機械学習による
特許自動分類の試行

ＳＶＭとDeep Learnigによる特許分類

2018/11/27作成

JFEテクノリサーチ㈱
平川 雅彦

1. 調査の目的

自動ブレーキ特許

テキストマイニング

解決手段　距離マップ

(KHCoder)

周辺技術：
歩行者認識
自転車認識

SVM分析

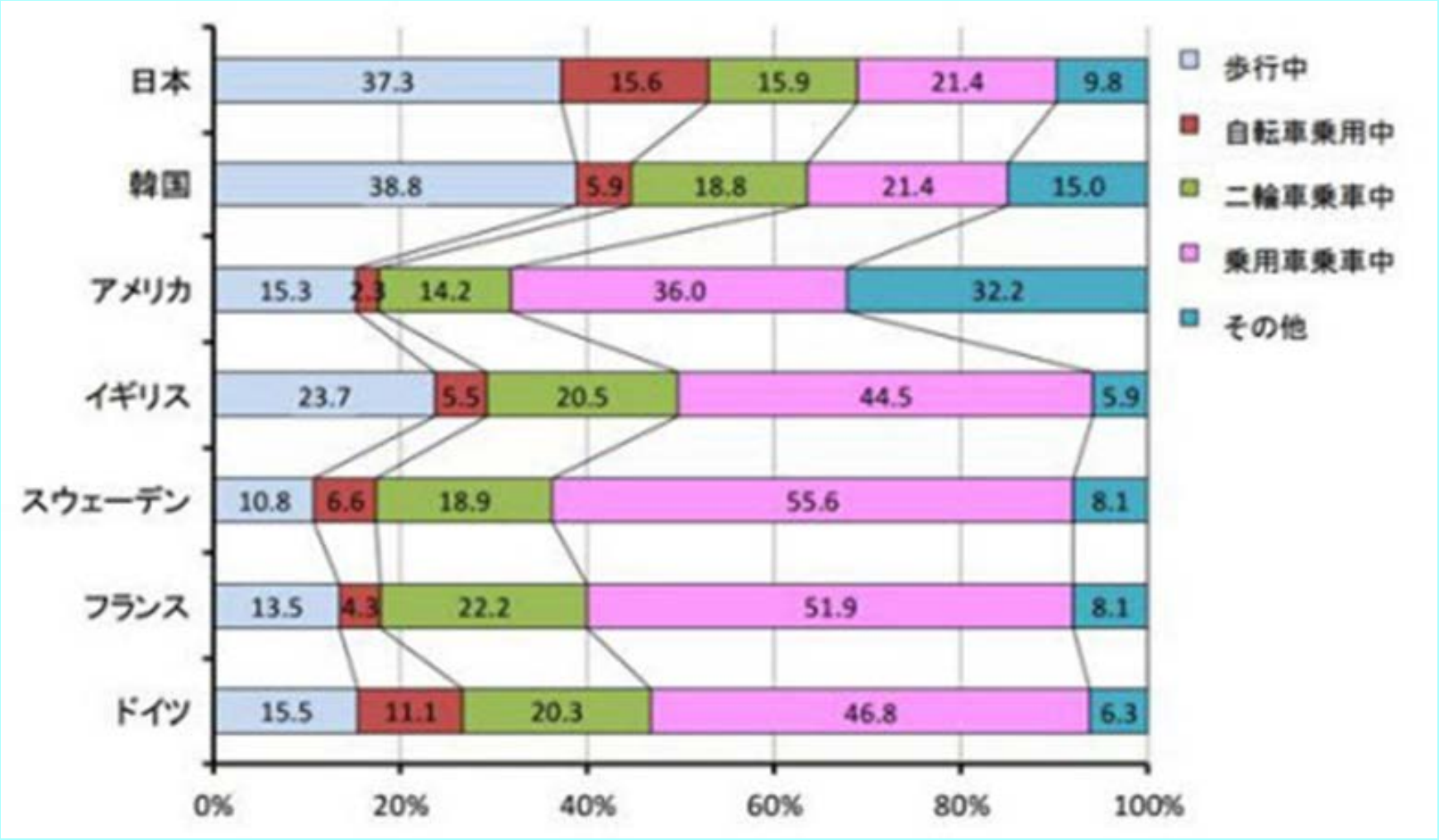
自動分類

Euro NCAP：
自動車安全評価項目

年	追加評価項目
2014	自動ブレーキ
2016	歩行者認識
2018	自転車認識 夜間の歩行者

交通事故死者数

自転車、二輪車の事故死者数は多い



出所：交通事故総合分析センター、「交通事故の国際比較（2015年）

2. 概念検索

検索質問

自動車に搭載したセンサを用いて自転車を認識する装置

単語分割 → 品詞で絞込み → 不要語の除去 → 重要度の計算

検索語リスト
(重要度付き)

認識	0.058
搭載	0.038
センサ	0.040
自転車	0.037
自動車	0.051
画像	0.042
装置	0.016

・質問中の出現度
・検索対象中の出現度

特許毎の類似度を計算

特許公報
全文索引

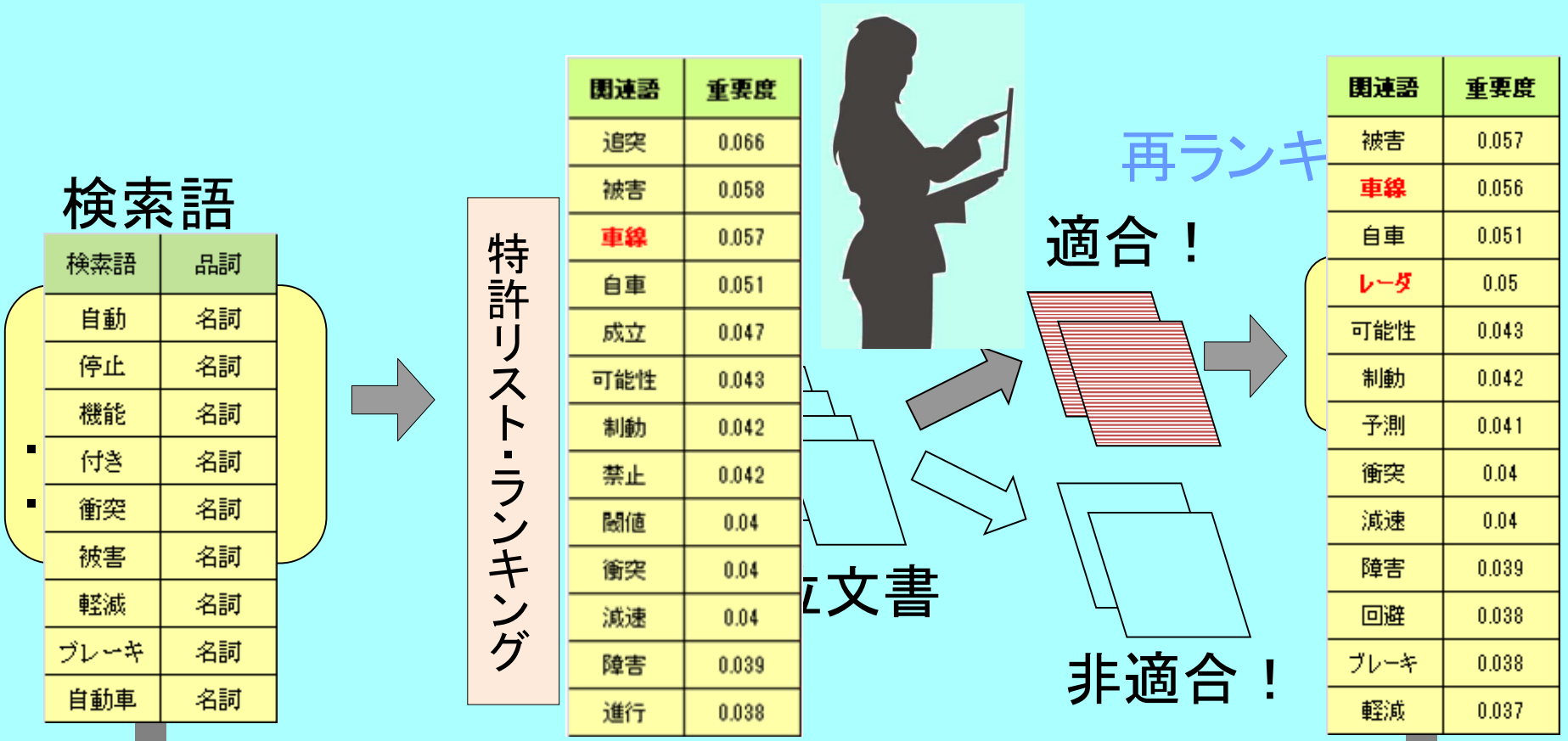
・検索語の重要度
・検索語の出現回数
・文書長の正規化

特許リストのランキング

順位	特許番号	重要度
1	2009-262700	0.348
2	2009-262699	0.253
3	2009-262698	0.245
4	2012-192842	0.218
5	2003-315452	0.187
6	2011-154580	0.175

HYPAT-i2による概念検索

対話型検索　適合性フィードバック



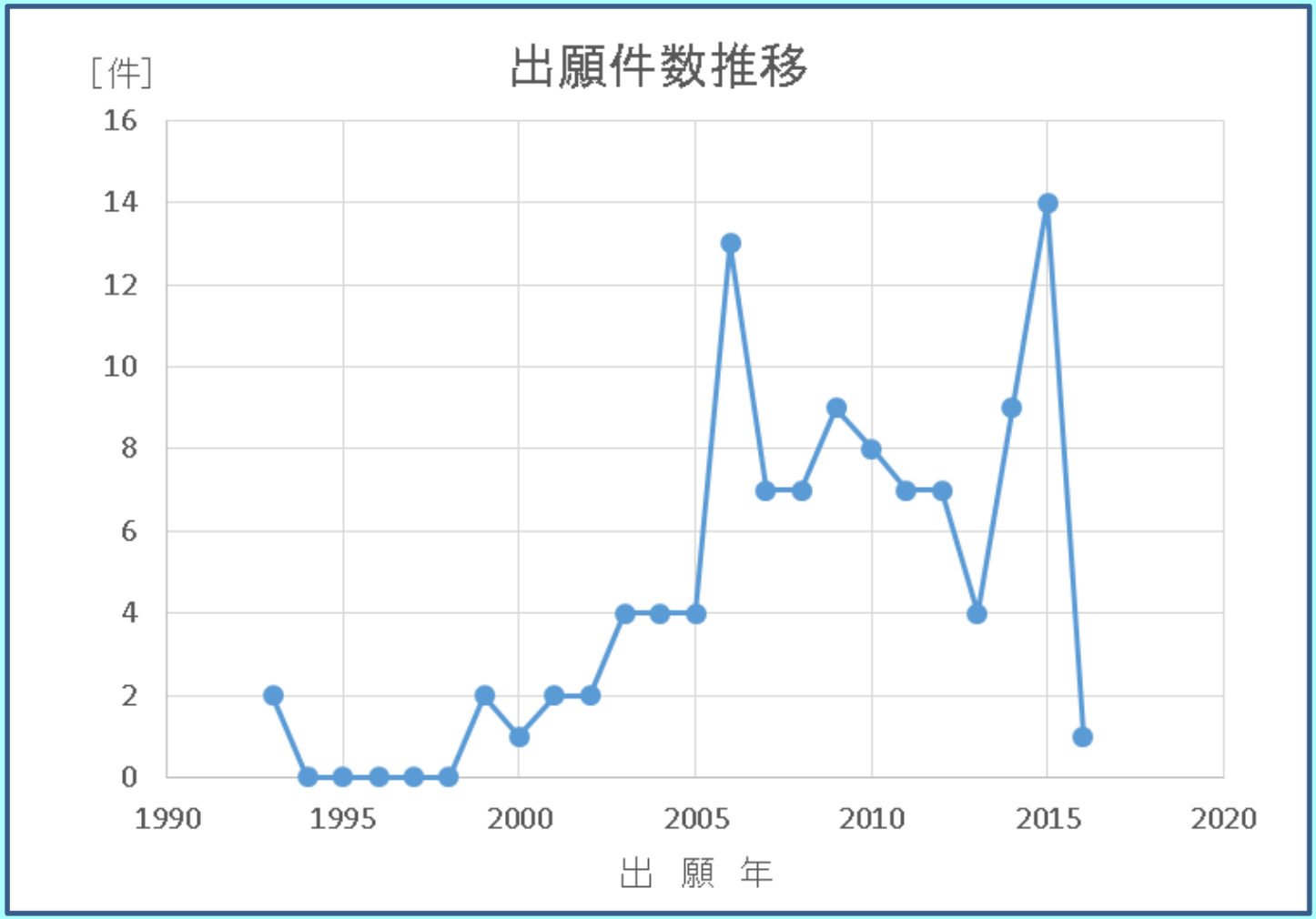
車線　　レーダ

3. 検索結果

自転車走行を認識する自動車に搭載するシステム

概念検索　＋　詳細検索　:対象107件

69件　　38件



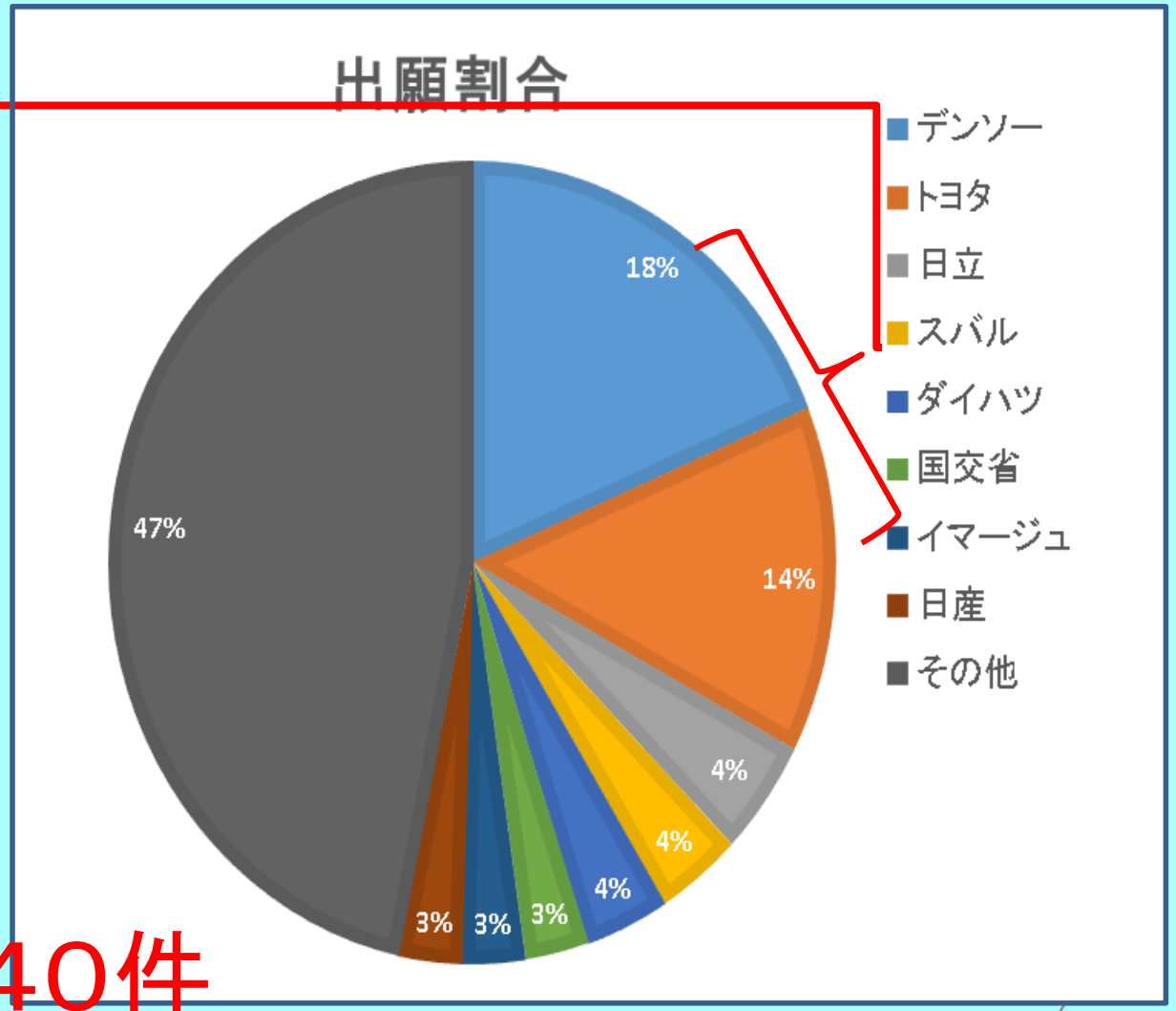
出願人の特徴

自動車メーカーが上位を占めている

①デンソー、トヨタで1／3占める

②電気関連：日立

③トヨタ以外の自動車メーカは低調



解析：上位3社

40件

上位3社のキーワードと頻度

キーワードだけでは全体象が見えない

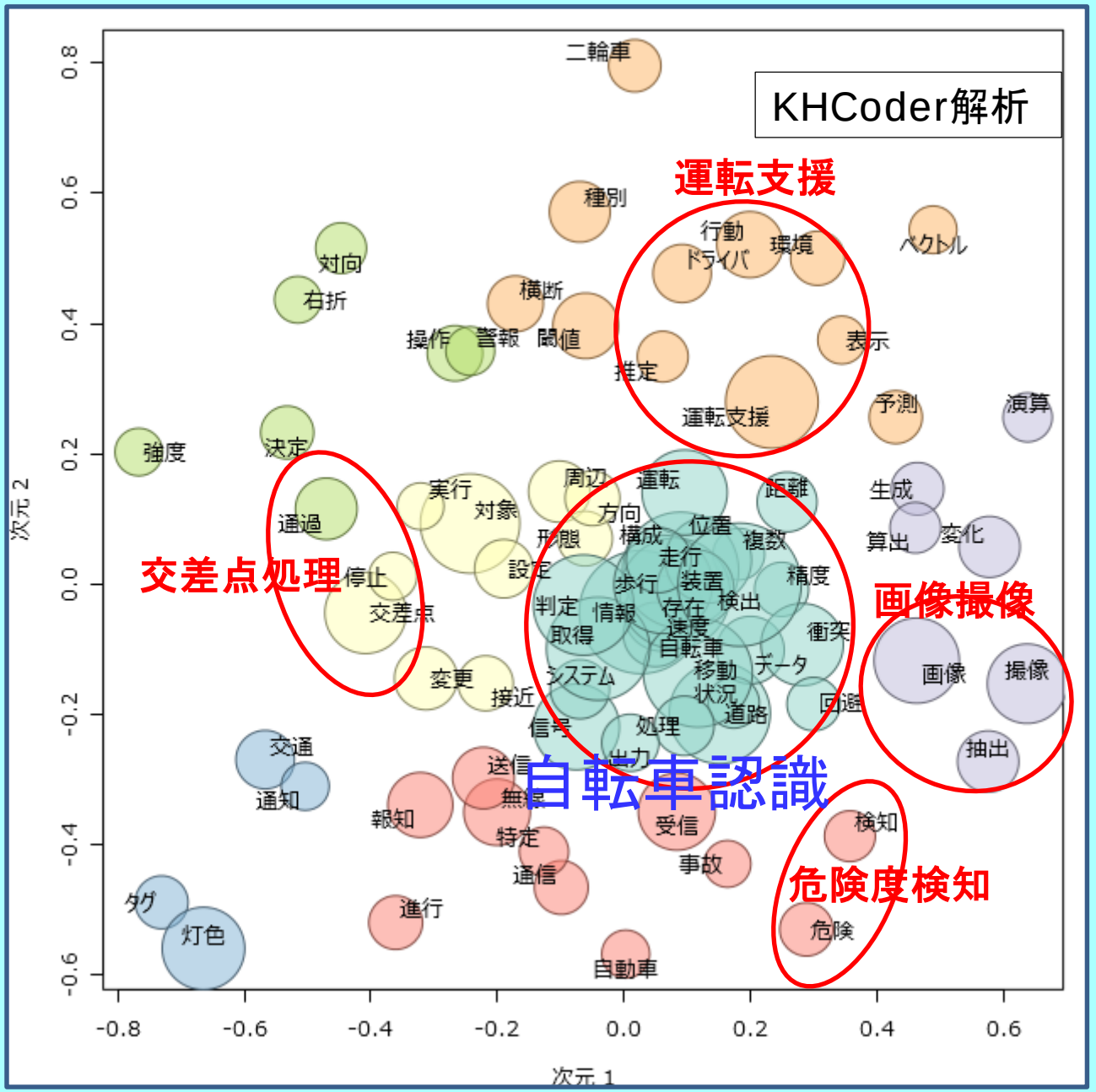
No.	Friq.	特許数	KW
1	1398	29	自車両
2	1114	37	情報
3	1107	35	判断
4	942	39	車両
5	910	36	歩行者
6	693	37	検出
7	631	36	ゾーン
8	597	19	物体
9	570	38	存在
10	568	38	自転車
11	531	13	移動体
12	502	16	運転支援
13	492	27	走行
14	449	35	取得
15	421	25	画像
16	417	20	交差点
17	386	21	閾値
18	373	32	設定
19	373	10	対象物
20	364	24	カメラ

No.	Friq.	特許数	KW
21	360	34	位置
22	328	29	出力
23	321	32	認識
24	311	18	衝突
25	311	10	ドライバ
26	310	26	道路
27	308	7	標
28	305	20	対象
29	286	24	受信
30	281	7	危険度
31	279	29	距離
32	267	22	送信
33	265	30	コントロール
34	261	15	ＥＣＵ
35	255	22	撮像
36	253	12	障害物
37	252	21	接近
38	244	21	変化
39	243	12	警報
40	233	20	運転者

4. 自転車認識技術の特徴

自転車認識の主要技術の周りに以下の技術が存在

- 運転支援
- 画像撮像
- 危険度検知
- 交差点処理



5. 特許の自動分類　5.1 SVM分類

自転車認識特許の抽出

課題解決文章の抽出

	灯色	危険	撮像	周期	通過	交差	物	衝突	lever
デン1	0	0.3	0	0	0	0	0.1	0	drvsapo
デン10	0	0	0	0	0	0	0.3	0.4	danglev
デン11	0	0	0	0	0	0.3	0	0	intsect
日立3	0	0	0.5	0	0	0	0.1	0	detct
日立4	0	0	0	0	0	0	0	0	detct
トヨ5	0	0	0	0.1	0	0.2	0.1	0.04	detct
トヨ6	0	0	0	0	0	0	0.02	0	drvsapo
トヨ7	0	0	0	0	0	0.08	0	0.04	drvsapo
トヨ8	0	0	0	0	0.1	0.1	0	0	intsect

n-gramのKW頻度

KW数の選択

評価ラベルの選択出願人／主題

学習／正解率評価

単wordのKW頻度

KW数の選択

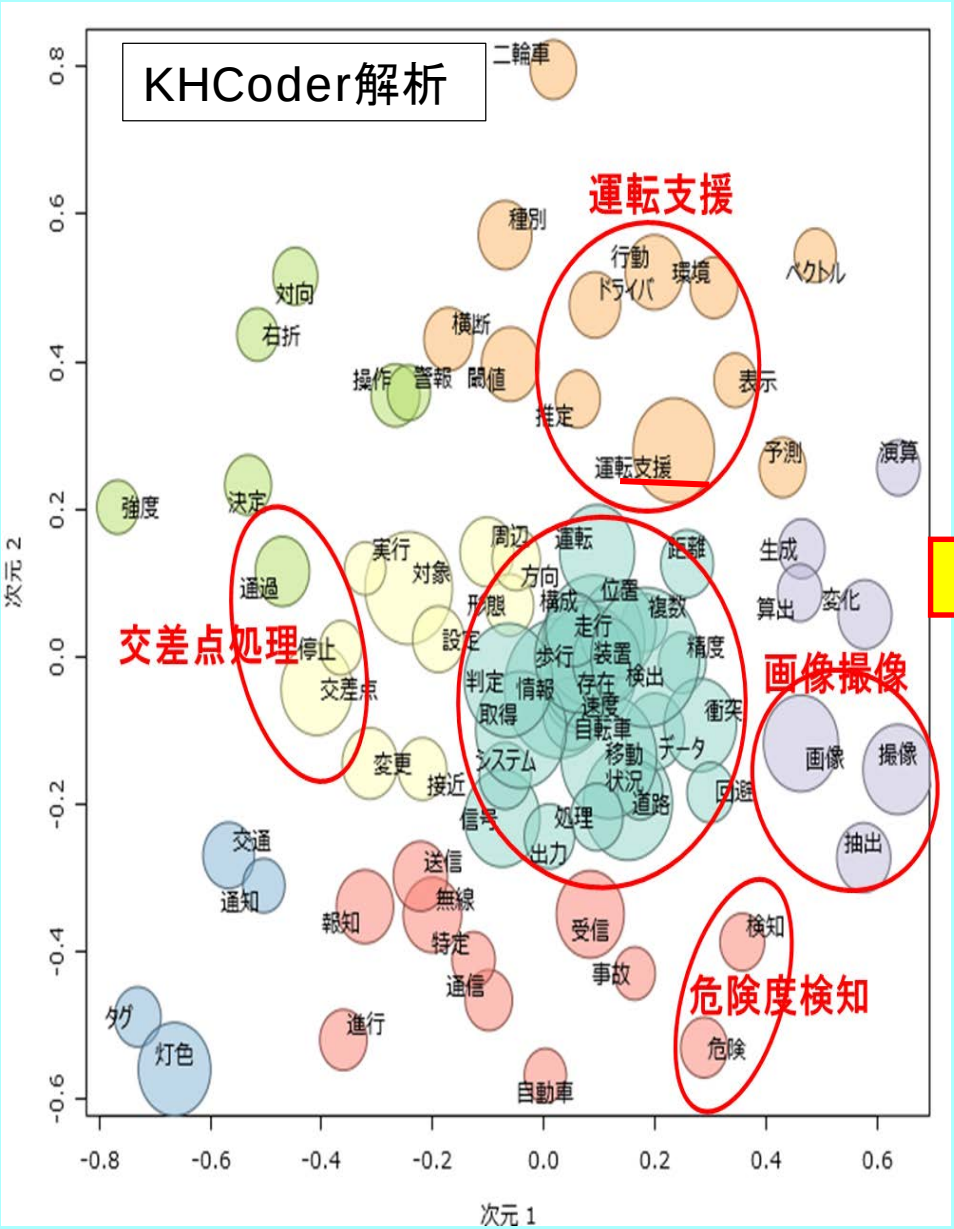
評価ラベル主題5種類/3種類

学習／正解率評価

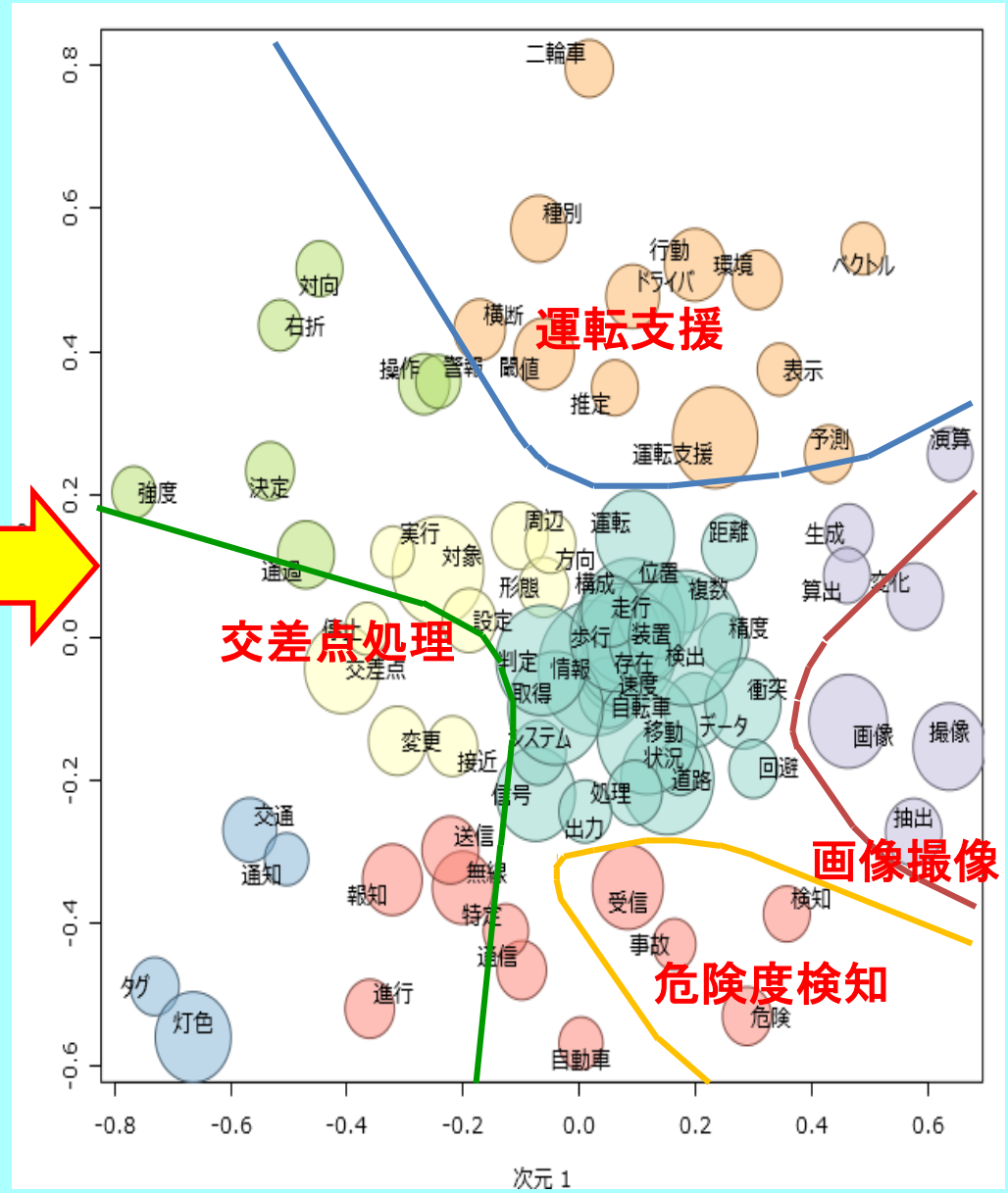
半分:学習　残り:推定用

分類モデルの決定

SVMによる特許分類イメージ

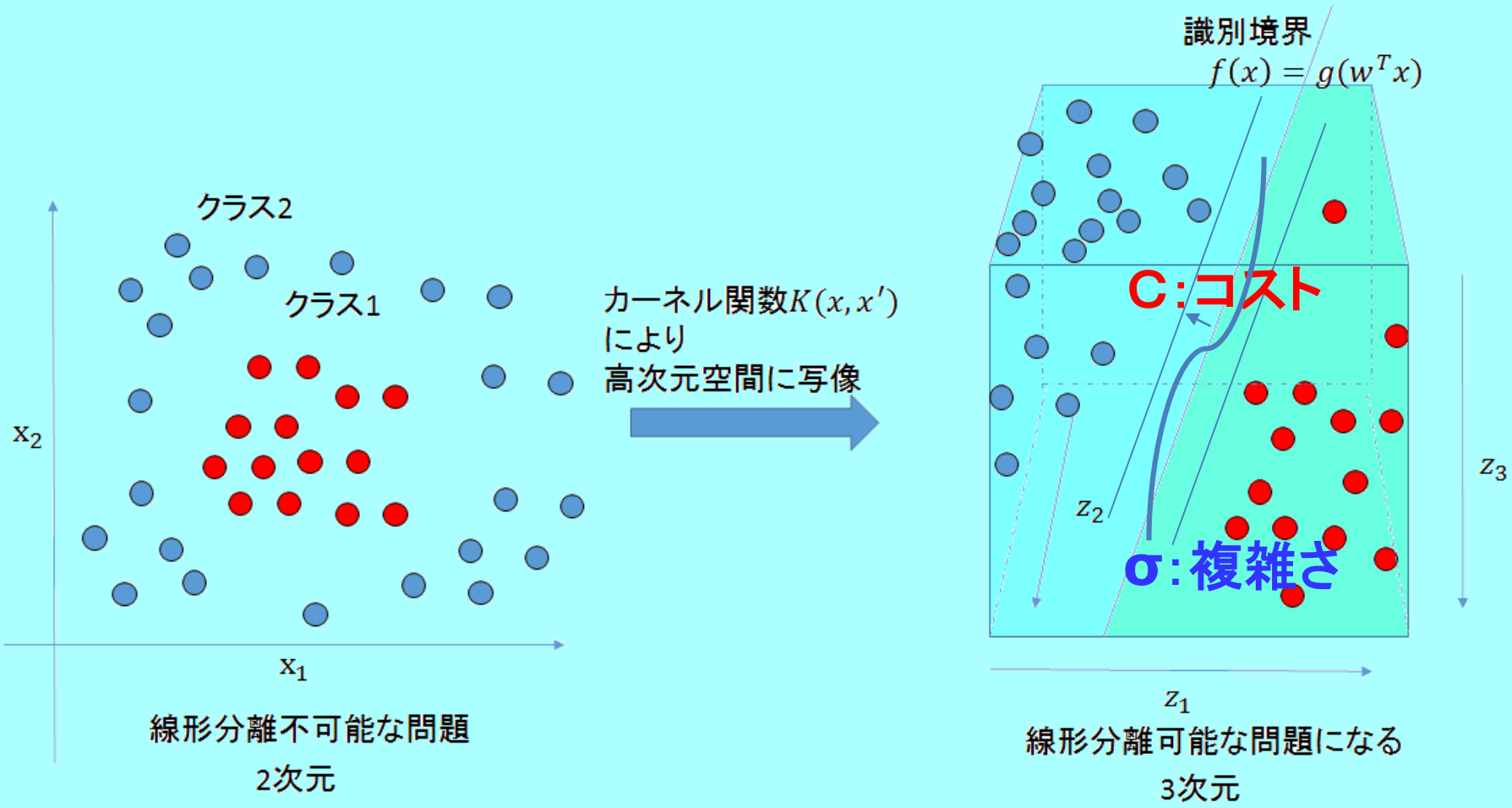


KWクラスタ分類



SVM分類

カーネル関数を用いて高次元化



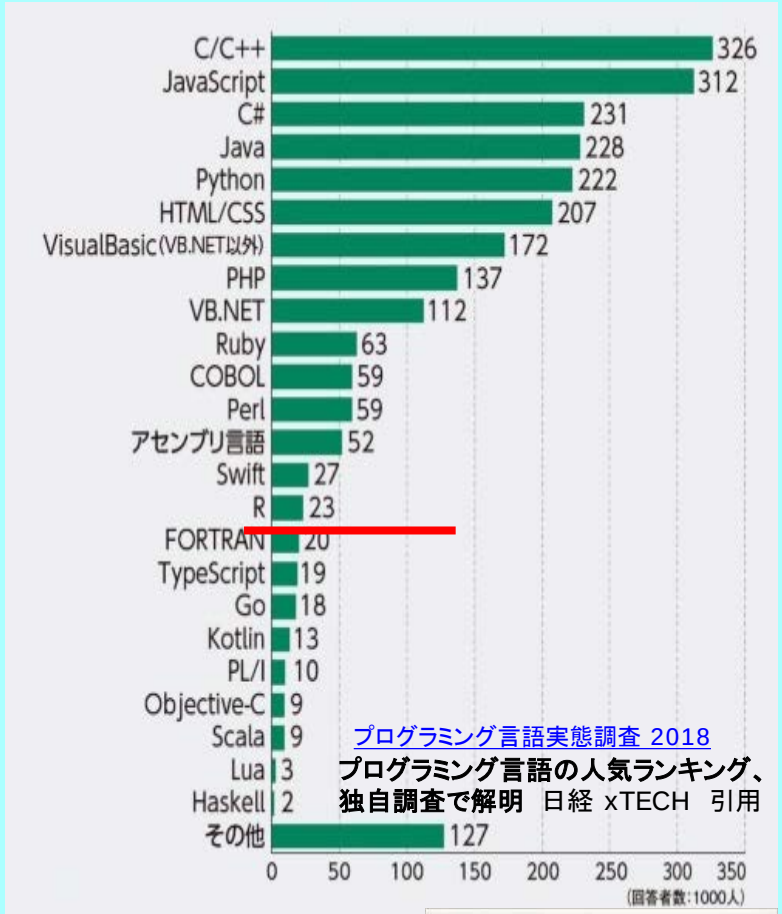
機械学習入門～ハードマージンSVM編～より引用

R言語とは

- 統計解析のフリーソフトウェア。
- いろんな人が便利な関数やパッケージを公開している。Rで大抵のことはできる。
- 形態素解析、機械学習(SVM,NN)を行うパッケージもある。 例 RMeCab,KSVM 等
- 多数の書籍が出版されている，マニュアルも充実。Rstudioによる Rの実行は便利：プロジェクト管理可能（Rによるテキストマイニング入門）。

  https://www.rstudio.com/

  を参照。



形態素解析の差異

- MeCab(フリーソフト)

  形態素|解析|と|分かち書き|の|違い|は|?
- Chasen(フリーソフト)

  形態素|解析|と|分かち書き|の|違い|は|?
- Kakasi(フリーソフト)

  形態素解析|と|分か|ち|書き|の|違い|は|?
- Text　Minig　Studio(有料ソフト)

  形態素解析と|分かち書きの|違いは|?

RMeCabによる頻度処理

RMeCab:RからMeCabを操作するパッケージ

・RMeCabC()関数 ： 短文の処理
・RMeCabText()関数 ： ファイルの解析結果をそのまま表示
・RMeCabDF()関数 ： データフレームの指定列を解析
・RMeCabFreq()関数 ： ファイルから頻度表を作成
・docMatrix()関数 ： 文書ターム行列(および重み付け), あるいはターム共起頻度行列を作成
・Ngram()関数 ： N-gram のカウント
・docNgram()関数 ： 指定されたディレクトリ内のすべてのファイルを対象に Ngramを抽出.
・NgramDF() 関数, NgramDF2() 関数 ： 基本的にNgram() 関数と同じであるが, N-gram を構成する各要素ごとに列に取ったデータフレームを出力.
・docDF()関数:指定したディレクトリのファイル全て、特定のファイル、あるいはデータフレームに対しNgramを作成してくれる関数

頻度の集計処理が1コマンドで可能

c2 <- docDF(“data2”,type=1,minFreq=5,N=3,pos=c(“名詞”))

N=3 が3-gram の設定パラメータ

R言語でのSVM実施例

- bic_svm<-ksvm(lever ~., data=bic_training,type="C-bsvc",kpar = list(sigma=0.1),C = 5)

bound-constraint svm classification

Support Vector Machine object of class "ksvm"

SV type: C-bsvc (classification)
parameter : cost C = 5

Gaussian Radial Basis kernel function.
Hyperparameter : sigma = 0.1

Number of Support Vectors : 20

Objective Function Value : -10.6803 -6.8296 -10.4272

Training error : 0

（1）KWの選定

A.　N-gram頻度

N-gram:
任意の文書や文字列などにおける任意のN文字が連続した文字列

/自/車両/と/対象/物/との/衝突/を/回避/する/装置

3-gram　⇩　名詞に限定

自・車両・対象　衝突・回避・装置

	TERM	サイクル. 内. 通過	ドライバ. 適切. 行動	移動. 通信. 端末	運転. 支援. 装置
1	densou1. csv	0	0	0	0
2	densou10. csv	0	0	0	0
3	densou11. csv	0	0	0	0
4	densou12. csv	0	0	0	0
5	densou13. csv	62	0	0	0

3-gram頻度

出願人による評価

①Min頻度:30　出願人分類　31変数

result_predict	dens	hita	toyo
dens	8	2	7
hita	2	1	0
toyo	0	0	0

平均正解率:45%

②Min頻度:15　出願人分類　175変数

result_predict	Dens	Hitch	Toyot
Dens	2	0	0
Hitch	4	0	0
Toyot	6	1	7

平均正解率:45%

主題による評価

主題による分類は正解率が低下する。

②Min頻度:15　出願人分類　175変数

result_predict	Dens	Hitch	Toyot
Dens	2	0	0
Hitch	4	0	0
Toyot	6	1	7

平均正解率:45%

③Min頻度:15　主題分類　176変数

result_predict	danglev	detct	drvsapo	intsect	pict
danglev	0	0	0	0	0
detct	0	5	4	2	1
drvsapo	0	5	2	1	0
intsect	0	0	0	0	0
pict	0	0	0	0	0

平均正解率:37%

B.　単wordKW頻度

KW数を減少すると精度UP　41%→39%→45%

①主題数5　kw数:175

result_predict	danglev	detct	drvsapo	intsect	pict
danglev	0	0	0	0	0
detct	2	4	5	1	1
drvsapo	0	3	3	0	0
intsect	0	0	0	1	0
pict	0	0	0	0	0

平均正解率:41%

②主題数5　kw数:73

result_predict	danglev	detct	drvsapo	intsect	pict
danglev	0	0	0	0	0
detct	2	8	7	3	0
drvsapo	0	0	0	0	0
intsect	0	0	0	0	0
pict	0	0	0	0	0

平均正解率:39%

単wordKW頻度Ⅱ

頻度の少ない危険度検知、画像処理を除くと精度UP

③主題数5　kw数:8

result_predict	danglev	detct	drvsapo	intsect	pict
danglev	0	0	0	0	0
detct	1	5	2	0	0
drvsapo	1	3	4	3	0
intsect	0	0	0	0	0
pict	0	1	0	0	0

平均正解率:45%

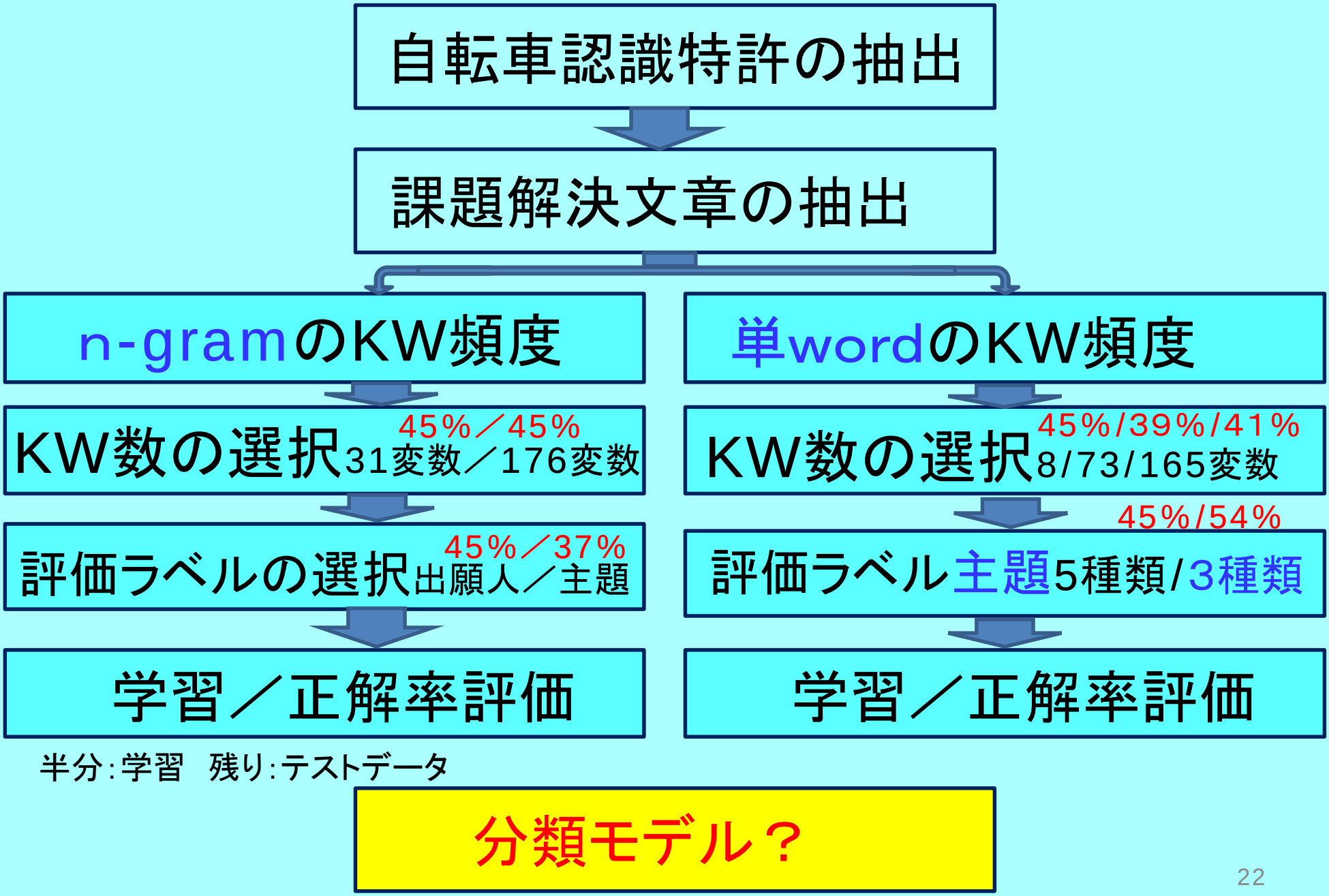
④主題数3　kw数:8

result_predict	detct	drvsapo	intsect
detct	9	1	1
drvsapo	1	4	2
intsect	0	0	0

瞬間高正解率:72%

⇒平均正解率:54%

5.2 SVM分類の結果

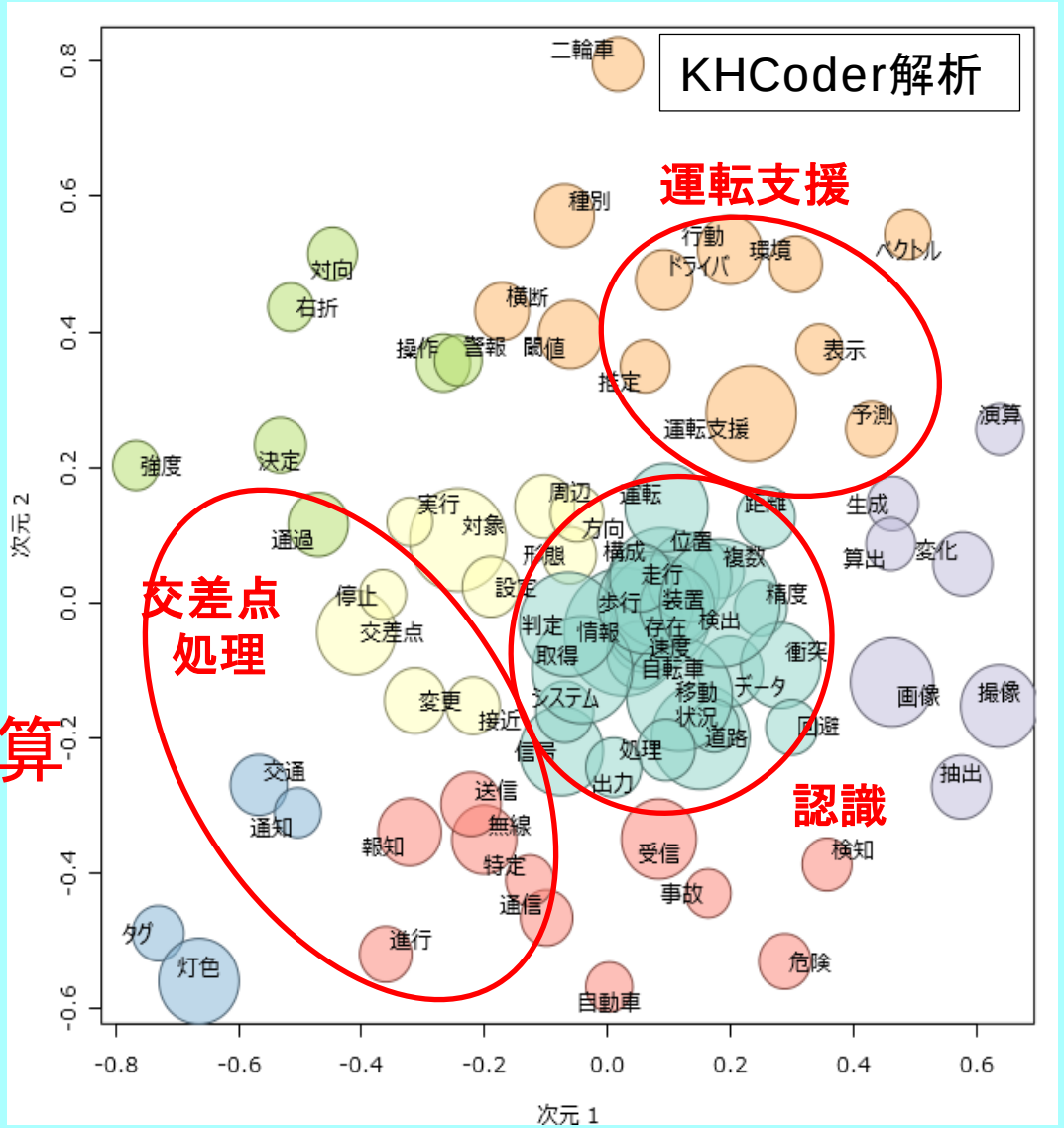

自転車認識特許の抽出
課題解決文章の抽出
n-gramのKW頻度
KW数の選択
45%／45%
31変数／176変数
評価ラベルの選択
45%／37%
出願人／主題
学習／正解率評価
単wordのKW頻度
KW数の選択
45%/39%/41%
8/73/165変数
評価ラベル主題5種類/3種類
45%/54%
学習／正解率評価
半分：学習　残り：テストデータ
分類モデル？

同義語処理

対象認識、交差点処理、運転支援に関する下位語をまとめる。

上位語	下位語	上位語	下位語	上位語	下位語
認識	歩行	交差点処理	交通	運転支援	推定
	走行		交差点		支援
	装置		進行		表示
	自転車		信号		ドライバ
	検出		検知		行動
	移動		通過		環境
	道路		通信		予測
	データ		無線		
	複数		送信		
	存在				
	状況				
	情報				
	構成				
	位置				
	速度				
	システム				
	処理				
	出力				
	記録				
	取得				
	精度				
	距離				

下位語の頻度を加算
方言のとりまとめ
距離の近いkwをGr化



同義語処理結果

正解率　max:50%　mean:34%

result_predict	danglev	detct	drvsapo	intsect
danglev	0	0	0	0
detct	2	9	5	0
drvsapo	0	1	1	0
intsect	0	0	2	0

認識に関する予想:7／16＝44%が外れ値
認識に関するkwの重要度が高いのでは？

SVM分類まとめ

- N-gramと単wordの頻度による分類では単wordの方が分類の正解率が高い。
- 特許のKW頻度では、分類正解率は40%～55%。同義語処理しても向上せず。
- KWの選択方法にはP値を使うことも考えられるが平均正解率は41%であった。

	danglev	detct	drvsapo	intsect	pict	Fisher.p値
灯色	0	0	0	214	0	1.29E-160
危険	116	0	13	0	0	9.95E-121
撮像	26	64	19	0	58	3.55E-80
周期	0	6	0	0	44	5.57E-76
通過	0	18	0	121	0	2.93E-74
交差点	0	34	13	149	0	4.35E-73
物	32	285	27	12	0	3.92E-72
衝突	93	122	25	0	0	3.43E-71

- KWの有／無よる分類を実施したが、平均正解率は52%で7%しか向上しない。

	灯色	危険	撮像	周期	通過	交差点	物	衝突	lever
densou1.csv	0	1	0	0	0	0	1	0	drvsapo
hitachi1.csv	0	0	0	0	1	1	1	1	detct
toyota1.csv	0	0	0	0	0	0	1	1	detct

result_predict	danglev	detct	drvsapo	intsect	pict
danglev	0	0	0	0	0
detct	1	3	2	0	0
drvsapo	1	5	4	0	1
intsect	0	0	0	3	0
pict	0	0	0	0	0

平均正解率:52%

特許頻度の特異性の回避策

P値
P-value
統計的仮説検定において、帰無仮説の元で検定統計量がその値となる確率のこと。P値が小さいほど、検定統計量がその値となることはあまり起こりえないことを意味する。

	danglev	detct	drvsapo	intsect	pict	Fisher.p値
灯色	0	0	0	214	0	1.29E-160
危険	116	0	13	0	0	9.95E-121
撮像	26	64	19	0	58	3.55E-80
周期	0	6	0	0	44	5.57E-76
通過	0	18	0	121	0	2.93E-74
交差点	0	34	13	149	0	4.35E-73
物	32	285	27	12	0	3.92E-72
衝突	93	122	25	0	0	3.43E-71

運転、車両、支援を追加

	ﾃﾞﾝｿ4	ﾃﾞﾝｿ5	ﾃﾞﾝｿ15	ﾃﾞﾝｿ20
運転	0.34	0.1	0	0.12
車両	0	0.07	0.08	0.05
支援	0.42	0.13	0	0

分類の正解率向上せず　47%

	灯色	危険	撮像	周期	通過	交差点	物	衝突
ﾃﾞﾝｿ12	0	0	0	0	0	0	0	0
ﾃﾞﾝｿ13	0.3	0	0	0	0.4	0.2	0	0
ﾃﾞﾝｿ15	0	0	0	0	0	0	0	0
ﾃﾞﾝｿ17	0	0	0	0	0	0	0	0
ﾃﾞﾝｿ18	0	0	0.5	0.6	0	0	0	0
ﾃﾞﾝｿ20	0	0	0	0	0	0	0	0
ﾃﾞﾝｿ4	0	0	0	0	0	0	0	0
ﾃﾞﾝｿ6	0	0	0	0	0	0	0.2	0.3
ﾃﾞﾝｿ8	0	0	0	0	0	0	0.4	0
日立1	0	0	0	0	0.2	0.1	0	0.3
日立2	0	0.2	0.3	0	0	0	0	0
日立5	0	0.5	0.2	0	0	0	0.1	0.4
ﾄﾖﾀ10	0	0	0	0	0	0	0	0
ﾄﾖﾀ12	0	0	0.1	0	0	0	0.3	0
ﾄﾖﾀ15	0	0	0	0	0	0	0.1	0
ﾄﾖﾀ3	0	0	0	0	0	0	0.2	0.1
ﾄﾖﾀ9	0	0	0	0	0	0	0.3	0
ﾃﾞﾝｿ5	0	0	0	0	0	0	0	0
ﾃﾞﾝｿ9	0	0	0	0	0	0	0.3	0
ﾄﾖﾀ13	0	0	0	0	0	0	0	0

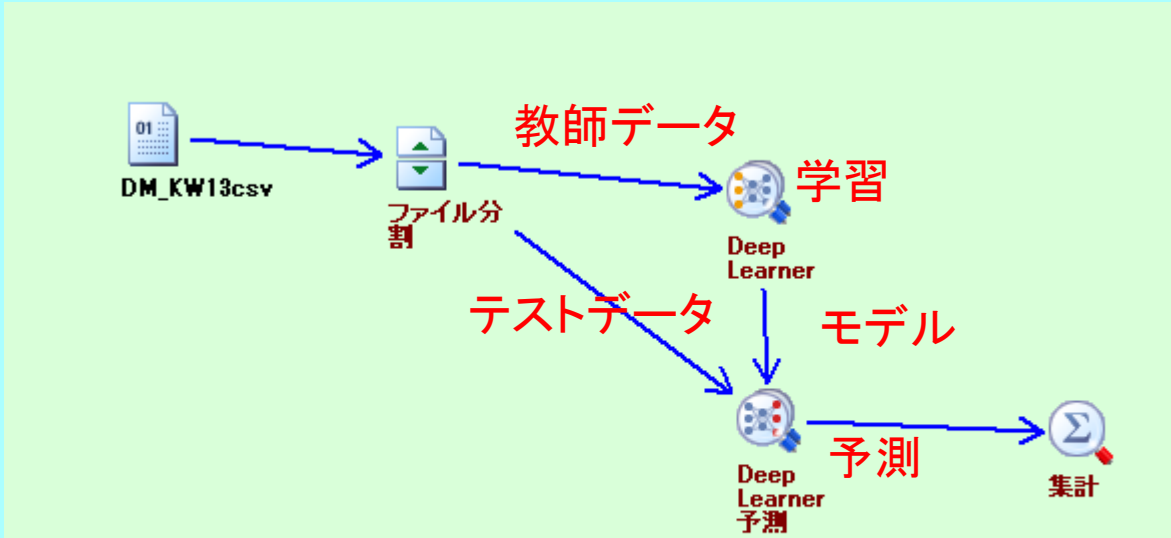
5.2 Deep Learnerによる分類

NTT数理システム製
Visual Mining
Studio上でDeep
Learnerを実行

40件の頻度テーブル

教師データ　:80%
テストデータ:20%

目的変数　:Lever
説明変数:単word
同義語処理



	認識	交差点	支援	車両	可能	形態	制御	判定	周辺	物体	画像	方向	衝突	lever
デン10	1	0	0	2	22	0	0	19	0	0	0	9	24	danglev
日立5	5	2	0	12	14	0	23	0	0	47	22	0	69	danglev
トヨ2	4	1	0	2	20	7	0	0	4	9	0	0	0	danglev
デン16	8	22	3	3	17	6	5	0	0	0	0	3	10	detct
デン17	8	7	1	7	13	0	3	32	0	0	0	0	6	detct
デン2	1	0	0	0	0	0	0	9	0	19	8	0	0	detct
デン1	2	0	0	6	4	0	0	0	0	0	0	0	0	drvsapo
デン12	4	3	0	5	0	0	4	0	0	0	0	0	0	drvsapo
デン15	5	1	0	6	3	0	8	10	0	0	0	13	0	drvsapo
デン11	5	5	1	12	4	3	0	0	0	0	0	8	0	intsect
デン13	13	32	4	32	61	7	4	68	9	6	0	0	0	intsect
デン14	8	14	2	16	28	10	50	7	4	0	0	0	0	intsect

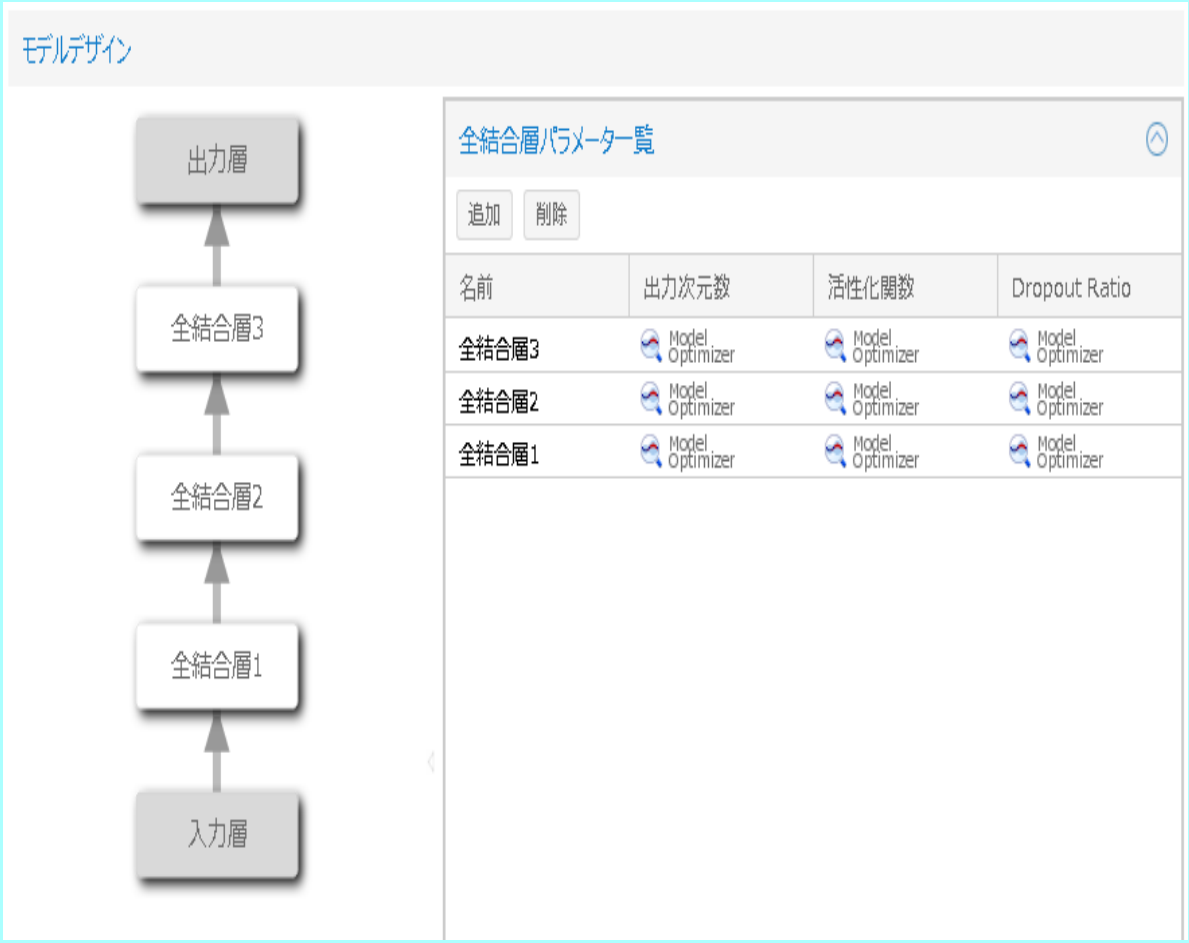
モデル化

パラメータ最適化で次元数×活性化関数×DropoutRatio
（13＊4＊3）＝156個のパラメータ設定は不要

教師データで学習し
3層ニューラルネット
ワークモデルを作成

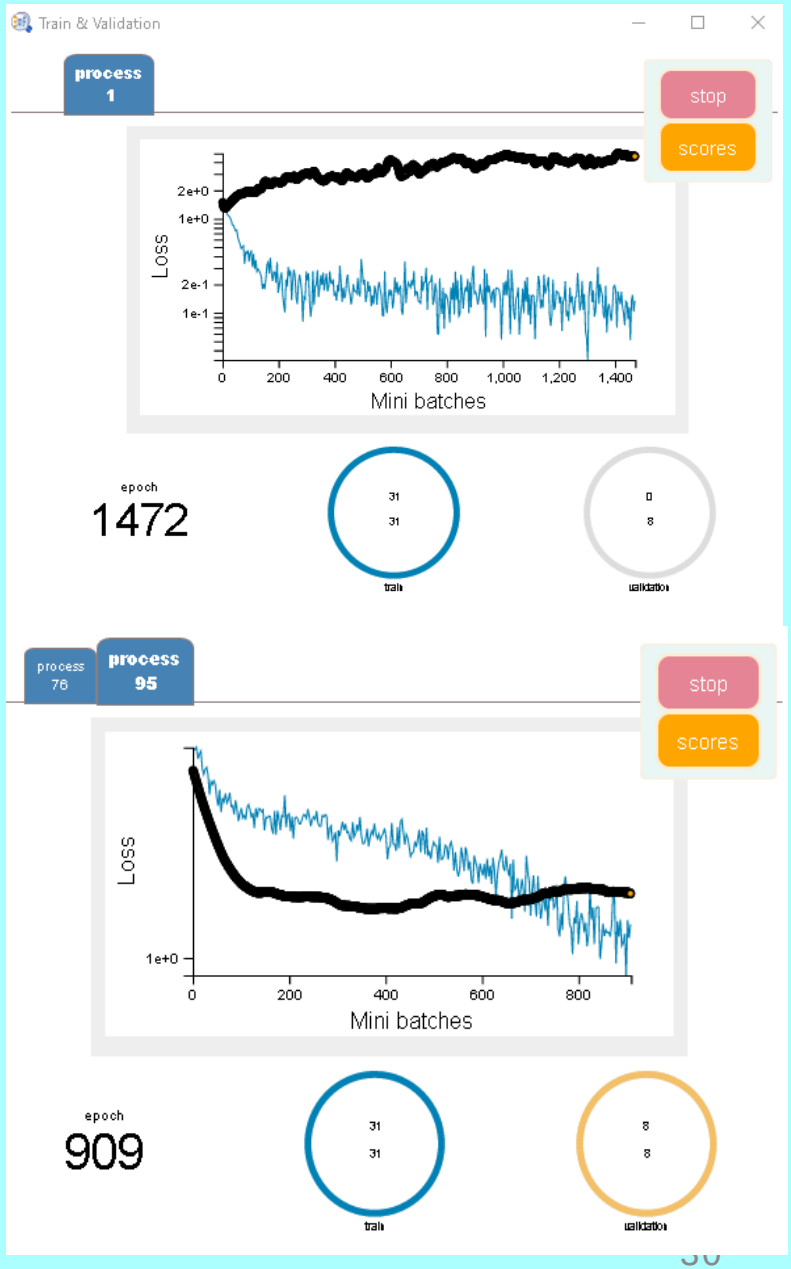
出力次元：13次元
↓
最適化
活性化関数：ReLu
softmax
↓
最適化

テストデータで予測
分類の正解率で評価



教師データのフィット率

A	認識平均	交差点平均	支援平均	車両平均	lever	lever.予測
densou10.	0.9	0	0	2	danglev	detct
hitachi5.cs	5.2	1.6	0.2	12	danglev	detct
toyota2.cs	4.2	0.8	0.1	2	danglev	detct
densou16.	8	22.4	2.8	3	detct	detct
densou17.	8.4	6.5	0.8	6.7	detct	detct
densou2.c	1.1	0	0	0	detct	detct
densou6.c	4	0	0	7	detct	detct
densou7.c	0.2	0	0	3.3	detct	detct
densou8.c	0.9	0.3	0	0	detct	drvsapo
hitachi1.cs	5.7	2.8	0.4	27.3	detct	detct
hitachi3.cs	0	0.4	0.1	1	detct	detct
toyota1.cs	2.4	0	0	6	detct	detct
toyota10.c	18.4	0.6	0.1	6.3	detct	detct
toyota12.c	10.3	0	0	18.3	detct	detct
toyota15.c	8.5	0	0	7.3	detct	detct
toyota5.cs	1.2	2.3	0.3	6	detct	detct
densou1.c	1.8	0.3	0	6.3	drvsapo	drvsapo
densou19.	0.4	0	0	0	drvsapo	drvsapo
densou20.	1.3	0	0	4.3	drvsapo	detct



当てはめの正解率：68%～85%

チューニングと分類精度

中間層を増やすと精度向上
- 2層:50%
- 3層:60%

エポックを増やすと精度向上
- 3層5000→10000
- 精度:60%→75%

出力次元数などを最適化設定
- エポック:5000
- 75%を達成

学習率:0.01→0.001
- 分類精度:75%→88%

但し、学習時間:
1時間→2時間

Case1: 学習率:0.01

	予 danglev	予 detct	予 drvsapo
danglev	1	0	0
detct	0	4	0
drvsapo	0	2	1

精度: 75%

Case2: 学習率:0.001

	予 _danglev	予 detct	予 _drvsapo
danglev	4	0	0
detct	0	3	0
drvsapo	0	1	0

精度: 88%

まとめ

小さなモデルを対象にすれば機械学習の流れと、ブラックボックス化した分類内容を少しは理解できる。

SVM分類

- 頻度による分類ではkw選定は、３－gramより単wordの方が分類精度は良い。
- 特許のKW頻度では、分類正解率は40%～55%。同義語処理しても向上せず。
- SVMはC-bsvcタイプで分類したが、線形／非線形で大差はなかった。

Deep　Learner分類

- 40件のデータによる分類試行結果、学習率などを調整すれば、精度：88%を達成できる