Desarrollo de nuevos modelos para la predicción de outliers en datos temporales

Fco. Javier Duque-Pintor[1], Alicia Troncoso[1]

[1]Depto. de Lenguajes y Sistemas Informáticos, Universidad Pablo de Olavide
{fjduqpin,ali}@upo.es

Resumen En esta tesis se estudian los modelos usados tradicionalmente en tareas de clasificación con el objeto de analizar la posibilidad de extenderlos para la predicción tanto de la ocurrencia de outliers como de su magnitud.

Key words: Clasificación no balanceada, outliers, series temporales.

1. Hipótesis

Los datos temporales son un tópico de investigación en diferentes disciplinas desde hace décadas. En la actualidad se trabaja con series de datos con centenares de atributos interrelacionados, a escala temporal que puede ser el segundo. Es preciso desarrollar técnicas que ayuden a descubrir las principales características de una serie temporal con el objetivo de predecir su comportamiento. Además, extender los modelos desarrollados para el análisis de series temporales a la predicción de datos atípicos supone un importante objetivo adicional.

Obtener una predicción aproximada de fenómenos futuros es crucial en actividades económicas (los errores conllevan pérdidas). Las técnicas de predicción se han basado en modelos estadísticos, como modelos AR, MA, ARMA, ARIMA o GARCH [4], o en redes neuronales, máquinas de vectores soporte, modelos basados en reglas [7] o modelos híbridos que intentan destacar las ventajas de cada uno de los métodos. Estas técnicas resultan insuficientes al resolver problemas complejos con datos del mundo real. En los últimos años se han aplicado técnicas de Minería de Datos [9,5], normalmente usadas para problemas de clasificación o clustering, a la predicción de datos temporales. Tras adaptarlas para resolver problemas de regresión concretos, tales como la predicción de la demanda y el precio de la energía en mercados eléctricos de diferentes países, han mostrado resultados muy competitivos.

La predicción de series temporales conduce inevitablemente a la predicción de datos atípicos que ocurren bajo circunstancias anormales, los llamados outliers. La detección de outliers antes de que ocurran es también una tarea crucial para la reducción del error de predicción en una serie temporal. Sin embargo, existe un gran vacío en la literatura ya que sólo se cuenta con técnicas estadísticas robustas [8] que permiten la detección de outliers una vez ocurridos, analizando

los valores de la serie de manera que los outliers sean detectados y considerados en la generación de modelos para la predicción de valores. Sin embargo, hasta donde conocemos, no existe ninguna técnica para la predicción a priori de estos datos atípicos, excepto algunas propuestas propias desarrolladas por el grupo de investigación del doctorando y, en particular, la directora de la tesis, basadas en el descubrimiento de motivos publicada en [5,6]. El problema de descubrir un fenómeno poco común y "aislarlo"para que no se use en la generación de un modelo de predicción, no condicionando así la predicción futura, es mucho más simple que predecirlo antes de que ocurra y, por supuesto, que predecir su magnitud.

Aunque la predicción de outliers es importante en cualquier serie temporal, adquiere especial relevancia cuando se trata de datos temporales con un gran impacto socio-económico como, por ejemplo, datos temporales de energía, de terremotos o datos temporales atmosféricos. La mayoría de los trabajos existentes para el análisis de este tipo de series temporales se basan en modelos estadísticos [2,1], modelos de regresión que normalmente exigen unas condiciones de distribución y no incluyen la detección a priori de los outliers. En esta tesis se pretende explotar nuevos modelos basados en técnicas avanzadas de la Inteligencia Artificial para el análisis de este tipo de series temporales, considerando la componente temporal del problema y obteniendo, además, patrones que permitan predecir cuándo una serie temporal tendrá un comportamiento anómalo antes de que éste ocurra.

2. Objetivos

El principal objetivo de esta tesis es el estudio de modelos usados habitualmente en tareas de clasificación para analizar la posibilidad de extenderlos para la predicción tanto de la ocurrencia de outliers como de su magnitud. Una relación de los objetivos se presenta a continuación.

- **Estudio de modelos basados en técnicas de clasificación para el análisis de datos temporales.** El primer objetivo consiste en un estudio exhaustivo de modelos existentes en la literatura para su posible adaptación a la predicción de la ocurrencia y magnitud de outliers. Se estudiarán en especial las técnicas de clasificación, debido a los resultados competitivos obtenidos de la aplicación de dichos métodos a problemas reales.
- **Diseño de nuevos modelos para predicción de comportamiento no usual en datos temporales.** La mayoría de los datos temporales presentan datos atípicos en su histórico y es interesante poder predecir cuándo esos valores van a ser datos atípicos y qué magnitud alcanzarán. Por tanto, se pretende diseñar modelos para predecir tanto la ocurrencia de un dato atípico como su magnitud. Estos modelos se diseñarán con un nuevo enfoque, original de este proyecto de investigación, basado en la resolución de un problema de clasificación multietiqueta bajo el paradigma del Aprendizaje Supervisado.

- **Obtención de modelos.** Una vez diseñados nuevos modelos para la predicción de outliers se aplicarán diferentes técnicas de clasificación a la resolución de dicho problema, haciendo especial énfasis en las de clasificación no balanceadas, ya que la ocurrencia de datos atípicos dentro de una serie temporal tiene una frecuencia mucho menor que el resto de valores de la serie. En efecto, cuanto mayor desbalanceo muestren los datos, mayor dificultad de resolución presentará el problema que se pretende abordar en esta tesis.

3. Metodología y plan de trabajo

Para asegurar la consecución de los objetivos propuestos, se desarrolla un plan de trabajo descompuesto en las tareas especificadas a continuación:

- **Revisión del estado del arte.** Se realizará un estudio de los distintos tipos de técnicas usadas para modelar comportamiento no usual. En particular, se estudiarán los trabajos publicados los últimos años en las revistas de impacto y en los congresos más prestigiosos dentro de la Inteligencia Artificial.
- **Diseño de modelos para la predicción de outliers.** Se acometerá el diseño de los modelos bajo un nuevo enfoque, formulando el problema como una clasificación multietiqueta. Inicialmente se definirán los parámetros necesarios para el diseño del modelo, esto es, los diferentes horizontes de predicción que se analizarán junto a la definición de los umbrales que determinarán los intervalos de predicción para la magnitud de los outliers. Estos intervalos dependerán de las características de los datos temporales bajo estudio, por lo que se analizará la posibilidad de desarrollar un método que, para cada conjunto de datos, determine automáticamente dichos umbrales. Estos parámetros tendrán una clara influencia en el grado de desbalanceo del problema y en el número de etiquetas que se usarán para su resolución.
- **Detección de outliers para el aprendizaje.** Para la fase de aprendizaje o ajuste del modelo es necesario conocer los outliers del conjunto de datos de entrenamiento. Por tanto, hay que indicar qué puntos de la serie temporal se consideran outliers una vez ocurridos, es decir, a posteriori ("detección de outliers"). Para ello se implementarán métodos naïve basados en la desviación estándar y métodos estadísticos robustos de la literatura [3].
- **Obtención de modelos para la predicción de outliers.** Definida la parametrización, los datos serán etiquetados en función de la misma teniendo en cuenta los outliers ya encontrados en el conjunto de datos. Se generará una batería de datasets con distintos números de etiquetas y horizontes de predicción. Ésto producirá una colección de problemas de diversa dificultad puesto que, cuanto más pequeño sea el horizonte de predicción y más pequeños sean los intervalos para la predicción de la magnitud, mayor desbalanceo tendrá el problema.

  Una vez el problema se haya modelado como un problema de clasificación multietiqueta, donde las etiquetas son parametrizables en función de los umbrales que determinen la magnitud de los outliers y el horizonte de predicción,

en una fase posterior se analizará cuál es el mejor conjunto de variables para obtener dicha predicción en cada uno de los problemas bajo estudio.

- **Técnicas de clasificación y resultados experimentales.** En esta tarea se aplicarán técnicas de clasificación a los diferentes modelos obtenidos en la tarea anterior y, en base a los resultados obtenidos, analizaremos cuáles son las técnicas más adecuadas para cada tipo de problema y de qué forma el desbalanceo influye en la resolución del problema. Para evaluar los resultados de predicción obtenidos de la aplicación de las técnicas de clasificación se usará la detección de outliers llevada a cabo en una tarea anterior, pero aplicada al conjunto de datos que se vaya a usar para testear.

El plan de trabajo se resume en la Tabla 1.

Tabla 1. Descripción del plan de trabajo.

1er año	
Revisión del estado del arte	Estudio exhaustivo de métodos usados en la literatura para la predicción de outliers.
	Estudio de técnicas de clasificación, y en particular técnicas de clasificación para datos no balanceados
	Obtención de conjuntos de datos públicos que se adecuen al problema, para su posible utilización como fuente de datos para el aprendizaje
	Estudio de plataformas de software libre con algoritmos benchmarking
Diseño de modelos	Definición de diferentes horizontes de predicción
	Definición de diferentes umbrales para definir intervalos de magnitud de los outliers
	Desarrollo de métodos para determinar automáticamente dichos umbrales para cada conjunto de datos
Obtención de modelos	Detección de outliers en el conjunto de entrenamiento mediante métodos naïve basados en desviación estándar y métodos estadísticos robustos
	Generación de datasets etiquetados teniendo en cuenta la parametrización
2º año	
Técnicas de clasificación y resultados experimentales	Definición de diferentes horizontes de predicción
	Aplicación de técnicas de clasificación no balanceada a los modelos obtenidos
	Análisis de resultados para determinar la técnica más adecuada para cada problema
	Validación de resultados mediante la detección de outliers en el conjunto de test
	Publicación en congreso internacional y/o revista de impacto
3er año	
Obtención de modelos	Estudio de métodos de selección de atributos publicados en la literatura
	Mejora de los modelos obtenidos anteriormente mediante la selección del conjunto de variables óptimo para la predicción
Técnicas de clasificación y resultados experimentales	Definición de diferentes horizontes de predicción
	Aplicación de técnicas de clasificación no balanceada a los modelos obtenidos
	Análisis y validación de resultados
	Comparativa con los modelos anteriores que no consideraban el conjunto óptimo de variables
	Publicación en congreso internacional y/o revista de impacto

4. Relevancia

La importancia de la tesis viene determinada por tres factores: el propio objetivo perseguido (basta recordar que no se pretende la usual detección de outliers a posteriori, sino su predicción), la originalidad de la aproximación escogida y, por último, su gran aplicabilidad práctica.

Con respecto a la aproximación, basada en un marco de aprendizaje supervisado multiclase, resulta novedosa puesto que es la primera vez que se resuelve la predicción de outliers como un problema de clasificación multiclase. De hecho, que el problema de clasificación bajo estudio sea un problema de clasificación no balanceada es un reto por sí solo. Además, se estudia cómo afectan diferentes variables, como el horizonte de predicción y la magnitud de los outliers, al desbalanceo del problema.

Por último, se pretende aplicar estos modelos a problemas reales con impacto socio-económico (predicción de picos en el consumo eléctrico, de terremotos o de determinados niveles altos de contaminación atmosférica), con el objetivo de predecir anomalías antes de que sucedan. Esto permite, si es posible, evitarlas, planificar acciones preventivas o activar determinados protocolos.

Referencias

1. P. Bird and Z. Liu. Seismic hazard inferred from tectonics: California. *Seismological Research Letters*, 78(1):37–48, 2007.
2. J. A. Adame-Carnero et al. Surface ozone measurements in the southwest of the iberian peninsula (huelva, spain). *Environmental Science Pollution Research*, 17(2):355–368, 2010.
3. S. Gelper, R. Fried, and C. Croux. Robust forecasting with exponential and holt-winters smoothing. *J. Forecast*, 29(1):285–300, 2010.
4. Jan G. De Gooijer and Rob J. Hyndman. 25 years of time series forecasting. *International Journal of Forecasting*, 22(3):443–473, 2006.
5. F. Martinez-Alvarez, A. Troncoso, J. C. Riquelme, and J. S. Aguilar-Ruiz. Lbf: A labeled-based forecasting algorithm and its application to electricity price time series. 2008.
6. F. Martinez-Alvarez, A. Troncoso, J. C. Riquelme, and J. S. Aguilar-Ruiz. Motif-based outlier detection in time series. *Pattern Recognition Letters*, 32(1):1652–1665, 2011.
7. M. Martínez-Ballesteros, A. Troncoso, F. Martínez-Álvarez, and J.C. Riquelme. Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution. *Integrated Computer-Aided Engineering*, 17(3):227–242, 2010.
8. P. J. Rousseeuw and M. Hubert. Robust statistics for outlier detection. *IEEE Transactions on Knowledge Data Engineering*, 1(1):74–19, 2011.
9. A. Troncoso, J. M. Riquelme Santos, A. Gómez Expósito, J. L. Martínez Ramos, and J. C. Riquelme. Electricity market price forecasting based on weighted nearest neighbors techniques. *IEEE Transactions on Power Systems*, 22(3):1294–1301, 2007.