Abstract

Similarity measure design remains at the core of many important data mining methods. Distance-based methods such as clustering, classification and nearest neighbor searching use distance functions as a key subroutine in their implementation. Clearly, the quality of the resulting distance function significantly affects the success of the corresponding methods in finding results.

Designing similarity measures for complex data is a challenge because of their particular properties: poorness, heterogeneity, and complexity. These properties make measuring similarities between values or integrating similarity scores on attributes into similarities between objects become difficult tasks. Similarity measures for complex data often require particular designs that are suitable for these properties. This dissertation focuses on similarity measures for three data types: categorical data, heterogeneous data, and graph data.

For categorical data, we investigate characteristics and properties of measures borrowed from binary vector measures to see their advantages and disadvantages. We propose an association-based dissimilarity measure that bases on relations between attributes to measure the dissimilarity between categorical values. The main idea is to estimate the dissimilarity between two values from dissimilarities of probability distributions of attributes conditioned on these two values. Intuitively, the greater the dissimilarities of the probability distributions, the greater the dissimilarity between these two values. This measure does not only overcome poorness in values of binary vector-based measures but also boosts accuracy of the classification nearest neighbor in experiments for a large number of real-life databases.

For heterogeneous data, we report existing similarity measures and point out their advantages and disadvantages. We propose an ordered probability-based similarity measure that is based on order relations and probability distributions. The key idea is to estimate the similarity between two attribute values by the probability of picking up a value pair that is less similar than or as similar as. As similarity scores between attribute values are probabilities, they are then integrated into similarities between objects by methods for integrating probabilities. The main advantage of this measure is that it uses up all particular properties of data types meanwhile still keeps homogeneity of similarity scores on different data types. This measure avoids determining common factors/operators for all data types that is the main drawback of the existing measures.

For graph data, we report my survey on similarity measures and point out their advantages and disadvantages. We propose a nonoverlap connected subgraph-based measure that estimates the similarity between two graphs based on three factors: nodes, edges and connectivity of their common subgraphs. The main idea is that the larger the connected common subgraphs of two graphs, the greater the similarity between these two graphs. Using of these factors makes this similarity measure be suitable to 2D chemical structure data. Experiments with clustering and classifications disclose advantages of this measure in practices. The experiments also reveal interesting relations between compound structures and other chemical properties.

In short, we report similarity measures for complex data and point out their disadvantages and advantages in use. We propose three measures that are particularly designed for categorical data, heterogeneous data, and graph data. The merits of these measures are proven by both theories and experiments.

Key words: Similarity measures, complex data, categorical data, heterogeneous data, graph data, binary vectors, condition probability, order relations, graphs, common subgraphs, chemical structures.

Abstract in Japanese

類似度指標の設計は，重要なデータマイニング・アプリケーションの多くにおいて依然として中核をなす課題である．クラスタリング，分類法，最近隣探索などの多くの手法でもその実装における中心的な下部ルーチンとして距離関数を用いることから，距離関数から導出される結果の品質が対応するアプリケーションの探索の成否に大きく影響することは明らかである．

複合・複雑データに対する類似度指標を設計することは，こうしたデータの情報量の低さ，非均質性，そして複雑性といった特殊な性質のために，野心的な試みとならざるをえない．値以外のデータ型に関してその類似度を測定すること，あるいは複数の属性における類似度をまとめてオブジェクト間の類似度として統合することは困難な課題であり，複雑データに対する類似度指標を設計する際には，この特質に対応することが要求される．本論文では，種々の複雑データのうち，カテゴリデータ，タイプ混交データ，グラフデータの３タイプのデータについて，それぞれを対象とする類似度指標を研究範囲とする．

カテゴリデータに関しては，二進ベクトル指標を取り入れた類似度指標についてその特性を調査し，その長所および短所を明らかにした上で，属性間の関係に基いてカテゴリ値の間の差異度を測定する相関ベース差異度指標を新たに提案する．この指標の要点は，任意の属性における二値間の差異度を，他の属性の条件付確率分布の差異度の総計として推定することにある．直感的に確率分布の差が大きければ，比較対象の二つの値の差異度も大きいと見做せる．多数の実世界データベースに対して実施した最近隣分類の実験において，この指標が二進ベクトル指標で表現できる情報量の少なさを克服するだけでなく，精度をも向上させることを示した．

タイプ混交データに関しては，先行する各種類似度指標についてその長所および短所を指摘しつつ，順序関係および確率分布に則った順序確率ベースの類似度指標を提案する．この指標の要点は，各属性における二つの値の類似性の程度を，順序関係においてより近いその他の値のペアの確率によって推定することである．属性値間の類似性の得点を確率として表すことから，確率を統合する様々な手法を適用することで，二つのオブジェクトの類似度を個々の属性における確率を統合したものとして得ることができる．この指標の主な利点は個々のデータ型の特性を利用しつつも，データ型によらず類似度の構成上の等質性を保持できることにある．また，全データ型に共通する共通要素／操作の決定は既存手法の短所であったが，これが不要であることもこの指標の長所である．

グラフデータに関しても，既存指標を調査し，その長所・短所を報告する．グラフデータに大して新たに提案する指標は，互いに素となる連結部分グラフに基づくものである．この指標は２つのグラフ間の類似度を，その共通部分グラフの節，辺，連結の程度という３つの因子に基づいて推定するもので

あり，その要点は，二つのグラフ間で互いに素となる連結部分グラフが大きければ，グラフ同士の類似度も高いと見做すことにある．グラフの節・辺・連結の程度という因子を用いるため，この類似度指標は二次元データとして表現される化学的構造データに適したものとなっている．クラスタリングおよび分類器による実験は，この指標の実用上の利点を実証するとともに，また化合物の構造とその化学的な特性との間の興味深い関係を明らかにすることにも成功している．

以上，本論文では，複雑なデータに関する既存の類似度指標について，それらの指標を使用する際の長所・短所を報告した．また，複雑なデータの中でも特にカテゴリデータ，タイプ混交データ，グラフデータのそれぞれに特化して設計した新たな類似度指標を提案し，これらの指標が理論的に優位性を備えるだけでなく，実験においても優れていることを実証した．

キーワード：類似度指標，複雑データ，カテゴリデータ，タイプ混交データ（不均質データ），グラフデータ，二進ベクトル，条件付確率，順序関係，グラフ，共通部分グラフ，化学構造