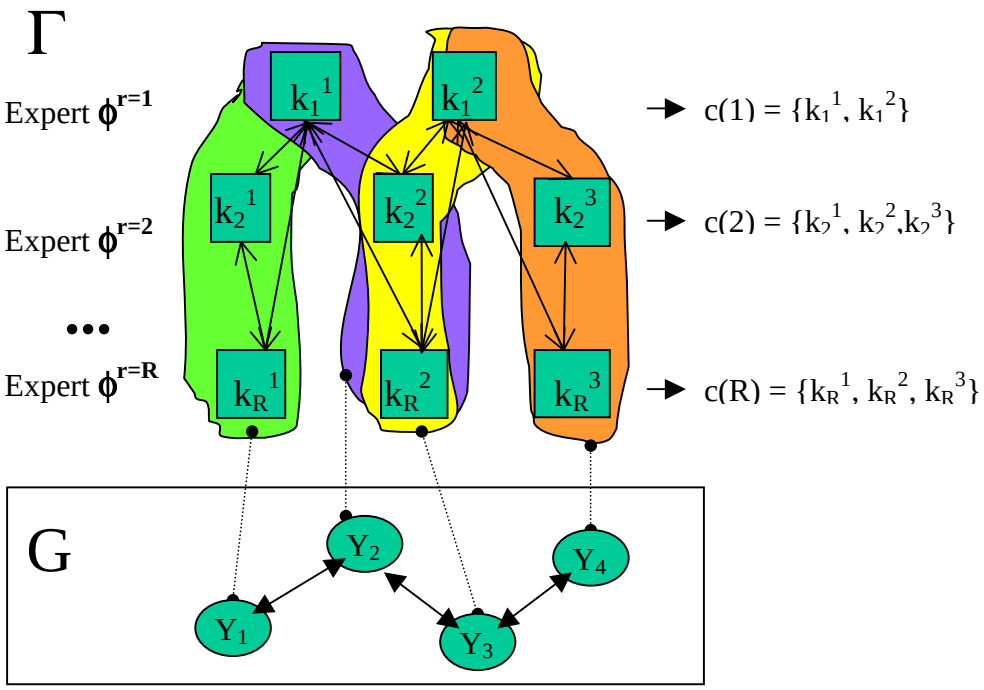


Figure 4.1: *Matching of R classifications going from the* Γ *graph to the dual* G *graph having 4 classes with* $c(Y_1) = \{k_1^1, k_2^1, \ldots, k_R^1\}$, $c(Y_2) = \{k_1^2, k_2^2, \ldots, k_R^2\}$, $c(Y_3) = \{k_1^2, k_2^2, \ldots, k_R^2\}$, $c(Y_4) = \{k_1^2, k_2^3, \ldots, k_R^3\}$.

4.2 Graph Measurement

The Bayesian formalism allows the process to determine the decision risk as measured by the fusion of the model data. The probabilistic space (Ω, Z, **μ**) associates an event and a measure **μ** for the L random variables. It is a Bayesian network if and only if there is a bijection between the events and the nodes of graph G with the following property:

$$\mu(\mathbf{Y_k}) = \prod_{r=1,\ldots R} P(k_r / V(k_r)) \qquad \text{with} \qquad \sum_{k=1,\ldots,K} \mu(\mathbf{Y_k}) = 1$$

where $V(k_r)$ is the neighbourhood topology performed by the parent nodes in Γ. According to the graph G, each chain $\mathbf{Y_k}$ is estimated by the marginal function:

$$\phi(x, Y_k) = \sum_{j=1,\ldots,M} (\tilde{N}_{kj} / N_{kj} - 1)^2 \qquad \leq \qquad M$$

The solution of the expert number r is given by the following mechanism: for two nodes $\mathbf{Y_k}$ and $\mathbf{Y_{k'}} \in$ SIL, if $k^h_r = k'^h_r$, then $\Lambda^r_{k,k'} = 1$. Therefore, each expert projection is calculated as:

(7) $$Y(r) = (Y_1, \ldots, Y_K)^t . \Lambda^r = \{Y_{\varsigma,h(r)}\}_{h(r) = 1,\ldots, Kr}$$

where $\varsigma_{h(r)}$ is the class that gathers all Y_k with the label k^j_r: $\varsigma_{h(r)} = \{ Y_k / k_r = k^{h(r)}_r \}$ ([4]). Then, the evaluation is done by the min-max filter in order to recalculate each group of K_r classes.

4.3 The SOFI Fusion Process

Fusion goal is to deliver a classification which minimises the **ϕ** function of the R aggregated classifications. Our work derives from [54], [55], [56], [57], [40] and [50]. Each expert delivers a classification Y(r) shown in (7) evaluated by the **ϕ** (x, y). Regarding the R classifications, an inference scheme is deduced from the research of

[4] Each classification can be viewed as optimising the mean-square error among all linear orthogonal transforms.

the distribution of P* which performs the minimum of the Kullback Leibler ([5]). The ϕ function is minimised by the following inference:

$$P^* = \text{argmin}_{P \in \boldsymbol{P}} [\, \phi\,(x,y) \,]$$

According to the projection formulated in (7), the following mechanism allows the distribution P(t) to be "visited" and estimated at the iteration t. We look for the stationary state of a Markov chain such as:

$$Y(0) \rightarrow Y(1) \rightarrow \ldots \rightarrow Y(T) = Y^*$$

$$P(Y(t+1)/Y(t)) = \exp\{- Y(t+1).(\Lambda^{(t+1)} - \Lambda^{(t)}) \}$$

with Y* involving K* classes minimising the function ϕ. Several strategies can be used to analyse the space of the solution. For our purpose, the optimal solution consists in discovering the K*-area of the graph G (see, for example, [58]). These problems are NP-complete and can only be resolved by stochastic strategies such as Monte Carlo or algorithms derived from it like simulated annealing. This algorithm allows the solution to be found for applications having a large number of nodes. The basic SOFI algorithm is:

1- Random initiation of the matrix $\Lambda(0)$ and estimation of temperature $1/\beta$ by the variance of the energy of the experts: let Y be the current classification
2- Choose randomly a label k^h_r and connect every point of the SIL having this label: perform $\Lambda(t)$ (D-separable process)
3- if $1/\beta(t) > 1$:
   Calculate the projection y(t) and $\phi\,(x, y(t))$
   If $\Delta\phi = \phi\,(x,y(t)) - \phi\,(x,y) < 0$, the new position is accepted: $y \leftarrow y(t)$
   If $\Delta\phi = \phi\,(x,y(t)) - \phi\,(x,y) > 0$, the new position is accepted with the probability
   $U < \exp(-\Delta\phi.\beta)$
   $1/\beta(t) \rightarrow 0.99/\beta(t)$, go to 2
   otherwise Iterated Conditional Mode algorithm (ICM), then go to 2 if $1/\beta(t) > 10^{-10}$, else END.

Step 3 of the SOFI algorithm uses the ICM in order to improve the position for a $\beta(t) < 1$ after having found a global solution for a $\beta(t) \geq 1$. We have tested numerous possibilities to analyse the solution space (step 2 in the SOFI algorithm) as a kind of k-nearest neighbour algorithm (used to analyse the K combinations). It is beyond the scope of this article to go into this in more detail, but the above algorithm has shown robustness and a minimum value function that was always better than with other strategies.

5 APPLICATION OF THE TIME SERIES PROBLEM

In this paper, we examine the medium-term forecasting ability of several alternative models of currency volatility. The reader can find in [6] the description of the data and the economic impact of forecasting volatilities. The data period covers more than eight years of daily observations, March 1991 - December 1998, for the spot exchange rate, 1 and 3-month volatility of the DEM/JPY, GBP/DEM, GBP/USD, USD/CHF, USD/DEM and USD/JPY. Comparing the results of 'pure' time series models, we investigate whether market implied volatility data which could add value in terms of medium-term forecasting accuracy. Based on the over 34000 out-of-sample forecasts produced, evidence tends to indicate that no single volatility model emerges as an overall winner in terms of forecasting accuracy. For our purpose, the problem is posed as following:

- How do the models fit out-of-sample and is there, currency by currency (in fact, we should say volatility by volatility) a better (or several better) forecasting model(s)?
- Do market implied volatility data and model combinations each add value in terms of forecasting accuracy?
- Finally, are some currency volatilities 'easier' to forecast than others?

[5] This is given by Shore and Johnson axioms: Unit Solution (convexity function), Independence between the Scale (every kind of SIL space), Independence between Knowledge (separable law), Partition Conserved (every part of the SIL space).

This is a typical Knowledge Discovery and data mining problem resolved here by the use of both the Track Toolkit managing the database and the SOFI model merging 7 experts. This experimental part contains both the data presentation and manipulation, and the forecasting models presentation. After this, we focus our experimentation on the aggregation of the knowledge in order to merge characteristics and find models could be considered either good or bad.

5.1 Volatility Data Processing

The return series we use for the 6 major exchange rates selected, DEM/JPY, GBP/DEM, GBP/USD, USD/CHF, USD/DEM and USD/JPY were extracted from a historical exchange rate database provided by Datastream. Logarithmic returns, defined as $\log(P_t / P_{t-1})$, are calculated for each exchange rate on a daily frequency basis. We multiply these returns by 100, so that we end up with percentage changes in the exchange rates considered, i.e. $s_t = 100.\log(P_t / P_{t-1})$.

As we are interested in analysing whether market implied volatility data can add value in terms of forecasting realised currency volatility, we must adjust our statistical computation of volatility to take into account the fact that, even if it is only the matter of a constant, in currency options markets, volatility is quoted in annualised terms. As we also wish to focus on medium-term volatility forecasts (i.e. 1 and 3-month out), taking, as is usual practice, a 252-trading day year (and consequently a 21-trading day month and a 63-day trading quarter), we compute the 1-month and 3-month volatility as the moving annualised standard deviation of our logarithmic returns and end up with the following historical volatility measures for the 1-month and 3-month horizons:

$$\text{HVOL}\mathbf{21}_t = 252^{1/2} \Sigma_{t-20,\ldots,t} |s_t|$$

and:

$$\text{HVOL}\mathbf{63}_t = 252^{1/2} \Sigma_{t-62,\ldots,t} |s_t|$$

where $|s_t|$ is the absolute currency return ([6]). For our experience, $\text{HVOL}\mathbf{21}_t$ and $\text{HVOL}\mathbf{63}_t$ are the realised 1-month and 3-month currency volatilities that we are interested in forecasting as accurately as possible, either for risk management or portfolio management purpose. Some summary statistics for these series over a restricted sample through 31 December 1998 are shown in tables 5.1 and 5.2 below and figure 5.1 showing the logarithm returns for the 6 historical exchange rates.

	DEM/JPY	GBP/DEM	GBP/USD	USD/CHF	USD/DEM	USD/JPY
Mean	8.353	5.449	6.697	8.812	7.972	8.113
Std. Deviation*	3.543	2.522	2.846	2.850	2.802	3.394
Skewness	1.845	1.479	1.213	1.104	1.125	1.800
Kurtosis	8.982	7.208	4.841	4.518	4.754	7.355
Jarque-Bera	4155.66	2225.60	779.94	603.86	685.01	2685.84
Probability	0.00	0.00	0.00	0.00	0.00	0.00

Heteroskedasticity-consistent standard deviations

Table 5.1: *Summary statistics of 1-month historical volatility (2 January 1991 - 31 December 1998)*

	DEM/JPY	GBP/DEM	GBP/USD	USD/CHF	USD/DEM	USD/JPY
Mean	8.312	5.432	6.702	8.794	7.968	8.080
Std. Deviation*	2.809	2.062	2.454	2.319	2.260	2.728
Skewness	1.338	0.634	0.931	0.926	0.674	1.543
Kurtosis	5.466	3.128	3.252	3.577	3.164	6.128
Jarque-Bera	1113.64	136.82	296.91	316.72	155.31	1624.60
Probability	0.00	0.00	0.00	0.00	0.00	0.00

Heteroskedasticity-consistent standard deviations

Table 5.2: *Summary statistics of 3-month historical volatility (2 January 1991 - 31 December 1998)*

[6] We use absolute returns as a measure of currency standard deviations: our currency returns have zero unconditional mean enables us to use squared returns as a measure of their variance.

As we can see from tables 5.1 and 5.2, all series are non-normally distributed and often fat-tailed. Further statistical tests of autocorrelation, heteroskedasticity and non-stationarity (not reported here in order to conserve space) show that all 1-month and 3-month historical volatility series exhibit strong autocorrelation and heteroskedasticity but, whereas all 1-month volatilities are stationary in levels, all 3-month volatilities are only stationary when first differenced.

Looking at the volatility curves in figure 5.1, it is clear that the exchange rates volatilities do not have the same behaviour. The exchange rate volatilities are, at the 21-day horizon, not as smooth as the 63-day horizon. Furthermore, those referring to Japanese currency exhibit erratic behaviour. These remarks will contribute to build the expert a priori knowledge.

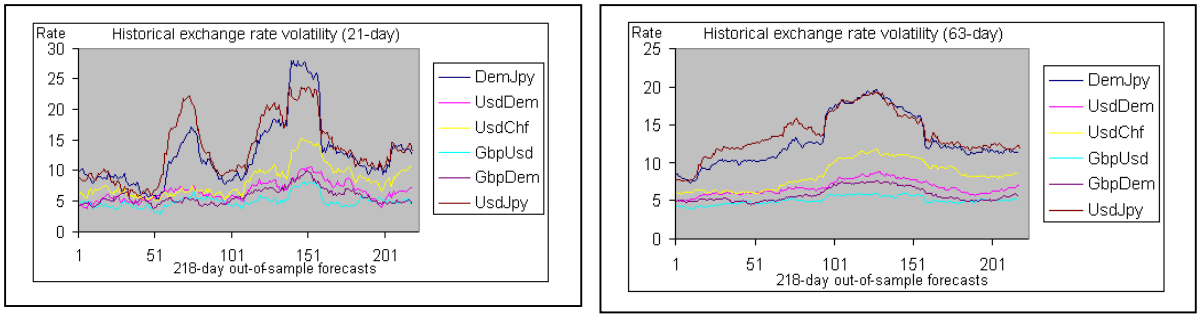


Figure 5.1: *KDD Application for studying 6 times series according to two horizons: 21-day and 63-day.*

5.2 Volatility Models Processing

The predictive time series models of currency volatility used in this article are the ARMA (AR), GARCH (or the Generalised Autoregressive Conditional Heteroskedastic model [59], [60]), the Stochastic Variance (SV, [61], [63]) and some hybrid model combinations. Most of the retained modelling approaches are well documented in the literature. The exact specification is provided in [62] with an exhaustive review of these models. The models are listed, both linear and non-linear, that we have used for each time horizon considered. Each original time series model is complemented by a 'mixed' version counterpart integrating the added information provided by the relevant implied volatility data. The models are classified a priori into four categories:

1) Models noted '*lower*': 4 models using 'little information', i.e. the time series models are based on a statistical analysis of past historical volatility. The general form is such that, at each period ***t***, for both forecasting horizons n=21 days and n=63 days, a function ***f*** delivers an estimation h_{t+n}: $h_{t+n} = \boldsymbol{f}\,(\,HVOL_{n,t}\,)$

   - Model **Garch (1,1)**: based on variance of log-returns;
   - Model **AR(10)**: based on squared log-returns;
   - Model **AR(10)**: based on absolute log-returns;
   - Model **SV(1)**: based on squared log-returns;

2) Models noted '*upper*': the 4 above models using 'more information', i.e. the implied volatility data noted by IMP. The general form is such that, at each period ***t***, for both forecasting horizons n=21 days and n=63 days, a function ***f*** delivers an estimation h_{t+n} : $h_{t+n} = \boldsymbol{f}\,(\,HVOL_{n,t}\,,\,IMP_t)$

   - Model **Garch (1,1)**: based on variance of log-returns + implied volatility;
   - Model **AR(10)**: based on squared log-returns + implied volatility;
   - Model **AR(10)**: based on absolute log-returns + implied volatility;
   - Model **SV(1)**: based on squared log-returns + implied volatility;

3) Models noted *'naïve'*: 2 models assuming that the past available forecast prevailing volatility level. The data is normalised to give a variance level at each period ***t***, for both forecasting horizons n=21 days and n=63 days.

- Model **Naive-1**: Actual annualised historical volatility calculated by real value normalisation such that $h_{t+n} = \alpha.HVOL_{n,t}$;
- Model **Naive-2**: Implied volatility normalised such that $h_{t+n} = \beta.IMP_t$;

4) Models noted *'Sup'*: 3 model combinations which average the results of the previous models and therefore encapsulate even more information. The general form is such that, at each period ***t***, for both forecasting horizons n=21 days and n=63 days, a function ***f*** delivers an estimation h_{t+n}: $h_{t+n} = \otimes_m \boldsymbol{f}\ (h_{t+n}{}^m\ ,\ HVOL_{n,t}\ ,\ IMP_t)$.

- **Average-1:** model average of all previous models;
- **Regression-weighted** average of all previous models;
- **Average-2:** model average of all previous models except 'worst' model;

There are 13 volatility models (noted VMs) times 6 currency volatilities, i.e. 78 VMs per forecasting horizon. We resort to data mining tools to classify the 156 VMs, as we try to identify those VMs that minimise volatility estimation errors and/or volatility trend estimation errors. To allow for comparisons, the 78 VMs are normalised by the estimation error $E_{1,\,t}$ and the trend estimation error $E_{2,\,t}$ using the currency volatility HVOLn such that:

$$E^m_{1,\,t} = HVOLn_{t+n} - h_{t+n}$$

$$E^m_{2,\,t} = \tfrac{1}{2}\,[(HVOLn_{t+1+n} - HVOLn_{t-1+n}) - (h_{t+1+n} - h_{t-1+n})]$$

with m={1, …, 78} (for each VM at the 21-day or 63-day horizon), t={1, …, 218} (for each out-of-sample forecasting step) and n={21, 63} (for each forecasting horizon). We then look for the set M_{opt} of the VMs which minimise the range of values for $E^m_{1,\,t}$ and $E^m_{2,\,t}$ and maximise the confidence interval:

$$M_{opt,j} = \{M = \{m \in VM\} \mid \min_M[\max_{t=1,\,\ldots,\,218,\ m \in M} E^m_{j,\,t} - \min_{t=1,\,\ldots,\,218,\ m \in M} E^m_{j,\,t}]\}$$

where j={1, 2}. The four charts in figure 5.2 show for each forecasting horizon the estimation error $E_{1,\,t}$ and the trend estimation error $E_{2,\,t}$ of the 156 VMs. These charts illustrate how the models overlap. We will note by EE the estimation error $M_{opt,1}$ and by TE the trend error $M_{opt,2}$.

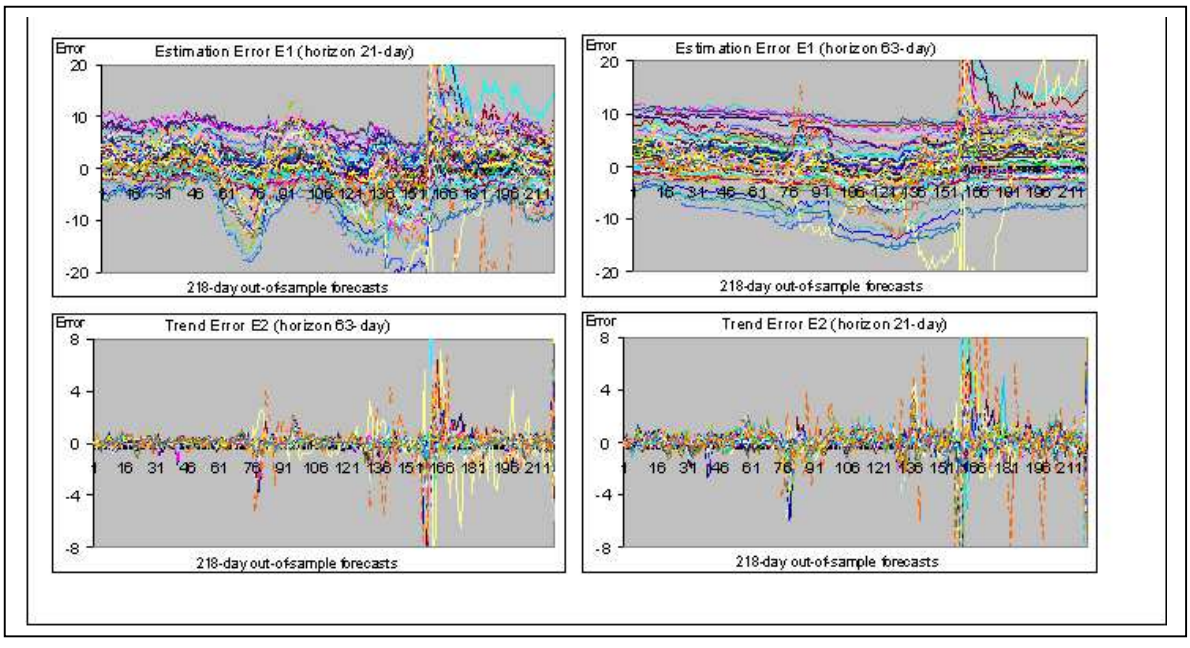


Figure 5.2: *Both Estimation Error (E_1) and Trend Error (E_2) for the 78 VMs according to the two horizons: 21-day and 63-day.*

5.3 Data Analysis, Expert Construction Procedure and Fusion

Having documented the 13 different models for the two time horizons considered, and produced daily out-of-sample estimations for the period 2 March 1998-31 December 1998 ([7]), the Track Toolkit is used to manage the time series data. The first step of our analysis concerns the building of forecasting accuracy measures using the Temporal Data Management Module (TDMM), which filters the stream of rows. The second step is the labelling process to integrate the a priori knowledge.

5.3.1 Data Sources: The Error Time Series Measured

Using the TDMM module of the Track data mining platform, we compute M=8 basics and well-known statistical measures involving scalar values and scoring the accuracy level of the 156 VMs. We also use several other 'intermediate' variables, such as the variance of these errors, their absolute value, etc (not reproduced here to conserve space). We use the following standard measures taken from the statistical literature:

Error Measures	**Opt. Value**	**21-day horizon**		**63-day horizon**	
		Min	Max	Min	Max
Root Mean Square Error (RMSE)	0	1.41	19.40	1.03	158.1
Mean Absolute Error (MAE)	0	1.14	11.37	0.9	23.27
Mean Absolute Percentage Error (MAPE)	0	0.16	0.78	0.14	1.91
Mean Square Error On Realisation (Theil U)	0	0.11	0.44	0.100	0.91
Estimation Error (EE)	0	-9.19	7.76	-9.57	9.89
Trend Error (TE)	0	-0.047	0.02	-0.08	0.01
Student Test (ST)	0	0.000	109.2	0.003	458.7
Fisher-Snedecor Test (FST)	1	1.010	1497	1.062	1702

***Table 5.3.1**: Measures provided by statistical filters scoring the accuracy levels of the 156 VMs.*

Each one of these variables gives us a basis upon which the volatility forecasts are compared across the different models that we use. For the 8 error statistics chosen, the lower the output, the better the forecasting accuracy of the model concerned (the optimal values shown in table 5.3.1 indicate a perfect fit).

5.3.2 Expert Analysis

For our application, we use the three following tools: simple experts, decision trees (ALICE), and a clustering tool. The use of each tool depends on the complicated questions involved in the depth-process ***m*** of knowledge: a simple expert for a specific question (***m***=1), the researched questions delivered by ALICE with a depth-process 1< ***m*** <M, and finally general questions using the clustering tool with ***m*** virtually equal to M. The clustering tool involves a classification constructed from a principal component analysis. Therefore, an expert is the outcome provided by a data mining tool that is guided by specific parameters.

The application uses 7 experts. The SOFI module consults each expert and evaluates its information contents. Each expert is associated with the evaluation measure noted Energy. Each expert is designed for discovering a specific knowledge:

- **Expert-Info** classifies the 13 models according to its complexity. It is gathering the 4 basic models are noted '*lower*', the 4 models using more information, i.e. the implied volatility data are noted '*upper*', the 2 models are noted '*naïve*' ("inert" models) and, finally, the 3 model combinations are noted '*sup*'. These models are grouped independently of the currency volatilities. We look for groups of models which have a particular and common behaviour.
- **Expert Money** gathers all models according to the 6 currency volatilities, especially regarding the Japanese exchange rate.
- **Expert-Model** consults the VMs on a model by model basis.
- **Expert-Alice1** pools the VMs according to the MAE, RMSE, MAPE, and Theil U when associated with the EE variable. The ALICE decision tree, assisted by the users, provides different solutions for each horizon.

[7] This produces a total of 34 008 forecasts, i.e. 13 volatility models times 2 forecasting horizons times 6 currency volatilities times 218 forecasting steps.

- **Expert-Alice2** gathers the VMs according to the EE variables compared with the TE variables. The ALICE decision tree produces different solutions for each horizon.
- **Expert-Cluster3Cl** delivers 3 classes with the aim to explain them with the 8 errors measures.
- **Expert-Cluster4Cl** delivers 4 classes with the aim to explain them with the 8 errors measures.

Table 5.3.2 summarises the result of the evaluation of each expert associated with the respective energy value, the number of the class and shows the depth-process number ***m*** of variables which have the most powerful discrimination value. The respective energy measurement is calculated from the **ϕ** function (this value is reported in the table of each expert). Assuming that the **ϕ** function measures the dependence between classes, a class becomes a rule if the variables explain that class. Therefore, a class exhibits a particular VM behaviour associated with its Label Meaning and the variables that explain the class. The tables give, for each label class, the variables accepted by the Chi-Square test and their probability measure. The probability is given by the percentage of identifiers in the class compared with Identifiers into the Min-Max truncature, $\forall j \in \{1,\ldots,M\}$, $\tilde{N}_{kj} / N_{kj}$ (see formula (5)).

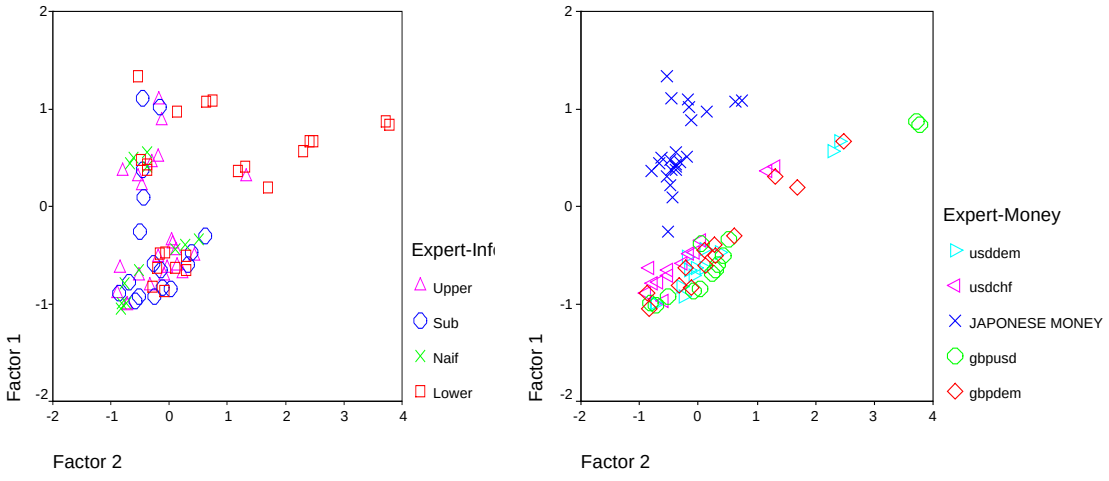
Expert	21-day horizon			63-day horizon		
	Class Number	Energy	Depth-Process	Class Number	Energy	Depth-Process
Expert-Info	4	45.84	1	4	45.86	1
Expert-Money	5	33.56	1	5	33.44	1
Expert-Model	13	38.33	1	13	40.37	1
Expert-Alice1	7	27.48	4	4	34.37	4
Expert-Alice2	5	29.86	5	6	32.80	4
Expert-Cluster3Cl	3	24.88	3	3	23.77	3
Expert-Cluster4Cl	4	25.65	3	4	27.17	3
SOFI	5	19.57	5	3	19.82	6

Table 5.3.2: *For the two horizons, the SOFI process combines the 7 experts to find the energy minimum. For the two applications,* ***Expert-Cluster3Cl*** *produces the classification with the most powerful discrimination (energy 24.88 and 23.77). The* ***SOFI*** *result provides a classification which minimises the* ***ϕ*** *function (19.57 for the 21-day horizon and 19.82 for the 63-day horizon) and greater the value of* ***m****.*

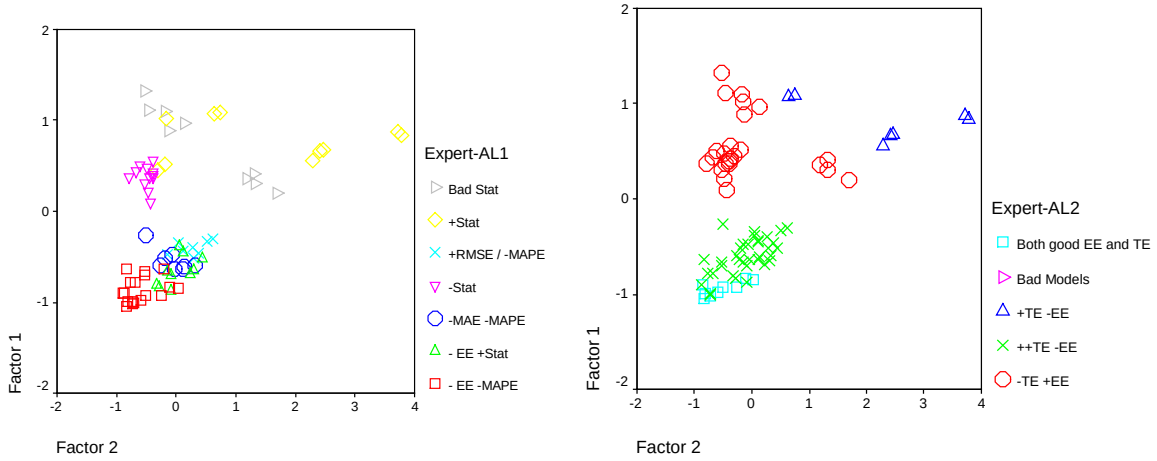
5.3.3 A Priori Knowledge Data: Expert Analysis

The SOFI output associates SPSS graphical view of each expert-classification. That is done by the 2-dimension plan coming from the first and second components of Principal Component Analysis. The expert model is not reported because gives not visual information. Where data is approximately close to the null vector, better is the solution and prediction of the model. For the 21-day horizon, the twice components explain 68% of the state projection and the 72% of the state projection for the 63-day horizon. The following part details the expert-classifications for the two horizons.

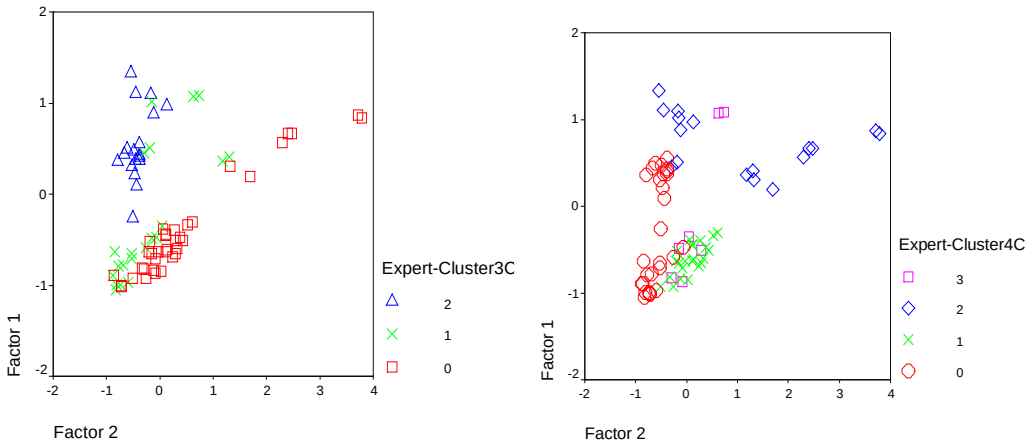
5.3.3.1 21-day horizon:



The Expert-Info shows that the “Lower” models forecasts bad the currency volatilities. The Expert Money produces what was expected for the Japanese currency, e.g. the models are not accurate in face of erratic volatilities. The data confirms the a priori knowledge. The other currency volatilities do not exhibit any particular behaviour when they are taken separately.

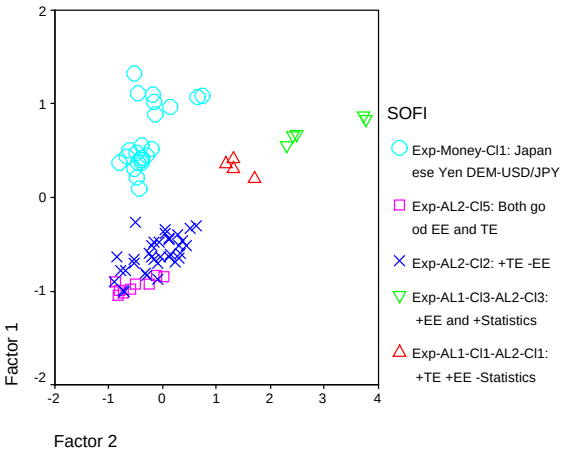


Expert Alice 1 gathers into 7 classes of which 5 classes have significant discrimination measurements. The label class indicates the meaning (“+EE” means large Errors Estimation) of each class according to the accuracy level of the statistical measures. Class “-EE+Stat/-MAPE” shows 18 models with particularly good statistical measures that are close to zero. Expert Alice 2 gathers into 5 classes, which have significant discrimination measurements. Class “-Both EE and TE” shows 9 models with particularly good measures closed to zero.



Into 3 or 4 classes, its experts separate the states according to the whole of the errors measures. The low energy value of this expert is explained by a discrimination over all the 8 errors measures (not reported here in order to conserve space).

For the 21-day horizon, fusion provides 5 classes which are shared between these following experts: Expert-Money, Expert-Alice-1 and Expert-Alice-2. The following tables and figure detail the classification results. A range of variables that are associated with high probabilities explains each class. Each SOFI class is totally or partially matched with a class of an expert. The percentage appearing in the table represents the amount of an expert label belonging to a given class.



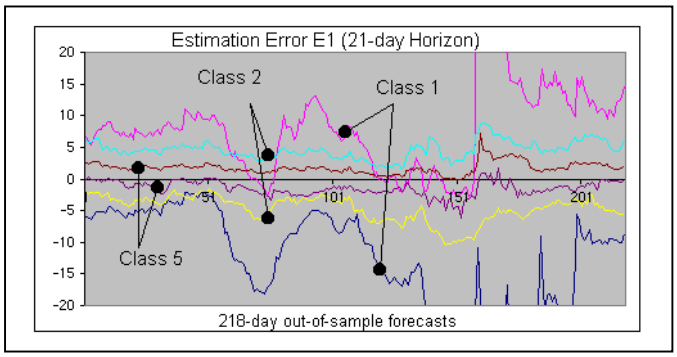
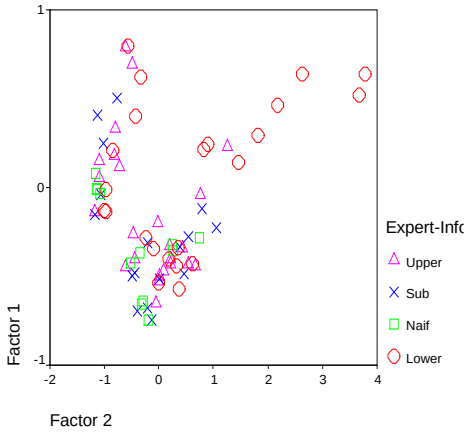


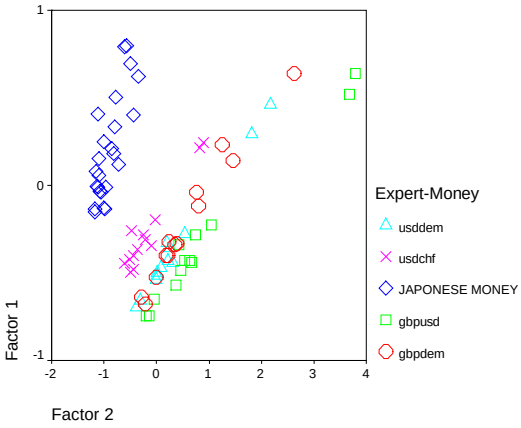
Figure 5.3.3.1: the Min-Max E1 truncature is given for the three classes which can be grouped into three behaviours: good, poor and other models.

5 Classes	**21-day horizon: Energy** = 19.57			
Label Class	Variable	Probability	Min	Max
CLASS 1: Expert Money **99.8% Class 1** Japanese Yen DEM/JPY and USD/JPY	RMSE MAE	0.743 0.743	5.049 4.4126	19.40 11.3797
CLASS 2: Expert Alice 2 **100% Class 2** 35 VMs: Good TE and Poor EE	RMSE MAE EE	0.854 0.854 0.603	1.60730 1.27480 -4.1634	5.04360 4.22530 3.57880
CLASS 3: Hybrid Expert Alice 1 (Cl 1) and Alice 2 (Cl 1) 4 VMs: Poor TE, Poor EE and Poor Statistics	MAPE Theil U EE	1.00 0.667 1.00	0.90480 0.29160 5.47500	1.03390 0.33480 6.62960
CLASS 4: Hybrid Expert Alice 1 (Cl 3) and Alice 2 (Cl 3) 5 VMs: Poor TE, Poor EE and Poor Statistics	RMSE MAE MAPE EE	0.833 1.00 1.00 1.00	7.26380 7.12950 1.22100 7.16220	7.85270 7.73130 1.65520 7.76680
CLASS 5: Expert Alice 2: **100% Class 5** 9 VMs: both Good EE and TE	Theil U EE	0.627 0.50	0.12060 -0.9551	0.20770 1.24450

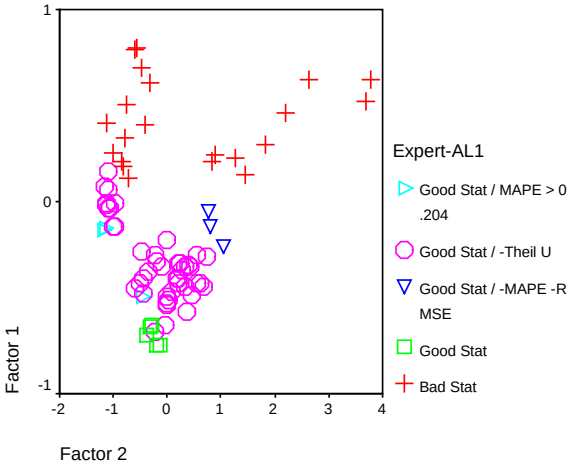
For the 21-day horizon, erratic currencies cannot be "explained" by a small number of rules. Class 3 and class 4 emphasise two particular groups explained by the models AR(10): "Lower" and "Upper" which are considered poor. Classes 1, 2 and 5 gather three particular behaviours: Class 1 concentrates Japanese Yen volatilities, class 2 of the poor VMs for currencies other than the Japanese Yen, and class 5 which concentrates 9 good VMs.

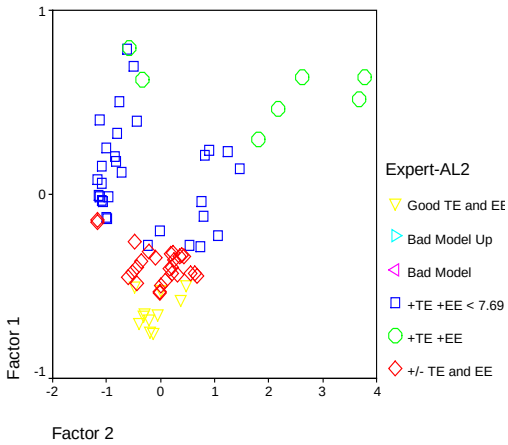
5.3.3.2 63-day horizon:



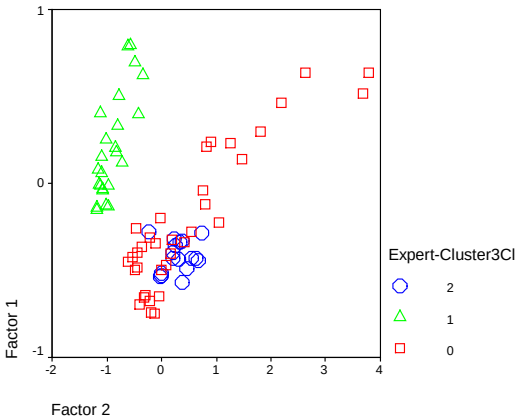


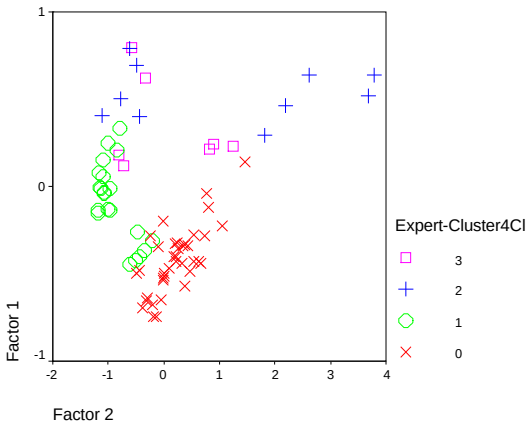
The Expert-Info shows that the “Lower” models forecasts bad the currency volatilities. The Expert Money produces what was expected for the Japanese currency, e.g. the models are not accurate in face of erratic volatilities. The data confirms the a priori knowledge. The other currency volatilities do not exhibit any particular behaviour when they are taken separately.





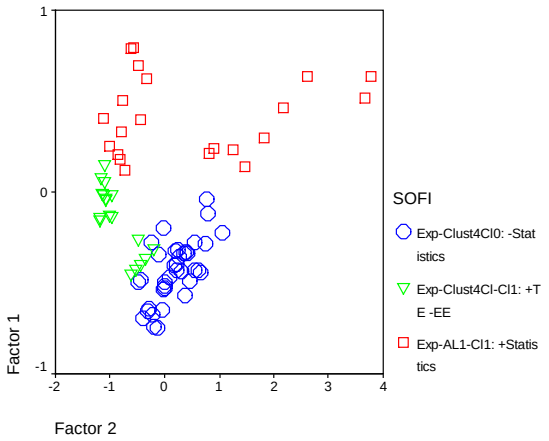
Expert Alice 1 gathers into 6 classes of which 5 classes have significant discrimination measurements. The label class indicates the meaning (“+EE” means large Errors Estimation) of each class according to the accuracy level of the statistical measures. Class “Good Stat” shows 11 models with particularly good statistical measures that are close to zero. Expert Alice 2 gathers into 5 classes, which have significant discrimination measurements. The label class indicates the meaning of each class according to the accuracy level of the estimation measures. Class “Good TE and EE” shows 9 models with particularly good measures that are close to zero.





Into 3 or 4 classes, its experts separate the states according to the whole of the errors measures. The low energy value of this expert is explained by a discrimination over all the 8 errors measures (not reported here in order to conserve space).

For the 63-day horizon, fusion provides 3 classes shared between two experts: Expert-Cluster4Cl and Expert-Alice-1.



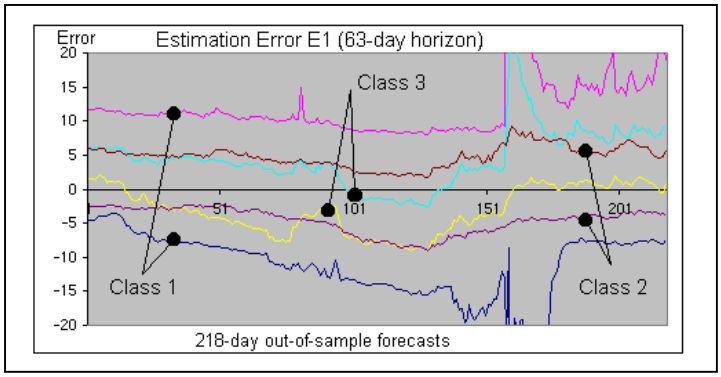


Figure 5.3.3.2: *the Min-Max E1 truncature is given for the three classes which can be grouped into three behaviours: good, poor and other models.*

3 Classes	**63-day horizon: Energy** = 19.82			
Label Class	Variable	Probability	Min	Max
CLASS 1: Expert Alice 1 **100% Class 1** 23 VMs: Poor Statistics	RMSE MAE	1.00 0.852	6.2233 4.1342	158.198 23.2794
CLASS 2: Expert Cluster4Cl **99% Class 1** 18 VMs	MAPE Theil U EE	0.681 0.621 0.694	0.20930 0.11810 -1.0346	0.39340 0.19590 2.45330
CLASS 3: Expert Cluster4Cl **99% Class 0** 37 VMs	RMSE MAE EE ET ST FST	0.712 0.6.61 0.627 0.617 0.578 0.507	1.0300 0.9000 -4.5506 -0.0078 0.9416 1.0763	4.94410 4.61890 4.64020 0.01480 23.3389 52.4310

For the 63-day horizon, the VMs are grouped into three classes. Good models for the GBP/DEM, GBP/USD and USD/DEM volatilities (Class 3: 37 VMs), poor models for other volatilities (Class 1: 23 VMs) and good models for the Japanese Yen and USD/CHF volatilities (Class 2: 18 VMs).
We conclude looking at the results according to the 2 horizons. For the 21-day horizon, the best group involves an effective set of models, which could be used to forecast volatilities. The classification at the 21-day horizon is highly dependent on the VM behaviours and the Japanese Yen specificity. The classification at the 63-day horizon is heavily relies on the currency dimension: the best group discriminates well between currencies, also it is not accurate for forecasting purpose.

6 CONCLUSION

In this paper, we have presented the SOFI module. It is a software which can be integrated on every data mining platform gathering several data mining tools. The SOFI module is a multi-classifier merging classifications in an associative way. It is especially designed for the Knowledge Discovery in Databases problem for which a priori information is integrated step by step. After accumulating it, the user has quick access to a synthetic result and can rebuild a complete analysis.

For the application above, the fusion process is immediate (a few seconds using a standard PC with a Pentium processor). Having both the source data and a priori data, the SOFI model makes the link between the 78 label classes (or nodes) defined by the 7 experts and the VMs data. Section 5.3 above has shown the informational quality integrated and distributed over the experts. With a logical labelling of the classes, the SOFI model allows for a methodical search of the knowledge hidden in the data.

For the two horizons, the stochastic process minimises the function and provides the classification that has the best discrimination measure. The fusion process starts with the estimation of the temperature (see SOFI algorithm, section 4). Figure 6.1 shows the energy minimisation steps of the fusion process for the two horizons: energy equal to 19.57 for the 21-day horizon and equal to 19.82 for the 63-day horizon.

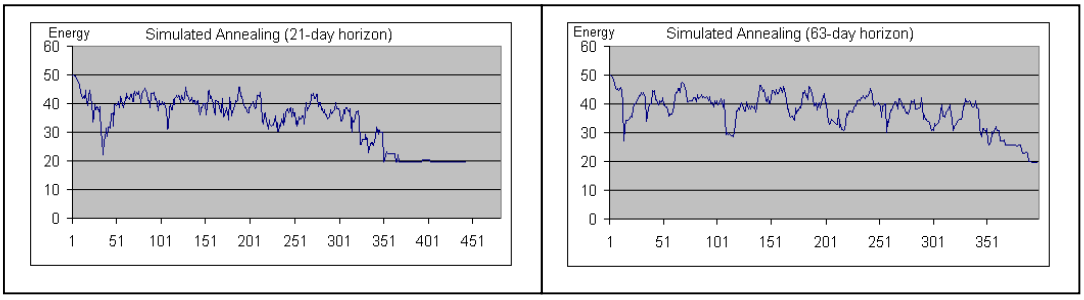


Figure 6.1: *The SOFI process produces the minimum of energy after 432 steps for the 21-day horizon and 398 steps for the 63-day horizon.*

We have seen that the SOFI process is unsupervised and non-parametric. Each classifier is directly estimated from the data as proposed in [9]. The algorithm shows a lower computational processing time. The minimisation procedure depends on the strategy to research the optimal solution. In a future, we will show how to perturb the R classifications. We have tested several strategies such as the K-means regarding node aggregation. The scheme called Expert-Class aggregation, presented here, provides results that always outperform other strategies. Here, the robustness of SOFI for our time series application has been evaluated by the fusion of other experts (random classification, expert clustering with 5 and 6 classes) which have not modified the final result.

In this article, the fusion module has been used to classify forecasting models. It has also been used by financial institutions to combine experts in order to optimise the resolution of Automatic Teller Machine (ATM) problems, score clients for marketing and analyse temporal customer characteristics. All these applications allow us to evaluate SOFI performances which are reported below in table 6.1.

NB of Variables	NB of Hoxels	DECAlpha 8200	Pentium MMX 233	Nb Processes
6	30	1 seconds	2 seconds	1 process
8	57	3 seconds	8 seconds	1 process
20	57	11 seconds	21 seconds	3 processes
50	57	53 seconds	102 seconds	3 processes
200	57	2 minutes	5 minutes	3 processes

Table 6.1*: Performance model for SOFI process. This table reports CPU time for the Unix System (DECAlpha 8200) and on PC (Pentium MMX 233). The time resource is collected according to the number of variables which must be optimised and the number of Hoxels (High Order of Pixels and the number of nodes of graph G). Also, the number of processes needed to have the global minimum of energy is reported.*

The good CPU time performances are due to the D-separable process used to find the minimum amount of energy. The D-separable property barely affects the number R of experts, and, consequently, the number of the nodes in graph G. Effectively, we have seen that in some application, few rejected case of data stream are rejected after the optimisation phase. Then, we project to observe precisely the result by measuring the ignorance and the noising classes into the set of the data with the Belief Network formalism (APRIOU).

The fusion module is useful when there is a large volume of data but also with an increasing number of analyses. Then, the key is to know how to combine different points of view taken from the data in order to produce a global and a synthetic view. Fisher has shown that a set of information accumulated reduces the risk decision. The biological and artificial real example intuitively show that the functionality process calculates and

combines both data volume and compatible knowledge in order to increase the validity of a decision. The SOFI model does that according to this definition of fusion, “the human system which calls upon its different senses, its memory and its reasoning capabilities to perform deductions from the information it perceives” [37].

SOFI has be integrated into the Track Toolkit Platform and it will be study the integration for OLAP data mining platform. Future works will be around, first, to text mining application where SOFI process would be able to help for the user-end integrating the a priori knowledge they have about the documents contents. The second kind of application will be dedicated for the implementation for real-time processing integrating time series analysis for detecting at each time the best predicting model set and estimating the risk performance.

References

[1] J. Pearl, *On Evidential Reasoning in a Hierarchy of Hypotheses*, Artificial Intelligence 28, Elsevier Science Publishers, 9-15, Amsterdam, North-Holland, 1986.

[2] S. Chauvin and L. Jañez Escalada, *Tracking Knowledge Data Bases: Fusion of Data Software*, International Conference of Model Recognising Shape, C.I.R.M. institution France, Marseille 1997.

[3] S. Thrun, C. Faloutsos, T. Mitchell and L. Wasserman, *Automated Learning and Discovery: State-Of-The-Art and Research Topics in a Rapidly Growing Field*. Report of CONALD meeting June 11-13, Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213, 1998

[4] M. C. Mozer, R Wolniewicz, D. B. Grimes, E. Johnson and H. Kaushansky, *Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry*, IEEE Transactions on Neural Networks, 2000.

[5] Kamran Parsaye and Mark Chignell, *Intelligent database tools & applications : hyperinformation access*, data quality, visualization, automatic discovery. Wiley, New York, 1993.

[6] T. W. Lee, *Independent Component Analysis*, Kluwer Academic Publishers, Boston, 1998.

[7] Pedro A.d.F.R. Højen-Sørensen, Ole Winther and Lars Kai Hansen, *Mean Field Approaches to Independent Component Análisis*, In preparation, Department of Mathematical Modelling, Technical University of Denmark B321, DK-2800 Lyngby, Denmark, 2001.

[8] K. Tumer and J. Ghosh, *Error Correlation and Error Reduction in Ensemble Classifiers*, Connection Science, Special issue on combining artificial neural networks: ensemble approaches, volume 8, No 3/4, pp 385-404, December 1996.

[9] V.N. Vapnick, *Statistical Learning Theory,* John-Wiley & Sons, New-York, USA, 1998.

[10] Jie Cheng and Russel Greiner, *Comparing Bayesian network classifiers*. In Kathryn B. Laskey and Henri Prade, editors, Morgan Kaufmann Publishers Proceedings of the Fifteenth Conference on Uncertainty in Articial Intelligence, UAI'99, 101-108, Stocholm, 1999.

[11] F.V. Jensen, *Introduction to Bayesian Networks,* Springer-Verlag, New York, 1996.

[12] A. Becker and P. Naïm, *Les Réseaux Bayésiens: Modèles Graphiques de Connaissance*, Éditions Eyrolles, 1999, ISBN 2-212-09065-X, Paris.

[13] J. Shafer, R. Agrawal, and M. Mehta. *SPRINT: A scalable parallel classifier for data mining*. In Proc. of the 22nd Int'l Conference on Very Large Databases, Bombay, India, September 1996.

[14] B.V. Desarathy, *Decision Fusion Strategies in Multisensor environments*, IEEE Transaction on Systems, Man and Cybernetics, Vol 21, No 5, 1140-1154, Sept. 1991.

[15] W.E. Weideman, M. Manry, H.-C. Yau, and W. Gong, *Comparisons of a Neural Network and a Nearest-Neighbor Classifier via the Numeric Handprint Recognition Problem*, IEEE Transactions on Neural Networks, Vol 6, No 6, November 1995.

[16] B. F. J. Manly, *Multivariate Statistical Methods*, second edition, Chapman & Hall, 1994.

[17] J.Y. Ching, A. K. C. Wong and K.C.C. Chan, *Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 18, No 7, 641-651, July 1995.

[18] T. Henderson, C. Hansen and B. Bhanu, *The Synthesis of Logical Sensor Specifications,* in Proc. SPIE, Vol 579, Intell. Robots and Comput. Vision, Cambridge, MA, 442-445, Sept. 1985.

[19] T. Henderson and C. Hansen, *Multisensor Knowledge Systems in Real-Time Object Measurement and Classification*, Springer-Verlag, A. K. Jain (Ed.), Berlin, Germany, 375-390, 1988.

[20] T.W. Prosser, *The Representation of Intelligence Fusion in the Joint Warfare System (JWARS) Analytic Model - Prototype Applications,* Joint Warfare System Office, TX/RX, 6634, London, July 1997.

[21] Y. Bar-Shalom, *Tracking and Data Association*, Boston, MA, Academic Press, 1988

[22] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, *Sonar Tracking of Multiple Targets Using Joint*

Probabilistic Data Association, IEEE J. Oceanic Eng, Vol 8, No 3, 173-184, 1983.
[23] D.G. Corr, A. M. Tailor, A. Cross, D. C. Hogg, D. H. Lawrence, D. C. Mason and M. Petrou, *Progress in Automatic Analysis of Multi-Temporal Remotely-Sensed Data*, International Journal of Remote Sensing, 10, 1175-1195, 1989.
[24] B.V. Desarathy, *Decision Fusion*, IEEE Computer Society Press, 1994.
[25] R.C. Luo and M.G. Kay, *Multisensor Integration and Fusion in Intelligent Systems*, IEEE Transactions on Systems, Man and Cybernetics, Vol 5, No 19, 901-931, 1989.
[26] I. Bloch, *Information Combination Operators for Data Fusion: A Comparative Review with Classification*, IEEE Trans. Systems, Man, and Cybernetics, Vol 26, No 1, 52-67, January 1996.
[27] S. Chauvin, *Fusion de Données par les Ensembles Flous Appliquée à l'Imagerie Satellitaire : Apprentissage et opérateurs*, LFA'95, Paris, 1995.
[28] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, 1976.
[29] J. Guan and D. A. Bell, *Evidence Theory and its Applications*, North-Holland, Amsterdam, 1992.
[30] S. Fabre, A. Appriou and X. Briottet, *Presentation and description of two classification methods using data fusion based on sensor management*, Information Fusion 2, pp 49-71, New York, August 2001.
[31] M. Chavent and E. Diday, *Probabilist Objects and Fusion*, IPMU, International Conference on Information Processing and Management of Uncertainty, Paris, 788-793, 1994.
[32] V. Radevski and Y. Bennani, *Combining Structural and Statistical Features for Handwritten Digit Recognition*, International Joint Conference of Information Sciences, 102-105, Research Triangle Park, NC, USA, 1997.
[33] S.B. Cho and J. H. Kim, *Multiple Network Fusion Using Fuzzy Logic*, IEEE Transactions on Neural Networks, Vol 6, No 2, 497-501, March 1995.
[34] D. Hall, *Mathematical Techniques in Multisensor Data Fusion*, Artech House, 1992.
[35] S. C. A. Thomopoulos, R. Viswanathan, and D. Bougoulias, *Optimal Distributed Decision Fusion*, IEEE Transactions on Aerospace and Electronic Systems, AES-25, Vol 24, No 5, 761-765, Sept. 1989.
[36] R. Tenney and N. Sandell, *Detection with Distributed Sensors*, IEEE Transactions on Aerospace and Electronic Systems, vol. AES-17, No 4, 501-510, July 1981
[37] L. Wald, *European Proposal for Terms of Reference in Data Fusion*, International Archives of Photogrammetry and Remote Sensing, Vol XXXII, Part 7, 651-654, 1998.
[38] Q. Zhang and P. K. Varshney, *Decentralized M-ary detection via hierarchical binary decision fusion*, Me Learning, 36, No ½, pp 105-139, Boston, July 1999.
[39] N.S.V. Rao, *Distributed Decision Fusion Using Empirical Estimation*, IEEE Transactions on Aerospace and Electronic Systems, Vol 33, No 4, 1106-1114, 1997.
[40] Jie Cheng, David A. Bell, and Weiru Liu, *Learning belief networks from data: an information theory based approach*. In Proceedings of the Sixth ACM International Conference on Information and Knowledge Management, 1997.
[41] E. Bauer and R. Kohavi, *An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants*, Machine Learning, 36, No ½, pp 105-139, Boston, July 1999.
[42] J. Pearl, *Fusion, Propagation and Structuring in Belief Networks*, Artificial Intelligence, 29, Elsevier Science Publishers, 241-288, Amsterdam, North-Holland, 1986.
[43] A.C. Kak and S. Chen, *Spatial Reasoning and Multi-Sensor Fusion*, Proc 1987 Workshop, Los Altos, CA: Morgan Kaufmann, 1987.
[44] R. Agrabal, T. Imielinski and A. Swami, *Database Mining: A Performance Perspective,* IEEE Transactions on Knowledge and Data Engineering, Special issue on Learning a Discovery in Knowledge-Based Databases, 743-769, 1993.
[45] S.S. Anand, D. A. Bell and J. G. Hughes, *Evidence-Based Discovery of Association Rules.* Internal Report, Faculty of Informatics, University of Ulster, 1996.
[46] Y.M. Sharaiha and N. Christofides, *A Graph-theoretic Approach to Distance Transformations*, IEEE of Pattern Recognition Letters, 1035-1041, Vol 15, 1994.
[47] K.C. Chang, C. Y. Chong and Y. Bar-Shalom, *Joint Probabilistic Data Association in Distributed Sensors Networks*, IEEE Trans. Automat. Contr., Vol AC-31, No 10, 889-897, 1986.
[48] M.I. Miller, A. Srivastava and U. Grenander, *Conditional-Mean Estimation Via Jump-Diffusion Processes in Multiple Target Tracking/Recognition*, IEEE Transactions on Signal Processing, Vol 43, No 11, 2678-2690, Nov 1995.
[49] J. N. Kapur, *Application of Entropic Measures of Stochastic Dependence in Pattern Recognition*, Pattern Recognition, Vol 19, No 6, 473-476, 1986.
[50] O. Catoni, *A Mixture Approach to Universal Model Selection*, Research Report, LMENS-97-22, Département de Mathématiques Appliquées, Paris, 1997.
[51] J-F Bercher, G. Le Besnerais and G. Demoment, *The Maximum Entropy on the Mean Method, Noise and*

Sensibility, 14th Coll. Maximum Entropy and Bayesian Methods, Cambridge University Press (Cambridge) 1994.

[52] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memory-less Systems*, Academic Press, New York, 1981.

[53] L. Devroye and L. Györfi, *Nonparametric Density Estimation: the L1 View*, John Wiley and Sons, New York, 1985.

[54] D.J.C. MacKay, *A Short Course in Information Theory*, TCM Seminar Room, Mott Building, Cavendish Laboratory, DJCM February 18, 1995.

[55] S.P. Luttrell, *An Optimisation of the Metropolis Algorithm for Multibit Markov Random Fields*, Inverse Problem, Vol 2, L15-L17, 15 Dec. 1995.

[56] P.S. Maybeck, *Stochastic Models, Estimation and Control*, Vols 1 and 2, Academic Press, New York, 1979 and 1982.

[57] E.T. Jaynes, *Information Theory and Statistical Mechanics*, Physics Review, Vol 106, No 4, 620-630, May 15, 1957.

[58] T. Uchiyama and M. A. Arbib, *Color Image Segmentation Using Competitive Learning*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 16, No 12, 1197-1206, Dec 1994.

[59] T. Bollerslev, *Generalised Autoregressive Conditional Heteroskedasticity*, Journal of Econometrics, 31, 307-27, 1986.

[60] S.J. Taylor, *Modelling Financial Time Series*, John Wiley & Sons, Chichester, 1996.

[61] A.C. Harvey, E. Ruiz and N. Shepherd, *Multivariate Stochastic Variance Models*, Review of Economic Studies, No 61, 247-64, 1994.

[62] C.L. Dunis, J. Laws and S. Chauvin, *The Use of Market Data and Model Combination to Improve Forecast Accuracy*, in C. L. Dunis, A. Timmermann and J. Moody [eds.], *Developments in Forecast Combination and Portfolio Choice*, John Wiley & Sons, Chichester, 2001.

[63] J. Hamilton, *Time Series Analysis*, Princeton University Press, Princeton, 1994.