Revista Cubana de Ciencias Informáticas
Vol. 10, No. Especial UCIENCIA, Noviembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 195-209

Tipo de artículo: Artículo original
Temática: Inteligencia Artificial
Recibido:17/08/2016 | Aceptado:10/10/2016

Aplicación de la minería de datos anómalos en organizaciones orientadas a proyectos

Outliers mining applications in project management organizations

Gilberto Fernando Castro Aguilar [1*], Iliana Pérez Pupo [2], Pedro Y. Piñero Pérez[2], Natalia Martínez [3], Yairilee Crúz Castillo[2]

[1]Universidad de Guayaquil, Ecuador. gilberto.castroa@ug.edu.ec, Universidad Católica de Santiago de Guayaquil, Ecuador. gilberto.castro@cu.ucsg.edu.ec, roberto.garcia@cu.ucsg.edu.ec
[2] Universidad de las Ciencias Informáticas. La Habana, Cuba, CP.:19370. {iperez,ppp, natalia,ycruz}@uci.cu

[*] Autor para correspondencia: gfercastro@gmail.com

Resumen

La minería de datos anómalos es un área de la minería de datos que aborda el problema de la detección de datos raros o comportamientos inusuales en los datos. Esta disciplina tiene una alta aplicabilidad en disímiles escenarios entre los que se destacan el aseguramiento de ingresos en las telecomunicaciones, la detección de fraudes financieros, la seguridad y la detección de fallas. En este trabajo los autores presentan distintos métodos para el descubrimiento de datos anómalos bajo un enfoque que agrupa las técnicas de minería de datos anómalos en: métodos supervisados, métodos no supervisados y métodos semisupervisados. Se presenta además la aplicabilidad de la minería de datos anómalos en la detección de errores y fallas, en la gestión de organizaciones orientadas al desarrollo de proyectos de software. En particular se presentan las técnicas de minería de datos anómalos asociadas a procesos de aseguramiento de ingresos, bajo un enfoque reactivo, en estas organizaciones. Se discuten en el trabajo los resultados de la comparación de varios algoritmos de minería de datos anómalos, tomando como datos para el experimento la base de datos de investigaciones para la planificación disponible en el Laboratorio de Investigaciones de Gestión de Proyectos de la Universidad de las Ciencias Informáticas. Finalmente se realizan pruebas estadísticas no paramétricas para la comparación de los resultados basados en la capacidad de detección de los algoritmos. Se arriban a conclusiones y se identifican que algoritmos presentaron los mejores desempeños.

Palabras clave: minería de datos anómalos, agrupamientos, gestión de proyectos, proyectos de software

Abstract

Revista Cubana de Ciencias Informáticas
Vol. 10, No. Especial UCIENCIA, Noviembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 195-209

Outliers mining are a data mining area of data mining; have to do with detecting rare data or unusual behavior data. This discipline has high applicability in dissimilar scenarios among which include revenue assurance in telecommunications, financial fraud detection and security. In this paper the authors present different methods for the discovery outlier's data on an approach that combines the techniques of outliers mining such as: supervised methods, unsupervised methods and semi-supervised methods for outlier's detection. The applicability of outliers mining is also presented in detecting errors and failures in the management of organizations oriented software development projects. In particular, these techniques could be used in revenue assurance process, in organizations oriented to software projects. Finally, authors presented results to apply, algorithm designed in "tasks and resources" data set published by Project Management Laboratory and generated GESPRO system. At the end, non-parametric statistical tests for comparing different algorithms, based on its detection performance. They arrive at conclusions and identify which algorithms presented the best performance.

Keywords: *outliers mining, clusters, project management, software projects*

Introducción

La minería de datos anómalos (*outliers mining*) es una disciplina dentro de la minería de datos definida por autores como Hawkins (Ben-Gal, 2005) que define "dato anómalo como una observación que se desvía mucho del resto de las observaciones apareciendo como una observación sospechosa que pudo ser generada por mecanismos diferentes al resto de los datos". Otros autores como Barnet and Lewis definen "dato anómalo, como una observación que se desvía marcadamente de otros miembros de la muestra en la cual se encuentra". Esta área de la minería de datos ha tenido aplicación en disímiles escenarios (Singh y Upadhyaya, 2012) entre los que destacan: la detección de fraudes en tarjetas de créditos, análisis de irregularidades en procesos de votación, en la detección de intrusos en redes, en la detección de anomalías en imágenes médicas, en la detección de fraude en las telecomunicaciones y la industria.

En las organizaciones orientadas a proyectos se planifican múltiples proyectos simultáneamente. Durante el desarrollo de los planes se introducen defectos que pueden afectar significativamente los costos previstos para la ejecución del proyecto provocados por errores humanos o enmascarando procesos de corrupción y fraude. Es un hecho que en el escenario de los proyectos de tecnologías de la información aproximadamente el 38 % son satisfactorios mientras que el resto son renegociados o cancelados (Standish Group International, 2015). La enorme cantidad de proyectos cancelados o renegociados tiene un elevado impacto negativo tanto social como económico. El objetivo de este trabajo es evaluar la aplicación de técnicas de minería de datos anómalos en la identificación de defectos en las

planificaciones de las organizaciones orientadas a proyectos y potenciar de esta forma el aseguramiento de sus ingresos.

En la detección de datos anómalos un elemento importante es la categorización de los tipos de datos anómalos en tres tipos de subcategorías. Los datos anómalos puntuales donde cada registro puede ser considerado de forma independiente como un registro anómalo respecto al resto de los datos (Barmade y Nashipudinath, 2014), (Zimmermann, 2014). La subcategoría datos anómalos colectivos que son aquellos que analizados de forma independiente no son anómalos pero vistos en una subcolección de los datos generales si lo son. Finalmente, la tercera subcategoría son los datos anómalos contextuales que se presentan en contextos donde existen atributos de contexto que pueden definir si determinado registro es o no un dato anómalo ante el cambio (Zolhavarieh, et. al., 2014), (Ren, 2013), (Hu y Bao, 2013). Un ejemplo, de este último caso son las series de tiempo donde la variable tiempo es un atributo contextual que puede determinar la ubicación de la instancia en el resto de la secuencia.

Respecto a la organización de las estrategias empleadas para la identificación de datos anómalos se identifican en la bibliografía diferentes enfoques. Algunos como Manish Gupta (Gupta, 2014) agrupan los métodos en supervisados, semisupervisados y no supervisados. Otros como Karanjit Singh (Singh, 2012) separan los métodos en los espaciales y los clásicos, donde ubican a los basados en distancia, los basados en densidad y los estadísticos. Por su parte Ben-Gal (Ben-Gal, 2005) los agrupa en métodos paramétricos y los no-paramétricos ubicando en esta última categoría los métodos basados en agrupamiento y los basados en distancia entre otros.

En el trabajo los autores siguen el enfoque de la propuesta de Manish Gupta (Gupta, 2014) y para la presentación de los resultados principales se organiza el trabajo de la siguiente forma. En la sección metodología computacional, por la importancia de este trabajo, se realiza un breve análisis de las principales técnicas para la minería de datos anómalos. Luego se presenta un algoritmo de detección de datos anómalos adaptado a las organizaciones orientadas a proyectos de software. En la sección de análisis de resultados se aplica el algoritmo propuesto en una base de datos de proyectos comparándose varios métodos y finalmente, se presentan las conclusiones del trabajo.

Metodología computacional

En el caso de las organizaciones orientadas a proyectos se identifican entre las principales causas del fracaso o la renegociación de proyectos (Standish Group International, 2015) los deficientes procesos de planificación, control y seguimiento, errores enmascarados en extensos y complejos cronogramas. Se identifica por los autores la necesidad

de emplear técnicas de detección de datos anómalos para la detección de fallas en los procesos de gestión de alcance, la gestión de logística, la gestión de los costos y en general de la gestión de integración de los proyectos. Para el desarrollo de la investigación se proponen los siguientes pasos:

1. Identificación de las principales técnicas para la detección de datos anómalos, considerando los enfoques no supervisados, semisupervisados y supervisados.
2. Propuesta de algoritmos basados en la bibliografía consultada para la detección de datos anómalos en los procesos de gestión de proyectos.
3. Evaluación de los diferentes algoritmos comparando estrategias tradicionales con las propuestas en el artículo tomando como indicador el porciento de detección correcta de datos anómalos.
4. Comparación de los resultados obtenidos empleando técnicas de validación cruzada y técnicas estadísticas que comparen si hay diferencias significativas en los resultados obtenidos por los diferentes algoritmos.
5. Análisis de los resultados finales y conclusiones.

Siguiendo el primer paso de la metodología propuesta se presenta un breve estudio del estado del arte de las técnicas de detección de datos anómalos agrupadas en tres subsecciones para los métodos no supervisados, los semisupervisados y los supervisados respectivamente.

Métodos para la detección de datos anómalos

Entre los métodos no supervisados para la detección de datos anómalos se encuentran los métodos basados en técnicas estadísticas, los métodos basados en la proximidad y los métodos basados en análisis espacial de los datos.

Los métodos basados en técnicas estadísticas se centran en detectar valores extremos apoyados fundamentalmente por técnicas de estadística descriptiva, el análisis de histogramas, el análisis basado en funciones de densidad probabilística, análisis de componentes principales y la regresión. La base de estos métodos se centra en los teoremas: desigualdad de Markov, desigualdad de Chevichev y en el teorema del Límite Central, (Zimek, et. al., 2012), (Li, et. al., 2014), (Deneshkumar, et. al., 2014). Entre algunas de las limitaciones de estos métodos se señalan: requieren para su funcionamiento el conocimiento de la función de densidad de los datos, con frecuencia son sensibles al aumento de la dimensionalidad, y son sensibles al ruido en diversos contextos (Zimek, et. al., 2012), (Henrion, et. al., 2012).

Los métodos basados en la proximidad agrupan los métodos basados en el agrupamiento, los métodos basados en distancia y los métodos basados en densidad. Uno de los métodos de este grupo que destaca por su facilidad de aplicación son los métodos basados en distancia que construyen un ranking de datos anómalos tomando como base el cálculo de distancias entre los k vecinos más cercanos (Ramaswamy, 2000). Entre los autores pioneros en el uso de

estos métodos se encuentran Knorr y Ng (Knorr y Ng, 2000) pero han sido empleados por numerosos autores (Ben-Gal, 2005), (Gupta, et. al., 2014), (Kuna, et. al., 2014), (Shpigelman, 2014), (Keogh, et. al., 2005). Variantes a estos métodos son presentadas en (Hautamaki, et. al., 2004) donde se introduce el concepto de número inverso de los vecinos más cercanos y por Prasanta Gogoi en (Gogoi, 2012) que se basa en el análisis de las simetrías de las relaciones entre los vecinos más cercanos. Los métodos basados en distancia reportan sus mejores resultados en escenarios con relativamente pocos datos a diferencia de los métodos basados en densidad y los basados en agrupamientos. Reportan buenos resultados por el nivel de granularidad que permiten en la búsqueda.

Otro grupo de métodos basado en la proximidad, son los métodos basados en agrupamientos. En estos métodos el ranking para la detección de los datos anómalos se forma a partir de la distancia de los datos a los centros de los grupos obtenidos por el propio algoritmo en un primer momento. Estos métodos por sus características se subdividen en agrupamiento jerárquico y los basados en particiones. Ejemplos de estos métodos son los algoritmos CLARA, CLARANS, K-means entre otros (Chandola, et. al., 2012), (Ben-Gal, 2005), (Pividori, et. al., 2015). Singh Vijendra y Pathak Shivani en (Vijendra y Pathak, 2013) Vinita Shah en (Shivani, et. al., 2015). Estos métodos permiten un análisis global de los datos y la detección de pequeños grupos de datos aislados.

Otro de los grupos de métodos relevantes dentro de las estrategias basadas en la proximidad son los métodos basados en análisis de densidad. Estos métodos se centran en segmentar regiones del espacio a diferencia de los métodos basados en agrupamientos donde se segmentan los puntos. Aquellos puntos que no estén cercanos a las regiones identificadas son considerados anómalos, son particularmente útiles por su interpretabilidad. Uno de los métodos basados en densidad más reconocidos es el método “Factor local de datos anómalos” (LOF). Otros ejemplos de combinaciones de estos métodos se encuentran en (Lee, 2015), (Lee, et. al., 2014).

Los métodos basados en el análisis espacial de los datos por su parte, son bastante cercanos a los métodos de agrupamientos, se basan en el principio de que un dato anómalo en el espacio es un objeto que al representar sus atributos en el espacio estos son significativamente diferentes de los objetos vecinos a él. Estos métodos son clasificados en dos subcategorías: métodos cuantitativos y métodos gráficos (Gupta, 2014). Algunos de estos métodos son: método basado en la profundidad, método basado en los ángulos y el método basado en proyecciones de subespacios dimensionales: Zhana Bao en (Bao, 2014), Karanjit Singh en (Singh, 2012). Tienen la dificultad de no encontrar los datos anómalos interiores en el conjunto de datos, solo determina con cierta facilidad solo los que se encuentran geométricamente por fuera del conjunto de datos.

Hay escenarios donde se conocen algunos datos anómalos o al menos donde se conocen datos anómalos de algunas clases, pero no de otras. En estos escenarios se conocen como semisupervisados y se recomienda el empleo de técnicas combinadas no supervisadas con las técnicas supervisadas, algunas de las cuales se describen a continuación.

Uno de los métodos semisupervisados reconocido en la bibliografía es el basado en la estrategia de descubrimiento de nuevas clases basada en descubrir datos que tengan un comportamiento que los diferencie de clases ya conocidas. Un ejemplo de aplicación de este método se presenta en (Al-Khateeb, et. al., 2012) donde se ha adaptado el método SVM para este fin. Mientras que otros ejemplos son presentados en (Masud, 2012).

Otro método semisupervisado que ha sido frecuentemente usado, es el método basado en el aprendizaje activo. En este método los datos con clasificados por iteraciones con la intervención de expertos humanos que clasifican o ratifican la clasificación realizada por los algoritmos de los datos analizados.

Finalmente, las técnicas de clasificación basadas en métodos supervisados también pueden ser empleadas en la detección de datos anómalos. Existen diferentes escenarios que caracterizan la aplicación de las técnicas supervisadas entre las que se destacan: escenarios con clases desbalanceadas, escenarios donde existen clases normales contaminadas y escenarios con información parcial para entrenamiento (Zhang, 2010). En la bibliografía consultada se identifican así mismo diversas técnicas para tratar estos escenarios, algunas de las cuales se referencian a continuación. A.M.Rajeswari, M.Sridevi y C.Deisy entre otros proponen el uso de sistemas basados en reglas (Rajeswari, 2014) y la combinación de estas con árboles de decisión (Aggarwal, 2013).

Otro enfoque empleado basado en métodos supervisados es el ensamblaje de técnicas que se basa en entrenar clasificadores con diferentes conjuntos de entrenamiento logrando una especialización en los mismos y luego conciliar los resultados por un método de votación.

En general los algoritmos de detección de datos anómalos pueden ser empleados de forma aislada o combinados. En las estrategias combinadas se destacan dos enfoques (Aggarwal, 2013): empleo combinado de algoritmos de manera secuencial y empleo combinado de algoritmos de forma independiente con una mezcla de los resultados finales.

Propuesta de algoritmo para la detección de datos anómalos en entornos orientados a proyectos.

En el caso de los entornos orientados a proyectos es preciso identificar diferentes situaciones propensas a que se presenten datos anómalos y en función de este conocimiento previo, proponer las técnicas o combinaciones de técnicas que pueden ser empleadas. Se muestran a continuación algunas de estas situaciones:

Revista Cubana de Ciencias Informáticas
Vol. 10, No. Especial UCIENCIA, Noviembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 195-209

Situación 1. Se consideran datos anómalos puntuales, aquellas tareas que no tengan asignados recursos con las competencias requeridas o que por su volumen requieran más recursos humanos de los asignados.

Situación 2. Se consideran datos anómalos puntuales, aquellas tareas cuya estimación de tiempo o costo esté por encima o muy por debajo de los valores previstos.

Situación 3. Son datos anómalos puntuales, aquellas tareas que no respetan en el cronograma la precedencia lógica o que tengan una holgura de espera excesivamente alta respecto a otras tareas.

Situación 4. Se consideran datos anómalos puntuales, aquellos requisitos en el EDT[1] del proyecto para los cuales no hay tareas en el cronograma del proyecto orientadas al desarrollo de los mismos.

Situación 5. Se consideran datos anómalos colectivos, aquellos proyectos que, aunque son similares a otros por su alcance pueden tener costos estimados muy por encima de la media.

Situación 6. Se consideran datos anómalos de contexto, la sobrecarga de recursos humanos o no humanos en el entorno de la gestión de múltiples proyectos.

Describimos a continuación el Algoritmo 1 para la detección de datos anómalos basado en la combinación de diferentes algoritmos adaptados a cada una de las situaciones explicadas. En cada iteración se aplica un algoritmo diferente enfrentando una situación diferente. La condición de parada está dada porque se ejecuten todos los algoritmos previstos y haya un nivel de captación de los resultados por los expertos humanos.

Algoritmo 1: Algoritmo propuesto basado en combinación de técnicas especializadas en cada tipo de problema

1. *AlgoritmoBasadoCombinacionMetodos (D, A)*

   Entradas:

   D: representa el conjunto de datos a analizar.

   A: representa el conjunto de algoritmos, siendo A_i un algoritmo y denotaremos $A_i \in A$

   A_{activo}: representa método de aprendizaje activo con la intervención de expertos.

   N: la cantidad de algoritmos.

2. *begin*

3. *i =1*

4. *$D_1 = D$*

5. *Mientras queden algoritmos sin aplicar hacer en caso contrario ir a la línea 10.*

Seleccionar el algoritmo A_i

6. *Seleccionar el conjunto de datos a partir del conjunto original D_i <- D*

7. *$P_i = A_i(D_i)$ // Detección de posibles datos anómalos*

[1] EDT: Estructura de desglose del trabajo.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. Especial UCIENCIA, Noviembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 195-209

8. $O_i = A_{activo}(P_i)$ // *Aplicación del aprendizaje activo en la verificación de los datos anómalos*
9. i ++
10. *Regresar a la línea 5*
11. $O = \cup O_i$ *//combinación de los datos anómalos detectados en cada iteración*
12. *end*

Se describen a continuación cada uno de los algoritmos empleados en el meta-algoritmo propuesto.

Algoritmo de aprendizaje activo: se ejecuta en cada iteración, garantiza una revisión por parte de expertos de los datos detectados como anómalos, quienes determinan si los datos detectados son realmente anómalos o no. Además, se guarda información que puede ser útil en sucesivos algoritmos de detección de datos anómalos. En las organizaciones orientadas a proyectos la detección de datos anómalos debe hacerse periódicamente considerando replanificaciones.

Algoritmo aplicado basado en distancias para detección de tareas anómalas respecto a la duración o el costo: actúa en dos pasos, primero aplica técnicas de agrupamientos basados en las características de las tareas. Aquellos grupos con pocas tareas pueden representar tareas muy específicas o tareas que se alejan por algunos de los campos de los grupos y se debe analizar si son grupos de tareas anómalas. En un segundo paso se determinan en cada grupo aquellas tareas que tienen diferencias significativas respecto a la duración o al costo con otras tareas de su mismo grupo, las cuales pueden ser identificadas como tareas anómalas.

Algoritmos basados en lógica borrosa para análisis de asignación de recursos y competencias: este algoritmo tiene como entrada el listado de los recursos humanos con sus competencias demostradas y evaluadas. Se conoce además el listado de tareas del proyecto, la cantidad de recursos humanos que requiere para su ejecución y las competencias esperadas. El algoritmo recorre todas las tareas determinando el factor de completamiento (FC) de la tarea y el factor de satisfacción de competencias (FS). Se define los valores de umbral μFS, μFC $\in\{0,1\}$. Aquellas tareas donde algunos de los dos factores, sobrepase el umbral establecido es considerada una tarea anómala. El factor de completamiento es determinado como, el cociente entre la cantidad de recursos asignados a la tarea y la cantidad en plan de estos recursos considerando las competencias requeridas (ver (**1**). El factor de satisfacción de las necesidades se determina como, la sumatoria de las áreas de los triángulos cuya base tiene como longitud la diferencia de los centros de los conjuntos borrosos triangulares que representan el nivel ideal esperado de la competencia para el desarrollo de la tarea y el nivel real de la competencia. En la

$$FC(T_p) = \sum_{i=1}^{|R|}\sum_{j=1}^{|C|}\left(1 - \frac{Cantidad\ recursos\ asignada_{ij}(T_p)}{Cantidad\ recursos\ \ plan_{ij}(T_p)}\right)$$

Revista Cubana de Ciencias Informáticas
Vol. 10, No. Especial UCIENCIA, Noviembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 195-209

$$FS(T_p) = \frac{1}{|R|}\sum_{i=1}^{|R|} A_i\left(\left|b_{C_{i-esperada}(T_p)} - b_{C_{i-real}(T_p)}\right|\right)$$

(**2** se denota como b_C, al componente central del número triangular que representa a la competencia C. Los niveles de las competencias están representados por una variable lingüística compuesta por siete conjuntos borrosos triangulares.

$$FC(T_p) = \sum_{i=1}^{|R|}\sum_{j=1}^{|C|}\left(1 - \frac{Cantidad\ recursos\ asignada_{ij}(T_p)}{Cantidad\ recursos\ \ plan_{ij}(T_p)}\right) \quad (1)$$

$$FS(T_p) = \frac{1}{|R|}\sum_{i=1}^{|R|} A_i\left(\left|b_{C_{i-esperada}(T_p)} - b_{C_{i-real}(T_p)}\right|\right) \quad (2)$$

Algoritmo para análisis de holgura y la precedencia entre tareas: se basa en la construcción del grafo de PERT orientado a nodos, con pesos en las aristas que indican la holgura de espera entre tareas. Aquellas aristas cuya duración sea mayor que un valor umbral y las precedencias no representadas son considerados datos anómalos.

Algoritmo basado en sistemas de información para detección de requisitos no cubiertos: basado en el sistema de información se identifican aquellos requisitos representados en el EDT del proyecto, para los cuales no hay representadas tareas en el cronograma.

Algoritmo para detectar proyectos anómalos respecto a los costos: aplica técnicas de agrupamientos sobre los proyectos considerando sus características fundamentales. Aquellos grupos con pocas tareas pueden representar proyectos muy específicos o proyectos que se alejan por algunos de los campos de los grupos y se debe analizar si son grupos de proyectos anómalos o no. En un segundo paso por cada grupo se analizan el área bajo la curva respecto a la "curva de la S" de los proyectos en un mismo grupo, determinándose que aquellos proyectos con diferencias significativas en el área bajo la curva son posibles proyectos anómalos.

Algoritmo basado en análisis de la sobrecarga de recursos: se basa en el análisis de uso de cada recurso, por unidad de tiempo. Se establecen indicadores asociados a la sobrecarga cuyo comportamiento histórico puede variar en determinados momentos del año.

Se presentan a continuación dos algoritmos básicos, empleados durante la aplicación de los algoritmos específicos explicados con anterioridad. Ver algoritmo basado en distancias (Algoritmo 2) y el algoritmo basado en agrupamientos empleado (Algoritmo 3).

Tabla 1. Presentación de los algoritmos básicos empleados para la detección de los datos anómalos

Revista Cubana de Ciencias Informáticas
Vol. 10, No. Especial UCIENCIA, Noviembre, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 195-209

Algoritmo 2: Algoritmo empleado basado en distancia.

Entradas: k, (vecinos más cercanos)
n, cantidad de datos anómalos a retornar
D, conjunto de datos
MaxDistancia(d,S) función de distancia máxima.
Vecindad(d, S, k) k elementos más cercanos a d.
PrimerOutlier(S, n) retorna primeros n elementos según la distancia a sus k vecinos más cercanos.
MaxUmbral(O) retorna la mayor distancia entre elementos

1. *c = 0 (umbral de corte)*
2. *O = {}*
3. ***Para cada d en D***
4. *Vecinos(d) = {}*
5. *Para cada b en D*
6. *Si | Vecinos(d) | < k ó Distancia(b,d)< MaxDistancia(d, Vecinos(d)) Entonces Vecinos(d) = Vecindad(d, Vecinos(d) U b, k)*
7. *Si | Vecinos(d) | > k y c > Distancia(b,d)*
8. *Entonces Volver a línea 5*
10. *Fin del ciclo línea 5*
11. *PrimerOutlier(O U b, n)*
12. *c= MaxUmbral(O)*
13. ***Fin del ciclo de línea 3***

Algoritmo 3: Algoritmo empleado basado en agrupamientos y combinado con distancia.

Entradas:
D, conjunto de datos
C cantidad de centros esperados.
Distancia(d,S) función de distancia de D a al conjunto de puntos S
c umbral de corte
O conjunto de datos anómalos

1. *O = {}*
2. *clusters = ClusterMethod(D, centers=C)*
3. *centros = clusters$centers*
4. *Para cada b en D*
   Si Distancia(b,centros) > umbral
   Entonces O = O U b
5. *Fin del ciclo línea 5*
6. *PrimerOutlier(O U b, n)*
7. *c= MaxUmbral(O)*
8. ***Fin del ciclo de línea 3***

Resultados y discusión

Se empleó en la experimentación la base de datos de "Asignación de Recursos y Tiempo" del repositorio para investigaciones de proyectos terminados del Laboratorio de Investigaciones en Gestión de Proyectos que contiene 9315 tareas agrupadas en 88 proyectos y con 4 tipos de recursos. En esta base de datos todas las tareas son correctas, pero incluye un conjunto de facilidades para la simulación con diferentes porcientos la cantidad de datos anómalos presentes en cada experimento y para modelar diferentes tipos de situaciones que provocan datos anómalos. Se desarrollaron dos experimentos: detección de datos anómalos asociados tareas con una duración diferente significativamente al resto y experimento asociado a proyectos con costos sobregirados.

Experimento 1: Detección de datos anómalos asociados a la situación de tipo 2: tareas con una duración excesiva respecto al resto. Se realiza una validación cruzada generándose datos para 10 corridas diferentes cada una de ellas

con un 4% de introducción de datos anómalos. Se comparan los resultados obtenidos usando el método basado en la distancia de Mahalanobis (llamaremos Mahalanobis), y el método basado en la distancia Euclidiana (llamaremos Euclidiana) con un algoritmo combinado de agrupamiento Kmeans con distancia Mahalanobis (llamaremos Kmeans.Mahalanobis). Se considera como criterio de calidad el porciento de casos identificados correctamente y se aplican técnicas no paramétricas para la comparación de los resultados.

En la comparación de los resultados se usó el test de Wilcoxon para comparación de dos muestras relacionadas y se obtuvieron los siguientes resultados:

Tabla 2 Resultados de la comparación de algoritmos usando test de wilcoxon muestras relacionadas en R

Comparación	Resultado
data: Kmeans.Mahalanobis and Mahalanobis V = 55, p-value = 0.001953 alternative hypothesis: true location shift is not equal to 0.	Se encuentran diferencias significativas siendo mejor el algoritmo Kmeans. Mahalanobis que representa la combinación de técnicas.
data: Euclidiana and Mahalanobis V = 9, p-value = 0.06445 alternative hypothesis: true location shift is not equal to 0	No se encuentran diferencias significativas entre los métodos basados en distancia empleados, aunque la distancia de Mahalanobis reportó mejores resultados en la detección.

En este experimento para poder aplicar los métodos basados en distancia todos los atributos considerados fueron ordinales o reales, y los ordinales fueron transformados en valores numéricos.

La distancia de Mahalanobis reportó mejores resultados que la distancia euclidiana. Se pudo constatar que la combinación de agrupamiento con un método basado en distancia reportó mejores resultados porque aprovechan las potencialidades de cada método, en particular la búsqueda global del método de agrupamiento y la búsqueda local y con alto nivel de granularidad que aportan los métodos basados en distancia.

Experimento 2: Detección de datos anómalos asociados a proyectos con costos sobregirados, puede haber presencia de datos anómalos colectivos, situación de tipo 6. Se realiza una validación cruzada generándose datos para 10 corridas diferentes cada una de ellas con 3 proyectos modificados con la introducción de datos anómalos en sus tareas. Se comparan los resultados obtenidos usando método tradicional basado en distancias con las distancias Mahalanobis y con el algoritmo propuesto anteriormente que combina agrupamiento Kmeans y la distancia Mahalanobis. Se considera como criterio de calidad el porciento de proyectos identificados como anómalos y se aplican técnicas no paramétricas para la comparación de los resultados.

En este caso al aplicar las técnicas de análisis descriptivo se obtienen los siguientes resultados:

Tabla 3 Resultados de la comparación de algoritmos con R. (*Wilcoxon signed rank test with continuity correction)*

Análisis descriptive (summary)	**Aplicación de test Wilcoxon muestras relacionadas**
Mahalanobis KmeansMahalanobisS Min.: 33.33 Min.: 33.33 1st Qu.: 41.66 1st Qu.: 66.67 Median: 66.67 Median :100.00 Mean: 60.00 Mean: 83.33 3rd Qu.: 66.67 3rd Qu.:100.00 Max. :100.00 Max. :100.00	data: KmeansMahalanobisS and Mahalanobis V = 23.5, p-value = 0.1241 alternative hypothesis: true location shift is not equal to 0

A partir del análisis de los resultados no se pudieron encontrar diferencias significativas entre las dos técnicas. No obstante, la combinación de Kmeans con distancia Mahalanobis reportó mejores resultados que el método basado en distancia de Mahalanobis aislado.

Conclusiones

Diferentes técnicas de minería de datos anómalos son aplicables en la gestión de múltiples organizaciones orientadas al desarrollo de proyectos de software. Pero, para su aplicación en este escenario, estas técnicas deben ser adaptadas a las características específicas del problema de planificación de proyectos de software. En el trabajo se aprecia que los métodos basados en la distancia de Mahalanobis reportan mejores resultados que los métodos establecidos en la distancia Euclideana, aunque en el escenario analizado de baja dimensionalidad y la prueba realizada no se pudieron encontrar diferencias significativas entre ambas técnicas. En el escenario real de planificación en una organización orientada a proyectos de software la combinación de técnicas soportadas por métodos de agrupamientos, métodos basados en distancia y métodos de aprendizaje activo reportan los mejores resultados. En particular se aplicó en el experimento una estrategia combinada basada en la aplicación aislada de diferentes métodos y posteriormente la agregación de los resultados. En el caso del problema de los datos colectivos vistos como tareas sobregiradas en un mismo proyecto se identificaron los mejores resultados a partir del uso de técnicas combinadas de agrupamiento y distancia que solo usando las técnicas basadas en distancia, aunque se debe señalar que las pruebas estadísticas aplicadas no encontraron diferencias significativas. Se recomienda además la experimentación con un número mayor de casos y corridas, aumentando la cantidad de métodos empleados para la detección de datos anómalos potenciando la comparación con otros métodos de detección de datos anómalos en particular con métodos basados en densidad.

Referencias

AGGARWAL, CH. Outlier Analysis. Springer, New York, USA, ISBN 978-1-4614-6395-5, DOI 10.1007/978-1-4614-6396-2, 2013.

AL-KHATEEB, T.; MASUD, M., et. al. Recurring and Novel Class Detection using Classbased Ensemble, ICDM Conference, 2012.

BARMADE, A.; NASHIPUDINATH, M. An Efficient Strategy to Detect Outlier Transactions. International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, 2014, 3(6).

BAO, Z. Two Phases Outlier Detection in Different Subspaces. DOI: 10.1145/2663714.2668046, 2014. Disponible en: http://www.researchgate.net/publication/268040848.

BEN-GAL, I. Outlier detection, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers, ISBN 0-387-24435-2. Department of Industrial Engineering, Tel-Aviv University, 2005.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly Detection for Discrete Sequences: A Survey. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(5): p. 823–839.

DENESHKUMAR, V.; SENTHAMARAIKANNAN, K.; Manikandan, M. Identification of Outliers in Medical Diagnostic System Using Data Mining Techniques. International Journal of Statistics and Applications, DOI: 10.5923/j.statistics.20140406.01, 2014, 4(6). Disponible en: http://www.researchgate.net/publication/274721695.

GOGOI, P., et. al. Outlier Identification Using Symmetric Neighborhoods. 2nd International Conference on Communication Computing & Security [ICCCS], doi: 10.1016/j.protcy.2012.10.029, 2012, 6, p. 239 – 246.

GUPTA, M.; GAO, J., et. al. Outlier Detection for Temporal Data. ISBN(paper): 9781627053754, ISBN(ebook): 9781627053761, 2014. Disponible en: www.morganclaypool.com.

HAUTAMAKI, V.; KARKKAINEN, I.; FRANTI, P. Outlier detection using k-nearest neighbor graph. International Conference on Pattern Recognition, 2004.

HENRION, M.; HAND, D., et. al. CASOS: A Subspace Method for Anomaly Detection in High Dimensional Astronomical Databases. Statistical Analysis and Data Mining, 2012. Disponible en: http://onlinelibrary.wiley.com/doi/10.1002/sam.11167.

HU, W.; BAO, J. The dato anómalo interval detection algorithms on astronautical time series data. Mathematical Problems in Engineering, ID 979035, 2013.

KEOGH, E.; LIN, J.; FU, A. HOT SAX: Finding the most unusual time series subsequence: Algorithms and applications. ICDM Conference, 2005.

KNORR, E.; NG, R.; TUCAKOV, V. Distance-based Outliers: Algorithms and applications, VLDB Journal, 2000, 8, p. 237–253.

KUNA, H. D., et al. Outlier detection in audit logs for application systems. Information Systems, DOI: 10.1016/j.is.2014.03.1, 2014, 44, p. 22–33. Disponible en: http://www.researchgate.net/publication/262915159.

LEE, CH.; LEE, H. Novelty-focussed document mapping to identify new service opportunitie. The Service Industries Journal, DOI: 10.1080/02642069.2015.1003368, 2015, 35(6): p. 345–361. Disponible en: http://dx.doi.org/10.1080/02642069.2015.1003368.

LEE, CH.; KANG, B.; SHIN, J. Novelty-focused patent mapping for technology opportunity analysis. Technological Forecasting & Social Change, DOI: 10.1016/j.techfore.2014.05.010, 2014, p. 355–365. Disponible en: http://dx.doi.org/10.1016/j.techfore.2014.05.010.

LI, Z.; ROBERT, J., et. al. A Unified Framework for Outliers Detection in Trace Data Analysis. IEEE Transactions on semiconductor manufacturing, DOI: 10.1109/TSM.2013.2267937, 2014, 27(1).

MASUD, M.; CHEN, Q., et. al. Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams. IEEE Transactions on Knowledge and Data Engineering, to appear, 2012. Disponible en: http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.109.

PIVIDORI, M.; STEGMAYER, G.; Milone, D. Cluster Ensembles for Big Data Mining Problems. 2015. Disponible en: http://www.researchgate.net/publication/281461828.

RAJESWARI, A. M.; Sridevi, M.; Deisy, C. Outliers Detection on Educational Data using Fuzzy Association Rule Mining. Int. Conf. on Adv. in Comp., Comm., and Inf. Sci. (ACCIS-14), 2014, p. 1–9. Disponible en: http://www.researchgate.net/publication/263814468.

RAMASWAMY, S.; RASTOGI, R.; SHIM, K. Efficient Algorithms for Mining Outliers from Large Data Sets. ACM SIGMOD Conference, 2000, p. 427-438.

REN, G. Detection of Outliers in a time series of available parking spaces. Mathematical Problems in Engineering, ID 416267, 2013.

SHIVANI, P.; SHAH, V.; VALA, J. Outlier Detection in Dataset using Hybrid Approach. International Journal of Computer Applications (0975 – 8887), 2015, 122(8).

SHPIGELMAN. A Unified Framework for Outlier Detection in Trace Data Analysis. IEEE Transactions on semiconductor manufacturing, DOI: 10.1109/TSM.2013.2267937, 2014, 27(1).

SINGH, K.; Upadhyaya, S. Outlier Detection: Applications And Techniques, IJCSI International Journal of Computer Science Issues, ISSN: 1694-0814, 2012, 9(3). Disponible en: www.IJCSI.org.

STANDISH GROUP INTERNATIONAL. The CHAOS Report. New York: The Standish Group International, Inc., 2015.

VIJENDRA, S.; PATHAK S. Robust Outlier Detection Technique in Data Mining: A Univariate Approach. Faculty of Engineering and Technology, Mody Institute of Technology and Science, Lakshmangarh, Sikar, Rajasthan, India 2013.

ZHANG, Y.; MERATNIA, P.; HAVINGA, P. Outlier detection for wireless sensor networks: A survey. IEEE Communications Surveys and Tutorials, 2010, 12(2).

ZIMEK, A., et. al. A Survey on Unsupervised Outlier Detection in High-dimensional Numerical Data. Journal on Statistical Analysis and Data Mining, 2012. Disponible en: http://onlinelibrary.wiley.com/doi/10.1002/sam.11161/abstract.

ZIMMERMANN, A. A feature construction framework based on dato anómalo detection and discriminative pattern mining, 2014. Disponible en: http://www.researchgate.net/publication/264049231.

ZOLHAVARIEH, S.; Aghabozorgi, S.; Wah, Y. A Review of Subsequence Time Series Clustering. The Scientific World Journal, ID 312521, 2014, 19. Disponible en: http://dx.doi.org/10.1155/2014/312521.