データ分析統合環境　PADOC/static　の概要

mabonki@ka3.koalanet.ne.jp

要旨

データ分析統合環境　PADOC/static は Windows PC 上で稼動し、対象データをロードして、編集、結合、分析、結果表示、一連の操作を繰返し行う事ができる環境である。分析環境としては２モードあり、コマンドベースとデータフロー図上がある。基本的な分析手法は一通り装備されている。

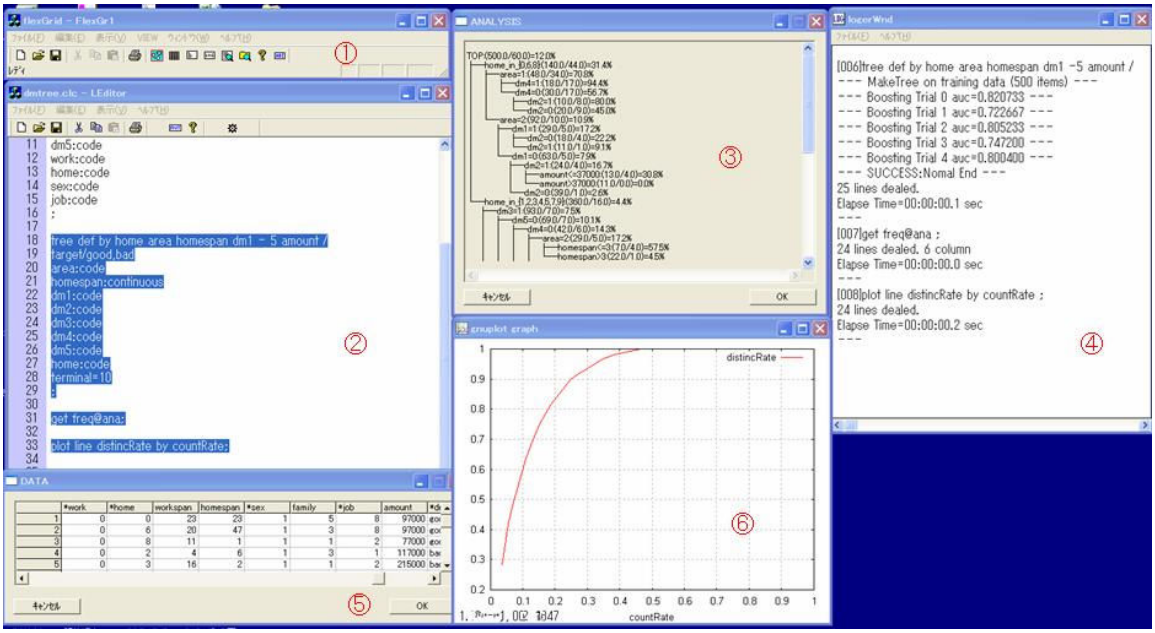
１．はじめに

主に SAS を使って 10 年以上データ分析を行ってきたが、SAS はデータ編集では便利だが分析手法の中身が開示さておらず、新しい分析モデルの提供も遅く、データを視覚的に捕らえるのに不便で古さが否めなかった。そこでデータ編集は SAS と同様なもので視覚的なデータ分析ツールを自己開発に至った。

２．コマンドモード

SAS と同様なデータ編集機能を一般的な C 言語仕様で記述できる様にした。分析機能は標準的な機能を実装した（表１参照)。実装して感じた事は、分析モデルの実装はネット上に豊富な文献があり思ったより困難ではないが、異常なデータや操作でも耐える様に汎用的に実装する方が困難であった。ツールの汎用性と頑健性の向上については数多くの人に使用した頂く必要がある。

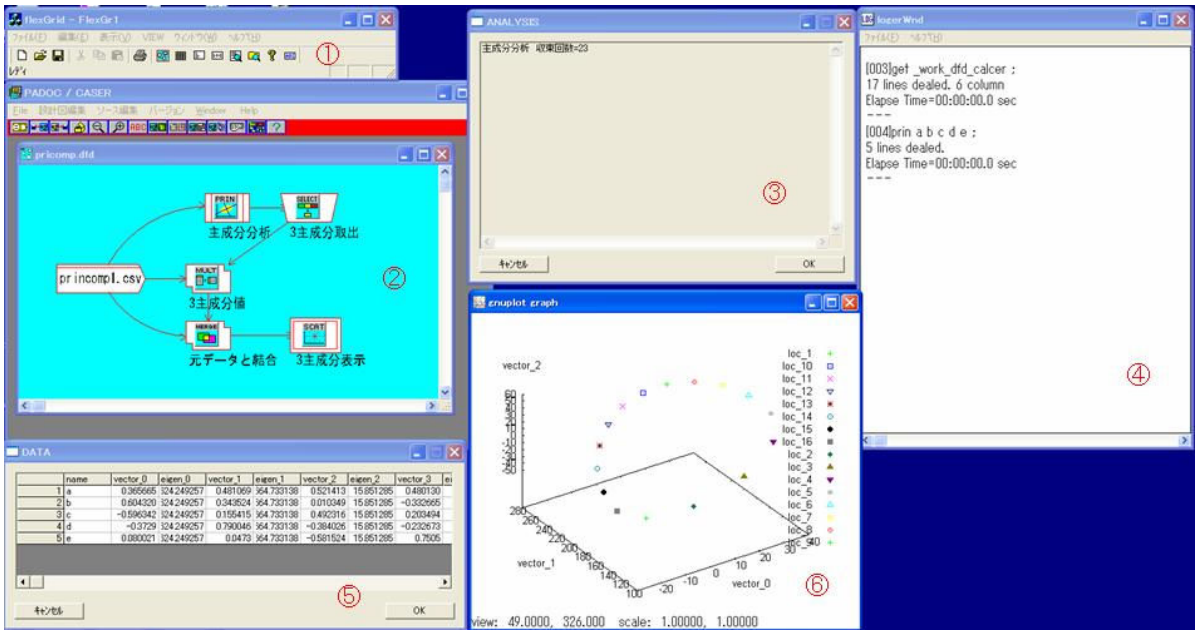
画面は SAS と同様な編成（コマンドバー　コマンドエディター、アウトプット、ログ）で構成さている。



① コマンドバー
② コマンドエディター（黒反転した部分が実行できる）
③ アウトプット
④ ログ（実行状態やエラーを表示）

３．データフローモード

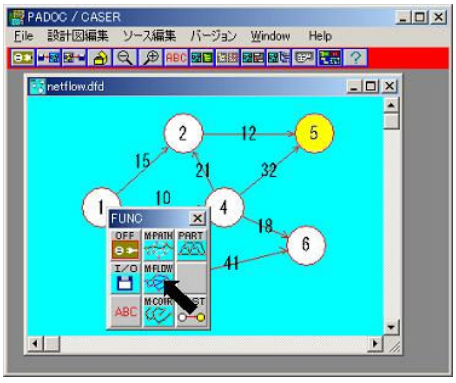
視覚的にデータ分析処理を行う目的でデータフロー図を採用した。これはトム・デマルコ『構造化分析とシステム仕様』に準拠しているので、矢印は曲線も使える。図形には処理に応じたアイコンが張られており視覚的にデータの処理過程が追える。



① コマンドバー
② データフローエディター
③ アウトプット
④ ログ

４．グラフィカルモデル

視覚的な環境では、ノードとパスで構成されるグラフィカルモデルの分析が可能であるので経路問題、ベイジアンネット、SEM を導入している。例として最大流出問題はノードとそのパス容量をデータフロー図で作成して以下の様に求めることができる。



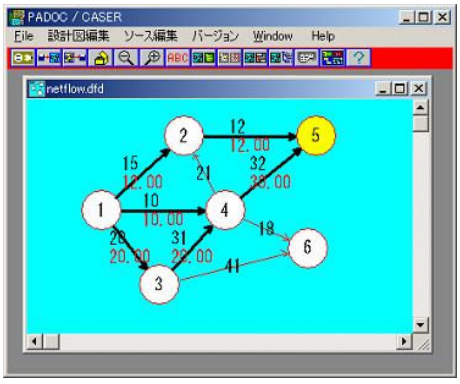


表１　分析機能一覧

種別	コマンド名	コマンド	説明	使用例
分析	AIC	aic	AIC表分析	aic flag by x y z;
分析	ANOVA	anova	分散分析	anova
分析	Association	assoc	アソシエーション分析	assoc
分析	Correlation	corr	相関分析（相関係数）	corr x y z;
分析	Frequency	count	区分数表	count x y z;
分析	conjoint	conj	コンジョイント分析	conj x y z by point;
分析	covariance	cov	相関分析（分散）	cov x y z;
分析	CoxHazard	cox	生存分析（比例ハザード）	cox flag by x y z;
分析	Dendrogram	dendro	樹形図分析	dendro x y z;
分析	EM	em	EMアルゴリズム	em x y z by seg;
分析	Factor	factor	因子分析	factor tag by x y z;
分析	GausianGraph	ggm	ガウシアン・グラフ	ggm x y z;
分析	groupLens	grouplens	協調フィルタリング	grouplens score by user item;
分析	K-nn	k_nn	N区分(近傍法)	k_nn tag by x y z;
分析	k-means	kmeans	N区分分析（KMEANS法）	kmeans tag by 5;
分析	LogitRegression	logit	ロジット分析	logit tag by x y z;
分析	Maharanobis	maha	2区分（マハラノビス法）	maha tag by x y z;
分析	*mcmc	mcmc	マルコフモンテカルロ法	mcmc x;
分析	minsqr	minisqr	最小二乗法（ハウスホルダー）	minsqr tag by x y z;
分析	NaiveBeyes	nbayes	ナイーブベイズ法	nbayes tag by x y z;
分析	neuro	neuro	ニューロ分析	neuro inp1 inp2 inp3 by out1 out2 ;
分析	Principle	prin	主成分分析	prin x y z;
分析	linerRegression	reg	重回帰分析	reg tag by x y z;
分析	RegTree	regtree	回帰木分析	tree tag by x y z;
分析	BasicStatic	static	基礎統計量	static x y z;
分析	som	som	自己組織化マップ	som x y z;
分析	Summary	sumup	サマリー処理	summary x y z;
分析	svm	svm	2区分（SVM法）	svm tag by x y z;
分析	DicsionTree	tree	判別木分析	tree tag by x y z;
経路	*anneal	anneal	焼鈍し法による巡回問題	–
経路	flow	netflow	最大流入問題	–
経路	*path	netpath	最小（長）経路問題	–
経路	*span	netspan	最大連結問題	–
計画	LinerPlan	lp	線形計画法	lp tag by x y z;
計画	nonLinerPlan	nlp	非線形計画法	nlp (filename)/
計画	integerPlan	intp	整数計画法	intp x1-6 by cond val;
計画	dynamicPlan	dynmp	動的計画法	dynmp f_1-3 by x;
時系列	kalman	kalman	カルマンフィルター	kalman x;
時系列	*covariance	tmcov	相関	tmcov x y z;
時系列	*crosscov	tmcrs	相互相関	tmcrs x y z;
時系列	*ar	tmar	自己相関モデル	tmar x
時系列	*ma	tmma	移動平均モデル	tmma x
時系列	*trend	tmtrend	トレンド	tmtrend x
時系列	*arma	arma	移動平均自己相関モデル	arma x
時系列	*average	tmave	平均化	tmave x y z;
時系列	*season	season	季節要因	season x y z;
時系列	*fourie	fourie	フーリエ変換	fourie x;
時系列	*calc	tmcalc	加減乗算	tmcalc add x y;
時系列	*adf	tmadf	ADF検定	tmadf x;
検定	average	test ave	検定（平均値）	test ave x y;
検定	rate	test rate	検定（比率）	test rate x y;
検定	wilcox	wilcox	検定（順位和）	wilcox x y;
検定	χ Squre	chisqt	検定（χ二乗）	chisqrt x1 – 5;
行列	mult	mxmult	行列の積	mxmult mx by vec_1 – 5;
行列	*reverse	mxrev	逆行列	maxrev vec_1 – 5;
行列	*add	mxadd	行列の和	mxadd mx by vec_1 – 5;
行列	*sub	mxsub	行列の差	mxsub mx by vec_1 – 5;
行列	*svd	svd	SVD	svd vec_1 – 5;