Академия народного хозяйства при Правительстве РФ

Кафедра системного анализа и информатики

Индуктивное моделирование: содержание и примеры применения в задачах обработки текстов

М. Александров
Академия народного хозяйства при Правительстве РФ
Автономный Университет Барселоны, Испания

Петербург 2010

Содержание

Введение

История

ИМСОМ = Индуктивный Метод Самоорганизации Моделей

- Был разработан в в 70-80 годы акад. А.Г. Ивахненко и его учениками
- Принадлежит к числу эволюционных алгоритмов Искусственного Интеллекта

Современность

В настоящее время говорят не столько об индуктивном методе, сколько об индуктивном подходе к процедуре моделирования.

Поэтому используются термины:

- индуктивное моделирование
- индуктивное порождение моделей

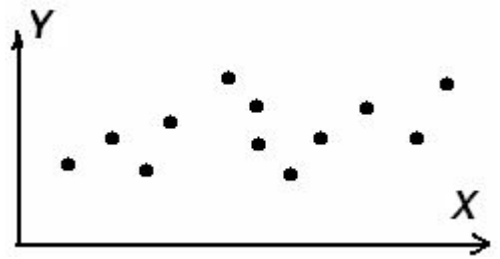
Они отражают развитие ИМСОМ

Введение

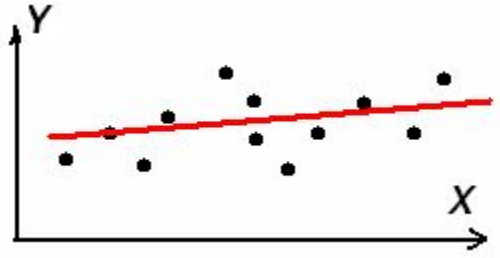
Что стоит за названием?

ИМСОМ = Индуктивный Метод Самоорганизации Моделей

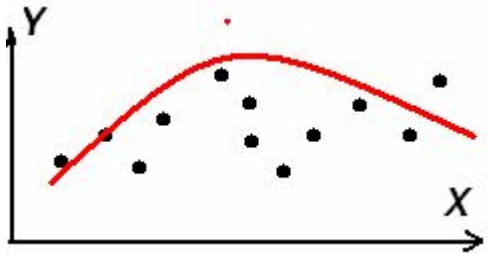
Понятие модели зависит от задачи. Если необходимо описать связь наблюдений (X_i, Y_i), то модель это зависимость $Y_i = F(X_i)$



Экспериментальные данные



Линейная модель - прямая



Квадратичная модель - парабола

Введение

В чем индуктивность?

Индукция = из частных случаев делают обобщенный вывод
Дедукция = из общего результата делают выводы о частных случаях

Пример индуктивного вывода – метод математической индукции

Индуктивность в нашем случае состоит в том, что мы рассматриваем конкретные частные модели, шаг за шагом усложняя их:
прямая, парабола, полином 3-го порядка, 4-го порядка, и т.д.

Но мы не рассматриваем одновременно все модели
(степенные и тригоном. полиномы, ряды экспонет, и т.п.)
или даже какой-то конкретный класс в целом

Введение

В чем самоорганизация?

Саморганизация системы – это изменение ее структуры/параметров под влиянием внешних условий

Самоорганизация у нас состоит в том, что модель меняется от простой к сложной в процессе автоматического перебора моделей, пока она не достигнет оптимальной сложности

Внешние условия – это данные наблюдений. Они отражают: как поведение объекта, так и шум

Тогда очевидно, что:
- Простая модель не реагирует на шум, но плохо отражает объект
- Сложная модель отражает объект, но чувствительна к шуму
Есть некоторый оптимум, который достигается в процессе перебора

Термин используется условно – у нас иллюзия самоорганизации. Ведь это мы меняем модель, а не она сама себя

Введение

Возможности

ИМСОМ позволяет выбрать модель оптимальной сложности из заданного класса моделей, чтобы описать ограниченный набор экспериментальных данных

Ограничения

ИМСОМ обладает преимуществами, когда отсутствует или почти отсутствует априорная информация о распределении параметров модели или даже о структуре модели в целом

Если такая информация имеется, или если данных достаточно много, чтобы такую информацию извлечь, то надо использовать другие подходы. *Они могут дать лучшие результаты !*

Введение

Терминология

Термин ИМСОМ был почти сразу заменен авторами метода на термин МГУА

МГУА = метод группового учета аргументов

GMDH = group method of data handling (англ.)

Приложения

- Аппроксимация функций
- Выбор вычислительной схемы
- Cluster validity
- Self-organizing Data Mining
- Обучение нейронных сетей

и т.д.

Содержание

Коллеги и соавторы

Pavel Makagonov

Titled Research Professor

Mixteca University of Technology, Mexico

Ex Vice-Director of Moscow Mayor Office

mpp2003@ inbox.ru

Автор первых приложений индуктивного метода к задачам обработки текстов

Автор модификации индуктивного подхода: селекция моделей вместе с селекцией данных

Коллеги и соавторы

Xavier Blanco

Titled Professor of French Philology Department

Universidad Autonoma de Barcelona, Spain

xavier.blanco@ uab.cat

Angels Catena

Coordinator of Master Program

Professor of French Philology Department

Universidad Autonoma de Barcelona, Spain

angels.catena@ uab.cat

Коллеги и соавторы

Alexander Gelbukh

Chief of NLP Laboratory

Center for Computing Research

National Polytechnic Institute, Mexico

gelbukh@ gelbukh.com

Natalia Ponomareva

Ph.D. student

Mathematician-Programmer

Wolwergampton University, UK

nata.ponomareva@ gmail.com

Содержание

Индуктивное Моделирование

Классы и сложность модели

ИМСОМ имеет дело с заранее фиксированным классом моделей.

Класс моделей зависит от рассматриваемой задачи.
Это могут быть:

- полиномы одной переменной
- линейные функции многих переменных
- кластеры объектов

и т.п.

Сложность модели – максимальное число параметров
при заданной структуре модели
В указанных выше случаях это:

- старшая степень полинома (+1)
- число переменных (+1)
- число кластеров

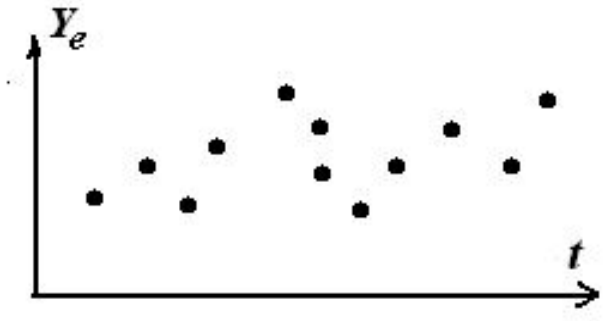
и .п.

Индуктивное моделирование

Каноническая проблема

Описание временного ряда некоторой формулой

Пусть начальная информация задана □



Экспериментальные данные

Заданная информация

- Тип зависимости (формула)
- Серия моделей из заданного класса и уровень шума

и т.п.

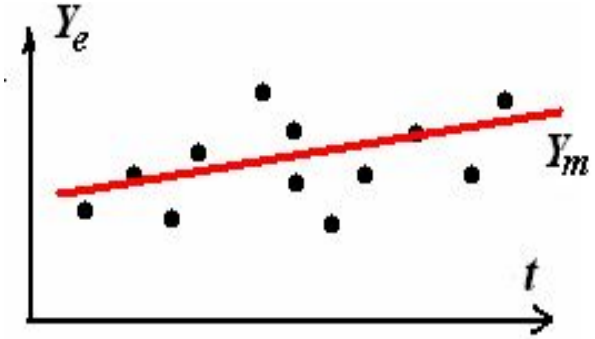
Индуктивное моделирование

Мы имеем начальную информацию

Заданная регрессионная модель

$Y_m = a_0 + a_1 t$ or $Y_m = a_0 + a_1 t + a_2 t^2$, *etc.*

$\| Y_m - Y_e \|$ => мин (используем МНК)



Точки – эксперим. данные

Красные линии – возм. модели

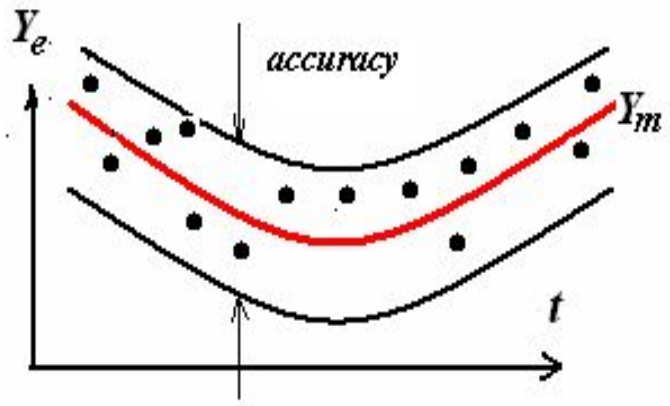
или

Серия моделей из заданного класса

с заданным уровнем шума

$Y_m = a_0 + a_1 t + a_2 t^2 + ...$

$\| Y_m - Y_e \|$ => ε^2 (используем МНК)



МНК = метод наименьших квадратов

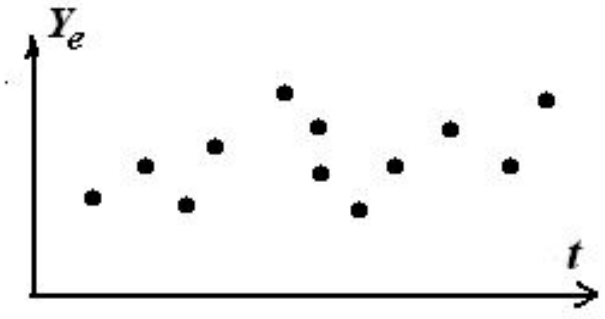
Корридор отражает ошибку ε

Индуктивное моделирование

Каноническая проблема

Описание временного ряда некоторой формулой

Пусть начальная информация отсутствует



Экспериментальные данные

У нас нет информации

В этом случае мы используем Индуктивное Моделирование

Для этого мы фиксируем класс моделей

Замечание: класс моделей должен отражать возрастающую сложность модели

Индуктивное моделирование

Принцип индуктивности

ИМСОМ не может найти самую оптимальную модель среди всех возможных! Он ищет оптимальную модель только в заданном классе

Пример класса моделей: полиномы одной пременной *(t)*

$Y_0 = a_0$

$Y_1 = a_0 + a_1 t$

$Y_2 = a_0 + a_1 t + a_2 t^2;$

.....

Пример класса моделей: линейные функции многих пременных (x_1, x_2..)

$Y_0 = a_0$

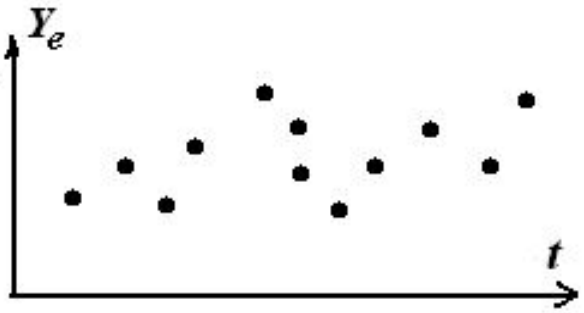
$Y_1 = a_0 + a_1 x_1$ $\quad Y_1 = a_0 + a_2 x_2$ .. $\quad Y_1 = a_0 + a_{10} x_{10}$ ..

$Y_2 = a_0 + a_1 x_1 + a_2 x_2$ $\quad Y_2 = a_0 + a_1 x_1 + a_3 x_3 \ldots$ $\quad Y_{2.} = a_0 + a_9 x_9 + a_{99} x_{99}$ ..

Индуктивное моделирование

Подход 1

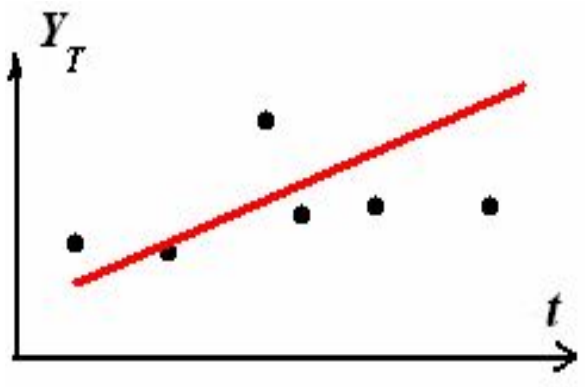
Мы должны обеспечить хорошее свойство прогнозирования, то есть ограниченную чувствительность к новым данным



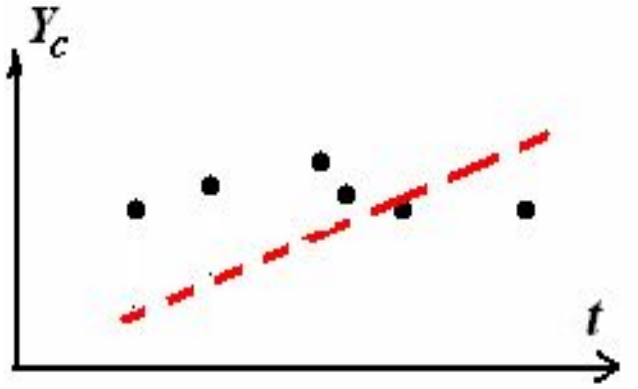
Экспериментальные данные

Критерий 1 (регулярность)

Модель, обученная на 1-м наборе данных должна давать хорошие результаты на втором наборе данных (***T*** обучение, ***C*** контроль)



Training – нечетные точки



Control – четные точки

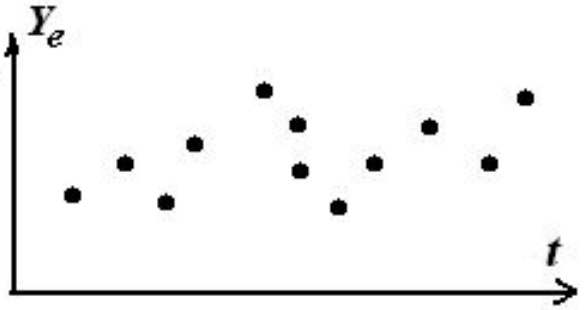
Индуктивное моделирование

Подход 2

Мы должны обеспечить хорошее

описательное свойство, то есть

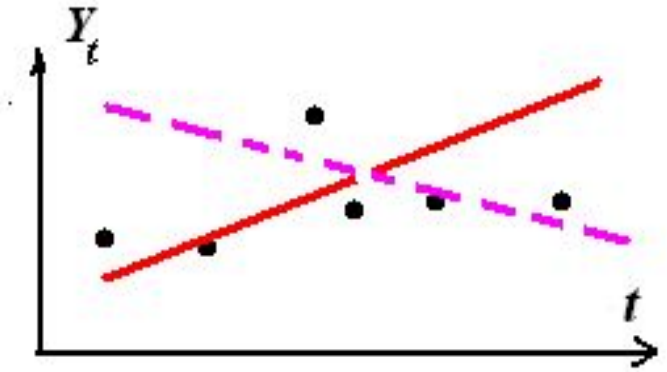
независимость описания от данных



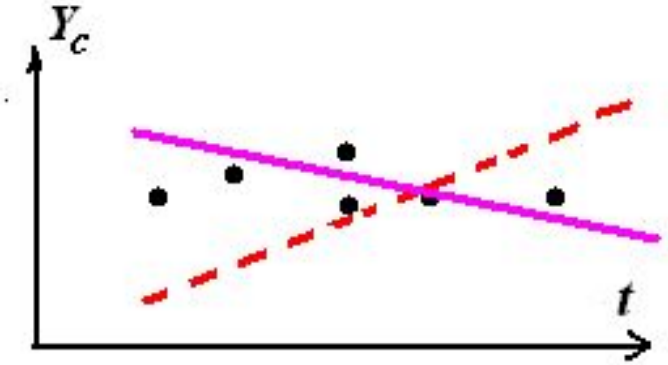
Экспериментальные данные

Критерий 2 (несмещенность)

Модель, обученная на 1-м наборе и проверенная на 2-м наборе (красная линия), должна быть подобна модели, обученной на 2-м наборе и проверенной на 1-м наборе (фиолетовая прямая)



Training – нечетные точки



Control – четные точки

Индуктивное моделирование

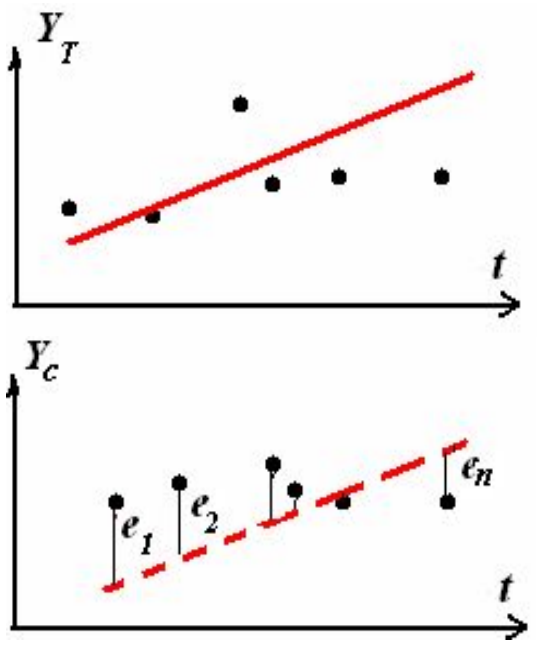
Формы внешних критериев

Качество модели оценивается внешними критериями

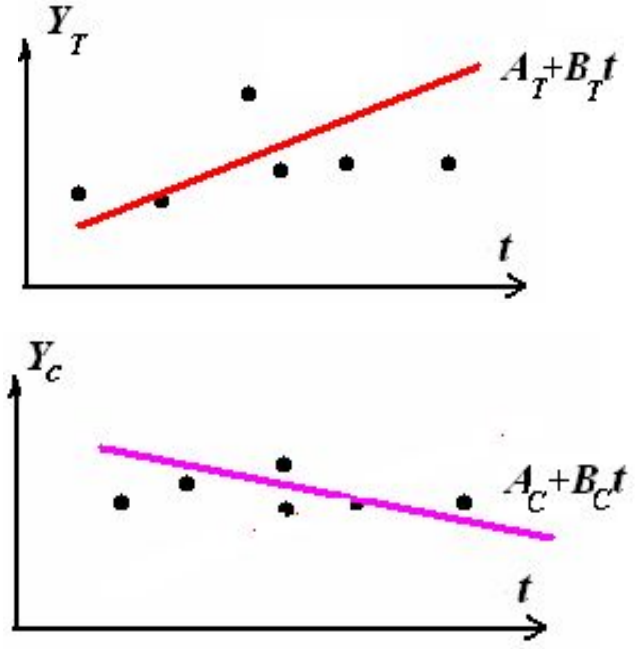
Каждый из критериев может быть представлен в одной из 2-х форм:

1) Критерии, ориентированные на данные

2) Критерии, ориентированные на модель



Критерий регулярности по данным: подсчет невязки ~ $\sqrt{\sum e^2_i}$



Критерий регулярности по модели: оценка близости

~ $|A_T - A_C| / A_T + |B_T - B_C| / B_T$

Индуктивное моделирование

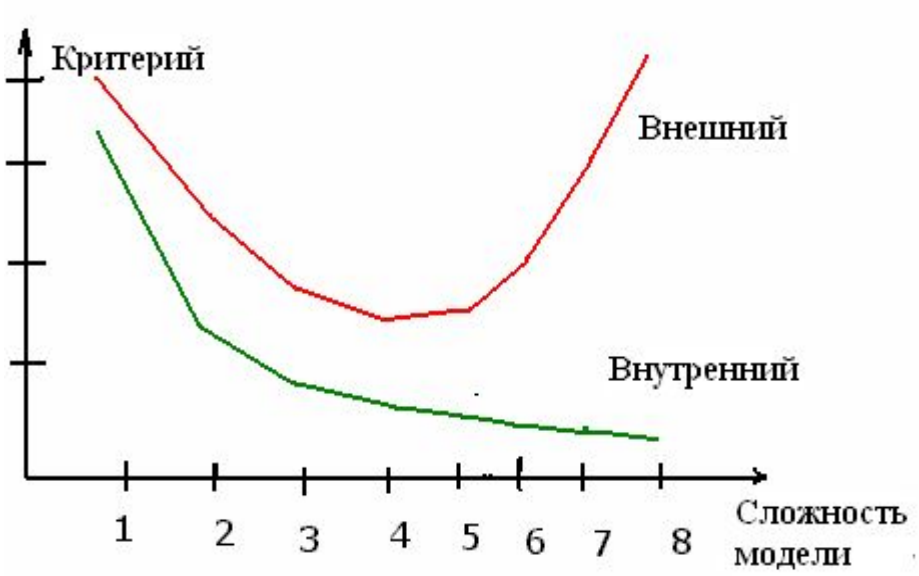
Шаги ИМСОМ <= !

1. Определить серию моделей

2. Экспериментальные данные =
   Данные для обучения +
   Данные для контроля

3. Для заданной сложности
   определяется лучшая модель
   для каждого набора, здесь
   используется *внутренний критерий*

4. Обе модели сравниваются
   с помощью *внешних критериев*
   (регулярность, несмещенность)

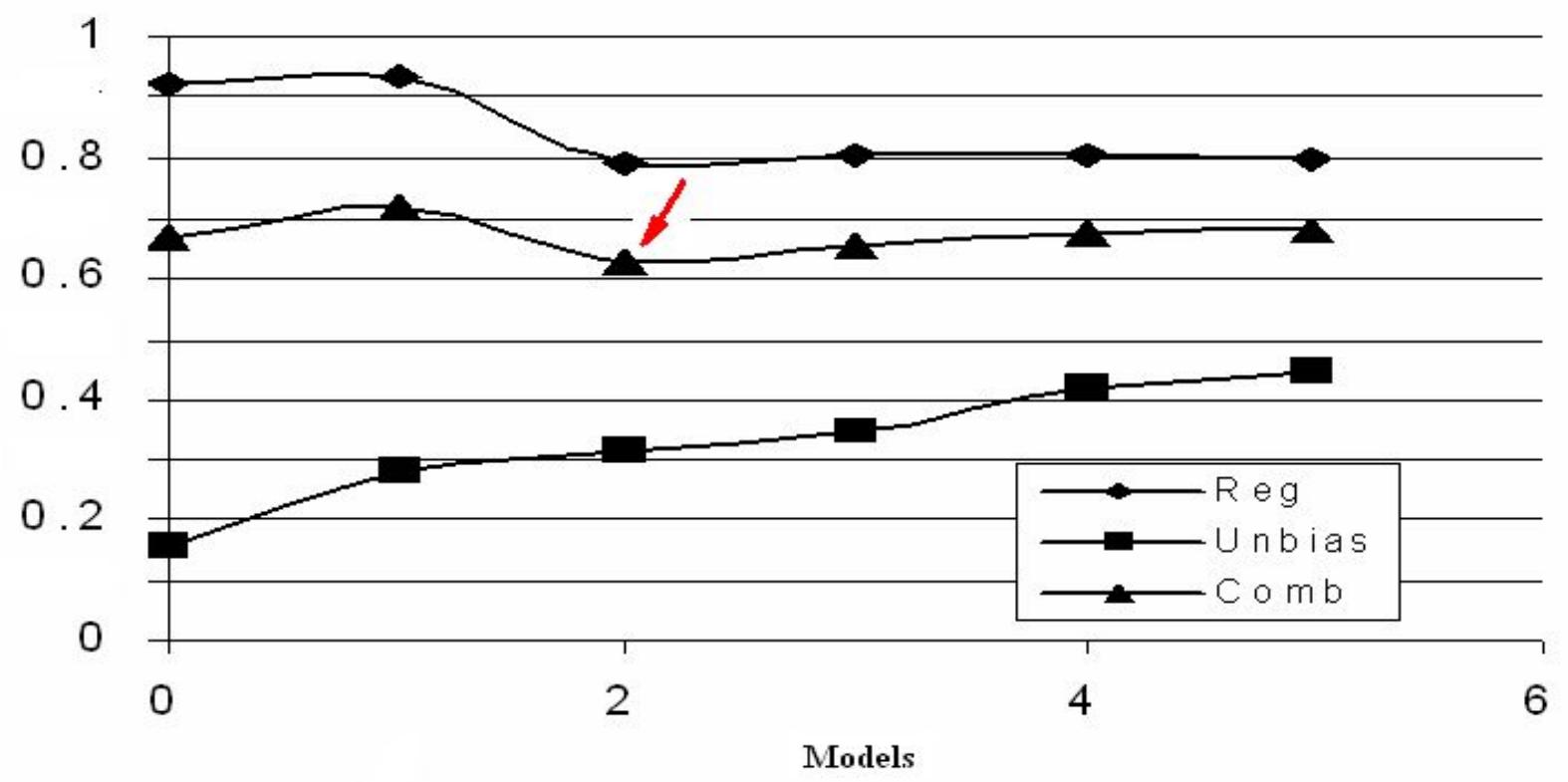


5. Если внешний критерий
   достигает минимума,
   то STOP, иначе
   увеличиваем сложность
   модели и идем на шаг 3

Индуктивное моделирование

Применение двух критериев, правило свертки

- Назначаются веса λ_1, λ_2: $\lambda_1 + \lambda_2 = 1$ и рассчитывается комбинированный критерий $K = \lambda_1 K_r + \lambda_2 K_u$
- Выбирается модель, лучшая по комбинированному критерию

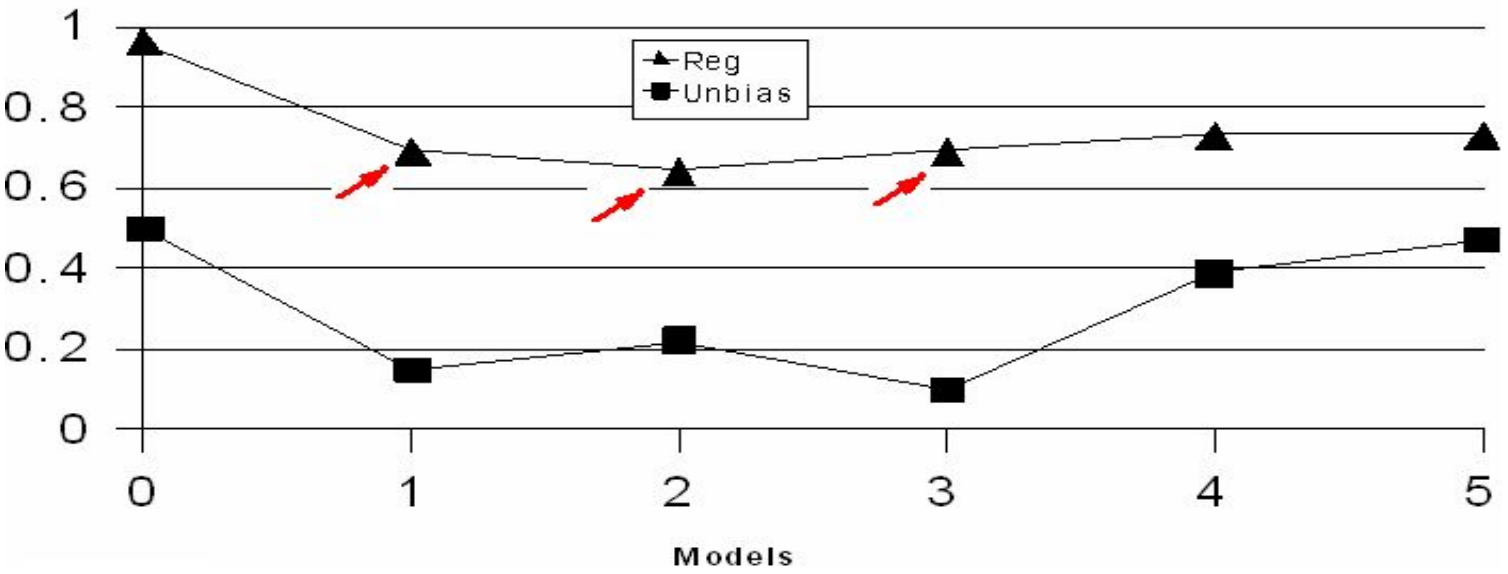


Индуктивное моделирование

Применение двух критериев, последовательный отбор

Вместо отбора модели по комбинированному критерию $K=\lambda_1 K_r + \lambda_2 K_u$ мы используем другую стратегию

- Выбираются лучшие модели по K_r
- Из них выбираются лучшие по K_u



В данном примере лучшими по K_r являются модели 1,2 и 3

Критерий K_u отбирает модель 3

Индуктивное моделирование

Подавление шума

Утверждение

Пусть имеем *N*-данных наблюдений $y_1, y_2, y_3, \ldots . y_N$

Пусть имеем *k*-параметров линейной регрессионной модели

$$F(t) = a_0 + a_1 t + a_2 t^2 + \ldots . a_{k-1} t^{k-1}$$

Число $n = N / k$ есть удельное число измерений

Тогда при расчете параметров модели шум будет подавлен в $\sqrt{n}$ раз

Пояснение

Пусть имеем с.в. ***X*** с отклонением σ. Пусть ***X*** = $x_1, x_2, x_3, \ldots .$ выборка

Известно, что среднее $\boldsymbol{M} = \sum x_i / n$ имеет отклонение $\sigma_M = \sigma / \sqrt{n}$

Утверждение (см.выше) следует из того, что коэффициенты регрессии также рассчитываются по формулам, связанным с усреднением

Индуктивное моделирование

Пример

Пусть мы должны восстановить параболу $Y(t)=a_0 + a_1 t + a_2 t^2$

Имеем $N=12$, $n=12/3=4$

Тогда подавление шума $\sqrt{4} = 2$

Экспериментальные данные

Следствие

Требование подавления шума определяет ограничения на необходимый объем данных измерений.

А именно, для подавления шума в 2-3 раза (это обычное требование) на один параметр должно приходиться 5-10 измерений

Индуктивное моделирование

Варианты ИМСОМ

Имеется два традиционных варианта:

1) Комбинаторный вариант ИСОМ - КОМБИ
Перебираются всевозможные модели в заданном классе от простых к сложным без селекции моделей. То есть, рассматриваются:

0-й порядок $Y_0 = a_0$,

1-й порядок $Y_1 = a_0 + a_1 x$

2-й порядок $Y_2 = a_0 + a_1 x + a_2 x^2$

3-й порядок $Y_3 = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ …..

2) Селективный вариант ИМСОМ - МГУА
В процессе перебора оставляют лишь наиболее значимые элементы модели, параметр селекции p задается. Пусть $p = 2$. Тогда имеем:

0-й порядок $Y_0 = a_0$,

1-й порядок $Y_1 = a_0 + a_1 x$

2-й порядок $Y_2 = a_0 + a_2 x^2$ $Y_2 = a_1 x + a_2 x^2$

3-й порядок $Y_3 = a_0 + a_3 x^3$ $Y_3 = a_1 x + a_3 x^3$ $Y_3 = a_2 x^2 + a_3 x^3$

Индуктивное моделирование

Пример применения селекции

Пусть имеем 20 точек наблюдений = 10 (обучение) + 10 (контроль)
Необходимо восстановить полиномиальную модель $F(t) = a_0 + \sum a_i t^i$
Пусть число параметров = 2, тогда шум подавляется в $\sqrt{(10/2)} \sim 2$ раза

1) Используем КОМБИ
   Тогда, очевидно, что максимально-допустимый порядок модели 1
   Мы сможем рассмотреть только $F_1(t) = a_0$ и $F_1(t) = a_0 + a_1 t$

2) Используем МГУА с селекцией 2-х параметрических моделей
   Тогда на каждом шаге отбираем модели с 2 параметрами
   В результате можем «добраться», например, до $F_5(t) = a_3 t^3 + a_5 t^5$

Почему называется МГУА = Метод Группового Учета Аргументов?

На каждом уровне сложности модели происходит селекция наиболее перспективных сочетаний аргументов (признаков)

Содержание

Постановка задачи

Предмет рассмотрения

Статистический стеммер.

Построение эмпирической формулы, обученной на примерах

Техника

Индуктивное моделирование

Постановка задачи

Стемминг

Состоит в выборе части слова,
отражающей основное значение слова

Примеры

sad, *sadly*, *sadness*, *saddened*, *saddened*
move, moving, moved, [*moveable* <= ?]

Применение

Индексация (параметризация) текстов,
где используются частотные списки слов

Постановка задачи

Проблема

Построить формулу для принятия решения
о подобии пары слов

Актуальность

Нам приходится обрабатывать многоязыковые корпуса и документов. Реальность: 25 официальных языков в Европе

Ограничения подхода

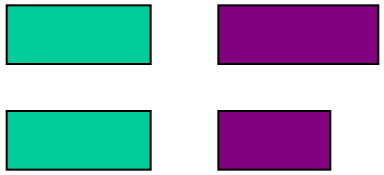
Только для флективных языков

Эмпирические формулы

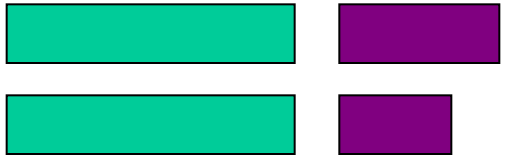
Параметры для сравнения пары слов

Мы будем обучать формулу на положительных примерах, то есть на парах подобных слов

1) ask***ing***
   ask***ed***
   y **= 3** ***n*** **= 5** ***s*** **= 11**



2) translat***ion***
   translat***ed***
   y **= 8** ***n*** **= 5** ***s*** **= 21**



Здесь:

y - длина общей части пары слов (***y*** - yes)
n **-** длина финальных частей (***n*** - no)
s - общая длина пары слов (***s*** - sum)
n/s - относительная доля несовпавших букв

Эмпирические формулы

Требования

Построенная формула должна отражать два обстоятельства:

1) Поддержать факт, что небольшое относительное число несовпавших букв ***n/s*** есть индикатор подобия слов
   translat*ion* **translat*ed***

2) Провести дискриминацию длинных слов. А именно:
   чем слова длинее, тем менее вероятно, что они подобны при том же отношении ***n/s*** ratio.

Лингвисты полностью поддержали эти два требования

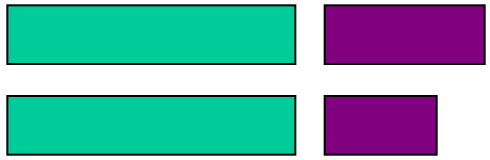
Эмпирические формулы

Модели для принятия решений

Какую формулу стоит настраивать
под примеры, заданные экспертом?

$n/s < C$
$n/s < F(s)$
$n/s < F(y)$
$n/s < F(y/s)$
etc.

translation
translated
$y = 8 \quad n = 5$
$s = 21$



Здесь:

C - константа
***F* (.)** - некоторая функция

Эмпирические формулы

Число степеней свободы

Как было указано выше, формула должна зависеть от:

- относительной доли числа несовпавших букв, то есть ***n/s***
- длины слов, то есть ***s*** или ***y***

Это определяет необходимость иметь 2 степени свободы

Рассмотрим:

n/s < ***C***	1 степень свободы ***n/s***
n/s < ***F*** (***s***)	2 степени свободы ***n/s*** и ***s***
n/s < ***F***(***y***)	2 степени свободы ***n/s*** и ***y***
n/s < ***F*** (***y/s***)	1 степень свободы, ***n/s***
etc.	

Коментарий: $y = (s-n)/2$, тогда $y/s = (s-n)/2s = 0.5\ (1- n/s)$

Поэтому: $n/s < F\ (y/s)$ имеет 1 степень свободы

Эмпирические формулы

Дискриминация длинных слов

yyynn *yyynn* ***n/s*** = 0.4

yyyyyynnnn *yyyyyynnnn* ***n/s*** = 0.4

<= Пусть они подобны

<= Меньшая вероятность, что они подобны

Объяснение

Финальная флективная часть в среднем имеет имеет ту же самую длину независимо от начальной основной части. Действительно:

-ing, -ly, -ingly, -al, -able, -ed, ...

те же самые как для длинных, так и для коротких слов

Эмпирические формулы

Таким образом, наше решение

$n/s < F(y)$, $F(y) = a_0 + a_1 y + a_2 y^2 + a_3 y^3 + \ldots a_k y^k + \ldots$

y **–** длина начальной общей части двух слов

n **–** общая длина их финальных несовпадающих частей

s – общая длина двух слов

Сложность модели $F(y)$ **?**

Чтобы определить сложность модели (степень полинома), мы используем ИМСОМ = Индуктивный Метод Самоопределения Модели

ИМСОМ позволяет построить подходящую модель при ограниченном наборе экспериментальных данных

ИМСОМ - Реализация

Подход

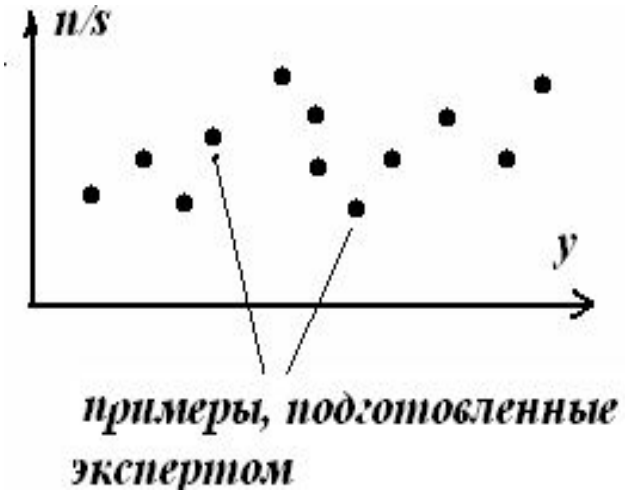
1) Мы рассматриваем

экстремальные случаи (равенство)

$n/s = a_0 + a_1 y + a_2 y^2 + \ldots$

2) Эксперт готовит «вручную»

пары подобных слов



Примеры

ask*ing* **ask*ed***

translat*ion* **translat*ed***

Перед началом эксперимента весь набор исходных данных (пары подобных слов) делится на обучающую и контрольную выборки

ИМСОМ - Реализация

Пример

$n/s = a_0 + a_1 y + a_2 y^2 + \ldots$

asking asked

$n = 5 \quad s = 11 \quad y = 3$

Spanish

N	Short words	N	Long words
1.	Celo	Celo*samente*	6.	Arrogan*te*	Arrogan*cia*
2.	Cazar	Caza*dor*	7.	Institucional	Institucional*mente*
3.	Arte	Artis*tico*	8.	Multiplica*dos*	Multiplica*ciones*
4.	Com*er*	Com*ida*	9.	Descentraliza*dos*	Descentraliza*bles*
5.	Alt*ura*	Alt*itud*	10	Caracteri*stica*	Caracteri*zaremos*

0-порядок

5/11 = a_0

Линейная

5/11 = $a_0 + a_1 3$

Квадратичная

5/11 = $a_0 + a_1 \mathbf{3} + a_2 9$

и т.д.

Решение

Для решения системы линейных уравнений мы используем МНК –
метод наименьших квадратов

Эксперимент

Внешние критерии

Регулярность K_r

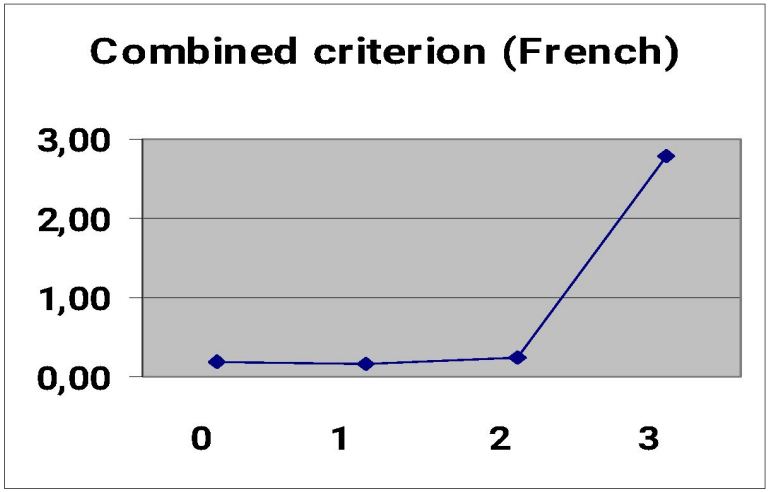
Несмещенность K_u

Комбинированный критерий

$K = 2/3\ K_r + 1/3\ K_u$

Динамика критерия *K*

Порядок	0	1	2	3
French	0.19	***0.15***	0.25	2.78
Italian	0.24	***0.17***	0.19	0.29
Portugal	0.25	***0.20***	0.22	2.30
Spanish	0.20	***0.16***	0.16	0.21



Результаты

Формулы (линейные модели)

French $n/s \leq 0.48 - 0.024\ y$

Italian $n/s \leq 0.57 - 0.035\ y$

Portugal $n/s \leq 0.53 - 0.029\ y$

Spanish $n/s \leq 0.55 - 0.029\ y$

Common $n/s \leq 0.53 - 0.029\ y$

Таким образом, общая формула
может быть записана так: $n/s \leq a - by$

Точность

Лемматизация 100%,~100%
Стемминг (Porter) > 90%
Эмпирическая формула ~ 80%-90%

Демонстрация

Пример

Начальный список	*1-й шаг*	*2-й шаг*	*3-й шаг*
transform (7)	transform (12)	transform (19)	transform (19)
transformed(5)	transformation(7)	translating (6)	translat (13)
transformation(7)	translating (6)	translator (7)	transport (11)
translating (6)	translator (7)	transport (11)	transported (2)
translator (7)	transport (11)	transported (2)	
transport (11)	transported (2)		
transported (2)			

Здесь:
Скобки содержат число появлений слов в тексте.
Значения сумируются, когда слова рассматриваются, как подобные

Дискуссия и выводы

Примеры Мартина Портера

Д-р Портер, автор знаменитого стеммера, реализованного на многих европейских языках, прислал письмо с примерами

1a. *bead, bear, beat* (короткий текст)
1b. *cat, cats* (короткий текст)

2a. ***bead,*** ..., *beagle, beagling,* ..., ***bear,***..., *beast, beastliness,* ..., ***beat*** (длинный текст)
2b. ***cat,*** *catalogue*, *cataplasm*, *catastrophe*, *catenary*, ***cats*** (длинный текст)

Примеры показывают влияние длины текста на работу стеммера:
- на коротких текстах различные по смыслу слова могут оказаться рядом и быть принятыми за сходные (см. 1a)
- на длинных текстах сходные по смыслу слова могут не оказаться рядом, и сходство не будет обнаружено (см. 2b)

Дискуссия и выводы

Недостатки стеммера

- Относительно низкая точность (80%-90%)
- Зависимость результатов применения от длины текста

Преимущества стеммера

- Языково независим – легко настраивается на язык и тему
- Простая настройка на заданный уровень ошибок 1 и 2 рода

Замечания

а) зависимость результатов от длины текста легко устраняется, если проверка сходства выполняется по всему списку, а не только для рядом стоящих слов

б) ошибки 1-го рода (пропуск сходных по смыслу слов) легко обнаруживаются и исправляются при ручном контроле результата

Содержание

Введение

Предмет рассмотрения

Subjectivity/Sentiment analysis.
Построение эмпирических формул для автоматической оценки вежливости, удовлетворенности и компетентности
на основе диалогов

Техника

Индуктивное моделирование

Введение

Subjectivity/Sentiment анализ это область обработки естественных языков (NLP), которая направлена на

автоматическую оценку

эмоций и мнений людей по отношению к некоторому объекту или событию. Эта тема стала популярной в эпоху Web 2.0

Это может быть интересно для таких бизнесс приложений, как оценка:

1. Удовлетворенности покупателей товарами и услугами
2. Мнений, относящихся к продуктам или событиям
3. Компетенции автора, использующего некий товар и услугу

Введение

Одно из приложений SSA это обработка диалогов.
В настоящее время такая обработка широко используется для построения автоматических диалоговых систем и для улучшения качества обслуживания.

В данной работе мы рассмотрим построение эмпирической формулы для оценки **вежливости** и приведем формулы для оценки **удовлетворенности** и **компетенции**

Будут использованы:

(i) объективные лексико-грамматические индикаторы, связанные с этими характеристиками

(ii) субъективные экспертные оценки диалогов

Исходные данные

Данные состояли из 85 диалогов между пассажирами и справочной железнодорожного вокзала Барселоны

Язык – испанский и каталанский

Распределение экспериментального материала

Обучение: 30 диалогов

Контроль: 30 диалогов

Экзамен: 25 диалогов

Пример: Вежливость

US: Good evening, *Could you* tell me the schedule of trains to Zaragoza for tomorrow?	DI: I will see, one moment. The next train leaves at 5-30
DI: For tomorrow morning?	US: 5-30
US: Yes	DI: hmm, hmm < SIMULTANEOUSLY >
DI: There is one train at 7-30 and another at 8-30	US: Well, and how much time does it take to arrive?
US: And later?	DI: 3 hours and a half
DI: At 10-30	US: For all of them?
US: And till the noon?	DI: Yes
DI: At 12	US: Well, *could you* tell me the price?
US: *Could you* tell me the schedule till 4 p.m. more or less?	DI: 3800 pesetas for a seat in the second class
DI: At 1-00 and at 3-30	US: Well, and what about a return ticket?
US: 1-00 and 3-30	DI: The return ticket has a 20% of discount
DI: hmm, hmm <SIMULTANEOUSLY>	US: Well, so, it is a little bit more than 6 thousands, no?
US: And the next one?	DI: Yes
	US: Well, *thank you very much*
	DI: Don't mention it, good bye

Вежливость, параметризация

А. Индиктор первого приветствия **(**g - greeting**)**

Имеет значение 1, при первом приветствии

«Could you please inform me ...»

и значение 0, если нетт:

«I need the information about ...»

Б. Вежливые слова (w-words**):**

«please», «thank you», «excuse me»

В. Вежливые грамматические формы (v-verbs**):**

глаголы в состагательном наклонении, то есть

«could», «would»,...

Вежливость, параметризация

Для числового представления ***v*** и ***w*** следует учесть

- длину документа
- информационный аспект появления слов

Поэтому вводятся:

нормировка на число фраз и подавление высоких частот

Вежливые слова: $w = Ln\,(1 + Nw/L)$,

Вежливые формы: $v = Ln\,(1 + Nv/L)$,

где ***Nw, Nv*** число вежливых слов и грамматических форм соответственно, и ***L*** число фраз.

Вежливость, параметризация

Ручные оценки учитывают только вежливость (но не грубость) по шкале:

0 - обычная вежливость

0.5 - повышенная вежливость

1 - чрезмерная вежливость

Примечание: опытные эксперты использовали шаг 0.25

Вежливость, параметризация

US: Good evening, *Could you* tell me the schedule of trains to Zaragoza for tomorrow? DI: For tomorrow morning? US: Yes DI: There is one train at 7-30 and another at 8-30 US: And later? DI: At 10-30 US: And till the noon? DI: At 12 US: *Could you* tell me the schedule till 4 p.m. more or less? DI: At 1-00 and at 3-30 US: 1-00 and 3-30 DI: hmm, hmm <SIMULTANEOUSLY> US: And the next one?	DI: I will see, one moment. The next train leaves at 5-30 US: 5-30 DI: hmm, hmm < SIMULTANEOUSLY > US: Well, and how much time does it take to arrive? DI: 3 hours and a half US: For all of them? DI: Yes US: Well, *could you* tell me the price? DI: 3800 pesetas for a seat in the second class US: Well, and what about a return ticket? DI: The return ticket has a 20% of discount US: Well, so, it is a little bit more than 6 thousands, no? DI: Yes US: Well, *thank you very much* DI: Don't mention it, good bye

First greeting ***g***	Number of polite words ***Nw***	Number of polite grammar forms ***Nv***	Indicator ***G***	Indicator ***W***	Indicator ***V***	Estimation
Yes	1	3	1	0.07	0.21	0.75

Вежливость, модели

Мы предположили, что зависимость между числовыми индикаторами и и уровнем вежливости может быть описана полиномиальной моделью.

Серия моделей увеличивающейся сложности:

Model 0: $F(\boldsymbol{g},\boldsymbol{w},\boldsymbol{v}) = A_0$

Model 1: $F(\boldsymbol{g},\boldsymbol{w},\boldsymbol{v}) = C_0\boldsymbol{g} + B_{10}\boldsymbol{w} + B_{01}\boldsymbol{v}$

Model 2: $F(\boldsymbol{g},\boldsymbol{w},\boldsymbol{v}) = C_0\boldsymbol{g} + B_{10}\boldsymbol{w} + B_{01}\boldsymbol{v} + B_{11}\boldsymbol{vw}$

Model 3: $F(\boldsymbol{g},\boldsymbol{w},\boldsymbol{v}) = C_0\boldsymbol{g} + B_{10}\boldsymbol{w}^2 + B_{01}\boldsymbol{v}^2$

Model 4: $F(\boldsymbol{g},\boldsymbol{w},\boldsymbol{v}) = C_0\boldsymbol{g} + B_{11}\boldsymbol{vw} + B_{20}\boldsymbol{w}^2 + B_{02}\boldsymbol{v}^2$

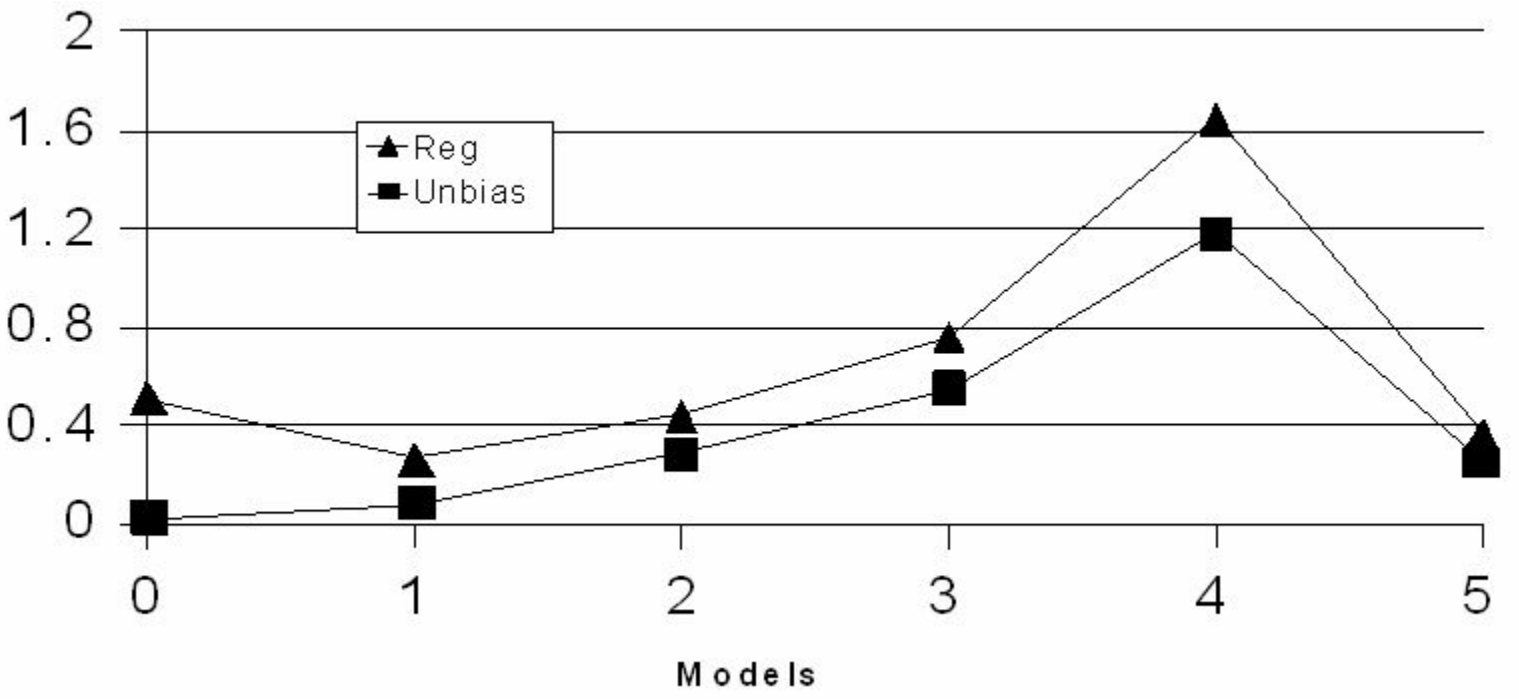
Model 5: $F(\boldsymbol{g},\boldsymbol{w},\boldsymbol{v}) = C_0\boldsymbol{g} + B_{10}\boldsymbol{w} + B_{01}\boldsymbol{v} + B_{11}\boldsymbol{vw} + B_{20}\boldsymbol{w}^2 + B_{02}\boldsymbol{v}^2$

Вежливость, модели

Подготовка данных для МНК

G	*W*	*V*	W^2	*WV*	V^2	**Manual estimation**
1	0.134	0.194	0.0178	0.0259	0.0377	1
0	0.111	0.057	0.0124	0.0064	0.0033	0.75
1	0.000	0.074	0.0000	0.0000	0.0055	0.25
1	0.000	0.031	0.0000	0.0000	0.0009	0
1	0.000	0.118	0.0000	0.0000	0.0139	0.75
1	0.043	0.043	0.0018	0.0018	0.0018	0.5
1	0.000	0.000	0.0000	0.0000	0.0000	0.25
1	0.043	0.083	0.0018	0.0035	0.0070	0.5
0	0.000	0.074	0.0000	0.0000	0.0055	0
1	0.134	0.069	0.0178	0.0092	0.0048	1

Вежливость, результаты



Наилучшая модель (по двум критериям)

$F(g,w,v) = 0.18g + 3.29w + 3.43v$ $\varepsilon = 0.16$

Удовлетворенность, параметризация

Серия моделей

Model 0: $F(b,f,q) = A_0$

Model 1: $F(b,f,q) = B_{100}b + B_{010}f + B_{001}q$

Model 2: $F(b,f,q) = B_{100}b + B_{010}f + B_{001}q + B_{110}bf + B_{101}bq + B_{011}fq$

Model 3: $F(b,f,q) = B_{200}b^2 + B_{020}f^2 + B_{002}q^2$

Model 4: $F(b,f,q) = B_{110}bf + B_{101}bq + B_{011}fq + B_{200}b^2 + B_{020}f^2 + B_{002}q^2$

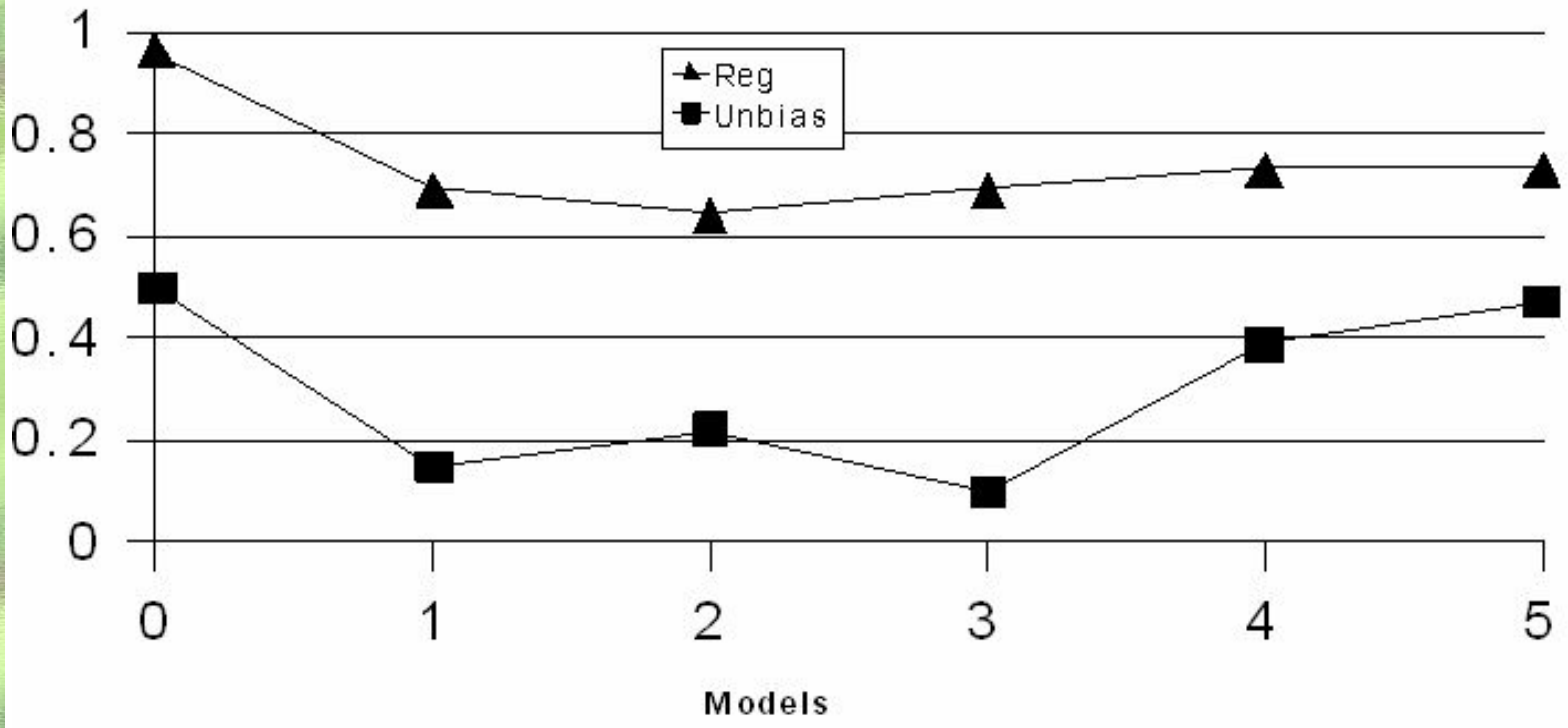
Model 5: $F(b,f,q) = B_{100}b + B_{010}f + B_{001}q + B_{110}bf + B_{101}bq + B_{011}fq + B_{200}b^2 + B_{020}f^2 + B_{002}q^2$

где:

b and ***f*** – положительная обратная связь с пользователем, в теле диалога (*'well', 'ok', 'all right', 'correct', 'splendid',* etc) и конце

q – вопрос-ответ, что отражает неудовлетворенность

Удовлетворенность, результаты



Наилучшие модели (по двум критериям)

$F(b,f,q) = 0.18b + 0.06f - 1.11q$ $\varepsilon = 0.35$

$F(b,f,q) = 0.20b^2 + 0.006f^2 - 1.78q^2$ $\varepsilon = 0.38$

Компетентность,параметризация

Серия моделей

Model 0: $F(b,f,q) = A_0$

Model 1: $F(b,f,q) = B_{100}l + B_{010}f + B_{001}q$

Model 2: $F(b,f,q) = B_{100}l + B_{010}f + B_{001}q + B_{110}lf + B_{101}lq + B_{011}fq$

Model 3: $F(b,f,q) = B_{200}l^2 + B_{020}f^2 + B_{002}q^2$

и т.д., как в предыдущем случае

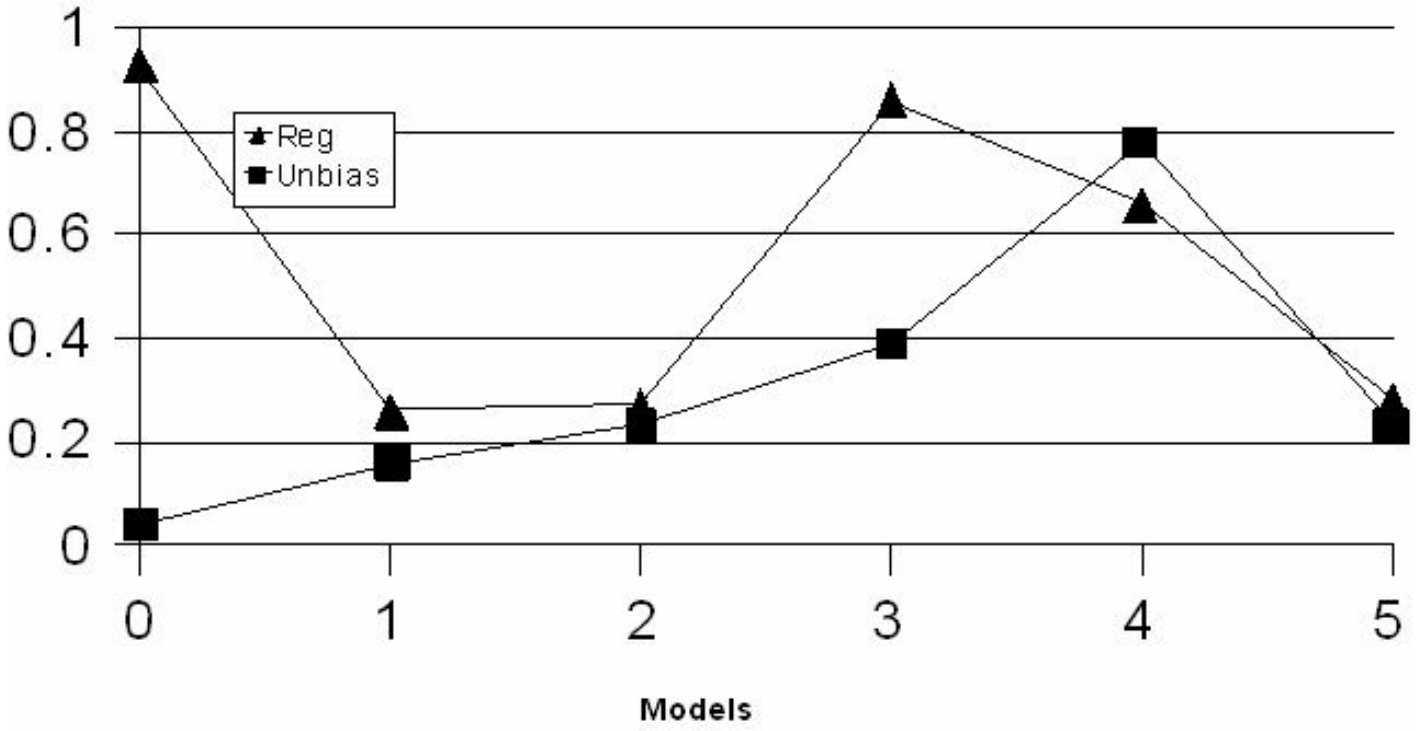
где:

f – уровень компетентности в первом вопросе пассажира ('any train to…?‘ vs. 'regional express at night to…?', etc.)

l –использованная специализированная лексика (номер поезда,..)

q – вопрос-ответ, который относится к вопросам пассажира и может отражать положительный и отрицательный ответ

Компетентность, результаты



Наилучшая модель (по двум критериям)

$F(f,l,q) = 0.52f + 0.19l + 0.16q$ $\varepsilon = 0.26$

Выводы

1. ИМСОМ обеспечивает методологию для автоматической оценки различных «размытых» характеристик диалога, имеющих высокий уовень субъективности

2. Построенные формулы правильно отражают вклад выбранных факторов в оцениваемую характеристику. Ошибки сравнимы с шагом ручной оценки диалога

Содержание

Введение

Предмет рассмотрения

Терминография.
Выявление гранулярности терминов заданной предметной области

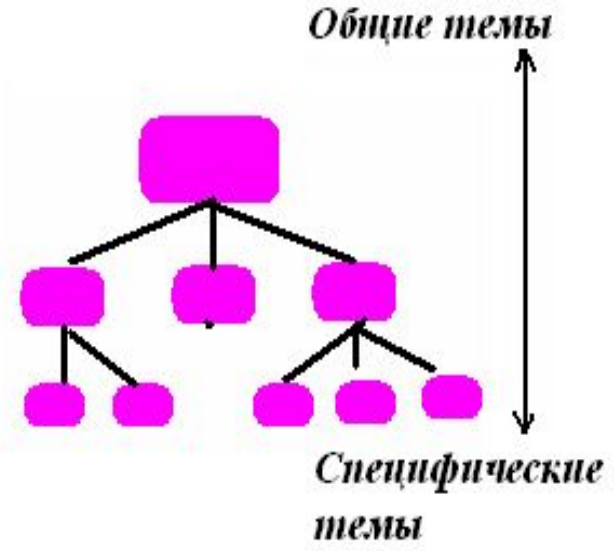
Техника

Идеи индуктивного моделирования

Введение

Под терминами будем понимать ключевые слова предметной области

Ключевые слова используются для:
А. **Суммаризации** документов
Б. **Кластеризация** документов
В. Построение **онтологии**

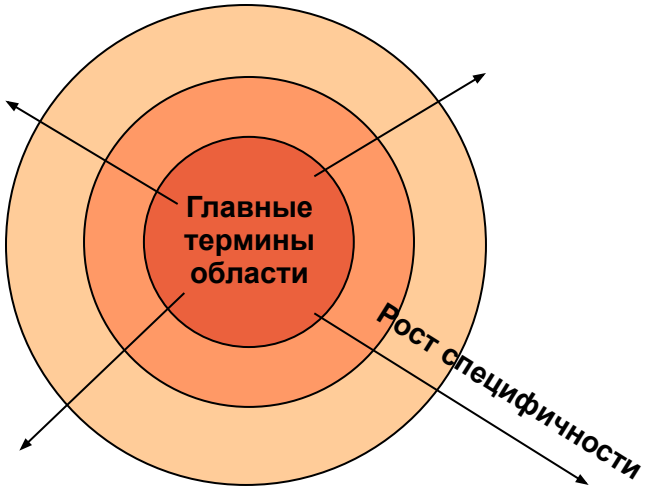


Мы полагаем, что
- корпус документов отражает несколько тем имеющих ***различную степень*** общности
- имеются ***слова*** ответственные за каждый уровень

Введение

Мы можем назвать общие термины предметной области как
coarse-grained terms
и очень специализированные термины как
fine-grained terms

Проблема: собрать вместе термины, связанные с одним уровнем гранулярности



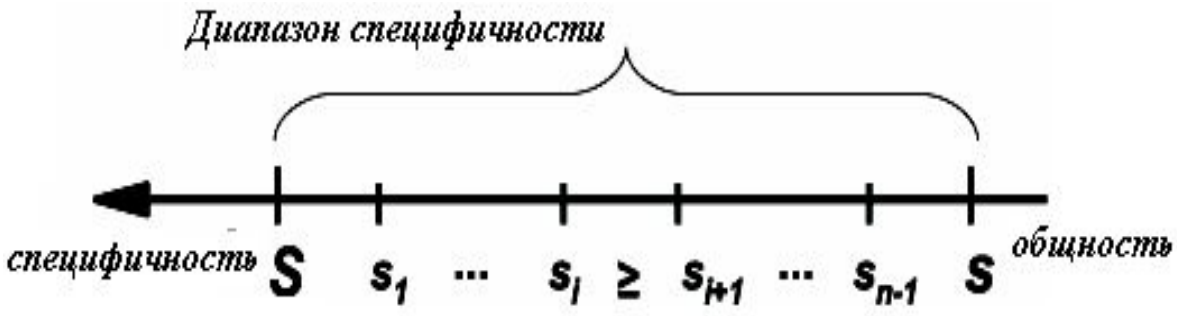
Применения гранулированных терминов:

1. Суммаризация документов по различным ***уровням детальности***
2. Кластеризация документов на различных ***уровнях детальности***
3. Построение онтологий для различных ***уровней детальности***

Определения

Определение гранулярности на основе корпуса текстов (corpus-based granularity):

Уровни гранулярности это классы терминов, имеющих близкие значения специфичности



Пусть $s_0, s_1, \ldots, s_n$ расщепление шкалы специфичности на n сегментов, так что $s_i \geq s_{i+1}$, $s_0 = S$, $s_n = s$.

Точки s_i называются **точками перехода** так как они находятся на границах между примыкающими уровнями гранулярности

Определения

Определение проблемы:

Выявление уровней гранулярности эквивалентно проблеме размещения точек перехода на оси специфичности.

Проблема гранулярности может быть разделена на две подпроблемы:

1. **Аппроксимация специфичности терминов предметной области посредством некоторой схемы взвешивания**
   – основанной на энтропии
   – основанной на стандартной девиации

2. **Размещение точек перехода на оси специфичности**
   Мы будем использовать **идеи** индуктивного моделирования

Аппроксимация специфичности

Пусть $D=(d_1, ..., d_N)$ коллекция документов и $X=(x_1, ..., x_N)$ частота слова w в документе d_i. Частоты нормированы на длину текста

1. Общность/гранулярность основанная на энтропии:

$$H(X) = -\sum_{i=1}^{N} p_i log(p_i)$$

где

$$p_i = Pr(d_i|X) = \frac{x_i}{\sum_{i=1}^{N} x_i}, \quad i = 1, ..., N$$

2. Общность/гранулярность, основанная на станд. девиации:

$$S(X) = \frac{m}{\sigma}$$

где

$$m = E(X) = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \sigma = \sqrt{E((X-m)^2)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2}$$

Описание ИМСОМ

Мы используем **ИМСОМ-подобный** алгоритм

Напомним основные шаги **ИМСОМ:**

1. Эксперт определяет ***последовательность моделей*** от простейшей к более сложным

2. Экспериментальные данные делятся на ***обучающую*** и ***контрольную*** выборки

3. Для заданного вида модели определяются наилучшие параметры на обучающей выборке с помощью некоторого ***внутреннего критерия*** (например МНК) **<= сейчас шаг исключен**

4. Полученная модель проверяется на контрольной выборке на основе некоторого ***внешнего критерия***. Глобальный минимум внешнего критерия определяет оптимальную модель

Метод

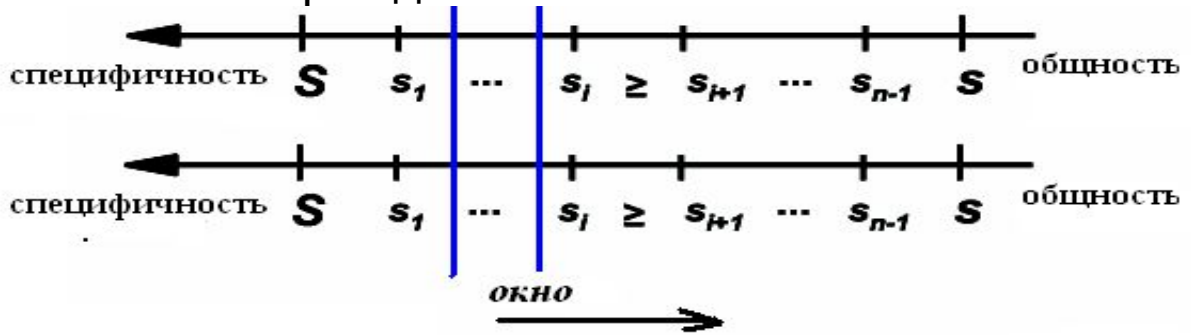
Первое, мы делим документы на **два набора.**
Оба они равноценны и называем их **Набор-1** и **Набор-2,**
а не обучающий и и контрольный, как в ИМСОМ
Затем мы упорядочиваем все слова согласно их **специфичности**:

Затем мы вычисляем расстояние между двумя наборами, в рамках **скользящего окна**

Метод

Шаги алгоритма

1. Фиксируем длину окна специфичности, прикладываем его к началу диапазона специфичности и берем термины внутри этого окна для обоих наборов данных



2. Вычисляем расстояние (**внешний критерий**) между распределениями специфичности обоих наборов данных.

В наших экспериментах мы используем относительную энтропию для специфичности, основанной на энтропии, и евклидово расстояние для специфичности, основанной на девиации.

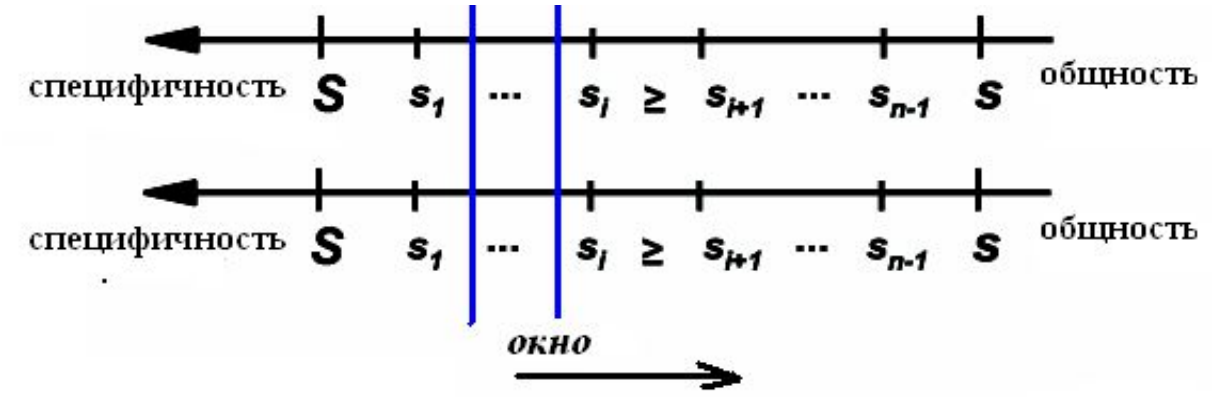
3. Перемещаем окно специфичности и переходим к Шагу 2.

Метод

Главная гипотеза:

Если распределения специфичностей внутри некоторого окна специфичности для обоих наборов данных **близки**,
то термины, содержащиеся в этом окне, принадлежат одному и тому же уровню гранулярности.

Окно, где два распределения достигают максимального расхождения, содержит точки неустойчивости, т.е. точки перехода от одного уровня гранулярности к другому.



Метод

Внешние критерии

Давайте зафиксируем одно и тоже окно ***Δs* = [*s1,s2*]** внутри диапазонов специфичности для каждого набора данных и давайте обозначим $p_t=p_t(\Delta s)$ и $p_c=p_c(\Delta s)$ распределения специфичности терминов, покрываемых этим окном для обоих наборов данных соответственно.

1. Относительная энтропия (или расстояние Кульбака-Лейбла):

$$K_1(\Delta s) = d(p_t, p_c) = \sum_{W_c(\Delta s)} p_t log\frac{p_t}{p_c} + \sum_{W_t(\Delta s)} p_c log\frac{p_c}{p_t},$$

2. Нормализованная версия Евклидова расстояния:

$$K_2(\Delta s) = d(p_t, p_c) = \sum_{W_c(\Delta s)} \frac{\sqrt{(p_t - p_c)^2}}{p_c} + \sum_{W_t(\Delta s)} \frac{\sqrt{(p_t - p_c)^2}}{p_t}$$

Характеристики корпуса

Мы используем корпус, названный hep-ex, изначально принадлежащий CERN -у.

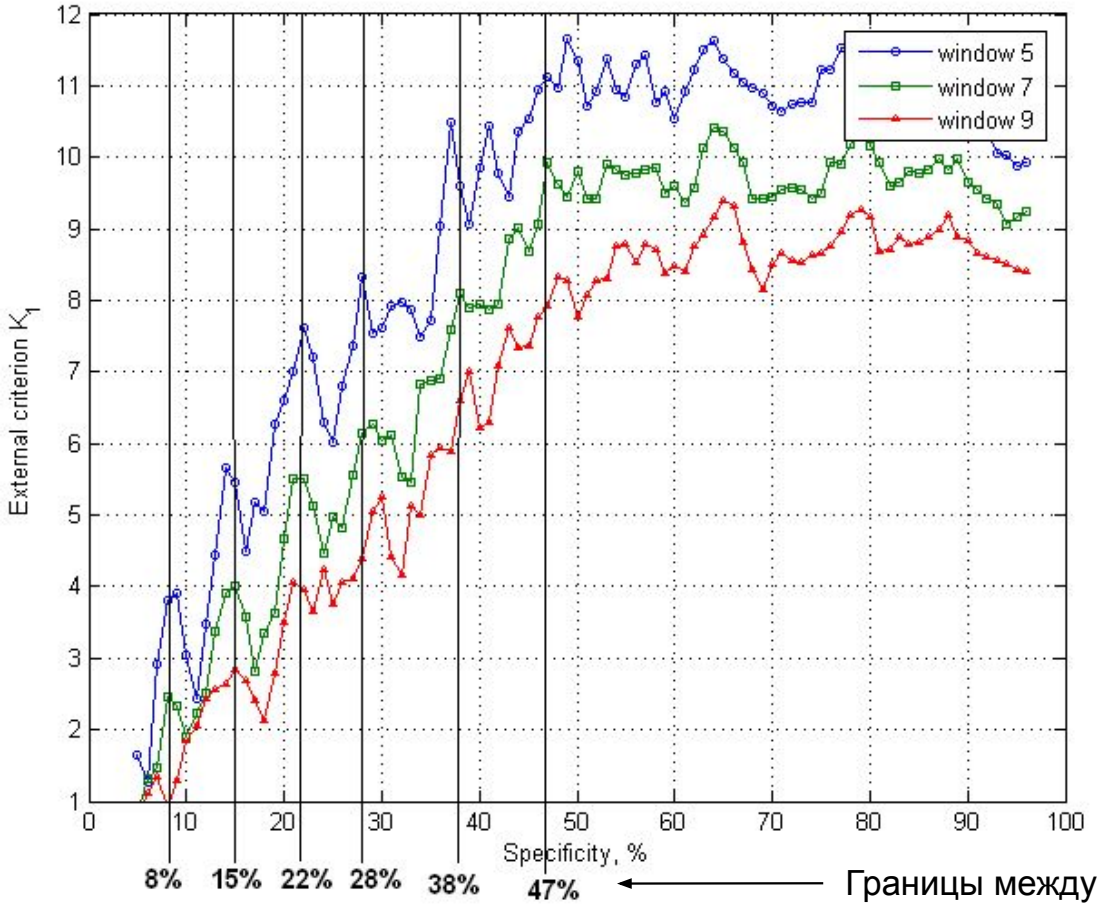
Он состоит из абстрактов статей, связанных с различными направлениями физики.

Техника препроцессинга: удаление стоп-слов и стемминг

Size of the corpus(byte)	962,802
Number of abstracts	2,922
Total number of terms	135,969
Vocabulary size	6,150
Term average per abstract	46.53

Результаты экспериментов (крит. K_1)

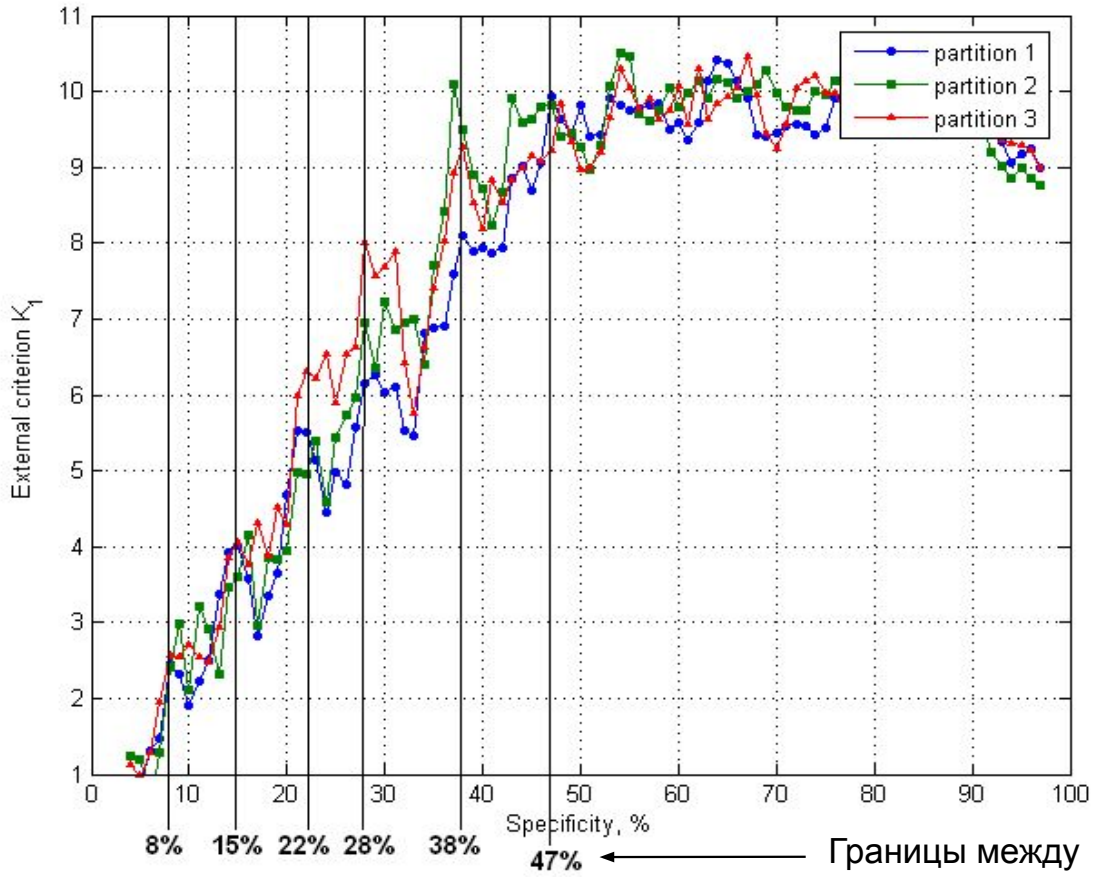
Поведение критерия K1 (основанный на энтропии) для различной длины окна



Границы между уровнями гранулярности

Результаты экспериментов (крит. K_1)

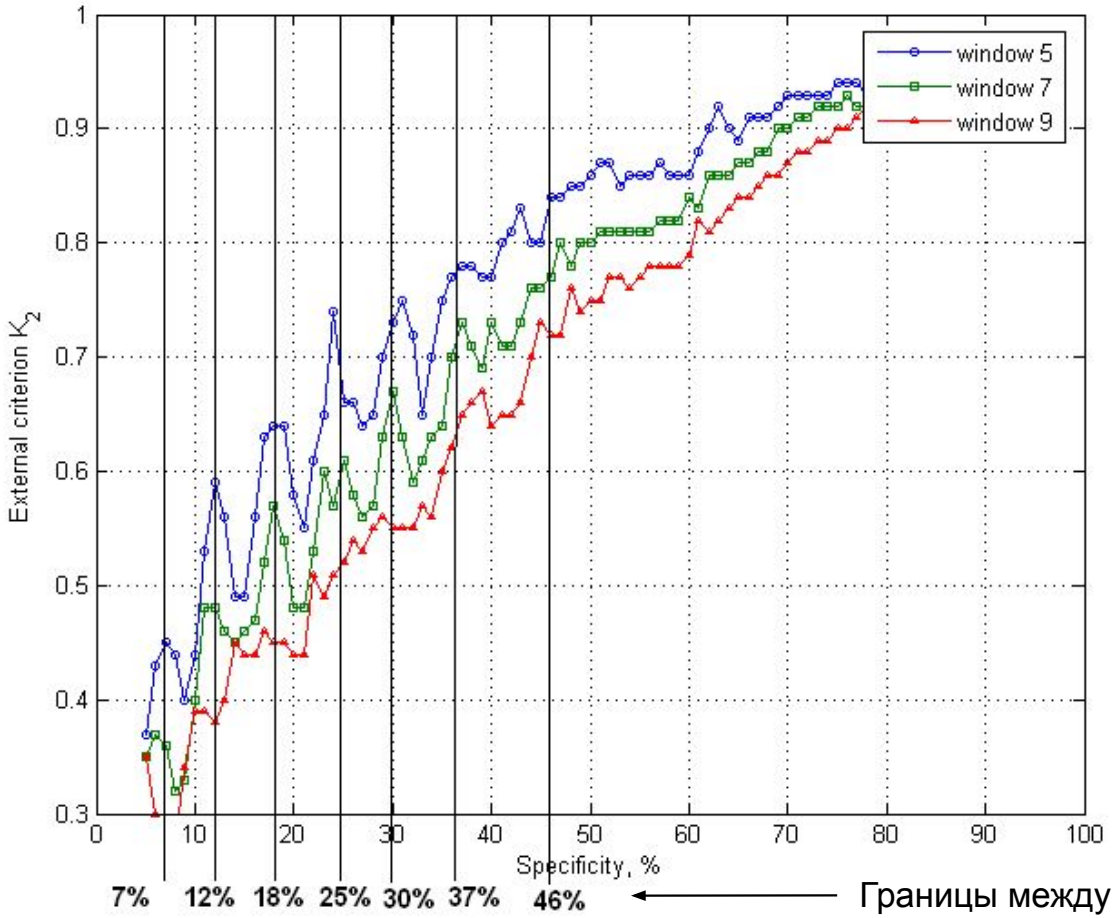
Поведение критерия K1 (основанного на энтропии) для различных разбиений корпуса



Границы между уровнями гранулярности

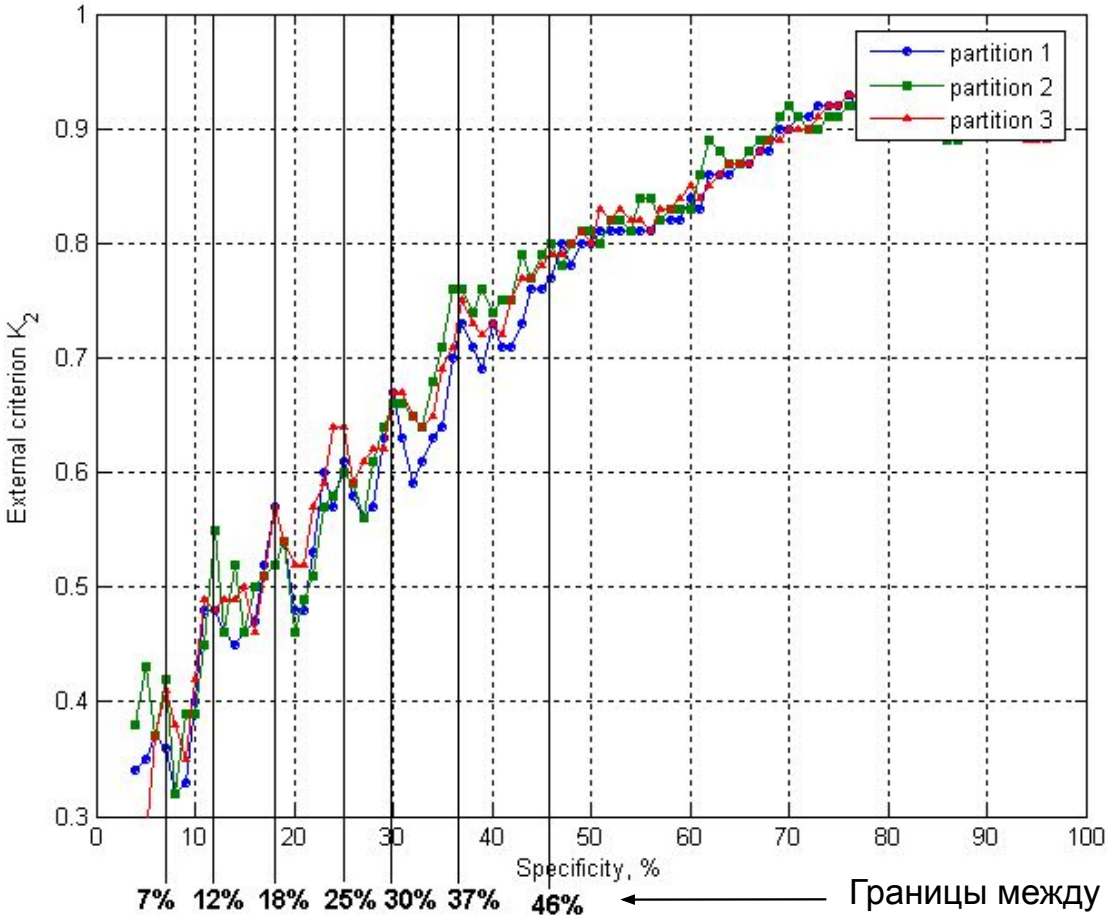
Результаты экспериментов (крит. K_2)

Поведение критерия K2 (основанного на девиации) для различных длин окна



Результаты экспериментов (крит. K_2)

Поведение критерия K2 (основанного на девиации) для различных разбиений корпуса



Примеры списков терминов

Слова в Таблице упорядочены согласно их специфичности

Entropy-based	Standard deviation-based
1. measur	1. us
2. us	2. measur
3. result	3. result
4. detector	4. present
5. data	5. detector
6. present	6. data
...	...
21. hadron	24. hadron
33. particl	43. particl
49. physic	47. physic
63. quark	71. electron
66. photon	78. photon
67. electron	81. quark
...	...

Выводы

1) Мы постарались формализовать ***понятие гранулярности*** для терминологии предметной области. Для этого мы ввели различные меры ***специфичности*** терминов и оределили класс гранулярности как множество терминов с близкой специфичностью

2) Был предложен ИМСОМ подобный алгоритм для выявления ***границ уровней гранулярностей***

3) ***Приблизительно 10%-15%*** слов были расположены не на своих местах. Таким образом, метод подходит для экспресс обработки, но должен быть улучшен для получения более точных результатов

Содержание

Ресурсы - Украина

(1) Международный центр информационных технологий и систем, НАН и МОН Украины, отдел информационных технологий индуктивного моделирования

http:// www.mgua.irtc.org.ua/ru/index.php?page=index

http:// www.gmdh.net/index.html

Основные направления научных исследований:

- теория ИМ сложных процессов по данным наблюдений
- создание интеллектуальных информ. технологий и инструментов моделирования и прогнозирования сложных процессов;
- решение прикладных задач моделирования и оптимизации экономических, экологических и технологических процессов

Заведующий отделом профессор, д.т.н. В.С. Степашко

Адрес: пр.Глушкова 40, Киев, 03680, Украина

Ресурсы - Украина

Поддержка сообщества ИМ

Отдел проф. В.С. Степашко организует:

1) Ежегодные Летние Школы по ИМ и смежным вопросам
   г. Жукин (Киев.обл.), июль, база ФМШ НАН Украины
2) Ежегодные Международные Конференции и Workshops по ИМ
   (чередуются конференции и workshops), Украина, Чехия, Польша

В текущем году:

Евпатория, май 2010 *http:// icim2010.felk.cvut.cz*

Ресурсы - Украина

(2) Компания Geos Research Group, Киев, Украина

http:// www.gmdhshell.com

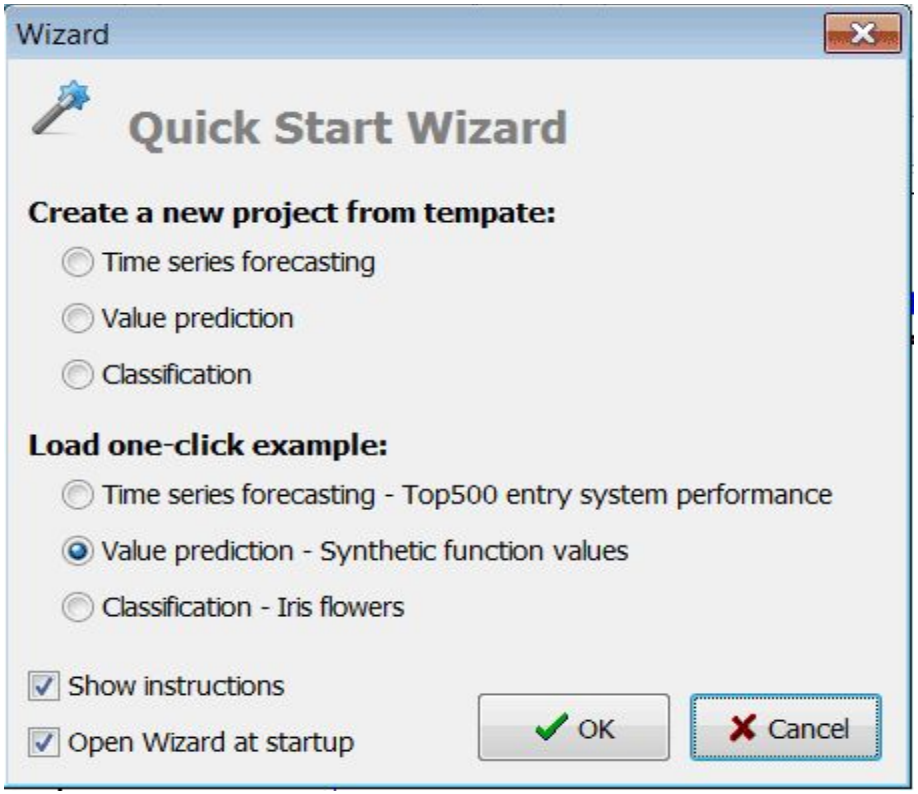
Компания разработала промышленную оболочку GMDH Shell, которая реализует ИМСОМ для решения задач Data Mining:

- прогноз временных рядов
- классификация
- визуализация результатов

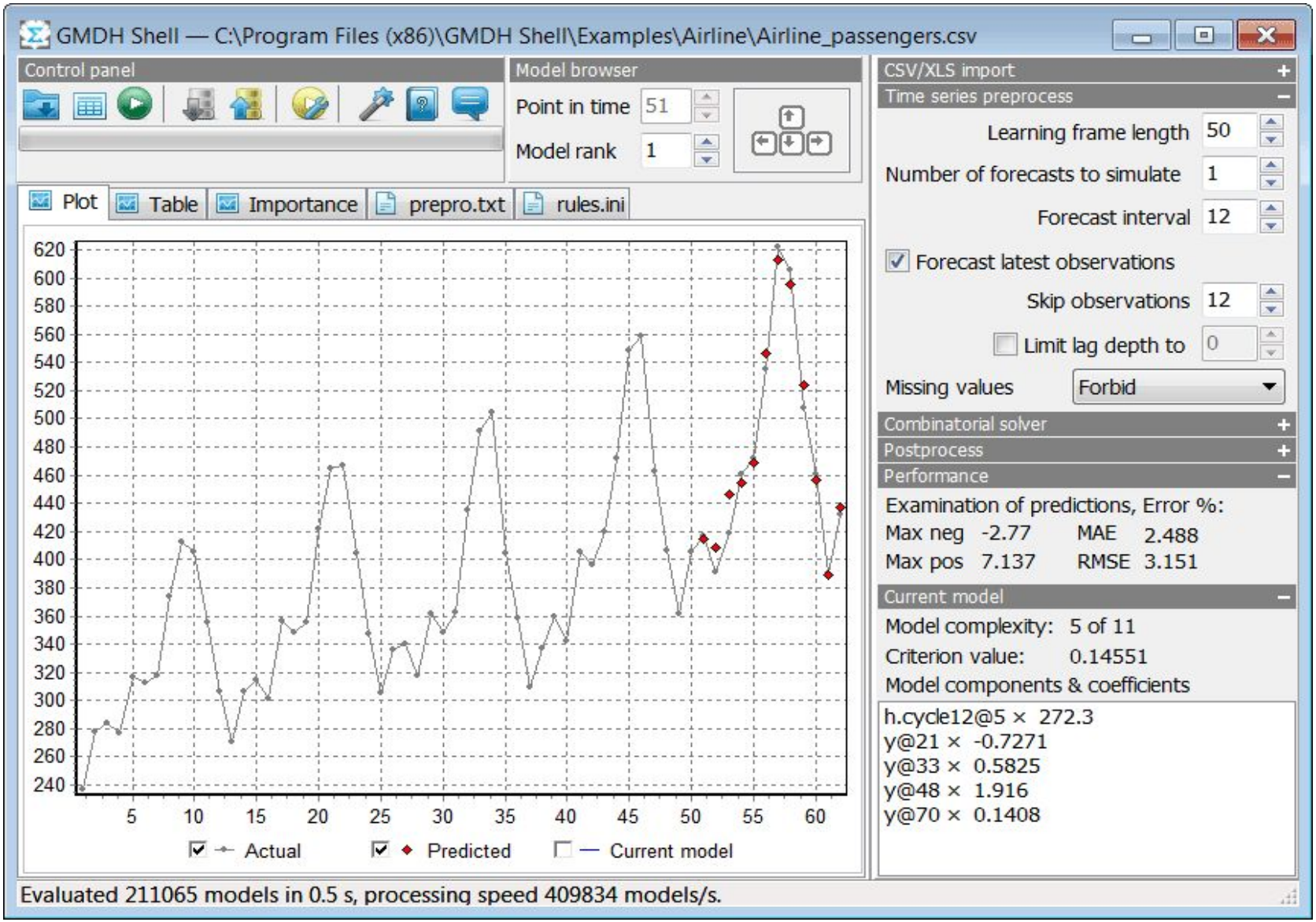
Используются:

- персональные ЭВМ
- кластерные системы

Для начинающих имеется Wizzard



Ресурсы - Украина



Пример работы оболочки GMDH-Shell

Ресурсы - Москва

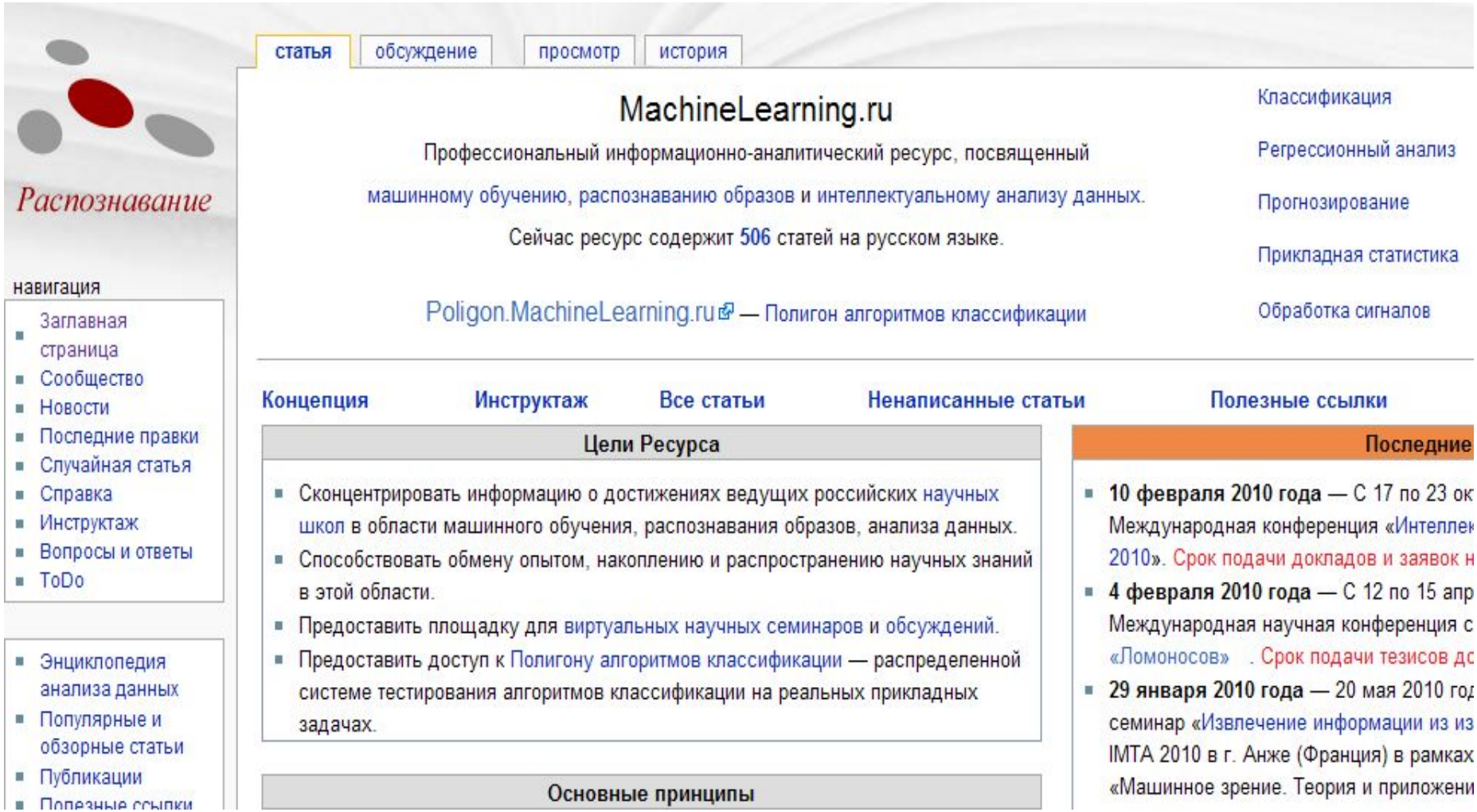
(1) Вычислительный Центр РАН
Московский физико-технический институт

http:// www.machinelearning.ru

Это *Wiki* подобный ресурс, связанный с Machine Learning
Содержит учебные и справочные материалы по ИМ: лекции, данные экспериментов, описание программ

Администратор ресурса д.ф.-м.н. К.В. Воронцов
Адрес: ул. Вавилова 40, Москва, 119333, Россия

Ресурсы - Москва



Часть главной страницы *Wiki* ресурса *http:// www.machinelearning.ru*

Ресурсы - Москва

(2) Компания Forecsys, Москва, Россия

http:// www.forecsys.ru/site/about/about/

Компания Forecsys — российский вендор BI-решений. Компания производит программное обеспечение и оказывает консалтинговые услуги в области анализа данных, прогнозирования, моделирования и оптимизации бизнес-процессов.

Одно из направлений: построение оптимальных регрессионных моделей

Подход: индуктивное порождение моделей (в т.ч. нелинейных)

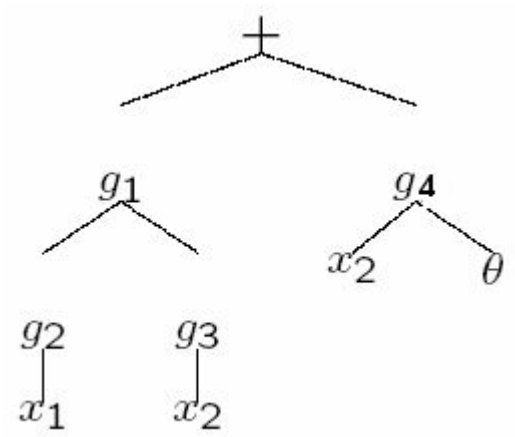
Техника: обучение/контроль, регуляризация

Реализация: программа MVR в форме открытого кода MatLab

Разработчик MVR: к.ф.-м.н. В.В. Стрижов

Ресурсы - Москва

Модель как произвольная суперпозиция



$f=g_1(g_2(x_1),g_3(x_2))+ g_4(x_2,\theta)$

Список порождающих функций

№	Функция	Описание	Параметры
Функции двух переменных аргументов, $g(\mathbf{b}, x_1, x_2)$			
1	plus	$y = x_1 + x_2$	–
2	times	$y = x_1 x_2$	–
3	divide	$y = x_1 / x_2$	–
Функции одного переменного аргумента, $g(\mathbf{b}, x_1)$			
4	multiply	$y = ax$	a
5	add	$y = x + a$	a
6	gaussian	$y = \frac{\lambda}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) + a$	λ, σ, ξ, a
7	linear	$y = ax + b$	a, b
8	parabolic	$y = ax^2 + bx + c$	a, b, c
9	cubic	$y = ax^3 + bx^2 + cx + d$	a, b, c, d
10	logsig	$y = \frac{\lambda}{1+\exp(-\sigma(x-\xi))} + a$	λ, σ, ξ, a

Процедуры генерации моделей, реализованные в MVR

Индуктивное моделирование: содержание и примеры применения в задачах обработки текстов

М. Александров

Академия народного хозяйства при Правительстве РФ
Автономный Университет Барселоны, Испания

MAlexandrov@ mail.ru

Петербург 2010