Revista de Matemática: Teoría y Aplicaciones 2009 **16**(2): 283–292
cimpa – ucr issn: 1409-2433

CLASIFICACIÓN AUTOMÁTICA SIMBÓLICA POR MEDIO DE ALGORITMOS GENÉTICOS*

Fabio Fernández–Jiménez† Alex Murillo–Fernández‡

Recibido/Received: 20 Feb 2008 — Aceptado/Accepted: 8 Dec 2008

Resumen

Se presenta una variante en los métodos de clasificación: un algoritmo genético para clasificación automática utilizando las herramientas del análisis simbólico de datos; esta implementación permite solventar los problemas de los métodos clásicos de clasificación: obtención de mínimos locales y dependencia de los tipos de datos con los cuales trabajan: continuos. El método fue programado en MatLab® y usa un operador interesante de codificación. Se comparan clases por su inercia intra-clases. Se usaron las siguientes medidas para datos del tipo simbólico: medida de disimilitud de Ichino-Yaguchi, medida de disimilitud de Gowda-Diday, diatancia Euclídea y distancia de Hausdorff.

Palabras clave: Clasificación automática, análisis simbólico, k-means, algoritmos genéticos, optimización.

Abstract

This paper presents a variant in the methods for clustering: a genetic algorithm for clustering through the tools of symbolic data analysis. Their implementation avoids the troubles of clustering classical methods: local minima and dependence of data types: numerical vectors (continuous data type).

The proposed method was programmed in MatLab® and it uses an interesting operator of encoding. We compare the clusters by their intra-clusters inertia. We used the following measures for symbolic data types: Ichino-Yaguchi dissimilarity measure, Gowda-Diday dissimilarity measure, Euclidean distance and Hausdorff distance.

*Este trabajo fue presentado en la 11^{th} Conference of the International Federation of Classification Societies, Dresden, Alemania, 13–18 Marzo 2009.

†Maestría en Matemática Aplicada, Universidad de Costa Rica, San José, Costa Rica. E-Mail: fabio.fernandez@gmail.com

‡Sede del Atlántico, Universidad de Costa Rica, Turrialba, Costa Rica. E-Mail: alex.murillo@ucr.ac.cr

Keywords: Clustering, symbolic analysis, k-means, genetic algorithm, optimization.

Mathematics Subject Classification: 91C20.

1 Introducción

Uno de los métodos de gran uso en la minería de datos es la **clasificación automática**. Este método permite encontrar grupos en la población inicial, dichos grupos son valiosos pues comparten una serie de atributos que los hace similares. Una vez identificados los atributos que los unen el analista puede realizar diversas acciones según el área de aplicación en la que se encuentre. Por ejemplo una aplicación de la clasificación automática es la segmentación de clientes, una vez segmentados los clientes podemos determinar aquellos clientes fuertes en movimientos (por ejemplo en una base de datos de compras) y ofrecerles ciertos productos. Realmente la aplicación del uso de la clasificación automática es muy amplia y permite atacar diversos problemas en múltiples áreas. Usualmente los métodos creados para solucionar el problema de clasificación automática trabajan con datos denominados clásicos, es decir solo con continuos [3].

El **K-Means** [5] es uno de los algoritmos de aprendizaje no supervisado más simples y conocidos para resolver el problema de la clasificación automática.

El procedimiento radica en clasificar un conjunto dado de datos en cierto número de grupos (clusters) fijados a priori (se asume k grupos). La idea principal es definir k centroides (núcleos), uno para cada grupo. Posteriormente se analiza cada individuo (punto en el conjunto de datos inicial) y se asocia con el núcleo más cercano.

Al asociar todos los individuos se deben recalcular los núcleos y nuevamente se repite el procedimiento de asociar los individuos al núcleo más cercano. Se detiene el algoritmo cuando ya no hayan cambios (los centroides no cambian). Finalmente, el algoritmo en lo que se enfoca es en minimizar una función objetivo, en este caso la inercia intraclase (la distancia de los individuos a su núcleo más cercano):

$$Inercia = \sum_{j=1}^{K} \sum_{i \in j}^{N} D(x_i, c_j)$$

Donde x_i representa al individuo i y c_j al centroide j. En este caso D es una distancia, usualmente Euclídea, y como para los datos clásicos cada una de las V variables corresponden a un valor numérico, basta con realizar restas para determinar la distancia de un individuo con respecto a otro:

$$D(x_i, c_j) = \sqrt{\sum_{v=1}^{V} (x_{iv} - c_{jv})^2} = \|x_i - c_j\|$$

La ventaja principal de este algoritmo es la rapidez con la que encuentra la clasificación óptima, no obstante la agravante es que dicho óptimo no es necesariamente el óptimo global, ya que por su estrategia de búsqueda el algoritmo cae en mínimos locales.

Si bien la clasificación automática es una buena herramienta para análisis de datos y en general para la minería de datos, el modelo clásico es muy restrictivo dado que solo permite realizar operaciones sobre datos continuos, esto realmente limita el dominio de variables a utilizar: no se puede utilizar de forma natural variables discretas con muchas modalidades.

Para mitigar el problema, la clasificación automática, utilizando **datos simbólicos** [2], amplía de sobremanera el conjunto de variables a utilizar: se pueden utilizar como tipo de dato los intervalos o las distribuciones de probabilidad, por ejemplo en vez de usar el promedio en el monto de compras de un individuo podemos utilizar un intervalo con el mínimo y máximo, lo cual es mucho mejor que el promedio, o mejor aún se puede utilizar un histograma con la distribución de compras del individuo (por monto).

Recientemente se ha incorporado un nuevo tipo de técnicas para tratar problemas de optimización y puntualmente la clasificación simbólica. Estas técnicas se conocen como heurísticas, y una de ellas son los algoritmos genéticos.

Un **Algoritmo Genético** (GA, por sus siglas en inglés) es una técnica de búsqueda usada en computación para encontrar la solución exacta o aproximada en problemas de optimización y búsqueda.

Los algoritmos genéticos están en la categoría de heurísticas de búsqueda global. Estos son una clase particular de algoritmos evolutivos, los cuales usan técnicas inspiradas en la evolución natural, tales como herencia, mutación, selección y cruzamiento (también llamada recombinación) [6].

2 Algoritmo genético para clasificación automática simbólica

Los algoritmos genéticos, como heurística, son una gran herramienta muy utilizada en la optimización, de tal forma que pueden aplicarse al problema de clasificación automática, el cual puede verse como el problema de minimizar la inercia intraclase o maximizar la inercia interclase (entre los grupos).

Estos algoritmos [4] ofrecen un mecanismo alternativo para resolver el problema de clasificación automática. En estos casos usualmente se plantea el problema desde el punto de vista de datos clásicos y se muestra como se encuentran buenos resultados al compararlos contra el K-Means, el cual es uno de los métodos más usados en Clustering. A diferencia del K-Means, los algoritmos genéticos hacen uso de todas sus propiedades para encontrar el óptimo global, para ello es fundamental, por ejemplo, la característica de la mutación.

Así, el objetivo de este trabajo es crear un algoritmo genético para resolver el problema de clasificación automática, pero en este caso los datos son simbólicos, y por ende la función a minimizar es distinta. El algoritmo propuesto para resolver la clasificación automática simbólica por medio de algoritmos genéticos se detalla en las siguientes secciones.

Primero se debe recalcar que la implementación es la de un GA simple, el detalle de la función objetivo y los operadores genéticos se detallarán con forme se avance en el artículo. Al inicio se detalla el formato de los datos de entrada, la codificación realizada, los operadores de cruce y mutación utilizados y el mecanismo de selección ejecutado.

2.1 La matriz de datos simbólicos

Dado que el problema principal es resolver la clasificación automática con datos simbólicos primero se definió la representación tabular de dichos datos simbólicos.

Se definió tres estructuras principales de datos:

- Lista de individuos.
  Vector de N entradas (N individuos).

- Lista de variables.
  Dicha lista consiste realmente en una matriz de $V \times 2$, donde V es el número de variables. En cada fila la primera entrada contiene el tipo de variable (1: continua, 2: intervalo y 3: histograma) y la segunda entrada define la cantidad de valores que representan a dicho dato, por ejemplo para los datos continuos solo se necesita un valor, para los intervalos dos (mínimo y máximo) y para los histogramas se quieren M valores, uno para cada una de las M modalidades.

- Matriz de datos.
  Es una matriz de $N \times T$, donde T depende de la cantidad de variables y el tipo de cada una de dichas variables. Cada fila representa a un individuo y los datos en dicha fila son los valores según el orden establecido en la matriz de variables.

2.2 Codificación

En las técnicas heurísticas modernas usualmente se deben realizar cambios según el problema que se trate de resolver, como lo suele ser la modelación de la función de generación de vecinos en el problema de la Búsqueda Tabú o en el Sobrecalentamiento Simulado, y en este caso los algoritmos genéticos no escapan a estas personalizaciones.

De tal forma, y enfocados en el problema que nos atañe que es la clasificación, la codificación de los individuos debe ser una forma que permita realizar cruces y mutaciones correctas y que logren una correcta representación de los grupos.

Inicialmente se realizó una codificación por pertenencia, no obstante esto presenta el problema que varias codificaciones representan la misma división de individuos, por ejemplo, para el caso de 7 individuos y 3 grupos se tienen las siguientes dos codificaciones que representan la misma partición: [1123133] y [3312322].

Debido al problema con este tipo de codificación se utilizó una codificación por filiación binaria [4], es decir cada individuo (partición) es un conjunto de K palabras (o vectores) de tamaño N; si el i-ésimo bit del j-ésimo vector está "encendido" significa que el objeto i pertenece al grupo j. Por ejemplo para la siguiente partición [23112231] ($N = 8$, $K = 3$), la codificación es la siguiente:

0 0 1 1 0 0 0 1
1 0 0 0 1 1 0 0
0 1 0 0 0 0 1 0

2.3 Operador de cruce

Para realizar el cruce (crossover en inglés) se construyó un operador que permita conservar los puntos comunes de los padres, esto es mantener la clase de los individuos que comparten el mismo grupo en cada padre. Para aquellos que no están en el mismo grupo en ambos padres se ubica aleatoriamente en alguno de los grupos.

Ejemplo: si se tienen los siguientes padres:
father = [2 3 1 1 2 2 3 1]
0 0 1 1 0 0 0 1
1 0 0 0 1 1 0 0
0 1 0 0 0 0 1 0

mother = [2 1 2 2 3 2 1 3]
0 1 0 0 0 0 1 0
1 0 1 1 0 1 0 0
0 0 0 0 1 0 0 1

El objetivo es conservar los individuos que están en la misma clase en ambos padres, en este caso los individuos 1 y 6 que pertenecen a la clase 2:

son = crossover(father, mother)
0 0 1 0 0 0 0 0
1 0 0 1 1 1 0 0
0 1 0 0 0 0 1 1
son = [2 3 1 2 2 2 3 3]

Efectivamente los individuos 1 y 6 continúan en clase 2.

2.4 Operador de mutación

En este caso la mutación es muy simple, se selecciona al azar un grupo del cual se moverá un individuo de un lugar a otro. Dicho individuo y el grupo destino también se escogen de forma aleatoria (distribución uniforme).

Ejemplo:
son = [2 3 1 2 2 2 3 3]
0 0 1 0 0 0 0 0
1 0 0 1 1 1 0 0
0 1 0 0 0 0 1 1

sonm = mutation(son)
0 0 1 1 0 0 0 0
1 0 0 0 1 1 0 0
0 1 0 0 0 0 1 1

En este caso se observa cómo se movió el individuo 4 de la clase 2 a la clase 1.

2.5 Función objetivo

En el caso de la clasificación simbólica el objetivo que se busca es reducir la inercia intraclase, para ello se implementó una función que precisamente mide la distancia de los individuos a cada uno de los núcleos/centroides respectivos[1].

Como se mencionó al inicio el método trabaja con datos simbólicos de tres tipos: datos continuos ($\mathcal{C}$), datos tipo intervalo ($\mathcal{I}$) y datos tipo histograma ($\mathcal{H}$); de tal manera la forma de calcular la distancia entre individuos y núcleos se realiza de forma distinta según el tipo de dato.

Para el caso de los datos continuos se realiza una resta (R), para el caso de intervalos se utiliza la métrica de IchinoYaguchi (IY), para el caso de los histogramas se implementó la similitud de Índice de Afinidad (AI) [2].

De tal forma, la distancia simbólica (D_s) utilizada para medir lejanía o cercanía del individuo x_i al centroide c_j es:

$$D_s(x_i, c_j) = \sum_{v=1}^{V} IY(x_{iv}, c_{jv}) 1_{\{v \in \mathcal{I}\}} + (1 - AI(x_{iv}, c_{jv})) 1_{\{v \in \mathcal{H}\}} + R(x_{iv}, c_{jv}) 1_{\{v \in \mathcal{C}\}},$$

donde:

- V es la cantidad de atributos (variables) de los individuos,
- IY es la métrica de IchinoYaguchi ($|\cdot|$ = cardinalidad = longitud del intervalo):
$$IY(x_{iv}, c_{jv}) = |x_{iv} \cup c_{jv}| - |x_{iv} \cap c_{jv}| + \gamma(2|x_{iv} \cap c_{jv}| - |x_{iv}| - |c_{jv}|),$$
- AI es la similitud de Índice de Afinidad (M es la cantidad de modalidades de la variable tipo histograma):
$$AI(x_{iv}, c_{jv}) = \sum_{m=1}^{M} \sqrt{x_{ivm} \times c_{jvm}},$$
- R es una resta de datos continuos:
$$R(x_{iv}, c_{jv}) = \frac{x_{iv} - c_{jv}}{c_{jv}}.$$

Si bien existen otras medidas a utilizar tales como Gowda-Diday para el caso de los intervalos o Euclídea para los histogramas, para el cálculo de inercia se utilizaron únicamente las mencionadas anteriormente. Es importante recalcar que en el método propuesto únicamente se trabajó con datos continuos, tipo intervalo y tipo histograma, no obstante pueden incorporarse otro tipos de datos como conjuntos, para lo cual se debe utilizar las disimilitudes correspondientes [2].

[1] Al igual que el caso clásico, cada centroide, en el caso simbólico, representa al promedio de los individuos que pertenecen a dicho grupo (para $\mathcal{C}$ promedio simple, para $\mathcal{H}$ promedio simple por modalidad, para $\mathcal{I}$ promedio simple en cada extremo inferior y superior del intervalo).

3 Pruebas experimentales

Se efectuaron pruebas con tres conjuntos de datos, uno conformado por datos tipo histogramas (tabla 1) y los otros dos para datos tipo intervalo. En la tabla 2 se detallan las tablas utilizadas. La primera de ellas es la conocida como *"Oils"* o de los aceites de Ichino.

Inds	Variable a			
Ind1	0.8212	0.1200	0.0460	0.0128
Ind2	0.8092	0.1250	0.0532	0.0126
Ind3	0.7769	0.1333	0.0699	0.0199
Ind4	0.8368	0.1106	0.0448	0.0077
Ind5	0.8310	0.1185	0.0415	0.0091
Ind6	0.8605	0.1057	0.0314	0.0024
Ind7	0.8535	0.1040	0.0345	0.0081
Ind8	0.8523	0.1095	0.0289	0.0094
Ind9	0.8336	0.1121	0.0429	0.0114
	Variable b			
Ind1	0.4831	0.1134	0.1247	0.2787
Ind2	0.4871	0.1275	0.1337	0.2517
Ind3	0.5747	0.0911	0.1145	0.2197
Ind4	0.4622	0.1306	0.1432	0.2641
Ind5	0.4615	0.1424	0.1310	0.2650
Ind6	0.4393	0.1499	0.1466	0.2642
Ind7	0.4540	0.1404	0.1407	0.2649
Ind8	0.4343	0.1386	0.1542	0.2728
Ind9	0.4892	0.1152	0.1317	0.2639
	Variable c			
Ind1	0.8212	0.1203	0.0457	0.0128
Ind2	0.8092	0.1257	0.0531	0.012
Ind3	0.7769	0.1337	0.0697	0.0197
Ind4	0.8368	0.1114	0.0442	0.0075
Ind5	0.8310	0.1191	0.0411	0.0088
Ind6	0.8605	0.1066	0.0305	0.0024
Ind7	0.8535	0.1043	0.0342	0.0081
Ind8	0.8523	0.1108	0.0276	0.0094
Ind9	0.8336	0.1126	0.0427	0.0111

Tabla 1: Datos generados para histogramas.

Otra de las tablas utilizada con intervalos es la referente a hongos [2], la que puede descargarse desde: `http://www.mykoweb.com/CAF/species_index.html`, ya que por su tamaño no se incluye en el artículo.

Oil	Grav. Específica	Pto Cogelamiento	Yodo	Saponificación
linseed	[0.930, 0.935]	[-27, -18]	[170, 204]	[118, 196]
perilla	[0.930, 0.937]	[-05, -04]	[192, 208]	[188, 197]
cotton seed	[0.916, 0.918]	[-06, -01]	[99, 113]	[189, 198]
sesame	[0.920, 0.926]	[-06, -04]	[104, 116]	[187, 193]
camelia	[0.916, 0.917]	[-25, -15]	[80, 82]	[189, 193]
olive	[0.914, 0.919]	[00, 06]	[79, 90]	[187, 196]
beef	[0.860, 0.870]	[30, 38]	[40, 48]	[190, 199]
hog	[0.858, 0.864]	[22, 32]	[53, 77]	[190, 202]

Tabla 2: Datos de Aceites de Ichino.

Con las tablas de datos anteriores se realizaron 1000 ejecuciones, esto para cada una de las tablas y métodos. En cada ejecución se almacenó el valor óptimo encontrado.

Con el objetivo de comparar el rendimiento del algoritmo genético se tomó la inercia de las particiones generadas (en cada una de las 1000 ejecuciones) contra el resultado de un K-Means adaptado para uso de datos simbólicos y utilizando las mismas disimilitudes (este K-Means se ejecutó durante 1000 iteraciones). El algoritmo genético propuesto se ejecutó con dos valores para la población inicial: 4 y 8 individuos.

Datos	GA (4)	GA (8)	K-Means(1000)
Oils	91.6%	99.3%	83.9%
Hongos	43.7%	63.9%	32.7%
Histogramas	58.0%	80.0%	22.0%

Tabla 3: Porcentaje de veces en el cual el valor mínimo fue encontrado.

En la tabla 3 se muestra el resultado obtenido, tanto del GA como del K-Means, este resultado se interpreta como el porcentaje de veces que los algoritmos obtuvieron el menor valor (encontrado por alguno de ellos, y en este caso por ambos). En esta tabla, como es claro, se observa que el algoritmo genético encuentra el mínimo más veces que el K-Means.

Además hay que resaltar que los valores promedio y máximo de la función objetivo (a minimizar) siempre son mejores en el algoritmo genético que en K-Means; y de hecho al aumentar el tamaño de la población inicial se obtienen mejores resultados (ver ejemplo para datos de histogramas en la tabla 4). De tal manera, con el algoritmo genético no solamente se tiene mejores resultados en la cantidad de veces que se encuentra el mínimo, sino también en cuanto al promedio y máximo de inercia.

En síntesis, en ambos casos (GA) se encontró la mejor partición que arroja como óptima en promedio la clasificación por medio de K-Means, sin embargo se notó que según crece el número de individuos de la población inicial el método se vuelve lento.

Para el caso conocido de Oils, se encontró la mejor solución que también retorna K-Means y que representa fielmente los grupos que se pueden observar si se ejecuta un ACP (Simbólico).

Algoritmo	Mínima inercia	Promedio de inercia	Máxima inercia
GA (4)	**0.0014**	0.0033	0.0016
GA (8)	**0.0014**	**0.0021**	**0.0015**
K-Means(1000)	**0.0014**	0.0040	0.0021

Tabla 4: Valores de inercia para los datos de histogramas.

4 Conclusiones

- Con base en los resultados obtenidos el algoritmo genético arroja mejores resultados (en cuanto a la función objetivo) al compararlos contra con un método tipo K-Means.
- El enfoque del análisis simbólico amplía la variedad de datos que se pueden analizar, lo cual es de suma aplicación en los problemas del mundo real.
- La codificación utilizada representó una ventaja en la ejecución del método, gracias a su correcta representación de las particiones.
- Si bien los operadores de mutación y selección son simples, al combinarse con el operador de cruce planteado los resultados obtenidos son muy satisfactorios.
- Al agregar más individuos en la población inicial se obtuvo mejores resultados, no obstante hay que recordar que conforme se aumente dicha cantidad también el tiempo de ejecución aumentará.

5 Trabajo por hacer

Se está preparando el software para analizar datos más complejos y así para probar el algoritmo, y además se desea construir otros operadores genéticos para compararlos contra los desarrollados. También, junto con los datos más complejos, se probará el método con datos mixtos, es decir datos con variables de distintos tipos: intervalos, continuos e histogramas. Finalmente, se explorará el uso de otras distancias para datos simbólicos, como por ejemplo la expuesta en [1] (distancia de Wasserstein), durante la 11^a conferencia de la Federación Internacional de las Sociedades de clasificación.

Referencias

[1] Arroyo, J; Maté, C. (2009) "Descriptive distance-based statistics for histogram data", in: 11^{th} *Conference of the International Federation of Classification Societies*, March 13–18, Dresden: 105–106.

[2] Billard, L.; Diday, E. (2006) *Symbolic Data Analysis: Conceptual Statistics And Data Mining*. Wiley, New York.

[3] Castillo, W.; González, J.; Trejos, J. (2009) *Análisis Multivariado de Datos*. Manuscrito en preparación.

[4] Larrañaga, P.; Lozado, J. (2002) *Estimation of Distribution Algorithms*. Kluwer Academic Publishers, Dordrecht.

[5] MacQueen, J. (1967) "Some methods for classification and analysis of multivariate observations", *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, Vol. 1, University of California Press, Berkeley: 281–297.

[6] Pham, D.T.; Karaboga, D. (2000) *Intelligent Optimization Techniques*. Springer, London.