2F3-01

事例間類似度の帰納に基づく分類手法

Classification based on induced similarity among examples

小酒井 一稔 *1　犬塚 信博 *2　和田 幸一 *2
Kazutoshi Kozakai　Nobuhiro Inuzuka　Koichi Wada

*1名古屋工業大学電気情報工学科
Department of Electrical and Computer Engineering, Nagoya Institute of Technology

*2名古屋工業大学大学院工学研究科情報工学専攻
Department of Computer Science and Engineering, Guraduate School of Engineering, Nagoya Institute of Technology

Classification problems are to classify examples to given classes by using examples that are already classified. Instance-based methods are an approach to the problem, where examples are classified referring classes of neighbor examples. For this purpose a similarity metric is necessary. This paper uses a nobel similarity that is defined between two exapmles as probability of the examples being in the same class. An instance-based classification method with this metric equals to a posteriori probability classification. The probability or the similarity is calculated from a set of combined examples of pairs that include the information of two examples and the indication of agree or disagree of classes of the pair. In this paper, classification with the combination methods are demonstrated and compared with some conventional methods.

1. はじめに

既分類の事例に基づいて新規の事例の分類を予測する，事例に基づく分類では，類似尺度が必要である．数値属性の表現にもとづいたユークリッド距離などの類似尺度は，その表現形式に利用が制限される．また，対象が類似しているかどうかはその見方によって変わる（分類の場合であれば，どんな分類をしたいかによって変わる）ことから，事例の表現形式から一律に類似度が決まることにも問題がある．

これまでにも表現形式に依存しない類似尺度について研究されてきている．たとえば [Lin 98] は対象の共通性・差異性から情報理論的に類似度を与えた．[Martin & Moal 01] は概念の類似をその生起する確率の一致によって与えるが，確率的な概念の与え方に問題がある．[Kontkanen et al. 99] は事例を条件とした分類の事後確率の分布間の距離で類似尺度を与えている．そのほか，データマイニングの文脈でも多くの類似性尺度が各領域で研究されている．

山田らは二事例間の類似度を，それらの分類が一致する確率で定義することを提案し，制限された範囲でこれを求める手法を提案した [Yamada 02]．この類似度は確率的値で定義されており，データの形式依存しないことから，統一した理論的枠組となる可能性がある．また，分類（すなわち観点）が変われば類似度も変わるという性質も持つ．山田ら（[Yamada 02]）はこの類似度を用いて k-NN 法に類似する分類法（sim_{MAP} 分類法）を与え，一定の条件の下でその結果が最大事後確率分類と一致することが示した．

山田らの類似度は，形式に依存しない定義であるが，その分類手法で用いる類似度計算では，形式に依存した計算を行っている．すなわち，二事例が同分類となる確率を計算するために，事例が＜属性－値＞表現で表されることを仮定し，二事例間の関係をやはり＜属性－値＞表現で表すことで，この確率を与えるモデルを導出するものである．したがって，このような二事例間の形式を表現する方法と，そこから確率モデルを導出する形式が，類似度およびその結果の分類に大きく影響する．

本研究では，二事例間の関係を表す形式を新たに二種類導入する．これによって従来扱うことのできなかった，名義属性と実数値属性が混在する事例を扱うことができることを示す．また，これらの分類性能への影響を実験によって評価する．

山田らの手法では，二事例間の関係を一つの＜属性－値＞形式で表すため，単に値を併置する並列結合法や直積結合法の他に，二つの事例の属性ごとの値を汎化して表現する一致汎化結合法を与えている．本論文では，一致汎化結合法が各属性の値が一致したかどうかを表すことで多くの情報を失っていることに注目し，これを改良した二種類の手法を提案する．一つは単に一致したかどうかでなく，一致した場合はその値を残す最小汎化結合法である．もう一つは，その属性値が分類に与える影響を距離として算出し，距離幅ごとに汎化する距離汎化結合法である．距離汎化結合法は，距離の算出を工夫することで名義属性，連続値属性の両方に利用できる．

連絡先: kozakai@phaser.elcom.nitech.ac.jp

2. 分類確率に基づいた類似度と sim_{MAP} 法

まず [Yamada 02] に習って以下で用いる記号を導入する．

U を可能な全ての事例からなる集合，即ち事例空間，C を事例の可能な分類の集合とする．また，既分類の訓練事例は事例とその分類の対 $\langle x, c_x \rangle$ $(x \in U, c_x \in C)$，訓練事例集合は $S_{\mathrm{base}}(\subset U \times C)$ と書き，$U \times C$ の多重部分集合である．事例空間 U 上に一定の確率分布を仮定し，この分布に従って選ばれる事例を表す確率変数として X, Y を用いる．また，この事例が属する分類クラスを表す確率変数には C_X, C_Y を用いる．

2.1 分類確率に基づいた類似度

山田らは「互いに類似している事例は同じ分類に属す可能性が高い」という直感に基づき，次の類似尺度 $sim(x, y)$ を与え，これを用いた分類投票 $vote(x, c)$ を定義した．

定義 1 (類似性尺度) 事例 $x, y \in U$ に対する類似度 $sim(x, y)$ を次式で与える．

$$sim(x, y) = P(C_X = C_Y | X = x, Y = y) \qquad (1)$$

定義 2 (分類投票) 事例 $x \in U$ の分類 $c \in C$ に対する分類投票を次の通りとする．

$$vote(x, c) = \sum_{y \in U} sim(x, y) P(Y = y, C_Y = c) \qquad (2)$$

表 1: sim_{MAP} 法の概要

入力： 事例集合 S_{base}，クラスの集合 C，
未分類の新事例 x ，
出力： 事例 x の分類 $class(x)$

1 **foreach** $c \in C$
S_{base} でのクラスの頻度から $P(C=c)$ を推定
2 S_{base} から S_{cmb} を生成
3 S_{cmb} から確率モデルを学習
4 (分類投票 $vote(x,c)$ を求める)
4.1 $vote(x,c) := 0$ for all $c \in C$
4.2 **foreach** $\langle y, c_y \rangle \in S_{\mathrm{base}}$
4.3 結合事例 $cmb(x,y)$ を生成
4.4 $cmb(x, x_S)$ と確率モデルから $sim(x,y)$ を求める
4.5 $vote(x,c_y) := vote(x,c_y) + sim(x,y)$
5 x の分類 class(x) を次式で決定する
$class(x) := \mathrm{argmax}_{c \in C}\, vote(x,c)/P(C=c)$
6 **return** $class(x)$

特に，事例に対する分類が決定的である場合は，$P(C_Y = c|Y = y) = \delta(c, c_y)$ （ただし c_y は事例 y の分類，$\delta(\cdot,\cdot)$ はクロネッカーのデルタ）となるので，次式の通り変形できる．

$$\begin{aligned} vote(x,c) &= \sum_{y \in U} sim(x,y) P(Y=y) P(C_Y = c|Y=y) \\ &= \sum_{y \in U} sim(x,y) P(Y=y) \delta(c, c_y) \end{aligned} \quad (3)$$

山田らはある条件下で事後確率が次式の通り分類投票から導かれ，これを分類に利用することが妥当であることを示した．

定理 1 (類似度による事後確率導出定理) 決定的分類を持つ事例が，独立かつ同一の分布 *(iid)* にしたがって生起するとき，事例 $x \in U$ の分類 $c \in C$ に対する事後確率は投票 $vote(x,c)$ と分類の事前確率 $P(C=c)$ により，式 *(??)* で与えられる．

$$P(C_X = c|X = x) = \frac{vote(x,c)}{P(C=c)} \quad (4)$$

通常の重み付き k-NN 法は最大投票で分類予測するが，ここでは投票結果を事前確率で修正すべきであることを示している．

2.2 類似尺度に基づく分類：sim_{MAP} 法

前述の定理を分類に利用するには式 (1) で定義された類似性尺度をデータから計算する必要がある．山田らは，事例から確率モデルを学習し，これを利用して確率を与える方法を提案している．この方法では二つの事例を組合わせて，その二つが同分類であるか否かを表現する新しい事例を作り，そこから同分類である確率のための確率モデルを学習する．

結合事例は，事例の対を１つの事例とみなし，それらの分類の一致／不一致を分類として持つ事例である．事例 $x, y \in U$ に対し，そのを結合した何らかの表現を $cmb(x,y)$ と表し，次の集合を考える．

$$S_{\mathrm{cmb}} = \{\langle cmb(x,y), \delta(c_x, c_y)\rangle | \langle x, c_x\rangle, \langle y, c_y\rangle \in S_{\mathrm{base}}\} \quad (5)$$

事例とその分類が分かっているとき，新たな事例に対しその分類の確率を与える事後確率与える確率モデルを学習することができれば，上の結合事例集合から式 (1) の類似度を求めることができる．このような学習器を PPL (Posterior Probability Learner) と呼ぶことにする．実験で用いた PPL については実験に関する節で述べる．以上の手法を sim_{MAP} 法と呼ぶ．その概要を**表** 1 に示す．

3. 結合事例生成手法

本節では，分類の一致確率の導出するために必要な，結合事例を生成する手法について述べる．前述の通り分類確率に基づく類似尺度自身は事例の表現形式に依存しないが，結合法を用いてそれを算出する手法は形式に依存する．ここでは事例が＜属性-値＞のベクトル形式で表されると仮定する．即ち，各事例は定まった個数の属性 $A, B, \cdots$ を持ち，その各々が事例に固有の値を持つ．以下では，事例 x の属性 A における属性値を $x.A$ と表すことにする．

山田らが与えた結合法である並列結合法，直積結合法，一致汎化結合法を述べ，続いて本稿で提案する最小汎化結合法，距離汎化結合法を述べる．

並列結合法

二つの事例 x, y に対して，次のように属性値を併置して $cmb(x, y)$ を作る方法を並列結合法と呼ぶ．

$$cmb(x, y) = \langle x.A,\ y.A,\ x.B,\ y.B,\ \cdots \rangle$$

属性数は二倍となり，二事例の情報はそのまま保存される．

直積結合法

二つの事例の対応する属性値を組にして属性値とする結合法を直積結合法と呼ぶ．

$$cmb(x, y) = \langle (x.A,\ y.B),\ (x.B,\ y.B),\ \cdots \rangle$$

属性の数は変わらず，属性値が元の属性値の取りうる空間の直積になる．情報は並列結合法同様に保存されるが，値の対応が明確であること，連続的な値をこの方法で結合した場合，値の位相的関係が元とまったく変わることなどの違いがある．

一致汎化結合法

次の結合では，対応する属性値が等しいかどうかのみを属性値として持つ．

$$cmb(x, y) = \langle match(x.A,\ y.A),\ match(x.B,\ y.B),\ \cdots \rangle$$

$$match(x.a, y.a) = \begin{cases} agree & x.a = y.a \text{ のとき} \\ disagree & x.a \neq y.a \text{ のとき} \end{cases}$$

大きく情報が失われる一方，事例間の近さは明確になる．

各結合法を用いた例を次に示す．各事例は「天気」,「気温」,「風力」の三属性をもつ事例空間で，次の二つの事例を考える．

事例 A：$\langle \mathrm{fine}, \mathrm{cold}, \mathrm{weak} \rangle$
事例 B：$\langle \mathrm{rain}, \mathrm{cold}, \mathrm{strong} \rangle$

このとき，上の三種類の結合法では次の結合事例が生成される．

並列結合法： $\langle \mathrm{fine}, \mathrm{rain}, \mathrm{cold}, \mathrm{cold}, \mathrm{weak}, \mathrm{strong} \rangle$
直積結合法： $\langle (\mathrm{fine}, \mathrm{rain}), (\mathrm{cold}, \mathrm{cold}), (\mathrm{weak}, \mathrm{strong}) \rangle$
一致汎化結合法： $\langle disagree,\ agree,\ disagree \rangle$

本稿では，これらの結合法に加えて二つの結合法を提案する．二つはともに汎化結合法，即ち，情報をそのまま保存するのではなく，二つの事例の属性値の関係を何らかの方法で汎化する．一致汎化結合法は値の一致／不一致のみであったが，これを妥当な範囲で汎化度合いを緩めるものである．

最小汎化結合法

第一の方法は属性値が一致したかどうかのみではなく，一致したときはその値を残す結合法である．この方法は論理式の汎化に用いる最小汎化 (least general generalization) との類似から最小汎化結合法と呼ぶ．

$$cmb(x,\, y) = \langle lgg(x.A, y.A)),\ lgg(x.B, y.B),\ \cdots \rangle \qquad (6)$$

ただし，$lgg(\cdot,\cdot)$ は次の通り定める．

$$lgg(x.a, y.a) = \begin{cases} x.a & x.a = y.a \text{ のとき} \\ disagree & x.a \neq y.a \text{ のとき} \end{cases}$$

この手法を用いて 前述の事例を結合すると以下のようになる．

$$\langle disagree,\ \text{cold},\ disagree \rangle$$

距離汎化結合法

もう一つの提案は距離汎化結合法である．汎化結合法では属性値の類似を表わす情報によって事例を結合し，汎化を行っている．そこで，値の間の距離をそのための情報とすることが考えられる．即ち，次のように結合事例を生成する．

$$cmb(x,\, y) = \langle dist(x.A, y.A)),\ dist(x.B, y.B),\ \cdots \rangle \qquad (7)$$

ここで，$dist(a,\, b)$ は値 a と b の間の何らかの距離である．

本研究ではこの距離として HVDM（混合属性値間距離測度）[Wilson & Martinez 97] で用いられる属性値間距離を利用する．HVDM は名義属性と実数値属性が混在する事例に対し適正な距離を与えるために考案された距離尺度である．

事例は＜属性-値＞のベクトル形式を仮定しており，事例間の距離は属性毎に定義された距離の二乗和で与えられる．このため，各属性に対して定義された距離が不釣合いにならないようにする必要がある．そこで，名義属性と連続値属性に対してそれぞれ次の通り属性値間の距離を与えている．

名義属性 a に対しては，次の通り $x.a$ と $y.a$ の距離を定める．

$$dist(x.a, y.a) = \sqrt{\sum_{c \in C} |P(c \mid x.a) - P(c \mid y.a)|^2} \qquad (8)$$

即ち，属性値を条件とするクラス上の分布の差で与えている．

連続値属性 a に対しては，その差を値についての分散 σ_a で正規化したものを用いる．

$$dist(x.a, y.a) = \frac{|x.a - y.a|}{4\sigma_a} \qquad (9)$$

どちらかの属性値が欠損している場合は一律に 1 とする．

$$dist(x.a, y.a) = 1 \qquad (10)$$

HVDM はこれらの属性値間距離によって次の距離を定義する．

$$\text{HVDM}(x, y) = \sqrt{\sum_{a \in \mathcal{A}} dist(x.a, y.a)^2} \qquad (11)$$

ここで，$\mathcal{A}$ は属性集合である．

定義からわかる通り，HVDM はそれ自身属性の表現の違いに強く，またクラス分類の結果に依存して距離が変わる点で，本研究と目的が一致しいる．

距離汎化結合法は HVDM 距離自身ではなく，HVDM で用いられる属性値間の距離，式 (9)〜(10) を利用する．即ち，これらの式で与える距離 $dist(\cdot,\cdot)$ を式 (7) に当てはめることで結合事例を得ることが可能となる．

しかしながらこの距離は，二つの属性値のあらゆる組み合わせに対して異なる値を取る可能性が高く，結果として直積結合法が値の組み合わせをそのまま記憶したことと同じ効果を持つことになりかねない．直積結合法と異なり，距離は実数値となることから，その位相的性質を利用した確率モデルを利用すればこのままでも利用できる可能性があるが，後述の実験方法においては値の違いだけを利用するため，このままでは距離を利用した効果が期待できない．

そこで，距離が連続値として得られることを利用して，以下の式により距離の区間ごとに汎化行う．

$$dg(x.A, y.A) = \lfloor dist(x.A, y.A)/\text{interval} \rfloor \qquad (12)$$

interval は汎化パラメータ，これを区間として距離をまとめる．以上より，距離汎化結合法は以下のように事例を結合する．

$$cmb(x,\, y) = \langle dg(x.A, y.A)), dg(x.B, y.B),\ \cdots \rangle \qquad (13)$$

interval が大きいとき，距離の大きな違いも同じ値にまとめられる．この値を変化させることで生成される結合事例から学習することで，さまざまな学習が可能となる．

距離汎化結合法を前述の例に適用する．三つの属性の属性値間距離が 0.7, 0.0, 0.3 と計算されたとき，interval=0.3 とすると距離汎化結合法は次の結合事例を得る．

$$\langle 3,\ 0,\ 1 \rangle$$

4. 実験および考察

PPL 学習器

前節で与えた事例結合法を sim_{MAP} 法と組み合わせた分類手法を実験によって評価する．ここでは結合事例から分類クラスの一致する確率を求めるための PPL として，ナイーブベイズの仮定に基づく学習器を使用する．即ち，事例 $x =< x.a_1, x.a_2, \cdots >$ がクラス c をとる確率を，次の通り推定する．

$$P(c_j | x.a_1, x.a_2, \cdots) = \frac{P(c_j)}{P(x.a_1, x.a_2, \cdots)} \cdot \prod_{i=1}^{n} P(a_i | c_j) \qquad (14)$$

これを用いて類似尺度を次のように計算する．

$$\begin{aligned} sim(x, y) &= P(1 | cmb(x.a_1,\, y.a_1), cmb(x.a_1,\, y.a_1), \cdots) \\ &= \frac{P(\delta(c_x, c_y))}{P(cmb(x.a_1,\, y.a_1), cmb(x.a_1,\, y.a_1), \cdots)} \\ &\quad \cdot \prod_{a \in A} P(cmb(x.a,\, y.a) | \delta(c_x, c_y)) \end{aligned} \qquad (15)$$

ここで結合事例 $cmb(x,\, y)$ の各属性値を $cmb(x.a_1,\, y.a_1)$ 等と表記した．

上式で項 $P(cmb(x.a_1,\, y.a_1), cmb(x.a_1,\, y.a_1), \cdots)$ は直接求めず，$P(1 \mid cmb(x.a_1,\, y.a_1),\ cmb(x.a_1,\, y.a_1), \cdots)$ と $P(0 \mid cmb(x.a_1,\, y.a_1),\ cmb(x.a_1,\, y.a_1),\ \cdots)$ の和が 1 となるように正規化する．

評価実験

実験データには，UCI の機械学習データベース [UCI 98] より，属性表現されるデータを用いた．実験結果を表 2，4. に示す．この中には，名義属性のみ，連続的な属性のみ，またその両方を持つデータがある．[Yamada 02] による手法と最小汎

表 2: 距離汎化結合法の実験結果

DB 名	C4.5	5-NN	距離汎化結合法 interval= 0.001	0.01	0.1	1
bala. ∗	79.7 (±4.8)	82.1 (±4.1)	86.6 (±2.1)	85.6 (±2.5)	87.5 (±2.1)	79.8 (±3.0)
flar. ∗	71.2 (±5.4)	71.5 (±5.9)	60.3 (±4.4)	60.3 (±4.8)	61.3 (±4.8)	63.4 (±4.6)
haye. ∗	74.2 (±8.3)	81.8 (±3.9)	36.3 (±8.8)	55.3 (±14.9)	71.2 (±14.5)	39.4 (±13.6)
lymp. ∗	80.0 (±4.3)	81.2 (±10.5)	79.1 (±11.8)	82.4 (±5.6)	81.8 (±8.9)	80.4 (±12.4)
monk1 ∗	85.4 (±6.6)	75.0 (±21.1)	67.7 (±19.7)	67.7 (±19.7)	67.7 (±19.7)	73.4 (±20.3)
monk2 ∗	60.9 (±14.9)	58.1 (±19.7)	58.6 (±13.2)	58.6 (±13.2)	58.0 (±13.5)	62.2 (±12.1)
monk3 ∗	91.8 (±9.3)	85.2 (±6.7)	85.1 (±27.6)	85.1 (±27.6)	83.5 (±25.4)	77.7 (±20.3)
prom. ∗	81.1 (±7.3)	88.7 (±11.8)	73.4 (±3.0)	77.4 (±18.0)	84.9 (±8.7)	89.6 (±3.9)
vote ∗	94.3 (±3.4)	93.7 (±4.3)	90.7 (±3.0)	88.3 (±2.4)	90.7 (±2.7)	90.0 (±3.0)
zoo ∗	95.0 (±4.3)	96.0 (±11.1)	93.1 (±4.1)	92.1 (±4.4)	93.0 (±8.6)	93.1 (±4.1)
auto. ○	68.1 (±7.1)	66.3 (±5.6)	71.4 (±6.5)	71.9 (±7.8)	71.1 (±8.2)	65.5 (±7.0)
cpu. ○	68.9 (±6.6)	65.1 (±9.3)	55.9 (±12.1)	55.9 (±12.1)	55.9 (±12.1)	49.3 (±10.0)
ecoli ○	83.3 (±3.6)	83.7 (±4.3)	81.2 (±3.6)	81.5 (±3.5)	82.7 (±2.9)	42.3 (±4.6)
live. ○	66.4 (±5.7)	63.6 (±4.7)	59.6 (±5.2)	59.4 (±5.2)	59.4 (±5.2)	58.5 (±6.8)
wine ○	86.5 (±5.9)	92.7 (±8.5)	91.0 (±5.1)	92.7 (±6.3)	93.3 (±6.3)	89.3 (±5.7)
brig. ∗,○	53.7 (±10.5)	63.0 (±26.1)	60.2 (±17.4)	60.2 (±22.2)	64.8 (±15.9)	42.6 (±14.4)
flag ∗,○	73.7 (±9.1)	64.4 (±7.5)	70.1 (±6.1)	67.0 (±4.7)	59.8 (±4.8)	56.2 (±5.5)
hep. ∗,○	81.9 (±9.2)	83.2 (±3.4)	80.6 (±7.5)	82.6 (±6.1)	84.5 (±6.6)	81.9 (±6.1)

数字は k-交差検定で得られた精度 (%)，カッコ内は 95%信頼区間を示す．5-NN の距離空間は距離汎化結合法に用いたものと同じ HVDM である．
DB 名下の ∗ は名義属性を，○ は連続値属性を各々含むこと示す．

化結合法は名義属性のみのデータにしか適用できない．評価方法には k-fold cross validation 法を用い，分割数 k は最大値を 10 とし，各分割内の事例の数が 30 個以上とになるように選択した．結果には精度と 95% 信頼区間を示した．表 2 には，決定木学習木 C4.5 と 5-NN(最近傍法）の結果を含めた．このとき 5-NN では距離汎化結合法で用いたものと同じ距離尺度を利用した（sim_{MAP} 法では事例間距離そのものは利用していないことに注意）．距離汎化結合法では interval の大きさによって汎化の度合いを変化させることができる．これを変化した場合の実験結果を示している．

5. おわりに

本論文では，分類の一致する確率に基づいた分類法の実現について，二つの事例からその分類が一致する確率を計算するための事例結合法を研究した．このため，最小汎化結合法と距離汎化結合法を提案し，実装・評価した．この手法は，属性の種類ごとに定義される属性値間距離を用いて，事例を結合する手法である．この手法によって生成される結合事例は，属性値間距離を属性値としてもち，名義属性と連続値属性が混在した事例においても方法を適用することができるようになった．

実験においては，これらの方法によって必ずしも優れた結果を得ていない．距離汎化結合法では距離に変換した後，これを区間で区切り名義属性として扱っているため，十分に性質が活かされていないと思われる．これらの改良は今後の課題である．また，この interval の大きさはその値を大きくしたとき一致汎化結合法に近づき，小さくしたときに直積結合法に近づくこと，また，最適な値を頂点とする単峰性を予測できるが，このことは必ずしも実験結果から認められない．これらの検討とパラメータの決定法も今後に残る．関連研究，特に [Kontkanen et al. 99] との関連については今後検討したい．

表 3: 距離汎化結合法以外の結合法の実験結果

DB 名	sim_{MAP} 法 一致汎化	直積	最小汎化
bala.	90.2 (±2.2)	89.6 (±2.9)	86.4 (±3.5)
flar.	56.6 (±6.2)	63.7 (±4.5)	60.0 (±6.7)
haye.	66.7 (±18.9)	56.8 (±22.0)	70.5 (±12.7)
lymp.	82.4 (±10.8)	82.4 (±4.3)	80.4 (±4.1)
monk1	68.5 (±18.4)	70.2 (±19.8)	67.7 (±16.2)
monk2	61.6 (±11.7)	62.1 (±12.1)	62.2 (±12.1)
monk3	86.0 (±28.9)	93.4 (±11.5)	93.4 (±11.5)
prom.	79.2 (±11.4)	89.7 (±10.4)	92.5 (±10.5)
vote	89.3 (±2.9)	92.0 (±4.1)	92.0 (±3.9)
zoo	92.1 (±11.2)	97.0 (±7.5)	96.0 (±11.5)

参考文献

[Yamada 02] Yasuhiro Yamada, Nobuhiro Inuzuka and Hirohisa Seki: MAP Classification with a Similarity Measure, Proceedings of The IASTED International Conference on Artificial and Computational Intelligence, pp.155-160, 2002

[Wilson & Martinez 97] D. Randall Wilson and Tony R. Martinez : Improved Heterogeneous Distance Functions, Jounal of Artificial Intelligence Research, Vol.6, 1997.

[UCI 98] Blake, C.L. and Merz, C.J., UCI Repository of machine learning databases, `http://www.ics.uci.edu/~mlearn/MLRepository.html`, 1998, Irvine, CA: Univ. of California, Dept. Information and Computer Science.

[Kontkanen et al. 99] Petri Kontkanen, Jussi Lahtinen, Petri Myllymäki, Tomi Silander and Hennry Tirri, USING BAYESIAN NETWORKS FOR VISUALIZING HIGH-DIMENTIONAL DATA, in Proceedings of Pre- and Post-processing in Machine Learning and Data Mining: Theoretical Aspects and Applications, a workshop within Machine Learning and Applications (ACAI-99), pp.38-47, 1999.

[Lin 98] Dekang Lin, An Information-Theoretic Definition of Similarity, in Proceedings of the 15th International Conference on Machine Learning, pp.296-304, 1998.

[Martin & Moal 01] Lionel Martin and Frédéric Moal, A Language-Based Similarity Measure, in Proceedings of the 12th European Conference on Machine Learning, pp.336-347, 2001.