**Author (s): M. Hammad Patel, Mujtaba Ahmed Abassi, M. Seed and Shah Jamal Alam**

**Title of paper: A scheme to analyze agent-based social cimulations using exploratory data mining techniques**

**Journal: Complex Adaptive Systems Modeling**

**Pag: Art. Nro. 1**

**Year: 2018**

**PROBLEMAS QUE EL AUTOR DESEA RESOLVER**

Explorar y comprender los resultados de los modelos basados en agentes.

Los modelos basados ​​en agentes que simulan la realidad social generan **resultados que resultan de una compleja interacción de procesos** relacionados con las reglas de interacción de los agentes y los parámetros del modelo. A medida que tales modelos basados ​​en agentes se vuelven más descriptivos y se basan en la evidencia, se convierten en una herramienta útil para simular y comprender la realidad social. Sin embargo, el número de parámetros y las reglas de interacción de los agentes crece rápidamente. Tales modelos a menudo tienen parámetros no validados que deben ser introducidos por el modelador para que el modelo sea completamente funcional. Estos parámetros no validados a menudo son informados solo por la intuición del modelador y pueden representar vacíos en el conocimiento existente sobre el estudio de caso subyacente.

Los modelos basados ​​en agentes tienden a generar grandes volúmenes de datos simulados que son dinámicos y de alta dimensión, lo que los hace (a veces extremadamente) **difíciles de analizar.**

**DESCRIPCIÓN DEL APORTE DEL AUTOR**

En este documento se presenta una selección de **técnicas prácticas de exploración y extracción de datos** que pueden ser útiles para comprender los resultados generados a partir de modelos basados ​​en agentes. Se propone un **esquema simple** y se **demuestra su aplicación** en un modelo basado en agentes basado en la evidencia de las sociedades interétnicas (citas y matrimonios), llamado "DITCH". En el análisis reportado en este documento, se centra en la dinámica y la interacción de los parámetros clave del modelo y su efecto en los resultados del modelo.

**Técnicas prácticas de exploración y extracción de datos**

Varios artículos han propuesto y explorado técnicas de minería de datos para analizar simulaciones basadas en agentes. Uno de ellos es de Remondino y Correndo (2006), donde los autores aplicaron el 'ajuste de parámetros por ejecución repetida', es decir, una técnica en la que se realizan múltiples ejecuciones para diferentes valores de parámetros a intervalos discretos para encontrar los parámetros que resultan más influyentes. Los autores sugirieron diferentes técnicas de extracción de datos, como regresión, análisis de conglomerados, análisis de varianza (ANOVA) y reglas de asociación para este propósito. En otra obra, Arroyo et al. (2010) propuso un enfoque metodológico que implica un paso de minería de datos para validar y mejorar los resultados de un modelo basado en agentes. Presentaron un estudio de caso en el que se aplicó el análisis de conglomerados para validar los resultados de simulación del modelo "MENTAT". Su objetivo fue estudiar los factores que influyen en la evolución en una sociedad española de 1998 a 2000. Se encontró que los resultados de la agrupación eran consistentes con los datos de la encuesta que se utilizó para construir inicialmente el modelo.

Edmonds et al. (2014) utilizaron técnicas de agrupación y clasificación para explorar el espacio de parámetros de un modelo de comportamiento de los votantes. El objetivo de este estudio fue comprender los factores sociales que influyen en la participación electoral. Los autores utilizaron algoritmos de aprendizaje automático, como K-significa agrupación, agrupación jerárquica y árboles de decisión para evaluar los datos generados a partir de las simulaciones. Recientemente, Broeke et al. (2016) usaron el análisis de sensibilidad como la técnica para estudiar el comportamiento de los modelos basados ​​en agentes. Los autores aplicaron el análisis de sensibilidad OFAT ('Un factor a la vez'), global y basado en regresión en un modelo basado en agentes en el que los agentes extraen una fuente renovable difusora.

Las referencias citadas anteriormente no son de ninguna manera exhaustivas, pero proporcionan algunos ejemplos interesantes del uso de técnicas de minería de datos en el análisis de modelos basados ​​en agentes.

A partir de las técnicas prácticas de exploración y extracción de datos, los autores proponen un enfoque esquemático como un paso hacia la combinación de diferentes técnicas de análisis que se utilizan normalmente en el análisis de modelos basados ​​en agentes. Presentan un enfoque metodológico para utilizar técnicas exploratorias, estadísticas y de minería de datos para analizar las relaciones entre las entradas y los parámetros de salida de un modelo basado en agentes. Aplicar la técnica apropiada (o un conjunto de técnicas) para analizar el comportamiento de un modelo y la sensibilidad de los parámetros es la clave para validar y predecir cualquier fenómeno de palabra real en un modelo basado en agentes.

**Esquema Simple**

Primero, comenzamos con un análisis más amplio y exploratorio de las variables de entrada (parámetros) de un modelo seleccionado para comprender su efecto en los resultados del modelo basado en el agente dado. Esta es una forma típica de entender los modelos basados ​​en agentes, donde se explora una gama más amplia de parámetros para ver visualmente su relación con los resultados del modelo. A continuación, se realiza un análisis de sensibilidad del modelo. Con muchos parámetros de entrada, es difícil entender los resultados a través de los ojos. Por lo tanto, las técnicas como el coeficiente de correlación de rango parcial (PRCC) ayudan a medir las "relaciones monotónicas entre los parámetros del modelo y los resultados". El uso de técnicas de minería de datos permite además encontrar patrones en la salida generada en un rango más amplio de los parámetros de entrada de un modelo.

A continuación, presentamos una descripción general de algunas de las técnicas que se pueden aplicar:

1. **Análisis exploratorio de datos**

El análisis de datos en el análisis de datos exploratorios (EDA) es típicamente visual. Las técnicas EDA ayudan a resaltar características importantes en un conjunto de datos determinado. Elegir EDA como punto de partida en nuestro esquema propuesto proporciona una manera simple pero efectiva de analizar la relación entre los parámetros de entrada y salida de nuestro modelo.

En este documento, usamos el paquete ggplot2 en R para generar mapas de calor y diagramas de caja y bigotes para las variables de salida frente a los parámetros más influyentes que tienen variaciones.

1. **Análisis de sensibilidad**

El propósito de realizar un análisis de sensibilidad es estudiar la sensibilidad de los parámetros de entrada de nuestra ABM para generar las variables de salida y, por lo tanto, brindar una perspectiva más enfocada que las técnicas de análisis exploratorio. Se pueden utilizar varias técnicas para realizar análisis de sensibilidad.

En este documento, para los resultados realizamos múltiples técnicas de análisis de sensibilidad, como el método de importancia variable, el método de eliminación recursiva.

* 1. **Importancia Variable**

Para una variable de salida dada, la clasificación de cada variable de entrada (parámetro del modelo) con respecto a su importancia se puede estimar utilizando la información del modelo (conjunto de datos de entrenamiento). La importancia de la variable cuantifica así la contribución de cada variable de entrada (parámetro) para una variable de salida dada.

En nuestro caso, usamos el paquete caret en R, que construye un modelo lineal al dirigir un atributo dependiente contra el número de atributos de entrada y luego clasificarlos con respecto a su importancia estimada.

* 1. **Eliminación Recursiva**

El método de eliminación de características recursivas (también conocido como RFE) construye muchos modelos basados ​​en los diferentes subconjuntos de atributos utilizando el paquete caret en R. Esta parte del análisis se realiza para explorar todos los subconjuntos posibles de los atributos y predecir la precisión de los diferentes subconjuntos de atributos. Tamaños que dan resultados comparables.

1. **Usando minería de datos para analizar salidas ABM**

Existe un interés creciente en la simulación social en la aplicación de técnicas de minería de datos para analizar productos multidimensionales que se generan a partir de simulaciones basadas en agentes en un vasto espacio de parámetros. En esta sección, presentamos una descripción general de algunas de las técnicas comunes de datamining que se han utilizado para analizar los resultados de los modelos basados ​​en agentes.

* 1. **Clasificación y regresión de árboles**

Un árbol de clasificación / regresión se basa en un algoritmo de aprendizaje supervisado que proporciona representación visual para la clasificación o regresión de un conjunto de datos. Proporciona una forma efectiva de generalizar y predecir las variables de salida para un conjunto de datos determinado.

* 1. **K-medios de agrupación**

K-Means clustering es uno de los algoritmos de agrupamiento ampliamente implementado y se ha utilizado para analizar modelos basados ​​en agentes. La precisión del algoritmo de agrupación de K-medias depende de la cantidad de agrupaciones que se especifican en la inicialización; Dependiendo de la elección de los centros iniciales, los resultados del agrupamiento podrían variar significativamente.

**Demostración de aplicación**

Hemos utilizado el modelo basado en agentes DITCH (“Diversidad y matrimonio interétnico: confianza, cultura y homopía”) de Meyer et al. (2014, 2016) para nuestro análisis. Escrito en NetLogo, 2, el modelo simula las asociaciones interétnicas que conducen a matrimonios interétnicos informados en diferentes ciudades del Reino Unido y está basado en la evidencia.

Los agentes en el modelo DITCH se caracterizan por rasgos que influyen en sus preferencias para elegir socios adecuados en el transcurso de una ejecución de simulación. El modelo asume parejas / matrimonios heterosexuales dentro y entre diferentes etnias.

1. **Configuración experimental**

Inicialización de proporciones étnicas El modelo DITCH utiliza los datos del censo del Reino Unido de 2001 como base para el parámetro ethproportion.

Dadas nuestras limitaciones de recursos, realizamos los experimentos en dos fases: en la primera fase, examinamos la sensibilidad del modelo a la escala (en términos del número de agentes) y la medida en que los agentes buscan a sus socios potenciales en la red (es decir, , amor-radar). En la segunda fase, exploramos los parámetros del modelo específicos para expandir la red social de los agentes y aquellos relacionados con la compatibilidad de los agentes con sus socios potenciales.

Fase I: Primero exploramos el modelo variando dos parámetros con 10 repeticiones para un total de 600 carreras. Todos los demás parámetros se mantuvieron sin cambios. Cada simulación corrió para 120 garrapatas (10 años).

Fase II: En la segunda fase, mantuvimos la cantidad de agentes fijada en 3000. Luego modificamos los otros cinco parámetros del modelo para las configuraciones étnicas de las cuatro ciudades del Reino Unido (ver Tabla 1); Para un total de 9720 carreras. Cada simulación corrió para 120 garrapatas (10 años).

1. **Resultados de experimentos de simulación (Fase-I)**

En la Fase I, variamos el número de agentes y los tres valores para el parámetro del modelo de radar de amor. Para el resto de los parámetros, los valores predeterminados se utilizaron como se informó en Meyer et al. (2016). El propósito de realizar experimentos en la Fase I fue obtener una idea más amplia de los resultados del modelo, en particular, el resultado de interés, que es el porcentaje de matrimonios interétnicos que ocurren en un curso de 10 años. Principalmente, estábamos interesados ​​en probar la sensibilidad del modelo a la escala (el número de agentes) y la disponibilidad de socios potenciales una vez que la distancia social (parámetro del radar de amor) aumenta (Tabla 2).

Para resumir los resultados, generamos los gráficos de caja y bigotes y los mapas de calor (Janert 2010; Seltman 2012; Tukey 1977), para explorar la variación en la salida entre los dos parámetros variables y dentro de cada configuración de parámetros cuando se repite 10 veces. En particular, hay una fuerte disminución en el porcentaje promedio de matrimonios interétnicos cuando el número de agentes aumenta de 1000 a 2500, lo que es más evidente en el caso de Newham, donde la diversidad étnica fue mayor en contraste con el caso de Dover, donde el 98% de la población de agentes pertenecía al grupo étnico blanco.

Para un tamaño fijo de la población de agentes, el parámetro de radar de amor en el modelo DITCH influye en el porcentaje de matrimonios interétnicos para los cuatro casos (ciudades del Reino Unido). Esto no es sorprendente, ya que aumentar el valor de este parámetro permite a los agentes con un espacio de búsqueda más amplio encontrar socios potenciales y, por lo tanto, las posibilidades de encontrar un socio potencial que pertenezca a un grupo étnico diferente también aumentan. Sin embargo, la relación con el aumento de los valores del radar de amor en el modelo no es lineal para la variable de salida crossethnic para los cuatro casos

En Newham, que tiene la mayor diversidad étnica entre las cuatro ciudades consideradas, el porcentaje de matrimonios interétnicos aumenta a medida que aumenta la distancia social permitida (valor del parámetro del radar de amor), mientras que en el caso de Bradford y Dover, un aumento en el radar de amor de 1 a 2 da como resultado un aumento en el promedio de matrimonios interétnicos, pero un aumento adicional de 2 a 3 resultados de otra manera.

A partir de un análisis exploratorio de simulaciones de la Fase I, está claro que el modelo DITCH es sensible a la cantidad de agentes en el sistema. A medida que el efecto disminuye cuando la población de agentes aumenta más adelante, fijamos el número de agentes en 3000 para los experimentos de simulación en la Fase II. En el caso del radar de amor, la relación no lineal observada indica que otros parámetros del modelo que se mantuvieron fijos en la Fase I también contribuyen a la salida. Por lo tanto, a continuación se presenta una exploración adicional y un análisis más profundo de los cuatro parámetros del modelo.

1. **Resultados de experimentos de simulación (Fase II)**

En la Fase II, fijamos la población de agentes en 3000 y realizamos simulaciones a través de diferentes valores de los otros cinco parámetros del modelo, como se describe en la sección anterior. Aquí demostramos el uso de varias técnicas predictivas y de minería de datos que pueden ser útiles para explorar y analizar los resultados generados a partir de modelos basados ​​en agentes.

Primero, estimamos la "importancia" de los parámetros construyendo un modelo predictivo a partir de los datos simulados Brownlee (2016). El método clasifica los atributos por importancia con respecto a una variable dependiente. El método clasifica los atributos por importancia con respecto a una variable dependiente.

Nuevamente, el radar de amor y la posibilidad de nuevo enlace se encontraron como los más significativos (como las dos principales variables independientes). Habiendo identificado el radar de amor y la posibilidad de un nuevo enlace como los dos parámetros más importantes, exploramos la variación en el conjunto de datos generado para los cuatro casos (ciudades del Reino Unido) con respecto a estos dos parámetros, como se muestra en los diagramas de caja.

Aumentar el valor del parámetro de radar de amor da como resultado un aumento del promedio de matrimonios interétnicos en el modelo DITCH. El aumento de las posibilidades de formación de nuevos enlaces también contribuye, aunque de manera menos significativa. Las variaciones observadas en las gráficas de caja y bigotes también sugieren el papel de otros tres parámetros, que parecen jugar un papel cuando los valores de radar de amor y posibilidad de nuevo enlace aumentan.

Evaluación de los coeficientes de correlación de rango parcial También exploramos un subespacio del espacio de parámetros para identificar los parámetros más admisibles mediante la evaluación de los coeficientes de correlación de rango parcial (PRCC) para todas las variables de salida (Blower y Dowlatabadi 1994). El razonamiento detrás del cálculo de la PRCC es que para una salida particular, no todos los parámetros de entrada pueden contribuir por igual. Por lo tanto, para identificar los parámetros más relevantes, la PRCC podría ser útil. Una de las principales ventajas de identificar los parámetros más relevantes según el PRCC es que, dado un gran espacio de parámetros, si solo unos pocos parámetros de entrada tienen una contribución significativa para una salida particular, reduce significativamente la dimensionalidad del espacio de parámetros.

Siguiendo nuestro esquema propuesto, procedemos a generar un árbol de clasificación y regresión utilizando el generador de árboles de decisión de Weka

El árbol de decisión se construyó utilizando el algoritmo REPTree de Weka. Es un "aprendiz del árbol de decisión rápida y construye un árbol de decisión / regresión utilizando la ganancia de información / reducción de la varianza" (Hall et al. 2011). Como aquí estamos prediciendo el parámetro étnico cruzado, que es una variable continua, el algoritmo REPTree usa la reducción de varianza para seleccionar el mejor nodo para dividir.

Utilizamos los cinco parámetros variados para construir el árbol, en el que los parámetros del modelo DITCH love-radar, sd-education-pref, data-mean, new-link-chance, sd-dating fueron los predictores mientras que parámetro de salida interétnica fue la variable objetivo. Establecimos la propiedad minNum (el peso total mínimo de las instancias en una hoja) del clasificador en 200 para evitar el sobreajuste. El árbol resultante tenía las siguientes métricas de precisión / error en los datos de prueba / no vistos.

Error absoluto medio: 0.9582Root Error cuadrático medio: 1.2995

Una vez más, se encontró que el radar de amor era el segundo determinante más importante, especialmente en situaciones donde existía cierta diversidad étnica. Cuando el valor del radar de amor se estableció en 1 (es decir, solo se buscaron vecinos inmediatos en la red social), solo determinó el porcentaje de matrimonios interétnicos; sin embargo, para valores más altos del parámetro de radar de amor (es decir, 2 y 3), la salida fue influenciada por la posibilidad de enlace nuevo y, en otros casos, los parámetros relacionados con la datación de los agentes en la simulación.

Agrupación de K-Means en las 13 variables de salida de DITCH Ahora pasamos al algoritmo de agrupación de K-means para encontrar clústeres en el conjunto de datos generado. Realizamos el análisis de conglomerados en las 13 variables de salida del modelo DITCH que se registraron a partir de nuestros experimentos de simulación. Elegimos los datos de la Fase II, que involucraron cinco parámetros variados para cada área de muestra (una ciudad del Reino Unido) con 9720 carreras en total. Nuestro propósito de aplicar esta técnica fue agrupar las instancias de salida de naturaleza similar en grupos. Todas las variables de salida se normalizaron por primera vez antes de continuar con el siguiente paso para encontrar el número óptimo de grupos (k). Luego seguimos la técnica utilizada por Edmonds et al. (2014), en el que la suma de cuadrados dentro del grupo se calculó en función del número de agrupaciones para múltiples ejecuciones inicializadas aleatorias. El valor óptimo de los grupos en la gráfica podría identificarse como el punto en el que hay una curva o una curva similar a un codo.

El análisis de silueta también muestra que el valor óptimo para k es alrededor de 3 o 4 en este caso. Aquí la gráfica muestra una medida de similitud entre las instancias en cada grupo y, por lo tanto, proporciona una manera de evaluar parámetros como el número de grupos óptimos (Rousseeuw 1987). Los resultados de este análisis confirman que el número óptimo de clústeres debería ser alrededor de 4. Por lo tanto, ejecutamos el algoritmo de agrupamiento de K-medias para todas las trece salidas; los centroides de las cuatro agrupaciones de K-media. La partición de los datos en las cuatro agrupaciones proporciona una buena división entre los parámetros explorados.

la bondad del ajuste es alta (~ 87%) lo que indica que los grupos son distintos, con un número casi igual de instancias en los cuatro grupos. El porcentaje medio de matrimonios interétnicos fue mayor en el Grupo 2 (19.78%) y más bajo en el Grupo 4 (1.25%); mientras que se encontró que los grupos 1 y 3 estaban más cerca en términos de los matrimonios interétnicos promedio. Estos son los resultados que esperamos, ya que presentan una imagen bastante precisa de la distribución poblacional de las etnias en las cuatro ciudades del Reino Unido (Newham, Dover, Bradford y Birmingham).

**CONCLUSIONES**

A medida que los modelos de fenómenos sociales basados ​​en agentes se vuelven más complejos, explorar y analizar los datos generados se vuelve aún más difícil. Necesitamos un conjunto completo de análisis para analizar los datos que generan dichos modelos, incorporando análisis de redes sociales tradicionales o dinámicas, análisis espacio-temporales, aprendizaje automático o más recientes, como algoritmos de aprendizaje profundo.

Las técnicas analizadas en este documento no son de ninguna manera exhaustivas y la exploración de técnicas de análisis útiles para simulaciones complejas basadas en agentes es un área activa de investigación.

En este documento, se propone un enfoque simple paso a paso para combinar tres técnicas de análisis diferentes. Para ilustrar, seleccionamos un modelo basado en agentes basado en la evidencia existente de Meyer, llamado el modelo 'DITCH'. Como punto de partida, recomendamos el uso de técnicas de análisis de datos exploratorios (EDA) para analizar modelos basados ​​en agentes. EDA proporciona un conjunto simple y efectivo de técnicas para analizar la relación entre las variables de entrada y salida de un modelo. Las técnicas de análisis de sensibilidad (SA) siguen el espacio exploratorio y son útiles, por ejemplo, para clasificar los parámetros de entrada en términos de su contribución hacia una salida de modelo en particular. La aplicación de técnicas de datamining (DM) para analizar simulaciones sociales basadas en agentes es relativamente nueva.

Las técnicas tradicionales como EDA o SA (u otras técnicas estadísticas) son útiles, pueden fallar en capturar completamente un resultado complejo y multidimensional que puede resultar de simulaciones basadas en agentes.

El DM puede ser útil para proporcionar una mejor comprensión holística de los parámetros y procesos del rol en la generación de dicho resultado.