Un Entorno de Generalización basado en Distancias

Aristides Dasso[*], José Orallo[§]; María José Ramirez[§]; César Ferri[§]
[*]Universidad Nacional de San Luis, Argentina; [§]Universidad Politécnica de Valencia, España.
arisdas@unsl.edu.ar, {jorallo, mramirez, cferri}@dsic.upv.es

CONTEXTO

Este trabajo de investigación se encuentra enmarcado dentro del Proyecto de Incentivos código 22/F822: "Ingeniería de Software: Conceptos, Métodos y Herramientas en un Contexto de Ingeniería de Software", de la Universidad Nacional de San Luis, Argentina; en una de las líneas: "Métodos Formales y Prototipos Evolutivos", del mismo. Además se relaciona con el Grupo ELP del Departamento de Sistemas Informáticos y Computación de la Universidad Politécnica de Valencia, España. Dentro del contexto de desarrollo de herramientas, esta investigación tiene como objetivo el desarrollar Operadores de Generalización basados en Distancias, en un contexto de clasificación asistida.

RESUMEN

Los Operadores de Generalización producen patrones, aún cuando el patrón puede ser útil por sí mismo, esto no es suficiente en muchos casos donde el patrón no es claro para que los seres humanos puedan entenderlo. Es importante que el patrón sea lo suficientemente expresivo para que los seres humanos puedan entenderlo, ya que el patrón debería ser la explicación misma. Las Métricas (distancias métricas) son fáciles de entender y los patrones basados en distancias son útiles además de claras y se explican por sí mismas. Así poder relacionar operadores de Generalización, Distancias y Patrones nos daría las ventajas de todos ellos. En este trabajo se busca realizar ello en un contexto de Operadores de Clasificación Asistidos.

Palabras clave: *Minería de Datos. Generalización. Operadores de Generalización. Aprendizaje de Máquina. Distancias. Métricas. Operadores basados en Distancia.*

1. INTRODUCCION

Los patrones juegan un rol muy importante en el campo del aprendizaje de máquina (machine learning). Obtener un patrón a partir de un data set (conjunto de datos) es considerado un objetivo importante, sino el único objetivo en este campo. Los patrones generalizan regularidades encontradas en los datos. En consecuencia son considerados buenos predictores de las características y propiedades que los elementos comparten, tanto dentro como fuera del data set actual, pero todos cubiertos por el patrón.

La generalización está íntimamente relacionada con los patrones, en realidad se puede decir que ambos constituyen las dos caras de la misma moneda.

Los operadores de generalización obtienen patrones a partir de los data sets. Normalmente se dice que un operador de generalización "aprende" un patrón a partir de un data set y en consecuencia produce un patrón. Sin embargo no siempre la forma en que el patrón es expresado es fácilmente entendible para los seres humanos. Así, entonces, es importante al obtener un patrón poder explicarlo.

Normalmente la explicación está dirigida a seres humanos. Por otro lado, explicar el patrón con el patrón mismo es la forma más obvia y directa de explicarlo. Sin embargo muchas veces el patrón es expresado en un

lenguaje o formato que no resulta fácil de entender para los seres humanos, mientras que en otros casos el patrón es simplemente un mecanismo opaco con respecto a su funcionamiento interno.

Una de las últimas etapas en el proceso conocido como Knowledge Data Discovery (KDD, descubrimiento de conocimiento), que incluye la Minería de Datos, es la aplicación y diseminación del conocimiento adquirido. Dado que este conocimiento, sino en todos los casos en un número importante, se obtiene bajo la forma de un patrón, es crucial que el patrón sea lo más entendible posible para los seres humanos, así resultará más fácil de emplear.

Los espacios métricos y las funciones de distancia asociados a ellos son ampliamente empleados en el campo de machine learning. Hay numerosas técnicas de KDD que están basadas en distancias, tanto como hay distancias. Sin embargo es importante señalar que ellas juegan un rol central al aprender patrones y en proveer una base para producir reglas de similitud, ya que se asume que los elementos que se encuentran cercanos unos de otros comparten las mismas propiedades.

Por otro lado las distancias son fácilmente entendibles para los seres humanos.

Como una consecuencia de lo anterior, resulta conveniente basar lenguajes o métodos, para expresar patrones y generalizaciones, a partir de distancias, y en consecuencia, un objetivo importante pareciera ser el de tratar de relacionar distancias con patrones y generalizaciones. Para poder realizar exitosamente esta relación debemos encontrar un modo de expresarla.

En [Estr2008] esta relación se ha obtenido por medio de operadores de generalización mostrando las propiedades que un operador de generalización debe tener para que sea considerado un operador basado en distancia, y en consecuencia tener los patrones generados por dicho operador, también basados en distancia.

Esto ha sido mostrado para diferentes operadores, así como tipos de datos, para el caso de similarity clustering, pero no para métodos asistidos.

Como un ejemplo de los objetivos damos aquí las dos primeras definiciones:

Let:
- X a set of elements;
- (X, d) a metric space, d operates on X;
- C a finite set of class tags;
- E a data set, $E \subset X$;
- $F_c \subseteq E$;
- $F_c = \{e \in E \mid class_T(e) = c\}$;
- $class_T: X \to C$ a labelling function that given an element of X returns the class to which the element belongs according to a target concept T;
- $\mathcal{L}$ a pattern language;
- $\Delta : X \times X \to \mathcal{L}$ a binary generalisation operator;
- $Set : \mathcal{L} \to 2^X$ a function that returns the elements of X that are covered by a pattern. □

Podemos dar la definición de un operador de generalización [Estr2008]:

Definition 1. (Generalisation operator) For every finite set E of elements in X, a generalisation operator Δ is a function such that $\Delta(E) = p$ where $p \in \mathcal{L}$ and $E \subset Set(p)$. □

Basándonos en la definición anterior, ahora podemos dar la definición de un operador de generalización con respecto a una clase:

Definition 2. (Generalisation operator wrt a class) Given $c \in C$, for every finite set E of elements in X, a generalisation operator Δ wrt a class, is a function such that $\Delta(E) = p$ where $p \in \mathcal{L}$ and $E \subset Set(p)$ and $\forall\ x \in Set(p) \mid class_T(x) = c$. □

2. LINEAS DE INVESTIGACION y DESARROLLO

En este trabajo centramos nuestra actividad en las definiciones y propiedades de dichos operadores para adaptarlos a operadores de clasificación asistidos.

Hay tres propiedades básicas que un operador de generalización basado en distancia debe tener: reachability (alcanzabilidad), intrinsicability (intrisicabilidad) y minimality (minimalidad). Estas definiciones, deben ser aplicadas a distintos operadores y demostrar que los mismos las poseen o no. Asimismo esto debe hacerse para distintos tipos de datos.

Se ha aplicado a distintos tipos de datos, especialmente datos jerárquicos y en $\mathbb{R}^2$.

También trabajamos en diseñar una función de coste que permita evaluar cuan bueno es un operador en producir patrones. En otras palabras, pretendemos medir la calidad de una generalización dada.

Para realizar esto de una manera estricta es que introducimos el concepto de función de coste. Esta función ayuda a medir la calidad de una generalización evaluando tres variables: cuan complejo es el patrón, cuan bien el patrón se ajusta a los datos y por último cuan "puro" es el patrón.

De esta manera la función de coste tiene tres componentes (funciones) y cada uno de ellos evaluará cada uno de los factores mencionados. Damos más abajo la definición sin entrar en mayores detalles.

Definition 24. (Cost function) Let $k : 2^X \times \mathcal{L} \rightarrow \mathbb{R}^+ \cup \{0\}$. We will say that k is a cost function, if for every $E \in 2^X$, where E is finite, every pattern $p \in \mathcal{L}$ and $E \subset Set(p)$; *if* $Set(p) \neq X$ *then* there exits a constant $c > 0$ such that $k(E, p) < c$. □

La complejidad del patrón está directamente relacionada con la sintaxis del mismo y en consecuencia sigue los principios de MDL/MML (Minimum Description Length / Minimum Message Length), y es altamente dependiente tanto del tipo de los datos como del lenguaje de patrones elegido para su representación. [Rissan1978], [WaBo1968]

En cuanto al ajuste está directamente relacionado con la distancia empleada y basado con el concepto de frontera de un conjunto (border of a set). En otras palabras esto significa que dado un patrón p_1, éste se ajusta más a un conjunto $E = \{x_1, x_2, x_3\}$ que otro patrón p_2, si la frontera de p_1 está más cerca de E que la frontera de p_2. Dado que el concepto de frontera de un conjunto es intrínseco a los espacios métricos, ésta función no es tan dependiente de los datos.

La función que evalúa la pureza de un patrón está directamente relacionada con cuántos elementos de la clase de interés (los elementos "positivos") cubre el patrón y cuantos quedan fuera, así como cuantos elementos que no pertenecen al patrón (los elementos "negativos") son cubiertos por este y cuántos elementos negativos deja fuera.

El valor óptimo de esta función se obtiene cuando el patrón que está siendo evaluado cubre todos los elementos positivos y no cubre ninguno de los negativos.

Para esta función hay varios métodos candidatos, y estamos experimentando con varios de ellos, especialmente aquellos que son basados en distancias.

3. RESULTADOS OBTENIDOS/ESPERADOS

Se ha desarrollado un entorno teórico para operadores de generalización basados en distancia para aprendizaje supervisado. El entorno está siendo usado para identificar operadores de generalización para clasificación.

Asimismo este entorno general nos ha permitido continuar desarrollando otras líneas de investigación como las de evaluación de patrones basados en distancias.

Este trabajo de evaluación y desarrollo de funciones de coste está siendo llevado a cabo y actualmente estamos experimentando con varias funciones, algunas de las cuales muestran resultados interesantes, que pensamos pueden ser dignas de publicación en un futuro.

4. FORMACION DE RECURSOS HUMANOS

El trabajo aquí presentado se desarrolla en el marco de una tesis de doctorado conjunto entre la Universidad Nacional de San Luis,

Argentina y la Universidad Politécnica de Valencia, España, dentro del Programa AlfaLERNET de la Comunidad Europea.

5. REFERENCIAS

[Rissan1978] Rissanen, Jorma; Modeling by shortest data description. *Automatica,* vol. 14 (1978), pp. 465-471.

[WaBo1968] Wallace, C. S.; Boulton, D. M.; An information measure for classification. Computer Journal, Vol 11, No 2, August 1968, pp 185-194. Oxford University Press.

[Estr2008] Estruch, Vicent; Bridging the Gap between Distance and Generalisation: Symbolic Learning in Metric Spaces. PhD dissertation, Universitat Politècnica de València, Departement de Sistemes Informàtics i Computació. December 2008.