Aplicación de métodos agregados en la detección de puntos atípicos en series de tiempo meteorológicas

Application of ensemble methods in outlier point detection in meteorological time series

Luis-Alexander Calvo-Valverde[1], Nelson José Acuña-Alpízar[2]

Fecha de recepción: 30 de mayo de 2017
Fecha de aprobación: 3 de setiembre de 2017

Calvo-Valverde, L; Acuña-Alpízar, N. Aplicación de métodos agregados en la detección de puntos atípicos en series de tiempo meteorológicas. *Tecnología en Marcha*. Vol. 31-1. Enero-Marzo 2018. Pág 98-109.

DOI: 10.18845/tm.v31i1.3500

1 Doctorado en Ciencias Naturales para el Desarrollo (DOCINADE), Instituto Tecnológico de Costa Rica, Maestría en Computación. Programa Multidisciplinar eScience. Costa Rica. Correo electrónico: lcalvo@itcr.ac.cr
2 Maestría en Computación. Instituto Tecnológico de Costa Rica. Costa Rica. Correo electrónico: neacuna@gmail.com

Palabra clave

Valores atípicos; Métodos agregados; ARIMA; Regresión de soporte vectorial; SVR; Red bayesiana; Apilamiento; Bagging; AdaBoost.

Resumen

Para este trabajo de investigación, se estudió el desempeño de los métodos agregados en la detección de valores atípicos punto en series temporales uni-variables meteorológicas, utilizando la métrica F1 como medida de desempeño. Para esto se creó un programa que permite aplicar 3 clasificadores no agregados (regresión de soporte vectorial, ARIMA, redes bayesianas) y 3 clasificadores agregados (apilamiento, bagging y AdaBoost) a 3 conjuntos de datos de mediciones meteorológicas (precipitación, temperatura máxima y radiación solar).

Usando esta aplicación, se ejecutó un diseño experimental para comparar los clasificadores. En este diseño, primero se obtuvo el promedio de F1 de los clasificadores realizando múltiples pruebas en cada conjunto de datos. Luego, mediante una prueba estadística de hipótesis se compararon los promedios obtenidos por los clasificadores para determinar si las diferencias observadas eran significativas. Finalmente, se realizó un análisis de los resultados, enfocado en comparar el desempeño de los clasificadores agregados contra el desempeño del mejor clasificador no agregado en cada conjunto de datos.

En general se encontró que es posible mejorar significativamente el desempeño al detectar valores atípicos punto en algunas series temporales uni-variables utilizando métodos agregados. Sin embargo, para lograr esta mejora se deben reunir condiciones que, aunque varían dependiendo del método agregado, en general apuntan a mejorar la diversidad de los clasificadores base. Cuando no se reúnen estas condiciones, los métodos agregados no tuvieron una diferencia significativa en el desempeño con respecto al algoritmo no agregado que obtuvo el mejor desempeño en el conjunto de datos.

Keywords

Outliers; Ensemble methods; ARIMA; Support vector regression; SVR; Bayesian network; Stacking; Bagging; AdaBoost.

Abstract

For this research work, the performance of ensemble methods in the task of outlier points detection in meteorological univariate time series was studied, using the F1 metric to measure the performance. For this purpose, an application was created that allows applying 3 non-ensemble classifiers (support vector regression, ARIMA, bayesian networks) and 3 ensemble classifiers (stacking, bagging and AdaBoost) to 3 meteorological datasets (rainfall, maximum temperature and solar radiation).

Using this application, an experiment was executed to compare the different classifiers. In this experiment, first, the F1 average of the algorithms was obtained by executing multiple tests in each dataset. Then, using a statistical hypothesis test we compared the obtained averages to find out if the observed differences were significant. Finally, a result analysis was performed, focused on comparing the performance of the ensemble classifiers versus the performance of the best non-ensemble classifier for each dataset.

In general the results indicate that it is possible to significantly improve the performance in the outlier point detection task in some uni-variate time series by using ensemble methods. However, to obtain this improvement several conditions must be met. Although the conditions vary depending on the ensemble method, in general these conditions aim to improve the diversity in the base classifiers. When these conditions were not met, the ensemble methods didn't have a significant difference in the performance compared to the non-ensemble classifier that got the best performance in the datasets.

Introducción

Los puntos atípicos son observaciones que aparentan desviarse marcadamente de otros miembros de la muestra en la que ocurren [1]. Este trabajo se enfoca en la detección de valores atípicos en series de tiempo meteorológicas. En esta área de aplicación los valores atípicos pueden representar eventos meteorológicos como tornados, huracanes o incendios forestales [2]. También los valores atípicos pueden deberse a errores en los sensores o en el registro de la información y deben ser limpiados de los conjuntos de datos; este proceso de limpieza de los datos puede llevar a resultados más precisos ya que los valores atípicos pueden afectar significativamente el desempeño de los algoritmos de minería de datos [3].

Aunque la detección de valores atípicos se puede llevar a cabo de forma manual, también es posible llevarlo a cabo de forma automática usando enfoques estadísticos y de aprendizaje de máquina, los cuales son más rápidos que los enfoques manuales y obtienen buenos resultados [4].

Si bien es importante poder detectar los valores atípicos en el área de la meteorología, también es difícil debido a que las variables medidas corresponden a procesos caóticos. Al ser el comportamiento de los fenómenos meteorológicos tan variable, es difícil determinar qué corresponde a un comportamiento normal y qué corresponde a un comportamiento atípico de los datos. Debido a esta dificultad, los altos volúmenes de los conjuntos de datos y a la alta importancia de detectar los puntos atípicos, el lograr incluso mejoras modestas en el desempeño sigue siendo un problema abierto de investigación.

En este trabajo se buscó mejorar el desempeño, medido por la métrica F1, en la detección de valores atípicos punto en series temporales meteorológicas uni-variables. Para esto se desarrollaron clasificadores que detectan puntos atípicos utilizando métodos agregados. Los métodos agregados son algoritmos de aprendizaje que construyen un conjunto de clasificadores base y clasifican datos haciendo una votación ponderada de sus predicciones. Como resultado, a menudo el clasificador agregado se desempeña mejor que cualquiera de los clasificadores individuales [5]. Las condiciones necesarias y suficientes para que un método agregado sea más exacto que cualquiera de los miembros individuales es que los clasificadores base sean exactos y diversos, donde exactos quiere decir que tienen una tasa de error mejor que adivinar al azar y donde diversos quiere decir que los métodos cometen errores distintos entre ellos [6]. Los detectores de puntos atípicos desarrollados utilizan los algoritmos de Support Vector Regression (SVR) [7], ARIMA [8] y redes bayesianas [8] para los clasificadores base, y los métodos agregados de apilamiento [9], bagging [10] y AdaBoost [11].

Este trabajo tiene relación con una investigación doctoral en el DOCINADE que tiene relación con la aplicación del aprendizaje máquina en la predicción de cultivos agrícolas, y uno de los aspectos que trabaja esta investigación es la detección de valores atípicos en las variables de entrada, las cuáles son mayoritariamente meteorológicas. La selección de los algoritmos SVR, ARIMA y Redes Bayesianas se debe al deseo de analizar si estos algoritmos, utilizados en la predicción en el mundo agrícola, podrían combinarse usando métodos agregados para mejorar su capacidad de predicción.

Este trabajo se organiza de la siguiente forma. Primero, se provee la metodología utilizada. A continuación se muestran los resultados obtenidos en las distintas etapas del experimento y se analizan estos resultados. Por último, se presentan las conclusiones de la investigación.

Metodología

Para el trabajo de investigación, se estudió el desempeño de los métodos agregados (medido con la métrica F1) en la detección de valores atípicos punto en series temporales uni-variables meteorológicas, utilizando la métrica F1 como la medida de desempeño.

Para los distintos pasos del experimento se extrajeron los datos de 3 conjuntos de datos con distintas mediciones meteorológicas: Precipitación [12], Temperatura máxima [13] y Exposición solar [14]. Para estos conjuntos, se definieron 3 subconjuntos (disjuntos entre ellos); los subconjuntos de entrenamiento, validación y pruebas. En el cuadro 1 se resume el número de entradas en los subconjuntos.

Cuadro 1. Número de entradas en los subconjuntos de datos.

	Entrenamiento	Validación	Pruebas
Precipitación diaria	6000	3000	12000 (40 conjuntos de 300 valores)
Temperatura máxima	3000	1500	6000 (40 conjuntos de 150 valores)
Exposición solar	1300	650	2600 (40 conjuntos de 65 valores)

En algunas etapas del experimento se requiere que el conjunto de datos tenga valores atípicos. Ya que los conjuntos de datos utilizados no incluyen valores atípicos, se creó una copia de los conjuntos a las que se les agregó valores atípicos etiquetados. Con este propósito, cada valor se sustituyó por un valor atípico con un 5% de probabilidad. El valor atípico se seleccionó usando un método similar al utilizado en [4], en el que se busca que los valores atípicos se generen de manera aleatoria por debajo de o encima de ciertos percentiles. Para este trabajo los percentiles utilizados y sus respectivos valores para los distintos conjuntos de datos se indican en el cuadro 2. El valor atípico es generado de forma aleatoria, con una distribución uniforme en los rangos dados en el cuadro 2. Rangos de los atípicos por conjunto de datos..

Cuadro 2. Rangos de los atípicos por conjunto de datos.

Conjunto de datos	Rango percentil inferior	Rango inferior	Rango percentil superior	Rango superior
Precipitación diaria	-	-	98 - 100	20 – 66
Temperatura máxima	0 - 1	4,4 – 10,4	99 - 100	37,9 – 46,4
Exposición solar	0 - 1	0,3 – 2,7	99 - 100	33,1 – 34,5

Implementaciones de los algoritmos no agregados

La implementación de la regresión de soporte vectorial se hizo como una envoltura (o "wrapper" en inglés) sobre la implementación de la biblioteca scikit-learn [7]. Para utilizar SVR para series de tiempo y clasificación, se definió una ventana de tiempo de *N* días que se mueve en incrementos de 1 día, de forma similar a [15]. Para cada posición de la ventana, las variables independientes corresponde a los *N* valores dentro de la ventana de tiempo y la variable dependiente es el valor inmediatamente posterior a la ventana de tiempo. Para clasificar cada valor en la serie de tiempo como atípico o no atípico, el SVR entrenado realiza una predicción. Si el valor es esperado v_e es muy distinto al valor observado v_t el valor es clasificado como atípico.

La implementación del algoritmo ARIMA se hizo como un envoltorio que llama a la implementación de ARIMA en el paquete de R llamado forecast [8]. En este caso, no se necesitaron transformaciones en las series de tiempo ya que la implementación de ARIMA utiliza series de tiempo directamente sin modificaciones importantes. Para seleccionar los parámetros de ARIMA *p,d,q* se utilizó una funcionalidad disponible en la biblioteca para la búsqueda de estos parámetros de forma automática.

El detector de puntos atípicos basado en redes bayesianas se implementó utilizando la biblioteca Libpgm [16] con distribuciones de probabilidad discretas. Las mediciones continuas son convertidas a valores discretos, tomando el rango de posibles valores y dividiéndolo entre 5 y 10 categorías (dependiendo del conjunto de datos). La red bayesiana implementada consiste de 3 nodos x_1, x_2, x_3 que apuntan a un único nodo *y*. Para el entrenamiento, para cada valor v_t en la serie de tiempo, se asigna a los nodos x_1, x_2, x_3 los valores $v_{(t-3)}$,$v_{(t-2)}$,$v_{(t-1)}$, y al nodo *y* se le asigna el valor actual v_t. Para la detección de puntos atípicos, para cada valor v_t en la serie de tiempo, se hace inferencia para obtener la probabilidad de observar el valor actual dados los 3 últimos valores. Si la probabilidad obtenida es menor a un umbral dado en los parámetros del clasificador, el valor observado es clasificado como atípico.

Implementación de los métodos agregados

El método de apilamiento implementado utiliza 2 o 3 clasificadores base (indicado en los parámetros) correspondientes a los algoritmos no agregados mencionados en la metodología (SVR, ARIMA y redes bayesianas). Además, para combinar los resultados de los clasificadores base, se usa Support Vector Classfication (SVC) como meta-clasificador.

El entrenamiento del método agregado de apilamiento se ejecuta en dos etapas. Primero, los 3 clasificadores base se entrenan usando una porción de los datos de entrenamiento sin valores atípicos. Luego, en la segunda etapa, los clasificadores base clasifican otra porción de los datos de entrenamiento pero con valores atípicos etiquetados. Las predicciones de los 3 algoritmos, los valores observados y las etiquetas correspondientes se utilizan para entrenar al meta-clasificador SVC. Para realizar una predicción primero se obtienen las predicciones de los clasificadores base, y a continuación las predicciones y el valor observado se usan como entrada para el metaclasificador. La predicción del meta-clasificador es una etiqueta que indica si el valor es atípico o no.

El método agregado de bagging implementado utiliza el algoritmo introducido en [10], mientras que el de AdaBoost utiliza el algoritmo descrito en [11]. En ambos casos se usó la implementación de SVR descrita en la metodología para los clasificadores base ya que en las pruebas preliminares este fue el que obtuvo el mejor desempeño.

Selección de parámetros

Para la implementación de los distintos algoritmos, es necesario elegir los parámetros a utilizar en cada conjunto de datos de forma que se obtenga un alto desempeño. Para esto se

implementó una búsqueda en rejilla (“grid search” en inglés), seleccionando los parámetros que dieron como resultado el mayor F1. Se utilizó la métrica F1 ya que provee un balance entre la precisión y la exhaustividad de los algoritmos.

Cabe señalar que los parámetros seleccionados para los clasificadores no agregados, se reutilizan en los clasificadores base dentro de los métodos agregados. Así, la búsqueda en rejilla para el método agregado solo se realiza sobre los parámetros del algoritmo de método agregado. Esto para reducir significativamente el tiempo de ejecución de la búsqueda en rejilla para los clasificadores agregados.

Comparación de los algoritmos

Una vez seleccionados los parámetros a utilizar para cada algoritmo, se comparó el desempeño de los distintos algoritmos, esto medido por medio de la métrica F1.

Para realizar esta comparación se utilizó el subconjunto de pruebas del conjunto de datos. Cada algoritmo se ejecutó sobre 40 segmentos de igual tamaño del subconjunto de datos de pruebas y se obtuvo la métrica F1 sobre cada uno de los resultados.

Debido a la no normalidad de los datos resultantes, para la comparación estadística de los resultados se optó por utilizar el método no paramétrico de Kruskal-Wallis [17]; este pone a prueba la hipótesis nula de que las distribuciones de las mediciones F1 obtenidas por los algoritmos son todas iguales. Como prueba post-hoc para identificar cuales distribuciones de los F1 son significativamente diferentes entre sí, se usó la prueba de Nemenyi en pares [18] con la distribución de Tukey. La prueba de Nemenyi se hizo en dos niveles. Una a nivel de cada conjunto de datos y otra a nivel general usando las mediciones de los 3 conjuntos de datos. Para ambas pruebas, Kruskal-Wallis y Nemenyi, se utilizó un nivel de significancia (alfa) de 0.05.

Resultados

Ejecución de los algoritmos sobre los subconjuntos de pruebas

En el cuadro 3 se muestran los promedios F1 obtenidos en la clasificación de los subconjuntos de pruebas por los métodos agregados y el mejor clasificador no agregado en cada conjunto de datos. En el cuadro 3 se resaltó en negrita el mejor resultado obtenido.

Cuadro 3. Promedios F1 obtenido por los métodos agregados y el mejor clasificador no agregado.

Promedio de F1	Precipitación	Temperatura Máxima	Exposición Solar
Mejor clasificador no agregado	0.807 (SVR)	0.764 (ARIMA)	0.719 (SVR)
Apilamiento	0.780	0.885	0.683
Bagging	0.811	0.752	0.714
AdaBoost	0.807	0.763	0.711

En el conjunto de datos de precipitación el mejor promedio de F1 lo obtuvo el clasificador de bagging, aunque por un margen de apenas 0.004 sobre el mejor clasificador no ensamblado SVR. En el conjunto de datos de temperatura máxima, el mejor promedio de F1 lo obtuvo

el clasificador de apilamiento, con una diferencias de 0.121 con respecto al clasificador no ensamblado con el mejor promedio de F1, ARIMA. En el conjunto de datos de exposición solar, el clasificador no ensamblado SVR logró el mejor promedio de F1, superando apenas por 0.005 el promedio de F1 del clasificador de bagging.

Prueba de hipótesis

Como se mencionó antes, la prueba Kruskal-Wallis pone a prueba la hipótesis de que las distribuciones de las mediciones de F1 obtenidas por los algoritmos son todas iguales. El valor-p obtenido fue de 2.2e-16 el cual es menor al nivel de significancia (alfa) de 0.05 usado en las pruebas de hipótesis, por lo que se rechazó esta hipótesis.

Luego se realizaron las pruebas post-hoc de Nemenyi para comparar pares de algoritmos, en el cuadro 4 se resumen los resultados de esta prueba mostrando la comparación de los F1 del mejor método no agregado contra el mejor método agregado. Estos resultados muestran que sólo se obtuvo una diferencia significativa en el F1 en el conjunto de temperatura máxima.

Cuadro 4. Resumen de resultados de la prueba de Nemenyi.

	Precipitación	Temperatura Máxima	Exposición Solar
F1 Mejor Clasificador agregado	0.811 (Bagging)	0.885 (Apilamiento)	0.714 (Bagging)
F1 Mejor Clasificador no agregado	0.807 (SVR)	0.764 (ARIMA)	0.719 (SVR)
Diferencia F1	0.004	0.121	-0.005
Valor-p Nemenyi	0.9996	0.0084	1.0000
Diferencia significativa?	No	Sí	No

Análisis de los resultados

Como se observó en la sección anterior, solo en uno de los conjuntos de datos se observaron diferencias significativas entre el mejor promedio de F1 de los clasificadores agregados contra el de los clasificadores no agregados. Para entender estos resultados, y por qué razón los clasificadores agregados no se desempeñaron significativamente mejor que los clasificadores no agregados en distintos conjuntos de datos, se evaluaron varias posibles razones que se presentan a continuación.

La primera posible razón que se investigó fue que no se cumplieran las condiciones necesarias y suficientes para que los métodos agregados sean mejores que cualquiera de los miembros individuales. Estas condiciones son que los clasificadores base sean exactos y diversos [6].

Se consideró poco probable que los clasificadores base no cumplieran con la condición de exactitud, ya que los clasificadores no agregados que usan los mismos algoritmos utilizados en los clasificadores base obtuvieron una exactitud mejor que el azar.

Para evaluar la segunda condición, la de la diversidad de los clasificadores base utilizados por los distintos clasificadores agregados, se realizaron pruebas adicionales con el fin de recolectar información que permitiera medir la diversidad de los clasificadores base. Estas pruebas consisten en la re-ejecución de los clasificadores agregados sobre los subconjuntos

de datos de pruebas, recolectando información adicional necesaria para calcular el promedio de coeficiente de concordancia kappa (κ) de Cohen [19] [20] entre los clasificadores base de cada clasificador agregado.

Los valores resumidos se muestran a continuación, en el cuadro 5. El coeficiente kappa (□) de Cohen es igual a 1 si los clasificadores coinciden en cada ejemplar, es igual a 0 si la concordancia es la misma que la esperada al azar y valores negativos si la concordancia es menor a la esperada al azar. A mayor concordancia, menor es la diversidad.

Cuadro 5. Resumen de valores Kappa por conjunto de datos y método agregado.

Promedio de coeficiente kappa	Precipitación	Temperatura Máxima	Exposición Solar
Apilamiento	0.846	0.460	0.277
Bagging	0.964	0.938	0.837
AdaBoost	0.349	0.627	0.251

Uno de los resultados observados en el cuadro 5 es el alto promedio del coeficiente Kappa del clasificador de bagging en los 3 conjuntos de datos. En cuanto a la razón de esta falta de diversidad, es posible que se deba al uso de SVR para los clasificadores base. En [21] se muestra que las máquinas de soporte vectorial son estables; pequeños cambios en el conjuntos de datos de entrenamiento no producen grandes cambios en el clasificador. Posiblemente por este motivo los cambios en los conjuntos de entrenamiento realizados como parte del algoritmo de bagging no dieron como resultado clasificadores diversos. Esta falta de diversidad en el clasificador de bagging explica por qué este método en general obtuvo resultados similares al clasificador no agregado de SVR.

En cuanto al clasificador de apilamiento, el resultado de la métrica kappa varía según el conjunto de datos. En el conjunto de datos de precipitación, obtuvo un kappa promedio de 0.846 entre los clasificadores base; es decir, estos son poco diversos entre sí. Esto muestra que utilizar algoritmos distintos en los clasificadores base (que se supone tienen diferentes sesgos entre sí) no garantiza una alta diversidad en los algoritmos. Por otro lado, en los conjunto de datos de temperatura máxima y exposición solar se obtuvieron promedios de kappa bajos (0.460 y 0.277) lo que indica que en estos casos los clasificadores base son diversos entre sí.

Anteriormente, en el cuadro 1, se mostró que para el conjunto de datos de precipitación, el clasificador de apilamiento obtuvo un promedio de F1 similar (aunque menor) a SVR (el clasificador no agregado con el mejor promedio F1 en este conjunto de datos). El apilamiento obtuvo un promedio de 0.780 mientras que SVR consiguió un promedio de 0.807. Este resultado puede ser explicado por dos factores. Un factor es el menor número de ejemplos utilizados en el entrenamiento de los clasificadores base, debido a que se debe utilizar una porción del conjunto de datos de entrenamiento para entrenar el meta-clasificador. Esto posiblemente da como resultado que los clasificadores base tengan métricas menores que los respectivos clasificadores no agregados. El otro factor es la falta de diversidad entre los clasificadores base del apilamiento utilizados en el conjunto de datos de precipitación. Esta falta de diversidad limita la mejora en las métricas que se pueden obtener por la agregación de los clasificadores base.

Con respecto al clasificador de apilamiento en el conjunto de datos de temperatura máxima, en el cuadro 1 se mostró que este método obtuvo el mejor promedio de F1, superando incluso

a ARIMA (el clasificador no agregado con el mejor promedio F1 en este conjunto de datos). El método de apilamiento obtuvo un promedio de F1 de 0.885 mientras que ARIMA obtuvo un promedio de 0.764. Además, en el cuadro 2 se mostró que la diferencia en este conjunto de datos había sido significativa. Si bien es cierto, en este caso también es posible que los clasificadores base se vieran perjudicados por un menor conjunto de entrenamiento, se observa una alta diversidad entre estos, ya que el promedio de kappa es de 0.460. Esta alta diversidad muy posiblemente influyó en el resultado favorable obtenido en este conjunto de datos.

En cuanto al clasificador de apilamiento en el conjunto de datos de exposición solar, en el cuadro 1 se mostró que este tuvo un promedio de F1 similar (aunque menor) a SVR (el clasificador no agregado con el mejor promedio F1 en este conjunto de datos). El método de apilamiento obtuvo un promedio de 0.683 mientras que SVR consiguió un promedio de 0.719. En este caso, se obtuvo un promedio kappa de 0.277, por lo que no hubo falta de diversidad que perjudicara a este método agregado. Para explicar el resultado anterior se examinaron las métricas obtenidas por los clasificadores base. En estos resultados se puede apreciar que los clasificadores base del apilamiento obtuvieron una métrica F1 muy baja (0.451 y 0.301). Esto sucedió posiblemente debido al menor conjunto de datos de entrenamiento para estos clasificadores, con un efecto más pronunciado ya que este conjunto de datos es el más pequeño de los 3 utilizados. A pesar de esto, el método agregado obtuvo un F1 alto (0.744) comparado con los clasificadores base. Estos resultados parecen indicar que el método agregado si mejoró significativamente la métrica F1 de los clasificadores base (en parte gracias a la alta diversidad entre estos). Sin embargo, las métricas F1 de los métodos base en este caso son muy bajas, como para lograr superar el clasificador no agregado de SVR.

Finalmente, con respecto a la diversidad del clasificador de AdaBoost, se obtuvo una diversidad alta en los conjuntos de datos de precipitación y exposición solar (con un promedio de kappa de 0.349 y 0.251 respectivamente), mientras que en el conjunto de datos de temperatura máxima se obtuvo una diversidad media (con un promedio de 0.627). Además, los clasificadores se entrenan con conjuntos de datos de similar tamaño a los utilizados por los clasificadores no agregados. Ni la diversidad obtenida en estos casos, ni el tamaño del conjunto de datos de entrenamiento explican por qué AdaBoost no logró tener una diferencia significativa en el F1 con respecto a los clasificadores no agregados.

Para entender por qué el clasificador AdaBoost no logró superar al clasificador no agregado de SVR, se examinó la ejecución del AdaBoost con el fin de buscar otros factores que pudieron afectar negativamente el desempeño del algoritmo. Para esto, se volvió a ejecutar el algoritmo sobre los subconjuntos de datos de pruebas, pero almacenando para cada iteración *k* del entrenamiento del AdaBoost el error (ϵ_k), y el peso del clasificador generado en la iteración, calculados como se indica en el algoritmo de AdaBoost M1. Los resultados se muestran en el cuadro 6.

En el cuadro 6, se pueden observar un par de resultados importantes. El primero, es que debido al bajo error en el primer clasificador generado, el peso asignado a este es mayor al peso de los demás clasificadores combinados. Por esta razón el AdaBoost implementado siempre vota igual que el primer clasificador generado, y por lo tanto equivale al clasificador no agregado de SVR. Estos resultados se explican por la alta exactitud del SVR utilizado para los clasificadores base del AdaBoost; el peso del primer clasificador base generado es muy alto, peso que es muy difícil que los siguientes clasificadores puedan lograr superar. Estos resultados sugieren que esta alta exactitud en el SVR utilizado para los clasificadores base del AdaBoost es contraproducente. Es posible que utilizar clasificadores base con una menor exactitud hubiera producido mejores resultados.

Cuadro 6. Error y peso obtenido en las iteraciones de AdaBoost.

Conjunto de datos	Valor	Iteración			
		1	2	3	4
Precipitación	Error	0.017	0.083		
	Peso	4.033	2.405		
Temperatura máxima	Error	0.018	0.440	0.491	0.499
	Peso	3.990	0.243	0.038	0.003
Exposición solar	Error	0.020	0.284	0.493	
	Peso	3.870	0.924	0.028	

Conclusiones

La detección automática de puntos atípicos es útil en distintas aplicaciones, como la detección de eventos importantes o limpieza de conjuntos de datos, y permite que los puntos atípicos sean detectados de forma rápida y eficaz. En este trabajo se desarrollaron métodos agregados para la detección de valores atípicos punto en series de tiempo meteorológicas de forma automática y se comparó su desempeño contra los algoritmos no agregados. En general, se encontró que es posible mejorar significativamente el desempeño al detectar valores atípicos punto en algunas series temporales uni-variables utilizando métodos de aprendizaje máquina agregados. Sin embargo, para lograr esta mejora se deben reunir condiciones que, aunque varían dependiendo del método agregado, en general apuntan a mejorar la diversidad de los clasificadores base.

Como resultado de este procedimiento, se obtuvo una mejora significativa en el desempeño al usar el clasificador de apilamiento en el conjunto de datos de temperatura máxima con respecto al desempeño más alto obtenida por un clasificador no agregado. Sin embargo, en los otros conjuntos de datos y clasificadores agregados no se observaron diferencias significativas.

Se concluye que para el método agregado de bagging, el uso del algoritmo SVR para los clasificadores base no da buenos resultados. Esto se debe a que el algoritmo de SVR no es lo suficientemente inestable como para generar clasificadores diversos con los cambios introducidos por el bagging en los conjuntos de datos de entrenamiento.

Además, se encontró que la premisa que al utilizar algoritmos distintos en los clasificadores base en el método agregado de apilamiento se debería obtener una alta diversidad en los clasificadores, no se cumple en todos los casos. Como se observó, la diversidad obtenida en el método de apilamiento en el conjunto de datos de precipitación fue bastante baja.

En la mayoría de los métodos agregados implementados, los clasificadores base generados no fueron diversos entre sí. Esto dio como resultado que el desempeño de estos métodos no fuera significativamente distinto al desempeño obtenido por el mejor clasificador no agregado en los distintos conjuntos de datos.

En el método agregado de apilamiento se observó que se puede ver perjudicado por un conjunto de entrenamiento reducido, ya que el conjunto de entrenamiento se debe dividir para entrenar los clasificadores base y el meta-clasificador. Si los clasificadores base se entrenan con muy pocos ejemplos, estos podrían tener un desempeño muy bajo. Esto puede dar como resultado que ni con las mejoras logradas al agregar clasificadores diversos, se logre un mejor

resultado que con un algoritmo no agregado entrenado con la totalidad de los ejemplos de entrenamiento.

En el método de apilamiento, cuando los clasificadores base son diversos y tienen un desempeño suficientemente bueno, el desempeño del método agregado puede ser significativamente mejor al de un algoritmo no agregado.

Finalmente, se encontró que AdaBoost no solo puede tomar clasificadores base débiles (es decir, con exactitud apenas mejor que al azar) y combinarlos en un clasificador fuerte (con exactitud mucho mejor que al azar), sino que requiere que los clasificadores base sean débiles. De lo contrario, usar un clasificador muy fuerte puede ocasionar que el clasificador base generado en la primera iteración tenga un peso mayor al de los demás clasificadores base combinados. Como consecuencia, cualquier clasificación realizada por el AdaBoost es decidida únicamente por este primer clasificador base, la diversidad en los clasificadores base en realidad es nula y el AdaBoost equivale a un método no agregado.

Agradecimientos

El autor Calvo-Valverde agradece al Dr. Pablo Alvarado Moya, tutor de su tesis en el DOCINADE, por sus orientaciones respecto al presente trabajo. Y el autor Acuña Alpízar agradece a la Maestría en Computación del Instituto Tecnológico de Costa Rica por la excelente formación recibida en su proceso formativo.

Referencias

[1] F. E. Grubbs, «Procedures for Detecting Outlying Observations in Samples,» *Technometrics,* vol. 11, nº 1, pp. 1-21, 1969.

[2] C. T. Lu, Y. Kou, J. Zhao y L. Chen, «Detecting and tracking regional outliers in meteorological data,» *Information Sciences,* vol. 177, nº 7, pp. 1609-1632, 2007.

[3] V. R. Patel y R. G. Mehta, «Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm.,» *International Journal of Computer Science Issues,* vol. 8, nº 5, pp. 331-336, 2011.

[4] D. J. Hill, B. S. Minsker y E. Amir, «Real-time Bayesian anomaly detection for environmental sensor data,» de *Proceedings of the Congress-International Association for Hydraulic Research*, 2007.

[5] T. G. Dietterich, «Ensemble Methods in Machine Learning,» de *Proceedings of the First International Workshop on Multiple Classifier Systems*, 2000.

[6] L. K. Hansen y P. Salamon, «Neural network ensembles,» *IEEE Transactions on Pattern Analysis and Machine,* vol. 12, nº 10, pp. 993-1001, 1990.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot y É. Duchesnay, «Scikit-learn: Machine Learning in Python,» *Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[8] R. J. Hyndman, «R package version 6.1,» 2015. [En línea]. Available: http://github.com/robjhyndman/forecast.

[9] D. H. Wolpert, «Stacked generalization,» *Neural Networks,* vol. 5, nº 2, pp. 241-259, 1992.

[10] L. Breiman, «Bagging predictors,» *Machine Learning,* pp. 123-140, 1996.

[11] Y. Freund y R. R. E. Schapire, «Experiments with a New Boosting Algorithm,» de *International Conference on Machine Learning*, 1996.

[12] Australian Government Bureau of Meteorology, «Daily Rainfall Climate Data,» 2015. [En línea]. Available: http://www.bom.gov.au/climate/data/. [Último acceso: 29 Mayo 2015].

[13] Australian Government Bureau of Meteorology, «Maximum Temperature Climate Data,» 2015. [En línea]. Available: http://www.bom.gov.au/climate/data. [Último acceso: 29 Mayo 2015].

[14] Australian Government Bureau of Meteorology, «Daily Global Solar Exposure Climate Data,» 2015. [En línea]. Available: http://www.bom.gov.au/climate/data. [Último acceso: 29 Mayo 2015].

[15] Y. Radhika y M. Shashi, «Atmospheric Temperature Prediction using Support Vector Machines,» *International Journal of Computer Theory and Engineering,* pp. 55-58, 2009.

[16] CyberPoint International, LLC, 12 03 2015. [En línea]. Available: https://pypi.python.org/pypi/libpgm.

[17] W. H. Kruskal y W. A. Wallis, «Use of Ranks in One-Criterion Variance Analysis,» *Journal of the American Statistical Association,* vol. 47, nº 260, pp. 583-621, 1952.

[18] P. Nemenyi, «Distribution-free Multiple Comparisons,» Princeton University, 1963.

[19] J. Cohen, «A Coefficient of Agreement for Nominal Scales,» *Educational and Psychological Measurement,* vol. 20, nº 1, pp. 37-46, 1960.

[20] T. G. Dietterich, «An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees,» *Machine Learning,* pp. 139-157, 2000.

[21] O. Bousquet y A. Elisseeff, «Stability and Generalization,» *Journal of Machine Learning Research,* vol. 2, pp. 499-526, 2002.

[22] A. J. Smola, B. Sch y B. Schölkopf, «A Tutorial on Support Vector Regression,» *Statistics and Computing,* vol. 14, nº 3, pp. 199-222, 2004.