Tạp chí Tin học và Điều khiển học, T.28, S.2 (2012), 161–170

ĐỘ ĐO TƯƠNG TỰ HỖN HỢP CÓ TRỌNG SỐ MSM-R VÀ MỘT SỐ KẾT QUẢ THỰC NGHIỆM VỚI BÀI TOÁN PHÂN LỚP DỮ LIỆU

NGUYỄN TRUNG TUẤN

Viện Công nghệ thông tin kinh tế, Trường Đại học Kinh tế quốc dân

Tóm tắt. Độ đo tương tự hỗn hợp đóng vai trò quan trọng trong các bài toán phát hiện tri thức và khai phá dữ liệu dựa trên khoảng cách hay dựa trên độ tương tự giữa các đối tượng như bài toán phân lớp, phân cụm, ... Bài báo trình bày chi tiết hơn về độ đo tương tự hỗn hợp có trọng số được xác định tự động dựa trên lý thuyết tập thô (*Mixed Similarity Measure based on Rough sets theory - MSM-R*), phương pháp thực nghiệm và các kết quả thực nghiệm phân lớp sử dụng *MSM-R* trên một số tập dữ liệu mẫu, so sánh kết quả với kết quả phân lớp sử dụng độ đo của *Goodall*. Hai thuật phân lớp được sử dụng là *k-láng giềng gần nhất* (*k-nearest neighbors*) và cây quyết định C4.5. Các kết quả thực nghiệm cho thấy tính hiệu quả và khả năng áp dụng thực tiễn của độ đo *MSM-R* trong phân lớp dữ liệu.

Abstract. Mixed Similarity Measure plays an important role in the distance-based or similarity-based knowledge discovery and data mining problems such as classification, clustering... This paper aims to present more detailed studies on the Mixed Similarity Measure, which has attribute weights determined automatically and based on Rough sets theory (called Mixed Similarity Measure based on Rough sets theory - *MSM-R*). Moreover, the paper presents the experimental method and the experimental results for classification problem using *MSM-R* on some UCI datasets, comparing results with the results of classification using Goodall's measurement. Two proposed classification methods are k-nearest neighbors and decision tree (using C4.5 software). The experiment results show the effectiveness and practical applicability of the *MSM-R* in the real-world data classification problems.

1. GIỚI THIỆU

Độ đo tương tự hỗn hợp hay độ đo khoảng cách hỗn hợp giữa các đối tượng được biểu diễn bằng một tập các thuộc tính có kiểu dữ liệu khác nhau (bao gồm thuộc tính số - *numerical*, thuộc tính định danh - *nominal*, thuộc tính có thứ tự - *ordinal*) đóng một vai trò quan trọng trong các bài toán phát hiện tri thức và khai phá dữ liệu dựa trên khoảng cách hoặc độ tương tự. Chất lượng của các hàm độ đo này ảnh hưởng đáng kể đến sự thành công của ứng dụng tương ứng trong việc tìm kiếm các kết quả. Mỗi độ đo sẽ có hiệu quả nhất định đối với từng kiểu dữ liệu riêng biệt, nếu muốn kết hợp nhiều kiểu dữ liệu khác nhau trong cùng một độ đo người ta thường phải chuyển đổi chúng về cùng một kiểu dữ liệu. Tuy nhiên, việc chuyển đổi các kiểu dữ liệu khác nhau về cùng một kiểu dữ liệu sẽ có khả năng làm mất mát thông tin hoặc làm giảm hiệu quả của độ đo được sử dụng.

Hiện nay đã có một số nghiên cứu về độ đo tương tự hỗn hợp, đặc biệt là độ đo tương tự hỗn hợp do *Goodall* [3] đề xuất. *Goodall* đã đưa ra phương pháp tính độ tương tự hỗn hợp cho các đối tượng với cơ sở toán học chặt chẽ nhằm áp dụng cho bài toán phân loại thực vật. Cho đến nay, độ đo tương tự hỗn hợp này được đánh giá là độ đo có chất lượng tốt trên dữ liệu hỗn hợp [4, 5]. Tuy nhiên, nhược điểm của độ đo này là độ phức tạp tính toán lớn và không thoả mãn các tiên đề *metric*. Các tác giả trong [1, 2] đã có những nghiên cứu sâu và cải thiện tốc độ tính toán cho độ đo này với độ phức tạp tính toán tuyến tính. Trong [10], tác giả đã đề xuất phương pháp tính độ đo tương tự cho dữ liệu hỗn hợp, dữ liệu đồ thị (*graph*). Các tác giả trong [7] đã đề xuất những nội dung cơ bản về cách tính độ đo tương tự hỗn hợp có trọng số dựa trên lý thuyết tập thô (*Mixed Similarity Measure based on Rough sets theory - MSM-R*). Độ đo tương tự hỗn hợp *MSM-R* được tính gián tiếp thông qua độ đo khoảng cách hỗn hợp có trọng số của các thuộc tính hỗn hợp bao gồm cả thuộc tính số và thuộc tính định danh. Trong [8], các tác giả đã thực hiện thử nghiệm và đánh giá hiệu quả của độ đo *MSM-R* trong bài toán phân lớp dữ liệu kinh tế - xã hội thực tế của Việt Nam.

Bài báo trình bày những kết quả nghiên cứu chi tiết về độ đo khoảng cách hỗn hợp hay độ đo tương tự hỗn hợp có trọng số được xác định tự động dựa trên lý thuyết tập thô, đồng thời nhằm trình bày những phương pháp thực nghiệm và đánh giá hiệu quả, khả năng áp dụng của độ đo *MSM-R* trong bài toán phân lớp dữ liệu. Tiếp theo, Mục 2 trình bày chi tiết về độ đo tương tự hỗn hợp có trọng số được xác định tự động bằng lý thuyết tập thô. Mục 3 trình bày về phương pháp nghiên cứu thực nghiệm và các bộ dữ liệu được lựa chọn để thử nghiệm. Mục 4 trình bày các kết quả thực nghiệm trên các bộ dữ liệu khác nhau và bình luận về kết quả thực nghiệm. Cuối cùng là Mục kết luận và hướng nghiên cứu tiếp theo.

2. ĐỘ ĐO TƯƠNG TỰ HỖN HỢP CÓ TRỌNG SỐ CHO THUỘC TÍNH ĐƯỢC XÁC ĐỊNH TỰ ĐỘNG DỰA TRÊN LÝ THUYẾT TẬP THÔ

Nhằm xác định khoảng cách hỗn hợp giữa hai đối tượng hay xác định độ tương tự hỗn hợp giữa hai đối tượng thỏa mãn các tiên đề *metric*, một phương pháp tính trọng số cho các thuộc tính một cách tự động trong độ đo tương tự hỗn hợp với tiếp cận lý thuyết tập thô được đề xuất. Đây là một hướng mới về việc sử dụng lý thuyết tập thô trong các bài toán phát hiện tri thức và khai phá dữ liệu.

Giả sử các đối tượng trong một hệ quyết định được thể hiện bằng m thuộc tính $A = \{a_1, a_2, \ldots, a_m\}$, $a_{ik} \in dom\,(a_k)$ là giá trị trên thuộc tính k của đối tượng i, và thuộc tính quyết định hay thuộc tính phân lớp là d.

2.1. Khoảng cách cục bộ trên từng thuộc tính giữa hai đối tượng

Để tính khoảng cách cho hai đối tượng i và j, trước hết, ta tính khoảng cách cục bộ trên từng thuộc tính của hai đối tượng đó. Gọi g_{ijk} là khoảng cách giữa đối tượng i và j trên thuộc tính thứ k. Giá trị của g_{ijk} được xác định như sau:
- Với thuộc tính số, khoảng cách được chuẩn hoá là

$$g_{ij} = \begin{cases} 0, & max\left(\{a_k\}_1^{card(a_k)}\right) = min\left(\{a_k\}_1^{card(a_k)}\right) \\ \frac{|a_{ik} - a_{jk}|}{G_{kmax}}, & max\left(\{a_k\}_1^{card(a_k)}\right) \neq min\left(\{a_k\}_1^{card(a_k)}\right) \end{cases} \tag{2.1}$$

trong đó $G_{kmax} = max\left(\{a_k\}_1^{card(a_k)}\right) - min\left(\{a_k\}_1^{card(a_k)}\right)$

và $max\left(\{a_k\}_1^{card(a_k)}\right)$, $min\left(\{a_k\}_1^{card(a_k)}\right)$ tương ứng là giá trị lớn nhất và giá trị nhỏ nhất trong miền giá trị của thuộc tính thứ k trong hệ quyết định, các giá trị này thường được xác định trong bước tiền xử lý dữ liệu.

- Với thuộc tính định danh, khoảng cách đã chuẩn hoá là

$$g_{ij} = \begin{cases} \lambda_k, & a_{ik} \neq a_{jk} \\ 0, & a_{ik} = a_{jk} \end{cases} \tag{2.2}$$

trong đó λ_k có thể được tính theo một trong các cách sau:

1) Theo cách tính thông thường:

$$\lambda_k = 1. \tag{2.3}$$

2) Theo giá trị được chia đều cho số lượng giá trị của thuộc tính thứ k là $card\left(dom\left(a_k\right)\right)$:

$$\lambda_k = \frac{1}{card\left(dom\left(a_k\right)\right)}. \tag{2.4}$$

2.2. Tính trọng số cho các thuộc tính

Như đã biết, một đối tượng được xác định bởi một tập giá trị trên tập thuộc tính đặc trưng cho đối tượng, nếu nhiều đối tượng có cùng giá trị trên một thuộc tính điều kiện nào đó mà các đối tượng này lại cũng có giá trị trên thuộc tính quyết định hay phân lớp như nhau thì khi đó ta có thể coi là thuộc tính điều kiện có sự ảnh hưởng lớn đến thuộc tính quyết định hay phân lớp. Với tư tưởng trên và dựa trên lý thuyết tập thô, ta có thể xác định được mức độ ảnh hưởng của một thuộc tính a_k tới thuộc tính quyết định d (thuộc tính được sử dụng để thể hiện kết quả phân lớp)

$$\alpha_k = \frac{|POS_{a_k}(d)|}{|U|} = \sum_{X \in U/d} \frac{\left|\underline{a_k}(X)\right|}{|U|} \tag{2.5}$$

trong đó U là tập các đối tượng, $POS_{a_k}(d)$ là vùng dương được tính cho các phân hoạch của U dựa trên thuộc tính quyết định d, với a_k là thuộc tính điều kiện. Hệ số này thoả mãn điều kiện

$$0 \leq \alpha_k \leq 1. \tag{2.6}$$

Từ đó, ta đưa ra công thức xác định trọng số đã được chuẩn hoá cho thuộc tính k sau khi đã xác định được giá trị của α_k như sau

$$\beta_k = w_k^2 = \frac{\left(c^{\alpha_k}\right)^2}{\sum\limits_{k=1}^{m}\left(c^{\alpha_k}\right)^2} \tag{2.7}$$

với điều kiện $c > 1$. Ở đây, tác giả lựa chọn giá trị $c = e$ để tính toán và thực nghiệm trong

các phần sau này. Công thức (2.7) có thể viết thành

$$w_k = \frac{e^{\alpha_k}}{\sqrt{\sum_{k=1}^{m} (e^{\alpha_k})^2}}. \tag{2.8}$$

Các trọng số này thoả mãn điều kiện

$$w_k \geq 0, \forall k = \overline{1, m} \tag{2.9}$$

và

$$\sum_{k=1}^{m} w_k^2 = 1. \tag{2.10}$$

2.3. Khoảng cách giữa hai đối tượng

Từ khoảng cách cục bộ của các thuộc tính của các đối tượng và xuất phát từ công thức tính khoảng cách *Euclide*, tác giả sử dụng các trọng số đã được tính toán ở trên cho các thuộc tính tương ứng, khi đó khoảng cách giữa hai đối tượng được định nghĩa một cách tổng quát là:

$$G_{ij} = \sqrt{\sum_{k=1}^{m} (w_k g_{ijk})^2} = \sqrt{\sum_{k=1}^{m} w_k^2 g_{ijk}^2} \tag{2.11}$$

trong đó w_k là trọng số tương ứng với thuộc tính thứ k, được xác định theo (2.8).
Dễ thấy G_{ij} thoả mãn hoàn toàn các tiên đề *metric*.

2.4. Độ đo tương tự hỗn hợp có trọng số dựa trên lý thuyết tập thô (Mixed Similarity Measure based on Rough sets theory - MSM-R)

Từ các đề xuất trên, ta có thể tính độ đo tương tự hỗn hợp *MSM-R* cho hai đối tượng i và j một cách gián tiếp thông qua khoảng cách giữa hai đối tượng như sau:

$$S_{ij} = 1 - G_{ij} = 1 - \sqrt{\sum_{k=1}^{m} w_k^2 g_{ijk}^2} \tag{2.12}$$

2.5. Thuật toán xác định trọng số cho các thuộc tính

2.5.1. Thuật toán tính mức độ ảnh hưởng của thuộc tính a_k lên thuộc tính d

Thuật toán này lấy ý tưởng từ thuật toán tìm các lớp tương đương được trình bày trong [11]. Trước tiên sắp xếp dữ liệu trên thuộc tính a_k, sau đó lần lượt với từng giá trị trên a_k, đếm số lượng đối tượng bằng giá trị này trên a_k và số lượng có giá trị trên d bằng nhau tương ứng. Nếu hai số đếm này có giá trị bằng nhau thì các phần tử đó sẽ thuộc vùng dương của phân hoạch dữ liệu trên d đối với a_k. Độ phức tạp tính toán của thuật toán này tuỳ thuộc vào bước sắp xếp dữ liệu trên thuộc tính a_k, nếu ta chọn thuật toán sắp xếp *quick-sort* thì độ phức tạp tính toán là $O(nlog_2 n)$ trong đó n là số lượng đối tượng trong U, $n = |U|$. **Thuật**

toán: CalculateAlpha
Đầu vào:

- U : *các đối tượng*
- k : *thuộc tính cần xác định độ ảnh hưởng đến thuộc tính quyết định*
- d : *thuộc tính quyết định*

Đầu ra: α_k
Các bước:

1. Sắp xếp tất cả các đối tượng theo a_k
//Tính toán số lượng các phần tử trong mỗi lớp tương đương//
2. $i = 1$
3. count1=0
4. WHILE (i<=|U|)
  4.1. count2=0
  4.2. count3=0
  4.3. WHILE ((i+count2<=|U|) and (a_{ik}*==*$a_{(i+count2))}$*))*
    4.3.1. IF d_i*==*$d_{i+count2}$ *THEN count3=count3+1*
    4.3.2. count2=count2+1
  4.4. IF count2==count3 THEN count1=count1+count2
  4.5. i=i+count2
5. α_k *= count1/|U|*
6. RETURN α_k

2.5.2. Thuật toán tính trọng số cho các thuộc tính

Thuật toán: CalculateWeights
Đầu vào:

- U *: các đối tượng*
- d *: thuộc tính quyết định*

Đầu ra: $w2$ *: vector trọng số cho các thuộc tính*
Các bước:

1. tmp1=0
2. FOR $k = 1$ *TO* m
  2.1 tmp2 = CalculateAlpha(U,k,d)
  2.2 $w2_k = e^{2*tmp2}$
  2.3 tmp1 = tmp1+ $w2_k$
3. FOR k=1 TO m
  3.1 $w2_k = w2_k/tmp1$
4. RETURN w2

Trong thuật toán này, ta lần lượt tính α_k, $w2_k$ cho các thuộc tính, và tính tổng $w2_k$, để chuẩn hoá các trọng số, ta chia mỗi $w2_k$ cho tổng trọng số đã tính được. Độ phức tạp tính toán của thuật toán này tuỳ thuộc vào *CalculateAlpha*, nếu ta chọn thuật toán sắp xếp *quick-sort*

thì độ phức tạp tính toán là $O(mnlog_2n)$ trong đó n là số lượng đối tượng trong U, $n = |U|$, m là số lượng thuộc tính.

3. PHƯƠNG PHÁP NGHIÊN CỨU THỰC NGHIỆM

Thuật toán được lựa chọn để nghiên cứu thực nghiệm cho bài toán phân lớp dữ liệu là thuật toán k-láng giềng gần nhất (*k-Nearest Neighbor*), trong đó sử dụng phương pháp xác nhận chéo 10 lần (*10-fold cross validation*) cho mỗi bộ dữ liệu. Quá trình thực nghiệm cho từng bộ dữ liệu gồm các bước sau:

Bước 1: Tiền xử lý, làm sạch dữ liệu và loại bỏ dữ liệu khuyết.

Bước 2: Chia ngẫu nhiên bộ dữ liệu thử nghiệm U thành 10 bộ dữ liệu con với số lượng phần tử của mỗi bộ dữ liệu con là $|U|/10$.

Bước 3: Với 10 bộ dữ liệu con đó, lấy lần lượt từng bộ dữ liệu con làm dữ liệu kiểm tra (test) và 9 bộ dữ liệu con còn lại được sử dụng làm dữ liệu huấn luyện (train) cho thuật toán *k-NN*. Thực hiện:

Bước 3.1: Tính trọng số cho các thuộc tính dựa trên tập dữ liệu huấn luyện.

Bước 3.2: Phân lớp dữ liệu bằng thuật toán *k-NN*, sử dụng hàm độ đo tương tự là: *MSM-R*, độ đo tương tự hỗn hợp của Goodall. Với mỗi lần phân lớp, tác giả lựa chọn k cho thuật toán láng giềng gần nhất có giá trị lần lượt là $k = 1$, $k = 3$, $k = 5$, $k = 7$, $k = 9$, $k = 10$. Để chọn lớp cho mẫu cần phân lớp, tác giả dùng phương pháp chọn lớp xuất hiện nhiều nhất trong k láng giềng của mẫu cần phân lớp. Phân lớp dữ liệu bằng bằng cây quyết định, sử dụng phần mềm C4.5.

Bước 4: Lấy trung bình độ chính xác phân lớp của các lần thử.

Về dữ liệu được sử dụng để thực hiện thử nghiệm, tác giả lựa chọn 23 bộ dữ liệu từ kho dữ liệu mẫu phục vụ nghiên cứu tại *UCI Machine Learning Repository* [12], các bộ dữ liệu này được tiền xử lý để tạo thành 3 nhóm dữ liệu thử nghiệm:

- *Nhóm 1 - gồm 23 bộ dữ liệu:* Các thuộc tính có thứ tự trong các bộ dữ liệu được chuyển thành các thuộc tính định danh, các thuộc tính số và thuộc tính định danh trong bộ dữ liệu sẽ được giữ nguyên gốc.

  Nhóm 2 – gồm 23 bộ dữ liệu: Các thuộc tính có thứ tự trong các bộ dữ liệu được chuyển thành các thuộc tính định danh, giá trị của các thuộc tính số sẽ được rời rạc hóa thành các giá trị định danh tương ứng với thuộc tính đó bằng công cụ rời rạc hóa trong phần mềm *Rough Set Exploration System 2.2.2* [6], các thuộc tính định danh trong bộ dữ liệu sẽ được giữ nguyên gốc.

- *Nhóm 3 – gồm 4 bộ dữ liệu:* Các thuộc tính số và thuộc tính định danh được giữ nguyên gốc. Các thuộc tính có thứ tự được tiền xử lý bước đầu một cách đơn giản với phương pháp mã hoá các giá trị của thuộc tính này bằng các số trong đoạn $[0, 1]$, với khoảng cách chia đều là *1/(số lượng giá trị của thuộc tính - 1)*. Phương pháp mã hoá này sẽ đảm bảo được tính thứ tự của các giá trị thuộc tính, tuy nhiên vẫn chưa thể hiện được đầy đủ ý nghĩa của từng giá trị của thuộc tính. Ví dụ, một thuộc tính có thứ tự **"Sức gió"** có 3 giá trị là: *low*, *mid*, *high* khi mã hoá sẽ tương ứng với các giá trị: 0, 0.5, 1.

Bảng 3.1. So sánh độ chính xác phân lớp lớn nhất (%) bằng k-NN và cây quyết định

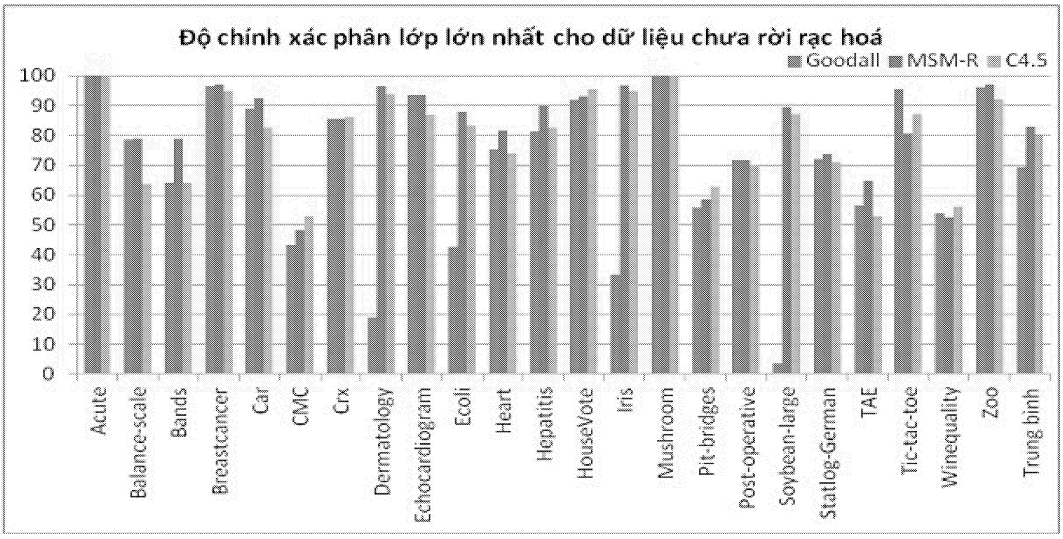
No	Dataset	Attributes		Objs	Data group 1			Data group 2		
		Nom	Num		k-NN		C4.5	k-NN		C4.5
					Goodall	*MSM-R*		*Goodall*	*MSM-R*	
1	Acute	6	1	120	100	100	100	100	100	100
2	Balancescale	5	0	625	78.74	78.88	63.85	78.74	78.88	63.85
3	Bands	20	19	277	64.21	78.99	64.21	64.21	73.19	64.21
4	Breastcancer	0	10	683	96.34	97.23	94.73	97.21	96.63	93.30
5	Car	7	0	1728	88.94	92.59	82.5	88.94	92.59	82.50
6	CMC	8	2	1473	43.12	48.07	52.8	48.78	46.64	48.64
7	Crx	10	6	677	85.59	85.57	86.37	85.00	86.22	86.35
8	Dermatology	33	2	366	19.25	96.35	93.81	28.76	96.63	94.10
9	Echocard	0	11	61	93.57	93.57	86.91	95.00	93.57	93.57
10	Ecoli	1	7	336	42.53	87.82	83.22	75.35	75.94	63.86
11	Heart	6	8	270	75.19	81.48	74.07	80.74	78.89	79.26
12	Hepatitis	14	6	80	81.25	90	82.5	88.75	85.00	78.75
13	HouseVote	17	0	232	91.88	93.13	95.61	91.88	93.13	95.61
14	Iris	1	4	150	33.33	96.67	94.67	90.67	97.33	94.67
15	Mushroom	23	0	5644	99.88	100	100	99.88	100	100
16	Pit-bridges	10	2	70	55.71	58.57	62.86	61.43	60.00	60.00
17	Postoperative	8	1	87	71.67	71.67	70	71.67	71.67	70.00
18	Soybeanlarge	36	0	266	3.7	89.5	87.28	3.70	89.50	87.28
19	StatlogGer	14	7	1000	72.1	73.8	70.9	72.30	71.50	70.00
20	TAE	4	2	151	56.34	64.82	52.77	67.59	66.52	50.89
21	Tic-tac-toe	10	0	958	95.41	80.68	87.15	95.41	80.68	87.15
22	Winequality	0	12	160	53.75	52.5	56.25	48.75	51.88	52.50
23	Zoo	17	0	101	96	97	92.18	96.00	97.00	92.18
	Average				**69.5**	**83**	**79.77**	**75.25**	**81.89**	**78.64**

Bảng 4.1. Kết quả phân lớp (%) với dữ liệu Nhóm 3

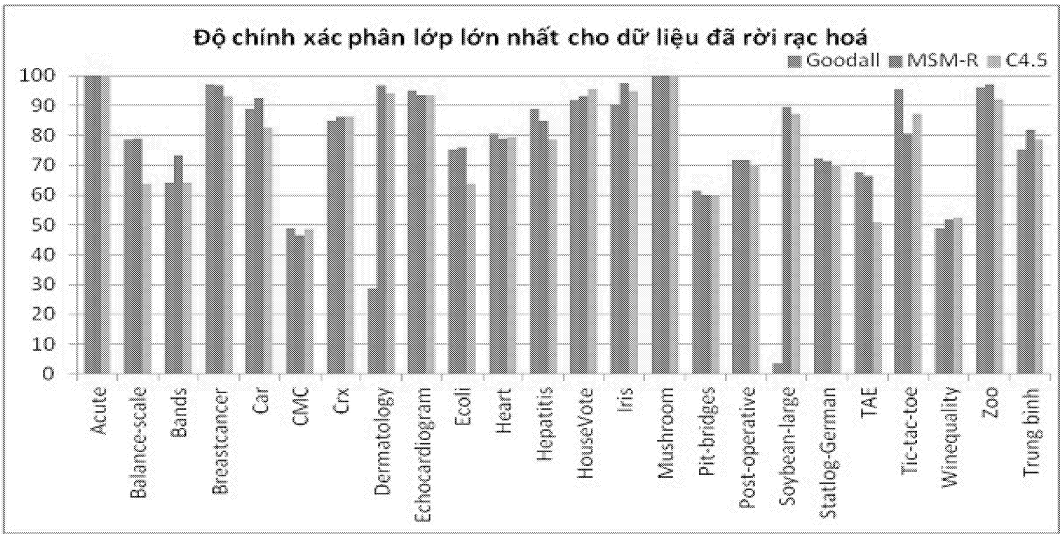
No	Dataset	Attributes			k-NN	
		Nom	*Ord*	*Num*	*Goodall*	*MSM-R*
1	Balance-scale	1	4	0	81.91	88.65
2	Car	1	6	0	84.83	70.03
3	CMC	4	4	2	42.58	48.55
4	Postoperative	1	7	1	70.56	71.67

4. CÁC KẾT QUẢ THỰC NGHIỆM

Độ chính xác phân lớp dữ liệu sau khi thực nghiệm với số lượng láng giềng khác nhau trong thuật toán *k-NN* được tổng hợp tại Bảng 3.1. Tại Bảng 3.1, Hình 4.1 và Hình 4.2 thể hiện sự so sánh giá trị lớn nhất của độ chính xác bằng thuật toán *k-NN* cho từng bộ dữ liệu với phương pháp tính độ tương tự theo *MSM-R* và theo *Goodall*, đồng thời so sánh với độ chính xác phân lớp bằng cây quyết định (sử dụng phần mềm C4.5). Ở đây, độ chính xác phân lớp lớn nhất (mặc dù có thể kết quả này chưa phải là kết quả tối ưu) cho mỗi bộ dữ liệu được tác giả xác định bằng cách tìm giá trị lớn nhất trong các độ chính xác phân lớp khi thực hiện với $k = 1$, $k = 3$, $k = 5$, $k = 7$, $k = 9$, $k = 10$, mỗi bộ dữ liệu sẽ đạt được độ chính xác phân lớp lớn nhất tại giá trị k khác nhau tuỳ thuộc bản chất của dữ liệu. Từ các kết quả này tác giả có nhận xét như sau:



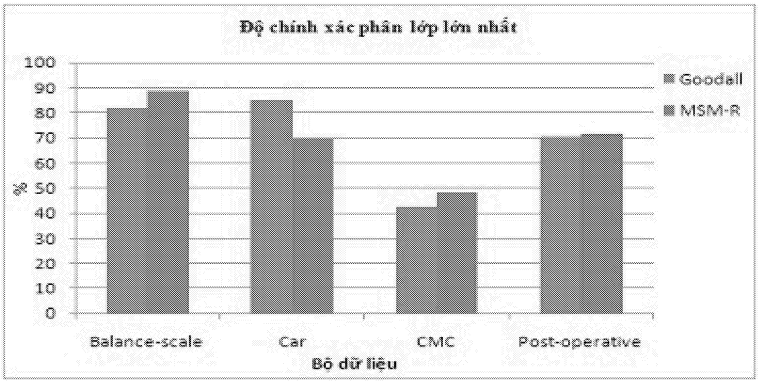
Hình 4.1. So sánh độ chính xác phân lớp lớn nhất cho dữ liệu Nhóm 1



Hình 4.2. So sánh độ chính xác phân lớp lớn nhất cho dữ liệu Nhóm 2

- Với mỗi giá trị khác nhau của k, độ chính xác phân lớp cho các bộ dữ liệu sử dụng độ đo hỗn hợp có trọng số được xác định tự động dựa trên lý thuyết tập thô *MSM-R* nhìn chung tương đương hoặc cao hơn so với kết quả phân lớp khi sử dụng độ đo tương tự hỗn hợp do *Goodall* đề xuất (20/23 bộ dữ liệu thuộc nhóm 1, 14/23 bộ dữ liệu thuộc nhóm 2 có độ chính xác phân lớp cao hơn hoặc tương đương).

- Độ chính xác phân lớp lớn nhất cho các bộ dữ liệu khi sử dụng độ đo *MSM-R* cũng tốt hơn khi sử dụng độ đo của *Goodall*. Khi so sánh độ chính xác phân lớp lớn nhất cho các bộ dữ liệu sử dụng độ đo *MSM-R* với độ chính xác phân lớp bằng cây quyết định,



Hình 4.3. Kết quả phân lớp với dữ liệu Nhóm 3

ta thấy rằng số lượng bộ dữ liệu có độ chính xác phân lớp dùng *MSM-R* cao hơn so với dùng cây quyết định chiếm số lượng nhiều hơn (17/23 bộ dữ liệu thuộc nhóm 1, 17/23 bộ dữ liệu thuộc nhóm 2). Như vậy, có thể đánh giá bước đầu rằng độ đo tương tự hỗn hợp *MSM-R* cho kết quả khá tốt trong bài toán phân lớp dữ liệu.

- Các kết quả phân lớp dữ liệu nhóm 3 thể hiện trong Bảng 4.2 và Hình 4.3 cho thấy các bộ dữ liệu có chứa thuộc tính có thứ tự khi áp dụng phân lớp có sử dụng độ đo *MSM-R* và mã hoá dữ liệu theo phương pháp đơn giản như đã mô tả ở trên cũng cho kết quả khả quan nhưng chưa đạt hiệu quả tuyệt đối khi so sánh với kết quả phân lớp sử dụng độ đo của *Goodall* (có 3/4 bộ dữ liệu cho kết quả phân lớp tốt hơn khi sử dụng độ đo *MSM-R*).

Qua quá trình nghiên cứu thử nghiệm và đánh giá các kết quả thu được, tác giả thấy rằng độ đo *MSM-R* có thể áp dụng một cách có hiệu quả trong bài toán phân lớp dữ liệu cho các đối tượng được thể hiện bằng tập các thuộc tính có kiểu dữ liệu khác nhau mà không yêu cầu chuyển đổi các loại thuộc tính đó về cùng một kiểu dữ liệu. Độ đo tương tự hỗn hợp *MSM-R* được tính gián tiếp thông qua độ đo khoảng cách hỗn hợp có trọng số được tính tự động dựa trên lý thuyết tập thô thoả mãn các tiên đề *metric*. Các trọng số này được xác định một cách tự động dựa trên chính các thông tin nội tại của tập dữ liệu mà không cần thiết phải phục thuộc vào kinh nghiệm và sự can thiệp của các chuyên gia trong lĩnh vực nghiên cứu [8].

5. KẾT LUẬN

Bài báo đã trình bày về độ đo tương tự hỗn hợp dựa trên lý thuyết tập thô (*MSM-R*), trình bày phương pháp nghiên cứu thực nghiệm phân lớp dữ liệu với độ đo *MSM-R* và các kết quả thu được từ bài toán phân lớp dữ liệu áp dụng độ đo này. Lý thuyết độ đo khoảng cách hỗn hợp có trọng số được xác định tự động dựa trên lý thuyết tập thô thoả mãn các tiên đề *metric* cũng đã được chứng minh. Thông qua việc phân tích, so sánh và đánh giá các kết quả thu được, đã làm rõ tính hiệu quả và khả năng áp dụng thực tiễn của độ đo này trong bài toán phân lớp dữ liệu. Hướng nghiên cứu tiếp theo để hoàn thiện độ đo *MSM-R* là tìm phương pháp kết hợp có hiệu quả hơn cho dữ liệu có thứ tự và tiếp tục mở rộng nghiên cứu

để kết hợp các loại dữ liệu đặc thù khác trong cùng một độ đo hỗn hợp như dữ liệu hình ảnh, âm thanh, văn bản... Việc thực nghiệm cũng cần phải tiếp tục thực hiện đối với nhiều bộ dữ liệu khác nhau, với các bài toán dựa trên khoảng cách khác nhằm tìm kiếm các qui luật và phương pháp áp dụng hiệu quả độ đo này.

TÀI LIỆU THAM KHẢO

[1] N. N. Binh, H.T. Bao, A mixed similarity measure in linear time and space computation for distance-based methodss, *Lecture notes in computer science* **1910** (2000).

[2] N. N. Binh, H.T. Bao, T. Morita, Study of a mixed similarity measure for classification and clustering, *3th Pacific-Asia conference on knowledge discovery and data mining*, Springer, 1999.

[3] D.W. Goodall, A new similarity index based on probability, *Biometrics* **22** (1966) 882–907.

[4] C. Li, G. Biswas, Conceptual clustering with numeric and nominal mixed data - A new similarity based system, *KDD: Techniques and Application*, World Scientific, 1997.

[5] C. Li, G. Biswas, Unsupervised learning with mixed numeric and nominal data, *IEEE Transactions on Knowledge and Data Engineering* **14** 4 (2002) 673–690.

[6] Logic Group, Institute of mathematics, Warsaw University, Poland, Rough Set Exploration System 2.2.2 (RSES), http://logic.mimuw.edu.pl/ rses/

[7] Nguyễn Trung Tuấn, Nguyễn Ngọc Bình, Huỳnh Quyết Thắng, Tự động xác định trọng số trong độ đo tương tự hỗn hợp với tiếp cận lý thuyết tập thô, *Chuyên san "Các công trình Nghiên cứu, Phát triển và Ứng dụng Công nghệ thông tin và Truyền thông"* **V-1** 2 (22), 2009.

[8] Nguyễn Trung Tuấn, Ngô Văn Thứ, Ứng dụng độ đo tương tự hỗn hợp trong phân lớp và dự báo dữ liệu kinh tế xã hội, *Tạp chí Kinh tế và Phát triển* **II** (158) (tháng 8/2010).

[9] Z. Pawlak, *Rough sets - Theoretical Aspects of Reasoning About Data*, Kluwer, 1991.

[10] L.S. Quang, "Similarity measures for complex data", Doctor of Philosophy Thesis, School of Knowledge Science, Japanese Advanced Institute of Science and Technology, 2005.

[11] N. H. Son, N. S. Hoa, Some efficient algorithms for rough set methods, *Proceedings IPMU'96,* Granada Spain (1541–1457).

[12] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.

Ngày nhận bài 11 - 4 - 2012
Nhận lại sau sửa 7 - 6 - 2012