Bull. Comput. Appl. Math 7(1): 43-50

Comparison of clustering quality indices for various methods using small soybean data set

Comparación de índices de calidad de agrupación para varios métodos utilizando un pequeño conjunto de datos de soja

Vivas Garza[1], Andrés María[2], Nelly Rosa[3]
[1]*Department of Mathematics, Universidade Federal de Sao Paulo, Brazil*
[2]*Department of Mathematics, Universidade Federal de Santa Catarina, Brazil*
[3]*Department of Statistics,Universidade Federal do Parana, Brazil*

Abstracto
La primera etapa de la adquisición de conocimientos y la reducción de la complejidad de un grupo de objetos es dividirlos en grupos en función de sus atributos o características. Organizar los datos en grupos sensibles es uno de los modos más fundamentales de comprensión y aprendizaje.

Palabras clave: agrupamiento difuso, descenso de gradiente, categórico, agrupamiento nominal, modelos fundamentals

Abstract
The first stage of knowledge acquisition and reduction of complexity concerning a group of objects is to partition or divide the objects into groups based on their attributes or characteristics. Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning.

Keywords: Fuzzy clustering, gradient descent, categorical, nominal clustering, fundamental models

1.Introducción

La primera etapa de adquisición de conocimientos y reducción de complejidad relativa a un grupo de objetos es particionar o dividir los objetos en grupos según sus atributos o caracteristicas[1]. "Organizar los datos en grupos sensibles es uno de los modos más fundamentales de comprensión y aprendizaje ". “Una operación fundamental en la minería de datos es la partición de un conjunto de objetos representados por datos en grupos o agrupaciones homogéneas ”[2]. La identificación de tipos de objetos es uno de los primeros pasos de Adquisición de conocimientos.

La agrupación en clústeres enfoque popular para implementar la partición. Es una clasificación no supervisada, agregación y segmentación[3]. Es el problema de particionar un conjunto de objetos en clases (llamadas clústeres) para que (i) los objetos que pertenecen a la misma clase son similares y (ii) los objetos que pertenecen a diferentes clases son disímil[4]. Al dividir los objetos en grupos, interesante se pueden encontrar grupos como los grupos de consumidores tener una propiedad particular útil para el análisis de mercado[5]. Los objetos se pueden dividir en grupos (o llamados grupos) según su proximidad en cuanto a características. Clusters son entonces gránulos de información, con cada grupo equiparando a un gránulo, y por lo tanto se puede utilizar en computacionalmente sistemas inteligentes como unidades de aprendizaje y razonamiento. los la palabra "proximidad" se utiliza como un término general que representa similitud y disimilitud[6].

2.Revisión de literature

La agrupación en clústeres no supervisada se ha estudiado ampliamente en aprendizaje automático, bases de datos y estadísticas de varios perspectivas. Muchas aplicaciones de la agrupación en clústeres se han discutido y muchas técnicas de agrupamiento se han desarrollado[7]. Hay dos métodos principales de agrupación. Métodos de agrupamiento estadístico objetos de partición según algunas medidas de proximidad, mientras que métodos de agrupamiento conceptual agrupan objetos de acuerdo a los conceptos que llevan los objetos[8]. Para automático agrupamiento, tanto un método para determinar la proximidad (similitud o disimilitud) entre los vectores de características y un método para determinar representantes (prototipos) de Por lo general, se requieren grupos[9].

La proximidad de los objetos se basa generalmente en su característica. vectores. Estos valores o atributos de características pueden ser de proporción escala, escala de intervalo, escala ordinal o escala categórica. Agrupando los

dos primeros juntos, los valores de características también pueden ser dividido en 3 tipos: numérico, ordinal y categórico[10]. Los valores numéricos de las características se comprenden bien. Ejemplos de los valores de una entidad ordinal son etiquetas como Infant, Niño y Adulto que se puede solicitar. Un ejemplo de la característica categórica es el color con valores como rojo, azul y amarillo donde no es sensato realizar pedidos a menos que se base en posición en el espectro de frecuencias[11].

El problema de la agrupación en clústeres se vuelve muy desafiante cuando los datos son ordinales o categóricos, es decir, cuando no hay medida de proximidad inherente entre valores de datos. A menudo las características ordinales se tratan como numéricas o categóricas[12]. Ambos enfoques son incorrectos. Los valores numéricos de las características se manejan fácilmente en clústeres porque los valores se pueden sumar y comparar. Sumando y la división permite el centro (centroide o prototipo) de un conjunto de valores a calcular. La diferencia entre dos valores numéricos se pueden calcular fácilmente tomando la diferencia[13]. La proximidad es un término general para la similitud y disimilitud pero sin pérdida de generalidad lo haremos restringirnos a la disimilitud de ahora en adelante. Los valores de características categóricas no tienen ninguna relación excepto por la igualdad entre ellos, y de ahí el la disimilitud entre dos valores no se puede definir fácilmente excepto en términos de igualdad. No hay pedido[14]. los distancia entre dos valores diferentes, por ejemplo, masculino y Mujer, se puede definir como 1, y la distancia entre dos los valores idénticos se pueden definir como 0. Característica ordinal los valores son similares a los valores categóricos en que aritmética las operaciones no tienen sentido y un método para encontrar el centro de un conjunto de valores no es obvio[15].

3.Discusión

La agrupación puede ser nítida o borrosa. En la antigua En este caso, cada objeto se coloca en un solo grupo. En En el último caso, se asigna un objeto a todos los clústeres para grados variables. Este grado puede ser cercano a cero y incluso cero para algunos clústeres y vectores de características. UNA El método comúnmente utilizado de agrupamiento estrictamente nítido es el k-significa. Las versiones difusas del algoritmo k-means se deben a Ruspini Bezdek, donde cada objeto es permitido tener valores de membresía para todos los clústeres en lugar que tener una membresía distinta en exactamente un grupo[16]. La matriz de miembros es una generalización de la matriz de membresía obtenida en agrupación nítida en la que caso puede llamarse matriz característica. Si, despues de fuzzy agrupación, todavía se desea poner un objeto en un solo cluster el cluster / clase / grupo asignado a un objeto puede ser elegido para ser el grupo en el que el objeto tiene el máximo valor de membresía. El proceso es entonces el de fuzzy agrupación, pero el resultado son agrupaciones nítidas. Agrupación difusa en este caso se utiliza como un paso intermedio para reducir el efecto del ruido. Trabajando solo en límites de datos numéricos el uso de estos algoritmos de tipo k-medias en áreas tales como minería de datos donde se encuentran grandes conjuntos de datos categóricos encontrado con frecuencia. Los datos categóricos y ordinales son abundante en bases de datos del mundo real.

La representación de grupos requiere la determinación de grupos prototipos a medida que se realiza la agrupación. Usando clúster los prototipos tienen tanto una ventaja como una desventaja. los La ventaja radica en el hecho de que no se requiere que Se determinará la disimilitud entre todos los pares de objetos. los La desventaja es que una forma de agregar vectores de características es requerido. Por lo tanto, dos operaciones en lugar de una deben ser definido. Esto puede no representar un problema en caso de vectores de características donde todas las características son de la escala de razón o variedad de escala de intervalo. En ese caso, la suma de números es una forma lógica de agregación. En caso de ordinal y características categóricas sin embargo, podemos tener un problema y puede resultar útil eliminar el requisito de agregación. A veces, los vectores de características ni siquiera están disponibles. los Los datos en algunas aplicaciones psicométricas se recopilan como proximidades solamente. En ese caso, agrupación relacional se aplican los métodos. Convertir una matriz de características vectores en una matriz de proximidad requiere que la disimilitud entre valores de características individuales es medible. los disimilitud entre los vectores de características es entonces una agregación de las proximidades iniciales.

Después de esta introducción, el artículo comienza con un breve revisión de la literatura. A continuación hay una breve descripción de C-significa difuso. A esto le sigue una descripción del método propuesto. Una discusión sobre la calidad de la agrupación las medidas siguen a continuación. Resultados de simulaciones y Los experimentos se proporcionan a continuación, seguidos de una descripción. del trabajo futuro, conclusión y resumen.

Para permitir nombres de variables de más de una letra y por lo tanto, permiten que los nombres de las variables sean las operaciones se indican mediante operadores explícitos. La multiplicación, por ejemplo, se define explícitamente utilizando la operador ×. La multiplicación implícita no existirá y na por ejemplo, es solo un nombre de variable. Los

nombres de las matrices están en negrita. Las variables de matriz de rango 1 están en minúsculas y los nombres de las matrices de rango superior a 1 están en mayúsculas. La división y multiplicación entre matrices generalmente se entre los componentes de las matrices.

Se han desarrollado muchos algoritmos para agrupar categorías datos. Algoritmos para agrupamiento los datos categóricos incluyen métodos de agrupación jerárquica utilizando el coeficiente de proximidad de Gower u otro medidas de proximidad, el algoritmo PAM, el algoritmos estadísticos difusos, y los conceptos métodos de agrupamiento.

Parte del trabajo se describirá ahora brevemente como se indica por los autores en sus resúmenes. Ralambondrainy presenta un enfoque para usar el algoritmo k-means para agrupar datos categóricos. Su enfoque es convertir atributos de categoría múltiple en atributos binarios (usando 0 y 1 para representar una categoría ausente o presente) y tratar los atributos binarios como numéricos en las k-medias algoritmo. Un inconveniente es que la media del grupo, dada por valores reales entre 0 y l, no indican el características de los clusters. Huang en su artículo presenta dos algoritmos que amplían el algoritmo k-means a dominios categóricos y dominios con números mixtos y valores categóricos. El algoritmo k-modes usa un simple medida de disimilitud de coincidencia para tratar objetos categóricos, reemplaza los medios de los grupos con modos y utiliza un método basado en la frecuencia para actualizar modos en el proceso de agrupamiento para minimizar el agrupamiento función de costo. Con estas extensiones, los modos k algoritmo permite la agrupación de datos categóricos en un moda similar a k-means.

Los autores también describen extensiones para el fuzzy Algoritmo de k-medias para agrupar datos categóricos. Por utilizando una medida de disimilitud de coincidencia simple para objetos y modos categóricos en lugar de medios para clusters, se desarrolla un nuevo enfoque, que permite el uso del paradigma k-means para agrupar eficientemente grandes conjuntos de datos categóricos. Guan definir nueva distancia basado en la distancia mejorada de Levenshtein con el relación de tolerancia para datos categóricos incompletos, y una nueva estrategia de disimilitud para datos numéricos incompletos. Li presentan un algoritmo de agrupamiento novedoso para conjuntos de datos mixtos mediante la modificación de la función de costo común; el rastro de la matriz de dispersión dentro del conglomerado. Una genética El algoritmo (GA) se utiliza para optimizar la nueva función de costes. para obtener un resultado de agrupamiento válido.

En, la técnica de agrupamiento utiliza un método simple de tipo FCM algoritmo iterativo que incluye un paso de cuantificación. En el paso de cuantificación, las puntuaciones de la categoría se derivan que se adaptan a la agrupación de FCM considerando los centros de agrupación y membresías. Li en su trabajo de estudio el criterio basado en la entropía para agrupar datos categóricos. Muestran que el criterio basado en la entropía se puede derivar en el marco formal de los modelos probabilísticos de agrupamiento y establecer la conexión entre el criterio y el enfoque basado en coeficientes de disimilitud. Los autores introduce la agrupación en clústeres basada en datos comprimidos que es una extensión del algoritmo de Birch. Su principal características se refieren al hecho de que puede ser especialmente adecuado para bases de datos muy grandes y puede funcionar tanto con atributos categóricos y características mixtas. Los autores desarrollar un grupo de atributos mixtos basado en conjuntos modelo para bases de datos mixtas numéricas y categóricas basadas en el método de conjunto de grupos. El método tiene excelente escalabilidad según sus creadores.

Ramakrishna proponen un algoritmo, aunque jerárquico en esencia, eso evita la Propagación de errores característicos mediante reasignación y eliminación de clústeres defectuosos. También proponen nuevos índices para la validación de clústeres en conjuntos de datos categóricos, un área que es casi inexplorado.

Ahmad proponen un método para calcular distancia entre dos valores de atributo del mismo atributo para aprendizaje no supervisado. Este enfoque se basa en el hecho de que la similitud de dos valores de atributo depende de su relación con otros atributos. Utilizan el medida de distancia propuesta con agrupación en modo k algoritmo para agrupar varios conjuntos de datos categóricos.

El método tradicional de agrupamiento difuso se llama difuso c-significa (FCM). La agrupación difusa es menos sensible al ruido que la agrupación nítida y se puede utilizar incluso cuando el resultado deseado son racimos nítidos. Este método de la agrupación requiere que los atributos sean numéricos. FCM en el principal consiste en determinar repetidamente prototipos para el clústeres que se encontrarán y cálculo de valores de pertenencia para los vectores de

características en los clústeres. Cada valor de atributo en un prototipo es la media ponderada de todos los miembros de El conjunto de datos se agrupará con pesos iguales a una potencia. del grado en que un objeto pertenece a un clúster.

Formalmente, deje que los pesos o valores de membresía de los conglomerados se designarán mediante Mp, k; la membresía de la función vector p en el grupo k. También deje que Fp, a represente los parámetros para los propios valores de atributo dentro de los vectores de características.

El exponente m se relaciona con el grado de borrosidad deseado. La agrupación nítida se logra con m igual a 1. A El valor comúnmente utilizado para la agrupación difusa. Los valores de membresía se calculan en términos de diferencia. entre los vectores de características y los prototipos.

Dp, k es la distancia o disimilitud entre el vector de características py prototipo k. Agregación de vectores de características y por lo tanto, se requiere de valores de características; de este modo restringir el uso de FCM. Como se argumentó antes la agregación no siempre es posible. Lo requerido Las operaciones aritméticas pueden no estar definidas o la característica los vectores pueden no estar definidos y solo las diferencias entre los vectores de características se conocen. En este caso otro Hay que seguir un enfoque. Esta es la esencia de este papel y se explica a continuación.

Esta sección presenta un método para agrupar vectores de características categóricas que se basan en reemplazar el conjunto original de vectores de características que contienen tanto categóricas y características numéricas con otro conjunto de vectores en (ℜ +) q donde q es una fracción del número de componentes en los vectores de características originales. Este nuevo conjunto de vectores, que tiene solo componentes numéricos, luego se agrupa usando FCM. Aplicar FCM en esta coyuntura ahora es posible porque el nuevo conjunto de vectores, las filas de F`, solo tienen componentes numéricos. FCM es deseable ya que es muy efectivo si la entrada es de la forma correcta, es decir. todos los componentes son de escala de intervalo.

A menos que una matriz de disimilitud, D (F), para el conjunto original de los vectores de características que se van a agrupar están disponibles la distancia entre dos vectores de características mixtas se encuentra de la siguiente manera.

Los atributos numéricos deben procesarse previamente de manera que todos están en [0,1]. Esto es para que los valores del atributo numérico las diferencias serán del mismo orden de magnitud que valores de las diferencias de atributos categóricos. Una definicion No se requiere la agregación de valores de características categóricas aquí, ya que no es necesario determinar los prototipos.

Un método para determinar F 'a partir de F o de D (F), si es disponible, es utilizar el descenso de gradiente en la suma de cuadrados de los errores entre las dos diferencias matrices.

Dado que la función de distancia es simétrica y las distancias entre vectores de características idénticas son cero, en realidad solo debe preocuparse por la parte de las matrices anteriores la diagonal principal.
Para asegurar valores no negativos de los componentes de F ’ podemos definirlo en términos de una variable de matriz auxiliar X como F '= X 2.

El algoritmo se puede resumir como sigue. por controlar el número de iteraciones la raíz cuadrada media Se utiliza el error definido.

El número de componentes necesarios para F 'no es grande y 5 componentes son suficientes para capturar la diferencias entre los vectores de características como lo demuestra simulaciones.

Otro método para determinar vectores dado entre vectores distancias es una clase de métodos llamados multidimensionales escalamiento.

Se puede determinar la calidad de la agrupación en clústeres comparar el resultado de la agrupación en clústeres, la partición del clúster, con una partición correcta conocida, la partición de clase, si es disponible. Una clave para comparar dos particiones nítidas, P (1) y P (2), es la tabla de contingencia, C. La entrada Ci, j es la número de objetos en el subconjunto i de P (1) y el subconjunto j de P (2). Curiosamente, P (1) y P (2) también se pueden determinar a partir de C dentro de un etiquetado de los elementos.

Varias medidas bien conocidas para comparar dos las particiones se obtienen de la tabla de contingencia incluido el índice Rand, Rand ajustado índice e índice de Jaccard. Si uno de las particiones es una partición de clase correcta asumida y el otra partición es el resultado de la agrupación, entonces estos las medidas se convierten en medidas de calidad agrupadas. Son definido en términos de los siguientes parámetros como se muestra. Los índices de calidad se miden en términos de pares enlazados de elementos. Para unir dos elementos en un tabique significa que están en el mismo subconjunto de la partición.

El número total de pares de elementos desordenados, incluyendo tanto adheridos como no adheridos, es
Otra medida de calidad de agrupación comúnmente aceptada basado en comparar una partición de clúster con un supuesto la partición de clase correcta es el índice de pureza del clúster. Un clúster se define como puro si todos los elementos del clúster vienen de una clase. Si todos los grupos son puros, entonces cada columna en la matriz de contingencia tendrá exactamente una entrada distinta de cero.

La variable n es el número total de elementos. los El valor máximo para el CPI es 1 cuando el máximo en cada La columna es la única entrada distinta de cero. Esta medida asume que una de las particiones, la partición de clases, es correcta.

Esto será 1 si hay exactamente una entrada distinta de cero en cada fila y cada columna. El PI penaliza por número de conjuntos en las particiones no es igual como lo hacen los RI, ARI y JI.

Huang posteriormente Kim un uso a medida que es equivalente a Cluster Purity y se llama Precisión de Huang.

La clase correcta para un clúster se define como ser la clase con el número máximo de elementos en el racimo.
Un ejemplo de una matriz de contingencia para 71 objetos es que se muestra a continuación. En este caso, el IPC es igual a 34/56 o 0.607 si asumimos que P (1) es la partición de clases y que P (2) es la partición del clúster.

Hay más grupos que clases en el ejemplo y CPI no penaliza la situación donde el número de clusters no es igual al número de clases.

clusters no es igual al número de clases. En los autores recomendaron el índice de Rand ajustado como el índice de elección después de muchos índices diferentes fueron evaluado para medir la concordancia entre dos particiones en el análisis de conglomerados con diferentes números de conglomerados.

Se realizan simulaciones para determinar la efectividad del método propuesto que se aplica a ambos y conjuntos de datos reales. El método propuesto, Método 0, es en comparación con otros métodos implementados por el autor aquí, que se denominarán métodos 1 y 2. Método 1 consiste en reemplazar cada atributo categórico por un número de atributos binarios igual al número de categorías propuestas por Ralambondrainy. Método 2, el método ingenuo, consiste en sustituir categorías por números que luego se tratan como numéricos. Resultados demostrar que este método simple puede funcionar igual o mejor que algunos métodos más complejos y, por lo tanto, es incluido en la comparación.

Los resultados del método propuesto también se comparan con resultados obtenidos por otros investigadores utilizando sus propios métodos propuestos. Los métodos propuestos por estos investigadores no se implementan y solo sus valores para se utilizan ciertos índices que han proporcionado. Esta significa que los valores de algunos índices para algunos métodos son no disponible ya que no se proporcionan tablas de contingencia que los valores se pueden calcular.

El índice de borrosidad, m, utilizado en FCM, se mantiene en 2 en todo. Los conjuntos de datos reales son de la máquina UCI Repositorio de aprendizaje.

de datos artificiales. Este conjunto de datos es generado por primero produciendo 5 vectores con 20 producidos aleatoriamente componentes de enteros entre 0 y 19. Cada uno de estos Luego se copian cinco vectores 20 veces produciendo una matriz de dimensión 100 por 20 y 5 clases con 20 elementos en cada. A esta matriz de vectores de características se agrega un binario matriz de la misma dimensión. La probabilidad de un 1 en esta matriz es del 80%. La matriz resultante es el conjunto de datos que se utiliza con los números enteros tratados como categorías. Añadiendo

un 1 es tan drástico como sumar cualquier otro número excepto el 0 ya que no se consideran diferencias sino solo igualdad para medir la distancia entre valores categóricos en el método 0.

Para el aprendizaje de los vectores de características equivalentes a distancia, la tasa de aprendizaje fue de 0,001. El número total de iteraciones, para obtener F ', fue 1000. El número de componentes en el vectores de características originales era 20. El número de componentes en los vectores de características equivalentes a distancia es establecido en 5. La reducción en la raíz del error cuadrático medio sobre las iteraciones se muestran. La raíz cuadrada media El error es sobre las diferencias entre los componentes de las dos matrices de disimilitud definidas por.

A continuación, se encuentra la matriz de contingencia para el método 0 aplicado al conjunto de datos artificiales.

La matriz de contingencia para el método 1 y la misma entrada está.

Hay 3 máximo en la columna 4 correspondientes a la clase 4. La matriz de contingencia para el método 2 se encuentra. La matriz de contingencia es la base de toda la agrupación medidas de calidad aquí.

Una comparación de los 3 métodos aplicados por el autor es dado donde CPI significa pureza de racimo, RI es el índice rand, ARI es el índice rand ajustado y JI es el Índice de Jaccard.

Esto muestra que el Método 0 es muy superior a los otros dos métodos. Lo interesante es que el método ingenuo, 2, es superior al método 1 propuesto por Ralambondrainy en este caso.

La segunda simulación implica aplicar el método de este documento a un pequeño subconjunto de la soja original base de datos. Se utiliza el conjunto de datos sobre enfermedades de la soja. porque todos los atributos de los datos pueden tratarse como categórico. La base de datos es de la UCI Machine. Depositario de aprendizaje. El conjunto de datos tiene 47 registros, cada uno está descrito por 35 atributos. El número de los valores de atributo que faltan son cero. Cada registro está etiquetado con una de las cuatro enfermedades: Diapehhe Stem Canker, Pudrición carbonosa, pudrición de la raíz por Rhizoctonia y Phytophthora Putrefacción. Excepto por Phytophthora Rot, que tiene 17 registros, todas las demás enfermedades tienen 10 registros cada una.

Gráfica de un conjunto de vectores de características con dos componentes y la misma matriz de distancia que la matriz de distancia para el El conjunto original de vectores de características se muestra. Este 2- La gráfica dimensional también se obtiene mediante escalada. La trama muestra una agrupación inherente.
Para obtener un conjunto equivalente de vectores de características con solo 4 componentes reales, en lugar de 5, utilizando el método de este papel requería el programa de capacitación como se demuestra.

Figura 3 Programa de entrenamiento para determinar el conjunto de vectores de características equivalentes de distancia para la soja pequeña conjunto de datos.

El resultado después de que se aplicó FCM es la contingencia matriz. Esto muestra un resultado de agrupamiento idéntico a las clases. No se encontró que menos de 4 componentes Ser suficiente. La tasa de aprendizaje se obtuvo por ensayo y error.

Los diversos índices de agrupamiento para los 3 métodos son como en Cuadro. Otros tres métodos para los que solo se agrupa la pureza las medidas que se proporcionan se encuentran. Los kmodes difusos ha sido desarrollado por Huang y Ng en el papel mencionado anteriormente. En la tabla NA significa no disponible en la referencia.

Nuevamente, el Método 0 es superior en términos de todo el grupo índices de calidad. Los métodos 1 y 2 son similares en términos de los índices.

Los mismos métodos que arriba también se aplicaron a los grandes El conjunto de datos de soja también se encuentra en los datos de aprendizaje automático. base. El número de registros en este caso es 307. El los atributos son los mismos que antes. Hay 19 clases en este caso. El folklore parece ser que los últimos cuatro las clases no están justificadas por los datos, ya que tienen muy pocas instancias correspondientes. Faltan 705 atributos valores, que en los métodos de los autores, correcta o incorrectamente, se tratan como una categoría distinta. El diagrama de abajo es

un diagrama de vectores bidimensionales que tienen las distancias internas como los vectores de datos originales. Es obvio que los clusters no son muy definido en este caso.

El número total de iteraciones necesarias para obtener vectores de características equivalentes con 4 componentes es 5000. La tasa de aprendizaje fue 0,0001. El horario de aprendizaje requerido para obtener los vectores de características equivalentes distantes es como se muestra en la Figura 5. Uso de una característica equivalente a distancia vectores con menos de 4 componentes no se tener éxito.

Nuevamente, el método 0 es superior en todos los aspectos a los otros 2 métodos. El método ingenuo, el método 2, es superior a método 1 propuesto por Ralambondrainy en este caso.

Otro conjunto de datos para el que se basa el método de este artículo aplicado también es de la base de datos de aprendizaje automático. Es conocida como la base de datos del zoológico. El número de instancias. El número de atributos. Aunque uno de los atributos, el número de piernas, es numérico, las diferencias en este valor no son significativas en cuanto a clase se refiere y, por tanto, este atributo se trata como categórico. El número de clases de animales.

Para determinar los vectores equivalentes de distancia con 4 Se utilizaron componentes 5000 iteraciones. La tasa de aprendizaje utilizado fue 0,001. El número de grupos permitidos. La matriz de contingencia se encuentra.

Una comparación de resultados para 5 métodos, incluidos 2 proporcionada por otro autor se muestra.

En, NA significa no disponible. Los resultados en 3 y 4 son de papel en las filas 5-7 son de. Nuevamente, el método 0 demuestra ser mejor en términos de precisión de partición. Los métodos 1 y 2 son similares en términos de los índices. Referencia que describe un método basado en entropía, no tenía valores para RI, ARI, y JI ni los otros métodos.

Este conjunto de datos contiene resultados de solicitudes de tarjetas de crédito. Hay 690 instancias con 6 numéricas y 8 atributos categóricos. Hay 2 clases. Esta base de datos fue generado por Quinlan puede recuperarse de. Kim se refieren al crédito australiano base de datos en su prueba, pero no utilice el conjunto de datos completo, pero solo 202 solicitantes y 9 atributos frente a 690 instancias y 14 atributos. Los autores no describen cómo se obtiene este subconjunto. Los resultados están.

Un método para agrupar vectores de características con mezcla Se han descrito atributos numéricos y categóricos. En el método propuesto, un conjunto de vectores de características, de muy pequeña dimensión (3-5 componentes), equivalente a la conjunto original de vectores de características en términos de inter característica distancias vectoriales, pero tener solo componentes numéricos es generado y luego FCM se aplica al nuevo conjunto de vectores de características. El método que se utiliza para encontrar el El nuevo conjunto de vectores de características es el descenso de gradientes. El método ha demostrado ser muy eficaz en términos de partición de clúster resultante.

4.Conclusión

Sin embargo, existe potencial de mejora. La idea de mapeo a un conjunto de vectores de características agrupables de FCM parece ser sólido pero computacionalmente intensivo y es por tanto, conviene buscar otros medios para hacerlo. La función de error puede tener una gran cantidad de optima y, por lo tanto, otros métodos de optimización deben ser considerado. El principal esfuerzo consiste en determinar la vectores de características de una matriz de distancia que está en el dominio de escalamiento multidimensional.

Otro inconveniente del método propuesto es que cuando el número de registros es muy grande la disimilitud La matriz será extremadamente grande y los métodos de reducción de datos tendrá que ser considerado. Un método de reducción de datos consiste en agrupar y luego reemplazar los datos originales establecido por los centroides de los grupos. Sin embargo, esto nos lleva de regreso al problema original de encontrar centroides cuando el los vectores de características son una combinación de categorías y números.

Otra área de trabajo futuro es el desarrollo de métodos en el caso de que exista una combinación de características ordinales y categóricas. Vectores de características en general constan de los 3 tipos de características, incluidas las ordinales y categórico. Para ordinal ni la distancia ni la agregación son definido. Para características categóricas, la distancia está definida pero la agregación no está definida.

Por tanto, el trabajo futuro consistirá en aplicar una método de escalamiento multidimensional eficiente y de combinando el método para tratar con características ordinales propuesto por el autor en otros artículos con el método para tratar las características categóricas propuesto aquí.

Referencias

[1] Beg, I., Rashid, T. "A Clustering Algorithm Based on Intuitionistic Fuzzy Relations for Tree Structure Evaluation", (2017) International Journal of Applied and Computational Mathematics, 3 (4), pp. 3131-3145.

[2] Khan, A., Katanic, D., Thakar, J. "Meta-analysis of cell- specific transcriptomic data using fuzzy c-means clustering discovers versatile viral responsive genes", (2017) BMC Bioinformatics, 18 (1), art. no. 295.

[3] Chen, Y.-T. "Medical Image Segmentation Using Independent Component Analysis-Based Kernelized Fuzzy c-Means Clustering", (2017) Mathematical Problems in Engineering, 2017, art. no. 5892039.

[4] Gligorić, M.V., Gligorić, Z.M., Beljić, Č.R., Lutovac, S.M., Damnjanović, V.M. "Long-Term Room and Pillar Mine Production Planning Based on Fuzzy 0-1 Linear Programing and Multicriteria Clustering Algorithm with Uncertainty", (2019) Mathematical Problems in Engineering, 2019, art. no. 3078234.

[5] Benaichouche, A.N., Oulhadj, H., Siarry, P. "Multiobjective improved spatial fuzzy c-means clustering for image segmentation combining Pareto-optimal clusters", (2016) Journal of Heuristics, 22 (4), pp. 383-404.

[6] Viattchenin, D.A. "Heuristic possibilistic clustering for detecting optimal number of elements in fuzzy clusters", (2016) Foundations of Computing and Decision Sciences, 41 (1), pp. 45-76.

[7] Baykasoğlu, A., Gölcük, İ., Özsoydan, F.B. "Improving fuzzy c-means clustering via quantum-enhanced weighted superposition attraction algorithm", (2019) Hacettepe Journal of Mathematics and Statistics, 48 (3), pp. 859-882.

[8] Ferdaus, M.M., Anavatti, S.G., Garratt, M.A., Pratama, M. "Development of C-Means Clustering Based Adaptive Fuzzy Controller for a Flapping Wing Micro Air Vehicle", (2019) Journal of Artificial Intelligence and Soft Computing Research, 9 (2), pp. 99-109.

[9] Yousefian-Jazi, A., Choi, J. "Sequential integration of fuzzy clustering and expectation maximization for transcription factor binding site identification", (2018) Journal of Computational Biology, 25 (11), pp. 1247-1256.

[10] Sepúlveda, E., Dowd, P.A., Xu, C. "Fuzzy Clustering with Spatial Correction and Its Application to Geometallurgical Domaining", (2018) Mathematical Geosciences, 50 (8), pp. 895-928.

[11] Shaari, M.A., Samsudin, R., Ilman, A.S. "Forecasting drought using modified empirical wavelet transform-ARIMA with fuzzy C-means clustering", (2018) Indonesian Journal of Electrical Engineering and Computer Science, 11 (3), pp. 1152-1161.

[12] Dagher, I. "Fuzzy clustering using multiple Gaussian kernels with optimized-parameters", (2018) Fuzzy Optimization and Decision Making, 17 (2), pp. 159-176.

[13] Maguerra, S., Boulmakoul, A., Karim, L., Badir, H. "A distributed execution pipeline for clustering trajectories based on a fuzzy similarity relation", (2019) Algorithms, 12 (2), art. no. 29.

[14] Owsiński, J.W., Kacprzyk, J. "New developments in fuzzy clustering with emphasis on special types of tasks", (2018) Control and Cybernetics, 47 (2), pp. 115-130.

[15] Güler Dincer, N., Akkuş, Ö. "A new fuzzy time series model based on robust clustering for forecasting of air pollution", (2018) Ecological Informatics, 43, pp. 157-164.

[16] Rubio, E., Castillo, O., Melin, P. "Interval type-2 fuzzy possibilistic c-means clustering algorithm", (2016) Studies in Fuzziness and Soft Computing, 342, pp. 185-194.