The Journal of Systems and Software 84 (2011) 2374–2383

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Adjusting Fuzzy Similarity Functions for use with standard data mining tools

Avichai Meged, Roy Gelbard*

Information System Program, Graduate School of Business Administration, Bar-Ilan University, Ramat-Gan 52900, Israel

ARTICLE INFO

Article history:
Received 20 June 2010
Received in revised form 4 July 2011
Accepted 6 July 2011
Available online 23 July 2011

Keywords:
Data representation
Data mining
Data segmentation
Classification
Clustering
Similarity function
Similarity measure

ABSTRACT

Data mining is crucial in many areas and there are ongoing efforts to improve its effectiveness in both the scientific and the business world. There is an obvious need to improve the outcomes of mining techniques such as clustering and other classifiers without abandoning the standard mining tools that are popular with researchers and practitioners alike. Currently, however, standard tools do not have the flexibility to control similarity relations between attribute values, a critical feature in improving mining-clustering results. The study presented here introduces the Similarity Adjustment Model (SAM) where adjusted Fuzzy Similarity Functions (FSF) control similarity relations between attribute values and hence ameliorate clustering results obtained with standard data mining tools such as SPSS and SAS. The SAM draws on principles of binary database representation models and employs FSF adjusted via an iterative learning process that yields improved segmentation regardless of the choice of mining-clustering algorithm. The SAM model is illustrated and evaluated on three common datasets with the standard SPSS package. The datasets were run with several clustering algorithms. Comparison of "Naïve" runs (which used original data) and "Fuzzy" runs (which used SAM) shows that the SAM improves segmentation in all cases.

1. Introduction

Data mining techniques are used in diverse areas such as recommendation systems, medical and technical diagnostics, market segmentation, customer profiling and hazard detection, to name a few. Mining techniques such as clustering, classification and association help identify trends and segments in organizational datasets. Improving mining results is an ongoing process which involves improving mining algorithms (Han and Kamber, 2006). Standard commercial tools such as SPSS, SAS, Clementine or even freely available tools such as WEKA implement diverse mining techniques and algorithms, some of which are very sophisticated. Given the availability and the reliability of these tools, they are preferred by practitioners and researchers over proprietary self development tools.

Clustering is a popular data mining technique, for supervised problems (classification problems where the number of groups-clusters is predefined) as well as for unsupervised problems (Giannotti and Pedreschi, 2008; Manying, 2007). Different clustering algorithms can produce different segments for the same datasets, as shown in previous studies such as in Gelbard et al. (2007). Similarity measures are essential to all clustering algorithms. Hence researchers and practitioners look for the best similarity function that will reflect the best membership relations such that objects in one segment are more similar to each other and objects in different segments are less similar and therefore distinct. Studies have shown that different similarity functions yield different clustering results (Bardakh and Fyfe, 2008; Strehl et al., 2000) and learning techniques have been developed for purposes of selecting the best similarity function (Bilenko and Mooney, 2003; Cohen and Richman, 2002; Ristad and Yianilos, 1998; Schultz and Joachims, 2004; Xing et al., 2003). However, researchers and practitioners cannot always pinpoint the exact similarity ratio between any two objects, or any two attributes or any two attribute values. Even if they have this information at their disposal, there is no way to determine or control it when working with standard data mining tools, since these tools use a single similarity function for all objects, attributes and values. This limitation prevents standard data mining tools from achieving better results.

The current paper presents the Similarity Adjustment Model (SAM) that can control similarity relations between attribute values while using standard tools. The similarity control is applied to the input dataset, before the clustering algorithm. This is achieved by transforming the input dataset into a new representational format that is legitimate input to standard tools and embeds new knowledge: the similarity relations between attribute values. These similarities are controlled by Fuzzy Similarity Functions (FSF) which are adjusted in iterative process to produce a better segmentation regardless of the classification-clustering algorithm used.

* Corresponding author. Tel.: +972 3 5318917; fax: +972 3 5353182.
E-mail addresses: megeda@gmail.com (A. Meged), roy.gelbard@biu.ac.il, gelbardr@mail.biu.ac.il (R. Gelbard).

0164-1212/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jss.2011.07.009

Running this process on training data helps to choose the "right" FSF for the test data. The model draws on principles of binary databases models (Gelbard and Meged, 2008; Spiegler and Maayan, 1985). In these models, the data are represented in a matrix where the rows stand for the database entities (objects) and the columns stand for different attribute values. Each datum in the original dataset is transformed into a series of fuzzy numbers representing the degree of similarity between the original datum and its neighbors. These degrees of similarity are also found in models such as the Similarity-based Fuzzy Relational Data Model (Buckles and Petry, 1982) and are produced automatically by FSF as in the Possibility Distribution Model (Prade and Testemale, 1984). The format of the SAM output is matrix-like, in that the cells contain crisp data as required in all standard data mining tools. This representation makes it possible to represent almost unlimited similarity relations between attribute values.

In what follows, the SAM is illustrated and then evaluated on three well known datasets. Each dataset was represented in two forms: a "Naïve" form that contained the original values and a "Fuzzy" form, which was a transformation of the Naïve version generated by FSF. In each iteration the Function Shape (a parameter of the FSF) was adjusted, and an additional SAM dataset was generated. The segmentation experiments were run on a standard SPSS package. The results were evaluated based on precision and recall parameters over the training and test samples. The dataset versions were classified using three different algorithms: K-Means, Two-Step and Hierarchical. Comparison of the resulting groups derived from the "Naïve" versions of the datasets and the "Fuzzy" versions shows that the SAM improved segmentation in all cases (datasets) and for all algorithms.

2. Background

2.1. Clustering and similarity

Clustering is gaining in popularity as a data mining technique and is one of the most extensively studied areas in data mining research (Giannotti and Pedreschi, 2008; Manying, 2007; Ngai et al., 2006). It is used in diverse areas such as recommendation systems, medical and technical diagnostics, segmentation, profiling and detection.

Clustering algorithms reflect hypotheses regarding the assignment of different objects to groups and classes on the basis of the similarity between them. In hard clustering, an object belongs to exactly one cluster while in fuzzy clustering each object can belong to more than one cluster with a specific degree of membership in each cluster. Common clustering algorithms are described in several works such as Estivill-Castro and Yang (2004), Gan et al. (2007), Jain and Dubes (1988), Jain et al. (1999), Lim et al. (2000), Xu and Wunsch (2005) and Zhang and Srihari (2004). The Two-Step, K-Means and Hierarchical algorithms, all three of which are popular in standard mining tools such as SPSS, were tested in this study. The Two-Step algorithm is based, as its name suggests, on two passes of the dataset. The first pass divides the dataset into a coarse set of sub-clusters, and the second pass groups the sub-clusters into the desired number of clusters. The K-Means algorithm, which is one of the most frequently used investigatory algorithms in data analysis, is based on determining arbitrary centers for the desired clusters, associating the samples with the clusters using a predetermined similarity/distance measurement, iteratively changing the center of the clusters, and then re-associating the samples. Because of the good time and space performance of this algorithm, it is used on large datasets. Another popular clustering algorithm type is the Hierarchical algorithm which takes the dataset entities that need to be clustered and starts by classifying the dataset so that each sample represents a cluster. Next, it merges the clusters in steps: each step merges the two clusters that have the maximum similarity (minimum distance) into a single cluster, until there is only one cluster (the dataset) remaining. This algorithm calculates the similarity/distance between clusters in several ways. One example is Ward's Method that calculates the centroid for each cluster and the square of the likelihood measure of each sample in the cluster and the centroid. The two clusters which when united have the smallest (negative) effect on the sum of likelihood measures are the clusters that need to be united.

Similarity measures are essential to all clustering algorithms. There are common similarity measures for quantitative features such as the Minkowski distance, the Euclidean distance which is special case of the Minkowski metric and is the most commonly used metric, the Mahalanobis distance, the Pearson correlation and Cosine similarity, which is the most commonly, used measure in document clustering (Xu and Wunsch, 2005). For binary qualitative features there are other common similarity measures such as the Hamming distance (Illingworth et al., 1983) and the Dice metric (Dice, 1945); for arbitrary qualitative features other similarity measures exist (Grabmeier and Rudolph, 2002). The notion of similarity can vary depending on the particular domain, dataset, or task at hand. Consequently, a large number of functions that compute similarity between objects have been developed for different data types, and these vary greatly in their expressiveness, mathematical properties, and assumptions (Duda et al., 2001; Gusfield, 1997). Research has shown that selecting an appropriate similarity measure for clustering can have a significant impact on clustering results (Bardakh and Fyfe, 2008; Strehl et al., 2000). Therefore, although traditionally clustering has been viewed as an unsupervised learning problem there has been increasing attention to semi-supervised clustering, where limited supervision is provided to obtain a better grouping of the data (Ceccarelli and Maratea, 2008; Lung et al., 2006; Wagstaff et al., 2001; Xing et al., 2003). Researchers use training data to learn accurate similarity functions to capture the correct notion of distance for a particular task at hand in a given domain.

Similarity functions can be trained using pair-wise relations between instances as suggested in works such as Bilenko and Mooney (2003), Cohen and Richman (2002), Ristad and Yianilos (1998), Schultz and Joachims (2004) and Xing et al. (2003). These approaches have shown improvement over traditional similarity functions for different data types such as vectors in Euclidean space, strings, and database records composed of multiple text fields. For example, in Bilenko et al. (2004) employing learnable similarity (distance) functions in clustering led to the development of the MPCK-Means algorithm, which is a semi-supervised variant of unsupervised K-Means clustering. MPCK-MEANS utilizes training data in the form of pair-wise constraints in a unified framework that encompasses cluster initialization, constraint satisfaction, and learning individual parameterized Mahalanobis distances for each cluster. The findings indicate that similarity function learning contributes to improvement over unsupervised clustering.

Although there have been many attempts in the area of controlling similarity functions to improve clustering results, none provide the ability to control similarities using standard mining tools. Rather, all these works control similarity between objects and clusters via proprietary ad hoc algorithms. Researchers are thus prevented from taking advantage of the richness that already exists in standard tools since in these tools, the number of possible similarity functions that can be used is limited and worse, the same similarity function is used for all objects, attributes and values.

The current study overcomes this limitation and makes it possible to control similarites using standard tools. Furthermore, the controlled similarities are at the level of attribute values, the most detailed similarity level that can be expressed. These similarity