Estadística Predictiva para Horarios en el Transporte Urbano

Tomás J. Moreyra

FaMAF

Resumen En este trabajo presentamos una primera aproximación a la predicción de los horarios de llegada reales de transporte público. Hemos realizado diferentes experimentos basados en un pequeño conjunto de datos de una línea de colectivo urbano de la ciudad de Córdoba, aplicando técnicas de preprocesamiento de datos (selección de características, normalización, etc.). Hemos aplicado dos tipos de modelos predictivos: regresión lineal y clasificadores (árboles de decisión y máquinas de vectores de soporte).

Los resultados obtenidos indican que es factible predecir con fiabilidad signifcativa los horarios de llegada reales de los colectivos, y que una mayor cantidad de datos contribuiría a una sensible mejora en la precisión de los resultados.

1. Introducción

La espera del transporte público de pasajeros es un tema de interés para gran parte de la población. En particular en la ciudad de Córdoba, donde el funcionamiento del sistema de colectivos urbanos es objetivo de muchas críticas.

La incorporación de GPS en las unidades de transporte permite conocer con precisión y en tiempo real la posición del colectivo en determinados momentos. A partir de esto, una de las empresas que funciona en nuestra ciudad provee un servicio en el que, a través de un mensaje de texto o una consulta a una página web, se puede conocer el tiempo restante para que el colectivo de una determinada línea llegue a una parada. Actualmente sólo una empresa hace uso de esto, y por lo tanto, una de las principales motivaciones de este trabajo es encontrar un modelo que permita predecir el horario de los colectivos a partir de los horarios históricos y otros factores.

El resto del artículo se estructura como sigue. En la siguiente sección describimos los métodos y herramientas usados en este trabajo, y seguimos con una descripción de los datos de estudio. En la sección 4 presentamos una aproximación al problema usando regresión lineal. En la sección 5 presentamos la aproximación aplicando clasificadores, describiendo los efectos de la selección de características y de la escasez de datos. Finalmente, terminamos con algunas conclusiones y presentamos líneas de trabajo futuro.

2. Herramientas y Métodos

Todos los análisis se hicieron utilizando Weka [1], una plataforma de software que integra una colección de algoritmos útiles para aprendizaje automático y minería de datos.

Se usaron dos herramientas de preprocesamiento de datos. El algoritmo CfsSubsetEval[2] con búsqueda exhaustiva, el cual busca un subconjunto de atributos altamente correlacionados con la clasificación, pero no correlacionados entre sí. Se utilizó también el algoritmo InfoGainAttributeEval[10], que evalúa cada atributo midiendo la ganancia de información con respecto a la clase. El primer enfoque utilizado para analizar los datos fue el método de Regresión Lineal. Por otro lado se utilizaron tres métodos de clasificación. Uno de éstos fue J48, una implementación del algoritmo de generación de árboles de decisión C4.5[4]. Se utilizó también el algoritmo NaiveBayes[7], un clasificador probabilístico basado en el teorema de bayes. El tercer algoritmo de clasificación fue SMO[9], que utiliza Support Vector Machines. En este último caso se utilizaron 3 kernels: PolyKernel[3], Puk[13] y RBFKernel[8].

Para evaluar los resultados se utilizó el método de Cross-Validation[6] con 10 folds (exceptuando el análisis de influencia de la escasez de datos, donde el conjunto de pruebas se provee en un archivo). Se observaron en particular el porcentaje de instancias clasificadas correctamente, el estadístico Kappa[5] y la media armónica de la precisión y el recall. El coeficiente kappa es una medida estadística del nivel de acuerdo entre 2 clasificadores. La precisión para una clase C es el porcentaje de instancias bien clasificadas dentro de las clasificadas como C, mientras que el recall es la proporción de instancias de una clase que fueron clasificadas como tal. Se busca que ambos valores sean cercanos a 1 y en este sentido la media armónica es de utilidad, puesto que se acerca a 1 mientras ambos valores lo hagan, pero se aproxima a 0 cuando sólo alguno decrece.

3. El Conjunto de Datos

3.1. Datos Crudos

Los datos son registrados por la empresa Geosolution [12] y se obtuvieron como colaboración del área de Servicio de la empresa Coniferal SACIF[11]

Los datos obtenidos corresponden 2 paradas del recorrido de la linea C5, una cercana al inicio del recorrido, y la otra cercana al final. Los datos son de la siguiente forma:

```
6 ; 369T ; 187 ; 15:29 ; 15:28 ; -0.56 ; 16.07 ; 20.00 ; ;
```

- El primer campo es el subramal. Se descarta, porque todos nuestros datos son de la misma línea.
- El segundo indica el chofer y el turno de trabajo. En este caso, es el chofer 369 trabajando en el turno tarde. Los choferes están identificados con un número, y los turnos con las letras M, T o N, indicando mañana, tarde o noche.

- El tercer campo es el número de móvil, que identifica el vehículo.
- Los siguientes 2 campos son la hora estimada a la que tenía que llegar el colectivo (predefinida en la empresa), y la hora a la que efectivamente llegó.
- El sexto campo es el atraso (Calculado como $horareal - horaestimada$).
- Los siguientes dos campos se describen como frecuencia teórica y real.

3.2. Procesamiento de los Datos

De los datos descriptos anteriormente, se descartan además el subramal, el móvil, el atraso y las frecuencias. El número de móvil se consideró irrelevante al problema. El atraso se descarta puesto que determina directamente la hora real, y no es real que uno pueda contar con este dato antes de que el colectivo llegue.

El campo que indica el chofer y el turno se divide en 2 campos nominales distintos para cada una de estas características. Para facilitar los cálculos, las horas, que son cadenas de caracteres de la forma "12:34" se mapean a números que indican el minuto en el día. Es decír "12:34" $\rightarrow 12 * 60 + 34 = 754$. Por lo tanto, de ahora en más, todos los campos que hagan referencia a un horario, estarán expresados en minutos en el intervalo [0, 1439] (i.e. entre 00:00 y 23:59).

A estos campos se le agregan más características que se pueden calcular. Una de estas es el día. El cálculo del día es simple, dado que la frecuencia de los colectivos es muy distinta un día de semana y un domingo.

Otra característica que se agrega es el promedio hasta el momento de los valores absolutos de los atrasos en relación al día y al turno, es decir, a cada posible par (dia, turno) se le asocia el promedio de los errores absolutos calculados con las instancias hasta el momento que hayan tenido ese día y ese turno. En el archivo arff se lo trata como `acc_mae` (mean absolute error acumulado).

Se agrega también el atributo `anteriores` como la proporción de colectivos anteriores que llegaron en tiempo. Para esto, se toman los 6 colectivos anteriores. Esta característica busca reflejar la posibilidad de que los atrasos sean sistemáticos, por ejemplo, si los últimos 4 colectivos llegaron atrasados, es más probable que el siguiente también esté atrasado (Algo que, en principio, no sabemos si es cierto).

El atributo `parada` es un nominal que distingue de cuál de las 2 paradas viene el dato. Más adelante se hace un mejor análisis de cuán útil puede ser distinguir entre las paradas.

La hora real de llegada a la parada del colectivo anterior también se agrega como atributo. El sentido de éste se explica con la Regresión Lineal.

Finalmente se agrega la clase. Se decidió que la clase sea binaria, donde True va a indicar que el colectivo llegó en tiempo, y False, lo contrario. El significado de "llegar a tiempo" se define maś adelante.

Finalmente el conjunto de atributos usados se pueden ver en el Cuadro 1.

atributo	valores
turno	M, T, N
chofer	368,369,...,376
hora_estimada	real
dia	lun, mar,...,dom
acc_mae	real
anteriores	real
parada	p1, p2
hora_anterior	real
class	True, False

Cuadro 1. Atributos usados para describir los datos de frecuencia de colectivos.

4. Regresión lineal

4.1. Selección de Atributos

Tomando en cuenta la idea original de predecir el horario en base a los horarios históricos, se realizaron algunas pruebas sin tener en cuenta la hora estimada. La idea es en general utilizar la hora real como variable dependiente de la regresión, y por esto no tendría sentido usar los atributos class y atraso, ya que estos se calculan a partir de la hora real. Además, ya que se está proponiendo no usar la hora estimada, no tendría sentido usar los valores que son calculados a partir de ésta, es decir, `acc_mae` y `anteriores`. Una primera aproximación sería hacer una regresión lineal simple contra la hora en que pasó el colectivo anterior. Suponiendo, por ejemplo, un caso ideal en que los colectivos pasaran en todo momento exactamente cada 12 minutos, se obtendría el modelo

$$hora\_real = 1 * hora\_anterior + 12$$

Claramente, esto no es acorde a la realidad. El horario real se ajusta en base al día, el rango horario, etc. Los atributos seleccionados por CfsSubsetEval son `turno`, `chofer` y hora_anterior. A partir de algunas pruebas a mano con Weka, se prueba también el caso del conjunto de atributos `turno`, `chofer`, día, parada, hora_anterior. Además, para ver la incidencia del uso del atributo `chofer`, se hace otra prueba con los atributos `turno`, día y hora_anterior.

Haciendo uso de la hora estimada (y por lo tanto, también de las otras características que se habían sacado) la selección de CfsSubsetEval es `hora_estimada`, `anteriores` y hora_anterior. Se agregan también las pruebas anteriores, pero ahora con la hora estimada. Y con los mismos criterios que antes, se eligen otros conjuntos de prueba que se consideraron que podían ser convenientes.

4.2. Análisis de resultados

En los gráficos de este informe, los nombres de algunos atributos se abrevian de la siguiente forma: hora_anterior, como hant; hora_real, hreal; `hora_estimada`, hest; `acc_mae`, err y `anteriores`, ants.

En la figura 1, se muestran la correlación (×100), el error absoluto medio (MAE) y la raíz del error cuadrático medio (RMSE) para los conjuntos de atributos que no incluían la hora estimada, y los que sí. Notar que tienen una escala muy diferente. En el primero, el MAE es del orden de 50, mientras que en el segundo, a penas del orden de 2.

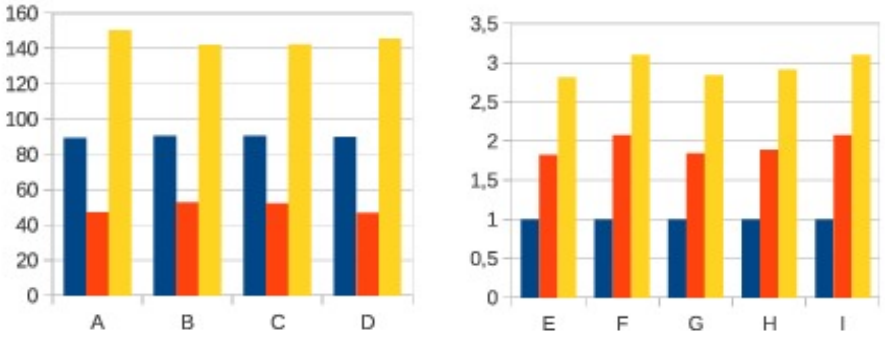

Figura 1. Correlación, MAE y RMSE. **A la izquierda**: Conjuntos de atributos que no incluyen `hora_estimada` A: hant; B: turno,chofer,dia,par,hant; C: turno,chofer,hant; D: turno,dia,hant. **A la derecha**:Conjuntos que sí contienen `hora_estimada` E: hest; F: turno,hest,dia,ants,hant G: turno,chofer,hest,dia,hant H: hest,ants,hant I: Todos los atributos

El origen de esta diferencia de puede notar en los graficos de la figura 2, donde se muestran las horas reales de llegada en función de la hora anterior, tal como están en los datos, en contraste con los modelos propuestos por la regresión lineal que toma como única variable independiente la hora anterior o la hora estimada, respectivamente. Los datos graficados corresponden a un solo día (un lunes) para que sea visible el error.

5. Clasificadores

5.1. Clases e Intervalos de Tolerancia

Para utilizar los clasificadores se deben definir las clases en las que se va a clasificar. Éstas van a estar definidas por intervalos de tolerancia en relación a la hora estimada de llegada del colectivo. Dada la tolerancia t y la hora estimada h, se dice que el colectivo llegó a tiempo si llegó en el intervalo $(h-t, h+t)$. Se llamarán clases True y False, si el colectivo llegó en horario o no, respectivamente.

Definir una tolerancia de 2 minutos genera un quiebre en la proporción de clases True y False. La figura 5.1 muestra el número de datos que pertenecen a cada clase si se toma una tolerancia de 2 o de 2.001 (que a fines prácticos, es

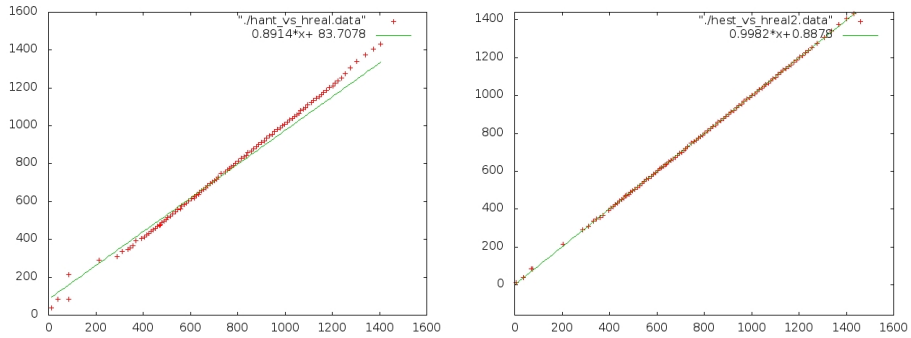

Figura 2. hora_anterior vs. hora_real; hora_estimada vs. hora_real

lo mismo). A la izquierda, tenemos que más del 60 % de los datos corresponden a la clase False, si se toma la tolerancia de 2; mientras que a la derecha, con tolerancia poco mayor a 2, sólo el 35 % de los datos quedan en esta clase.

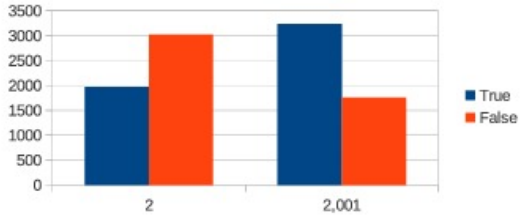

Figura 3. Comparación de intervalos de tolerancia.

5.2. Selección de atributos

Para esta clasificación, es necesario quitar 2 atributos: `hora_real` y `atraso`, ya que se busca es predecir si el colectivo va a llegar a horario y no tendría sentido pensar que contamos con la hora real de llegada o el atraso con el que va a llegar. El conjunto de atributos devuelto por CfsSubsetEval con búsqueda exhaustiva es `hora_estimada`, `acc_mae` y `hora_anterior`. Por otro lado, el ranking devuelto por InfoGainAttributeEval es el siguiente:

```
0.05283 hora_anterior
0.04821 hora_estimada
0.02656 acc_mae
0.01643 chofer
0.0098 anteriores
```

```
0.00974 dia
0.00402 turno
0.00269 parada
```

Teniendo en cuenta esto, se listan a continuación otros subconjuntos de atributos que, a partir de algunas pruebas con weka, se consideró que podían llevar a buenos resultados.

- CfsSubsetEval (`hora_estimada`, `acc_mae`, `hora_anterior`)
- CfsSubsetEval + `anteriores`.
- Todos los atributos
- Todos excepto el último (`parada`)
- Los 2 primeros (`hora_anterior`, `hora_estimada`)
- Los 4 primeros (hora_anterior, `hora_estimada`, `acc_mae`, `chofer`)
- `turno`, `chofer`, `hora_estimada`, `dia`, `acc_mae`, hora_anterior

5.3. Selección de la Clase

Los 2 intervalos de tolerancia mencionados producen 2 clasificaciones muy distintas para los clasificadores, pero iguales en la práctica. Con el fin de establecer si alguna es más conveniente, se corren todos los clasificadores propuestos con cada uno de los 7 conjuntos de atributos. Al comparar el porcentaje de instancias clasificadas correctamente, se observa un incremento de entre 4 % y 7 % con $t = 2,001$ respecto a $t = 2$

5.4. Análisis de resultados

En la figura 4 se muestra la performance de cada uno de los clasificadores usados con un subconjunto de características elegido y la clase con tolerancia 2.001. Las barras azules representan el porcentaje de datos clasificados correctamente y las barras rojas, el coeficiente Kappa (multiplicado por 100, a modo de que sea visible en el gráfico).

5.5. Performance J48 y SMO-PUK

Dado que las performances del J48 y del SMO con el kernel Puk se se distinguen por sobre las otras, los gráficos en la Figura 5 buscan comparar el funcionamiento de cada uno de los clasificadores en función de los distintos subconjuntos de atributos. Con las barras azules está la proporción de instancias clasificadas correctamente y en rojo el coeficiente kappa.

Los mejores resultados del J48 son el 70,3926 % de instancias bien clasificadas y kappa de 0,2822, mientras que los mejores del SMO-Puk están por debajo, con 69,1506 % y kappa 0,2506. Se probó variar el parámetro c para ver si era posible que este kernel superara el desempeño del J48, pero sólo se alcanzó un valor de kappa de 0,2619 con $c = 4$.

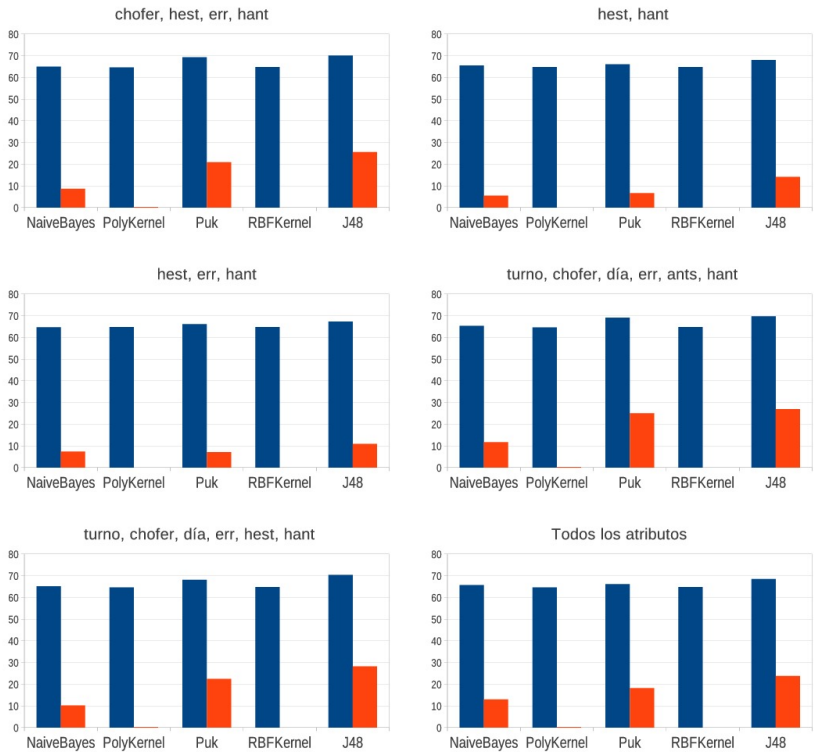

Figura 4. En cada grafico se muestra la performance de los 5 clasificadores elegidos con un determinado conjunto de atributos. Las barras azules representan el porcentaje de datos clasificados correctamente y las barras rojas, el coeficiente Kappa $\times 100$).

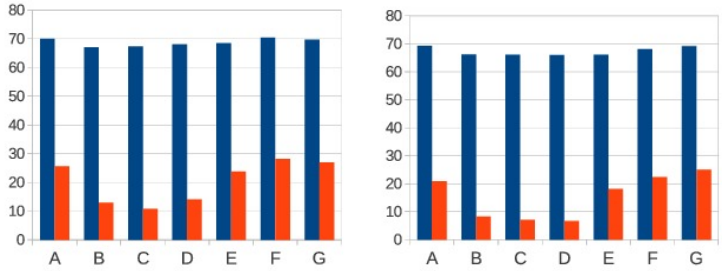

Figura 5. Performance J48 (izq.) y SMO con kernel PUK (der.) Los conjuntos de atributos son A: chofer,hest,err,hant; B: hest,err,ants,hant; C: herr,err,hant; D: hest,hant; E: todos los atributos; F: turno,chofer,hest,dia,err,hant; G: turno,chofer,hest,dia,err,ants,hant;

5.6. Influencia de la Escasez de Datos en los Resultados

Para tener una aproximación de cuánto podrían mejorar los resultados si se hubieran podido extraer más datos, se corren los clasificadores con los datos correspondientes a 1, 2, 3 y 4 semanas para analizar la mejora en los resultados. En la figura 6 a la izquierda se muestra en azul el porcentaje de clasificaciones correctas, y en rojo el kappa. A la derecha se tiene un grafico similar, pero teniendo en cuenta el factor temporal de la siguiente forma: Se construye el modelo con los datos de las primeras n semanas, y se utiliza como set de test la semana $n+1$ (para n en $1, 2, 3$). Se agrega en amarillo la media harmónica entre el promedio ponderado de la precisión y el recall de cada clase.

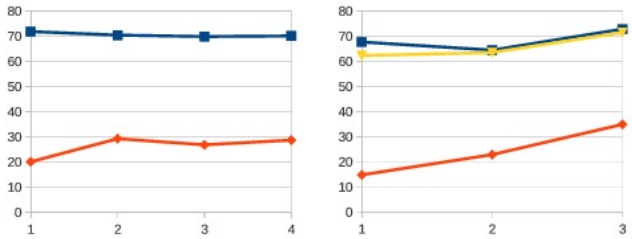

Figura 6. Curvas de performance para el J48

6. Análisis y Conclusiones

Respecto a la regresión lineal, se puede ver que la mejora es muy significativa tomando la hora estimada como variable independiente. Se puede decir entonces que la empresa hace una buena estimación. La pregunta es si se logró mejorar esta estimación usando otros parámetros. La mejor performance de la regresión se obtuvo utilizando todos los parámetros, con un error absoluto medio de 1,8253 y la raíz del error cuadrático medio de 2,8135. Calculando estos valores para la hora estimada contra la hora real, se obtienen 2,2169 y 3,1972. Por lo que se puede decir que hay una mejora en relación a la predicción que hace la empresa.

El clasificador con mejor desempeño fue el generador de árboles de decisión J48 con los atributos `turno`, `chofer`, `hora_estimada`, `dia`, `acc_mae` y `hora_anterior` con el 70,3926 % de instancias bien clasificadas y kappa de 0,2822. Una idea inicial era que el atributo `chofer` no sería significativo en los resultados, pero se puede ver que sucedió todo lo contrario. En los gráficos de performance del J48 y de SMO-Puk, el desempeño empeora visiblemente cuando no se utiliza este atributo.

Posiblemente el J48 arroje un mejor resultado porque captura mejor la topografía de la frontera de decisión, mientras que ninguno de los kernels usados proyecte hacia un espacio donde esta frontera sea lineal.

Respecto a los análisis para evaluar la mejora en función de la cantidad de datos, es posible que en el primer gráfico no se vea una mejora porque los intervalos de tiempo sean muy cortos y los datos no sean suficientes. Incluso, en algunos casos, el modelo pareciera funcionar mejor con menos datos, pero esto probablemente sea porque el modelo generado con menos datos se ajusta más a las características de un intervalo de tiempo.

6.1. Extensiones

Hubo algunos análisis posibles que se vieron limitados por la cantidad de datos. Una hipótesis es que el funcionamiento del transporte está afectado por el período del año, y esto podría verse con datos del transcurso de al menos uno o dos años. Además, con datos de períodos más largos de tiempo, permitiría analizar si algunas condiciones climáticas (lluvia, viento, temperatura, etc) pueden afectar los horarios.

Si se tuviera la posibilidad de conseguir datos de más paradas, se podrían interrelacionar, ya que si una unidad pasó o no en horario por paradas anteriores podría ser un factor importante para saber si el colectivo va a pasar en horario en una parada futura. Se podría también comparar la performance de modelos generados independientemente con los datos de cada parada y un modelo general que utilice un identificador de parada como atributo.

Referencias

1. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. *The WEKA Data Mining Software: An Update*. The university of Waikato, 2009.
2. Mark. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning Hamilton*. The university of Waikato, 1998.
3. Documentación de PolyKernel.
4. Wikipedia: C4.5 Algorithm.
5. Wikipeda: Cohen's Kappa.
6. Wikipedia: Cross-validation.
7. Wikipedia: Naive Bayes Classifier.
8. Wikipedia: Radial Basis Function Kernel".
9. Wikipedia: SMO.
10. Documentación de InfoGainAttributeEval.
11. Coniferal SACIF.
12. Geosolution: Soluciones para el transporte.
13. B. Ustun, W.J. Melssen, and L.M.C. Buydens. *Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel*. Elsevier, 2006.